Program Europskog odbora za zaštitu podataka *Support pool of experts*

Rizici za zaštitu podataka& mjere za njihovo ublažavanje

Veliki jezični modeli (LLM-ovi)

Autorica: Isabel BARBERÁ

U okviru programa *Support pool of experts* (SPE) Europski odbor za zaštitu podataka može angažirati stručnjake za izradu izvješća i alata o određenim temama.

Stajališta izražena u dokumentu su stajališta njihovih autora i ne odražavaju nužno službeno stajalište Europskog odbora za zaštitu podataka. Europski odbor za zaštitu podataka ne jamči točnost informacija. Ni Europski odbor za zaštitu podataka ni bilo koja osoba koja djeluje u ime Europskog odbora za zaštitu podataka ne mogu se smatrati odgovornima za upotrebu informacija sadržanih u dokumentu.

Pojedini dijelovi mogu biti zacrnjeni ili uklonjeni iz rezultata projekta ako bi njihova objava ugrozila zaštitu legitimnih interesa, uključujući, među ostalim, privatnost i integritet pojedinca u pogledu zaštite osobnih podataka u skladu s Uredbom (EU) 2018/1725 i/ili komercijalne interese fizičke ili pravne osobe.

Napomena: neslužbeni prijevod na hrvatski jezik. Dokument u izvorniku na engleskom jeziku je dostupan na: https://www.edpb.europa.eu/system/files/documents/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf

Sadržaj:

Izjava o odricanju od odgovornosti autora: Primjeri i upućivanja na trgovačka društva uključena u ovo izvješće navedeni su samo u ogledne svrhe i ne podrazumijevaju prihvaćanje niti sugeriraju da predstavljaju jedinu ili najbolju dostupnu opciju. Iako se ovim izvješćem nastoje pružiti temeljite i korisne informacije, ono nije iscrpno. Tehnološka analiza odražava stanje tehnike od ožujka 2025. godine i temelji se na opsežnim istraživanjima, navedenim izvorima i autorovoj stručnosti. Radi transparentnosti autor želi obavijestiti čitatelja da je LLM sustav korišten isključivo u svrhu poboljšanja čitljivosti i formatiranja dijelova teksta.

1. Kako koristiti ovaj dokument

Ovaj dokument pruža praktične smjernice i alate za programere i korisnike sustava koji se temelje na velikom jezičnom modelu (LLM) za upravljanje rizicima za privatnost povezanima s tim tehnologijama.

Metodologija upravljanja rizicima opisana u ovom dokumentu osmišljena je kako bi pomogla razvojnim programerima i korisnicima da sustavno prepoznaju, procijene i ublaže rizike za privatnost i zaštitu podataka, podupirući odgovorni razvoj i uvođenje LLM sustava.

Ovim se smjernicama podupiru i zahtjevi iz članka 25. Opće uredbe o zaštiti podataka, odnosno tehnička i integrirana zaštita podataka, i članka 32. Sigurnost obrade tako što se nude tehničke i organizacijske mjere za osiguravanje odgovarajuće razine sigurnosti i zaštite podataka. Međutim, svrha smjernica nije zamijeniti procjenu učinka na zaštitu podataka u skladu s člankom 35. Opće uredbe o zaštiti podataka. Umjesto toga, njime se dopunjuje postupak procjene učinka na zaštitu podataka rješavanjem rizika za privatnost specifičnih za sustave LLM-a, čime se povećava pouzdanost takvih procjena.

Struktura i pregled sadržaja

Dokument je strukturiran tako da čitatelje vodi kroz ključne tehnološke koncepte, postupak upravljanja rizicima, glavne rizike i mjere ublažavanja te praktične primjere. Cilj mu je podržati organizacije u odgovornom uvođenju sustava temeljenih na LLM-u, istodobno identificirajući i ublažavajući rizike za privatnost i zaštitu podataka pojedinaca.

U nastavku se nalazi pregled strukture dokumenta i tema obuhvaćenih u svakom odjeljku:

2. Kontekst
U ovom odjeljku predstavljeni su veliki jezični modeli, način na koji funkcioniraju i njihove zajedničke primjene. Također se raspravlja o mjerama procjene učinkovitosti, pomažući čitateljima da razumiju temeljne aspekte LLM sustava.

3. Protok podataka i povezani rizici privatnosti u LLM sustavima
Ovdje istražujemo kako nastaju rizici za privatnost u različitim modelima LLM usluga, naglašavajući važnost razumijevanja protoka podataka tijekom životnog ciklusa umjetne inteligencije. U ovom se odjeljku utvrđuju i rizici i mjere ublažavanja te se ispituju uloge i odgovornosti u skladu s Aktom o umjetnoj inteligenciji i Općom uredbom o zaštiti podataka.

4. Procjena rizika u području zaštite podataka i privatnosti: Identifikacija rizika
U ovom se odjeljku navode kriteriji za utvrđivanje rizika i primjeri rizika za privatnost specifičnih za LLM sustave. Razvojni programeri i korisnici mogu se koristiti ovim odjeljkom kao polazištem za utvrđivanje rizika u vlastitim sustavima.

5. Procjena rizika u području zaštite podataka i privatnosti: Procjena rizika & Evaluacija
Ovdje su navedene smjernice o tome kako analizirati, klasificirati i procijeniti rizike za privatnost, s kriterijima za procjenu vjerojatnosti i ozbiljnosti rizika. U ovom se odjeljku objašnjava kako izvesti završnu procjenu rizika kako bi se učinkovito odredili prioriteti napora za ublažavanje.

6. Zaštita podataka i kontrola rizika privatnosti
U ovom se odjeljku detaljno opisuju strategije postupanja s rizicima te se nude praktične mjere za ublažavanje uobičajenih rizika za privatnost u LLM sustavima. Raspravlja se i o prihvaćanju preostalog rizika i iterativnoj prirodi upravljanja rizicima u UI sustavima.

7. Procjena preostalog rizika
Procjena preostalih rizika nakon njihova ublažavanja ključna je kako bi se osiguralo da su rizici unutar prihvatljivih pragova i da nije potrebno daljnje djelovanje. U ovom se odjeljku opisuje kako se procjenjuju preostali rizici kako bi se utvrdilo je li potrebno dodatno ublažavanje ili je li model ili LLM sustav spreman za uvođenje.

8. Pregled & Nadzor
Ovaj odjeljak obuhvaća važnost preispitivanja aktivnosti upravljanja rizicima i vođenja registra rizika. Naglašava se i važnost kontinuiranog praćenja kako bi se otkrili rizici u nastajanju, procijenio stvarni učinak i razradile strategije ublažavanja.

9. Primjeri procjena rizika LLM sustava
Navedena su tri detaljna slučaja primjene kako bi se dokazala primjena okvira za upravljanje rizicima u stvarnim scenarijima. Ti primjeri pokazuju kako se rizici mogu utvrditi, procijeniti i ublažiti u različitim kontekstima.

10. Upućivanje na alate, metodologije, referentne vrijednosti i smjernice
U završnom odjeljku objedinjuju se alati, evaluacijski mjerni podaci, mjerila, metodologije i standardi kako bi se razvojnim programerima i korisnicima pružila potpora u upravljanju rizicima i evaluaciji učinkovitosti LLM sustava.

Smjernice za čitatelje

- Za razvojne programere: Upotrijebite ove smjernice za integraciju upravljanja rizicima za privatnost u životni ciklus razvoja i implementaciju vaših sustava temeljenih na LLM-u, od razumijevanja protoka podataka do provedbe mjera za identifikaciju i ublažavanje rizika.
- Za korisnike: Koristite ovaj dokument kako biste procijenili rizike za privatnost povezane s LLM sustavima koje namjeravate uvesti i upotrebljavati, pomažući vam da usvojite odgovorne prakse i zaštitite privatnost pojedinaca.
- Za donositelje odluka: Strukturirana metodologija i primjeri upotrebe pomoći će vam u procjeni usklađenosti LLM sustava i donošenju informiranih odluka na temelju rizika.

2. Kontekst

Koji su veliki jezični modeli?

Veliki jezični modeli (LLM) predstavljaju transformativni napredak u umjetnoj inteligenciji. Ti modeli opće namjene treniraju se na opsežnim skupovima podataka, koji često obuhvaćaju javno dostupan sadržaj, vlasničke skupove podataka i specijalizirane podatke specifične za određenu domenu. Njihove primjene su raznolike, u rasponu od stvaranja teksta i sažetka do pomoći kodiranja, analize sentimenta i još mnogo toga. Neki LLM-ovi su multimodalni LLM-ovi koji mogu obrađivati i generirati višestruke modalitete podataka kao što su slike, audiozapisi ili videozapisi.

Razvoj LLM-ova obilježen je ključnim tehnološkim prekretnicama koje su oblikovale njihov razvoj. Rani napredak u 1960-ima i 1970-ima uključivao je sustave temeljene na pravilima kao što je ELIZA, koja je postavila temeljna načela za simulaciju ljudskog razgovora kroz unaprijed definirane obrasce. U 2017. godini uvođenje transformatorskih arhitektura (vidi sliku 2) u osnovnom radu "Pažnja je sve što trebate"[1] revolucioniralo je polje omogućujući učinkovito rukovanje kontekstualnim odnosima unutar tekstualnih sekvenci. Naknadnim razvojem događaja, kao što su serija GPT OpenAI-ja i Googleov BERT (vidjeti sliku 3.), utvrđena su mjerila za obradu prirodnog jezika (NLP),[2]što je kulminiralo modelima kao što su GPT-4, LaMDA[3]i DeepSeek-V3[4] (vidjeti sliku 4.) kojima se integriraju multimodalne sposobnosti.

Kako funkcioniraju veliki jezični modeli?

LLM-ovi su napredni modeli dubokog učenja osmišljeni za obradu i generiranje jezika nalik ljudskom. Oslanjaju se na **arhitekturu transformatora**[5], koja koristi mehanizme pažnje za razumijevanje konteksta i odnosa između riječi. Iako se većina najsuvremenijih LLM-ova oslanja na transformatore zbog njihove skalabilnosti i učinkovitosti,[6] postoje alternative koje se temelje na RNN-u (ponavljajuće neuronske mreže), kao što su LSTM (dugoročna kratkoročna memorija) i druge koje se aktivno istražuju[7]. Transformatori zasad dominiraju jezičnim modelima opće namjene, ali inovacije u arhitekturama kao što su one uvedene modelima poduzeća DeepSeek mogu preoblikovati AI krajolik u budućnosti.

Razvoj[8]LLM-a može se podijeliti u nekoliko ključnih faza:

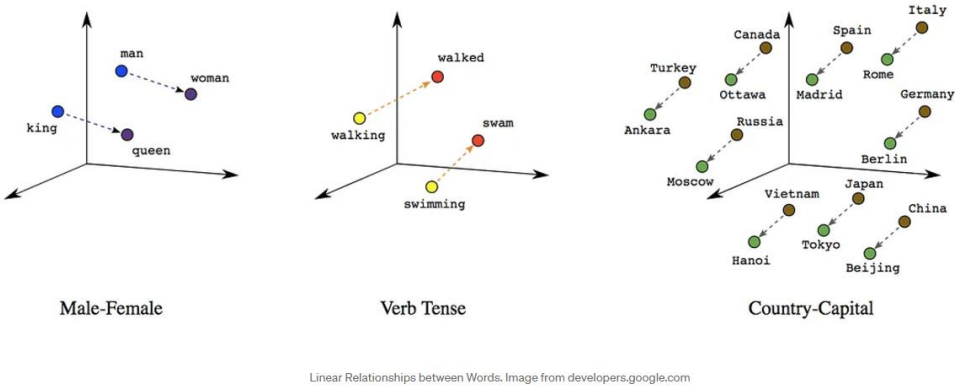
1. Faza osposobljavanja: Izgradnja modela

U ovoj fazi LLM-ovi uče obrasce, kontekst i strukturu u jeziku analizirajući ogromne skupove podataka.

1. **Prikupljanje skupova podataka:**
   Temelj LLM treninga leži u korištenju opsežnih skupova podataka (kao što su Common Crawl i Wikipedia) koji su pažljivo odabrani kako bi se osiguralo da su relevantni, raznoliki i visokokvalitetni. Filtriranje eliminira nekvalitetan ili suvišan sadržaj te osigurava usklađenost podataka za treniranje s predviđenim ciljevima modela.
2. **Predobrada podataka:**
   - Tekst se čisti i normalizira uklanjanjem nedosljednosti (npr. posebni znakovi) i nevažnog sadržaja, čime se osigurava ujednačenost podataka za treniranje.

[1] https://arxiv.org/pdf/1706.03762
[2] https://en.wikipedia.org/wiki/Natural_language_obrada
[3] https://blog.google/technology/ai/lamda/
[4] https://github.com/deepseek-ai/DeepSeek-V3
[5] https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)
[6] https://ai.stackexchange.com/questions/20075/why-does-the-transformer-do-better-than-rnn-and-lstm-in-long-range-context-depen
[7] Albert Gu, Tri Dao, „Mamba: Linear-Time Sequence Modeling with Selective State Spaces", 2024. https://arxiv.org/pdf/2312.00752, Peng et al, „RWKV: Reinventing RNNs for the Transformer Era", https://arxiv.org/pdf/2305.13048
[8] https://arxiv.org/html/2401.02038v2

- Tekstualni podaci se razbijaju u manje jedinice nazvane tokeni, koji mogu biti riječi, podriječi ili čak pojedinačni znakovi. Tokenizacijski algoritmi pretvaraju nestrukturirani tekst u sekvence pogodne za računalnu obradu.
- Tokeni se pretvaraju u numeričke identifikacijske oznake koje predstavljaju njihov položaj u rječniku. Te se oznake zatim pretvaraju u vektorske prikaze riječi (engl. *word embeddings*)[9] odnsosno guste vektorske reprezentacije koje bilježe semantičke sličnosti i odnose među riječima. Na primjer, semantički povezane riječi poput „kralj" i „kraljica" zauzimat će bliske položaje u prostoru vektorskih reprezentacija.



Slika 1. Izvor: https://towardsdatascience.com/a-guide-to-word-embeddings-8a23817ab60f

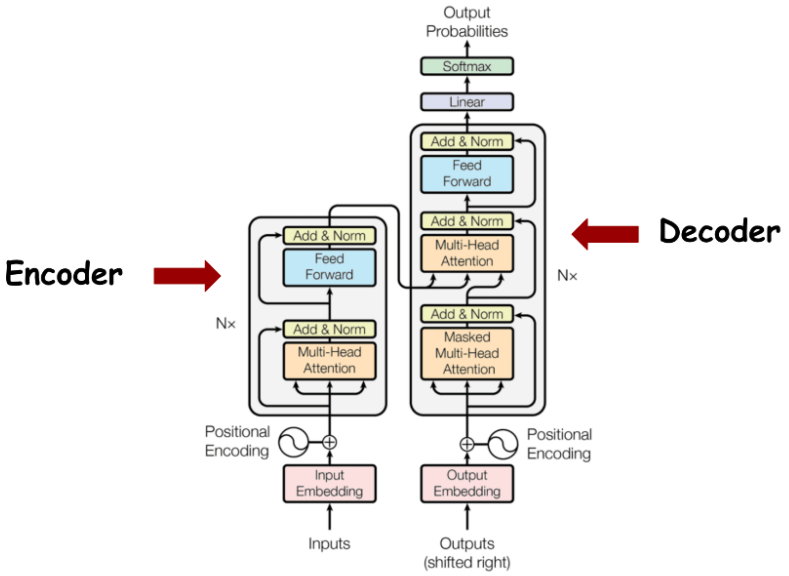
3. **Arhitektura transformatora:**[10]
   Arhitekture transformatora mogu se kategorizirati u tri glavne vrste: Samo koder, samo koder-dekoder i samo dekoder. Iako su arhitekture samo kodera bile temeljne u ranijim modelima, općenito se ne koriste u najnovijoj generaciji LLM-a. Većina najsuvremenijih LLM-ova danas koristi arhitekture samo za dekodere, dok se modeli kodera i dekodera još uvijek koriste u zadacima kao što su prevođenje i ugađanje uputa.

   - **Koder:**[11]
     Koder uzima ulazni tekst i pretvara ga u kontekstualizirani prikaz analizirajući odnose između riječi. Ključni elementi uključuju:
     - *Ugrađivanje tokena:* Tokeni se pretvaraju u numeričke vektore koji bilježe njihovo značenje.
     - *Pozicijski kodovi:* Budući da transformator paralelno obrađuje riječi, dodaju se pozicijski kodovi tokenskim ugradbama kako bi se prikazao redoslijed riječi, čuvajući strukturu ulaza.
     - *Mehanizmi pozornosti:* Koder procjenjuje važnost svake riječi u odnosu na druge u ulaznom slijedu, hvatajući ovisnosti i kontekst. Na primjer, pomaže u razlikovanju „parka" kao glagola i „parka" kao lokacije na temelju okolnog teksta.
     - *feed-forward mreža:* Niz transformacija primjenjuje se za poboljšanje kontekstualiziranih reprezentacija riječi, pripremajući ih za sljedeće faze.
   - **Dekoder:**[12]
     Dekoder generira tekst predviđanjem jednog tokena u isto vrijeme. Temelji se na izlazu kodera (ako se upotrebljava) i slijedu već generiranih tokena. Ključni elementi uključuju:

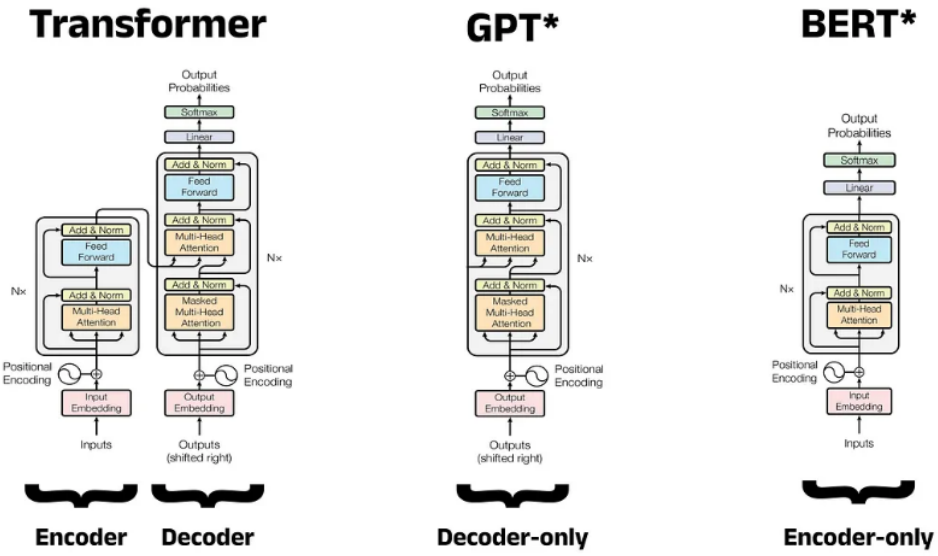
     - *Ulazni podaci:* Kombinira izlaze kodera s do sada generiranim tokenima.

[9] https://ingestai.io/blog/word-embeddings-in-nlp
[10] Vidjeti bilješku 1.
[11] https://www.geeksforgeeks.org/architecture-and-working-of-transformers-in-deep-learning/
[12] idem

- *Mehanizmi pozornosti:* [13] Osigurava da svaki token uzima u obzir prethodno generirane tokene kako bi se održala dosljednost i kontekst.
- *Mreža za prosljeđivanje sirovina (eng. Feed-Forward Network- FFN):* [14] Ovaj sloj razrađuje prikaze tokena kako bi se osiguralo da su relevantni i dosljedni.
- *Maskiranu pozornost:* Tijekom treninga, budući tokeni su skriveni od modela, osiguravajući da predviđa izlaze korak po korak bez "varanja".



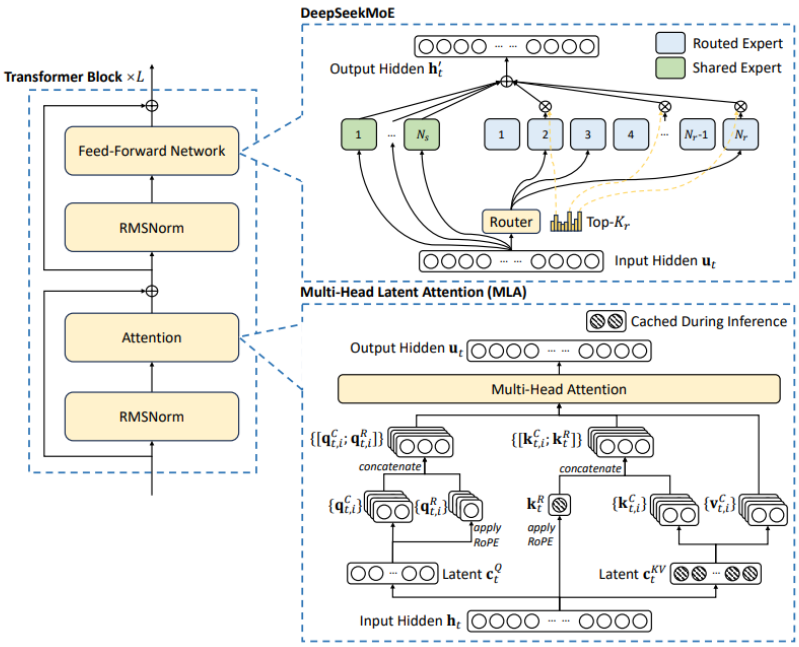
Slika 2. Arhitektura transformatora.
Izvor: https://arxiv.org/pdf/1706.03762



Slika 3. Usporedba arhitektura transformatora, GPT-a i BERT-a.
Izvor: https://towardsdatascience.com/a-complete-guide-to-bert-with-code-9f87602e4a11

[13] Arhitektura DeepSeek modela sadrži inovativni mehanizam pažnje nazvan Multi-head Latent Attention (MLA) koji komprimira Key/Value vektore koji nude bolju računalnu i memorijsku učinkovitost.
[14] DeepSeek modeli koriste DeepSeekMoE arhitekturu temeljenu na Mixture-of-Experts (MoE) uvođenjem više paralelnih stručnih mreža (FFN) umjesto jedne FFN.

Mixture of Experts (MoE) je tehnika koja se koristi za poboljšanje LLM-ova temeljenih na transformer arhitekturi, čineći ih učinkovitijim i skalabilnijim. Umjesto da se za svaki ulaz koristi cijeli model, MoE aktivira samo nekoliko manjih dijelova modela – takozvanih „eksperata" – ovisno o tome što je pojedinom ulazu potrebno. To znači da model u cjelini može biti znatno veći, no u svakom se trenutku koriste samo nužni dijelovi, čime se štedi na računalnoj snazi bez gubitka performansi.



Slika 4. Ilustracija osnovne arhitekture DeepSeek-V3 pod nazivom DeepSeekMoE na temelju mješavine stručnjaka (MoE).
Izvor: https://arxiv.org/pdf/2412.19437

4. **Trening / povratna veza & Optimizacija**[15]
   Faza osposobljavanja LLM-ova oslanja se na strukturiranu optimizacijsku petlju kako bi se poboljšala sposobnost modela da generira točne izlazne rezultate. Taj se iterativan postupak sastoji od sljedećih koraka:
   - *Izračun gubitka:* Nakon generiranja izlaznog slijeda, model ga uspoređuje s ciljnim slijedom ("točan odgovor"). Funkcija gubitka kvantificira pogrešku, pružajući brojčanu mjeru koliko predviđeni izlaz odstupa od željenog rezultata (gubitak).
   - *Backward pass:* Dobivena vrijednost gubitka koristi se za izračun gradijenata, koji pokazuju koliko je svaki parametar modela (npr. ponderi i pristranosti) pridonio pogrešci. Ti gradijenti ističu područja u kojima model treba poboljšati.
   - *Ažuriranje parametara:* Primjenom optimizacijskog algoritma, kao što su Adam ili SGD (Stochastic Gradient Descent), parametri modela prilagođavaju se. Ovim se korakom smanjuje pogreška za buduća predviđanja poboljšanjem pondera unutarnjeg modela.
   - *Ponavljanje:* Ovaj proces se ponavlja za tisuće ili milijune iteracija. Svakim se ciklusom postupno poboljšava uspješnost modela. Trening prestaje kada model postigne ravnotežu između točnosti podataka treninga i generalizacije do neviđenih unosa.

2. Kontinuirano poboljšanje – usklađivanje modela (nakon osposobljavanja)

Unaprijed istrenirani modeli, iako moćni, općenito nisu odmah korisni u svom sirovom obliku. Kako bi se ponašanje modela uskladilo s etičkim razmatranjima i korisničkim preferencijama, potrebno[16] ih je

[15] https://iifx.dev/en/articles/315715245
[16] https://cameronrwolfe.substack.com/p/razumijevanje i korištenje pod nadzorom

prilagoditi. Taj proces često uključuje primjenu tehnika kao što su nadzirano fino ugađanje podataka specifičnih za određeno područje ili jačanje učenja s povratnim informacijama ljudi (RLHF). Najčešće metode usklađivanja su:

- **Nadzirano fino ugađanje (SFT):**[17] Taj pristup uključuje osposobljavanje prethodno osposobljenog modela na označenom skupu podataka prilagođenom određenom zadatku, uz prilagodbe nekih ili svih njegovih parametara kako bi se poboljšala učinkovitost za taj zadatak.
- **Ugađanje uputa:**[18] Ova tehnika se koristi za optimizaciju LLM-a za praćenje korisničkih uputa i rukovanje konverzacijskim zadacima.
- **Učenje za jačanje s povratnim informacijama ljudi (RLHF):**[19]
  Ova metoda koristi ljudske povratne informacije za učenje modela nagrađivanja (RM), koji pomaže u usmjeravanju umjetne inteligencije tijekom procesa učenja. Model nagrađivanja djeluje kao strijelac, pokazujući AI koliko dobro funkcionira na temelju povratnih informacija. Tehnike poput optimizacije proksimalnih pravila (PPO) zatim se koriste za fino podešavanje jezičnog modela. Jednostavnim riječima, jezični model uči donositi bolje odluke na temelju signala nagrade koje prima. Direct Preference Optimization (DPO)[20] je pristup učenju pojačanja u nastajanju koji pojednostavljuje ovaj proces izravnim uključivanjem podataka o preferencijama korisnika u proces optimizacije modela.
  Dok RLHF ima za cilj uskladiti model s ljudskim preferencijama u različitim scenarijima pomoću ljudske povratne informacije, druga varijacija PPO tehnike pod nazivom Group Relative Policy Optimization (GRPO) koju su[21] uveli DeepSeek istraživači, zauzima drugačiji pristup. Umjesto da se oslanja na ljudske primjedbe, GRPO upotrebljava računalno generirane rezultate za automatizirano usmjeravanje procesa učenja i sposobnosti zaključivanja modela.
- **Učinkovito precizno podešavanje parametara (PEFT):**[22] Ova tehnika prilagođava unaprijed obučene modele novim zadacima trenirajući samo neke parametre modela, ostavljajući većinu unaprijed obučenog modela nepromijenjenom. Neke PEFT tehnike su adapteri, LoRA, QLoRA i brzo ugađanje.
- **Generacija proširena dohvaćanjem (eng. *Retrieval Augmented Generation*-RAG):**[23][24][25] Ova metoda poboljšava LLM-ove integriranjem mogućnosti dohvaćanja informacija, omogućujući im upućivanje na određene dokumente. Tim se pristupom LLM-ovima omogućuje da pri odgovaranju na upite korisnika uključe informacije specifične za određenu domenu ili ažurirane informacije.
- **Prijenos učenja:**[26][27] S ovom tehnikom, znanje naučeno iz zadatka ponovno se koristi u drugom modelu.
- **Petlje povratnih informacija:** [28] Povratne informacije korisnika iz stvarnog svijeta pomažu poboljšati ponašanje modela i omogućuju mu da se prilagodi novim kontekstima ili ispravi netočnosti. Povratne informacije mogu se prikupiti ponašanjem korisnika, na primjer na temelju zaključka o tome je li korisnik uključen u odgovor ili ga zanemaruje. Povratne informacije mogu se prikupljati i kada korisnici izravno daju povratne informacije o rezultatima modela, kao što su

[17] https://www.ibm.com/think/topics/fine-tuning
[18] https://www.ibm.com/think/topics/instruction-tuning
[19] https://arxiv.org/abs/2404.08555
[20] https://arxiv.org/abs/2305.18290
[21] https://arxiv.org/abs/2402.03300
[22] https://www.ibm.com/think/topics/parameter-efficient-fine-tuning
[23] https://aws.amazon.com/what-is/retrieval-augmented-generation/
[24] https://en.wikipedia.org/wiki/Retrieval-augmented_generation
[25] https://www.ibm.com/architectures/hybrid/genai-rag?mhsrc=ibmsearch_a&mhq=RAG
[26] https://dev.to/luxacademy/understanding-the-differences-fine-tuning-vs-transfer-learning-370
[27] https://en.wikipedia.org/wiki/Transfer_learning
[28] https://www.nebuly.com/blog/llm-feedback-loop

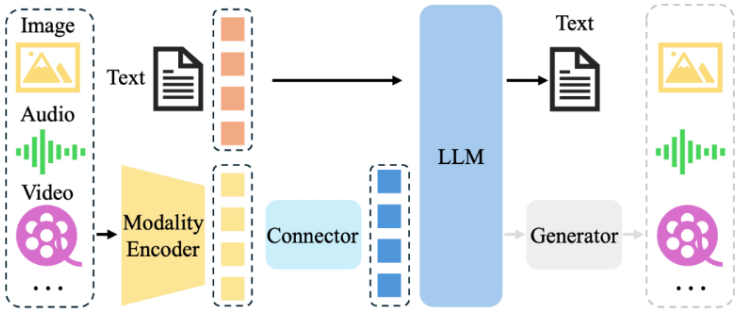
ocjene (palac prema gore / palac prema dolje, kvalitativni komentari ili ispravci pogrešaka. LLM se zatim dorađuje na temelju tih povratnih informacija.

3. Faza zaključivanja: Ostvarivanje rezultata

Nakon što se osposobi, model ulazi u fazu zaključivanja, gdje generira izlazne podatke na temelju novih ulaznih podataka slijedeći ove korake:

1. **Ulazni podaci (eng. *input*):** Korisnikov upit obrađuje se tokenizacijom i ugradnjom, pretvarajući ga u format koji model može razumjeti.
2. **Prerada (eng. *processing*):** Ulaz prolazi kroz arhitekturu transformatora, gdje mehanizmi pažnje i slojevi dekodera predviđaju sljedeće tokene u slijedu. Dekoder proizvodi vektor rezultata (zvan logits) za svaku riječ u vokabularu. Ovi rezultati se zatim prolaze kroz Softmax[29] funkciju, koja ih pretvara u vjerojatnosti. Model odabire najvjerojatniji token kao sljedeću riječ u slijedu, osiguravajući da je generirani tekst koherentan i kontekstualno relevantan.
3. **Izlazni rezultat (eng. *output*):** Model proizvodi vjerojatnosti za potencijalne sljedeće riječi, odabirom najvjerojatnijih opcija na temelju ulaza i konteksta. Ta se predviđanja kombiniraju kako bi se dobili usklađeni i relevantni odgovori.

Opisane tri ključne faze opisuju kako se razvija tradicionalni LLM samo za tekst. Multimodalni LLM-ovi slijede sličan proces, ali za rukovanje višestrukim modalitetima podataka uključuju specijalizirane komponente kao što su koderi specifični za modalitet, konektori i intermodalni fuzijski mehanizmi za integraciju različitih prikaza podataka, zajedno sa zajedničkim dekoderom za generiranje koherentnih izlaza kroz modalitete. Njihov razvoj također uključuje faze prethodnog osposobljavanja i fine prilagodbe; međutim, neke arhitekture grade multimodalne LLM-ove preciznijim prilagođavanjem već prethodno osposobljenog LLM-a samo za tekst umjesto da ga osposobljavaju ispočetka.



Slika 5. Tipična multimodalna LLM (MLLM) arhitektura.
Izvor: Yin, Shukang & Fu, Chaoyou & Zhao, Sirui & Li, Ke & Sunce, Xing & Xu, Tong & Chen, Enhong. (2024). A Survey on Multimodal Large Language Models (Istraživanje o multimodalnim velikim jezičnim modelima). 10.48550/arXiv.2306.13549.

U praksi LLM-ovi često čine dio šireg sustava: izravno su dostupni putem API-ja, ugrađeni unutar SaaS platformi, implementirani kao gotovi temeljni modeli prilagođeni (fine-tuned) za specifične slučajeve uporabe, ili integrirani u on-premise rješenja. Važno je napomenuti da, iako su LLM-ovi bitne komponente AI sustava, sami po sebi ne čine AI sustav. Da bi LLM postao dio AI sustava, potrebno je integrirati dodatne komponente, poput korisničkog sučelja, kako bi mogao funkcionirati kao cjelovit sustav[30]. U cijelom ovom dokumentu takve cjelovite sustave nazivat ćemo sustavima temeljenim na LLM-u, odnosno jednostavno LLM sustavima, kako bismo naglasili njihov širi kontekst i funkcionalnost.

[29] https://en.wikipedia.org/wiki/Softmax_function
[30] Uvodna izjava 97. Akta o umjetnoj inteligenciji

Ova je razlika ključna pri procjeni rizika povezanih s ovim sustavima, budući da LLM sustav sam po sebi nosi veće rizike zbog dodatnih komponenti i integracija u usporedbi sa samostalnim LLM-om.

Svaka faza razvojnog ciklusa LLM-a mogla bi dovesti do potencijalnih **rizika za privatnost** jer model stupa u interakciju s velikim skupovima podataka koji bi mogli sadržavati osobne podatke i generira izlazne rezultate na temelju tih podataka. Neke od ključnih zabrinutosti u pogledu privatnosti mogu se pojaviti tijekom:

- **Prikupljanje podataka:** Skup za treniranje, testiranje i validaciju mogao bi sadržavati osobne podatke koji se mogu identificirati, osjetljive podatke ili posebnu kategoriju podataka.
- **Zaključak:** Generirani rezultati mogli bi nenamjerno otkriti osobne podatke ili sadržavati pogrešne informacije.
- **Postupak RAG-a:** Možemo koristiti baze znanja koje sadrže osjetljive podatke ili osobne podatke koji se mogu identificirati bez primjene odgovarajućih zaštitnih mjera.
- **Petlje povratnih informacija:** Interakcije korisnika mogu se pohranjivati bez odgovarajućih zaštitnih mjera.

Tehnologije LLM-a u nastajanju: Uspon agentske umjetne inteligencije

Prema nedavnom izvješću Deloittea, očekuje[31] se da će do 2027. 50 % poduzeća koja se koriste generativnom umjetnom inteligencijom pokrenuti pilot-projekte ili provjere koncepta za uvođenje agentskih sustava umjetne inteligencije. Ovi sustavi su zamišljeni da funkcioniraju kao inteligentni pomoćnici, sposobni samostalno upravljati složenim zadacima uz minimalan ljudski nadzor.
Agenti umjetne inteligencije (eng. *AI agents*) autonomni[32] su sustavi koji se mogu izgraditi povrh LLM-ova i koji mogu obavljati složene zadatke kombiniranjem sposobnosti LLM-ova s sposobnostima zaključivanja, donošenja odluka i interakcije. AI agenti su proaktivni, sposobni za ponašanje usmjereno na ciljeve kao što su planiranje, izvršavanje zadataka i iteriranje na temelju povratnih informacija. Mogu djelovati neovisno i osmišljeni su za postizanje posebnih ciljeva organiziranjem više uzastopnih radnji. Mogu uključiti i povratne informacije kako bi s vremenom poboljšali svoje postupke ili odgovore. Napredni agenti umjetne inteligencije mogu integrirati sposobnosti drugih UI sustava, kao što su računalni vid ili audioobrada, za obradu različitih unosa podataka.[33]

Koncept agentske umjetne inteligencije i dalje je područje koje se razvija i još nije u potpunosti definirano. Različite organizacije i istraživači predlažu različita tumačenja onoga što čini agentski UI sustav. Na primjer, u Anthropicu[34]naglašavaju značajnu arhitektonsku razliku između tijekova rada i agenata:

- Tijekovi rada strukturirani su sustavi u kojima LLM-ovi i alati rade na unaprijed definiran način, slijedeći orkestrirane putanje koda.
- Za razliku od toga, agenti su dizajnirani da djeluju dinamički. Omogućuju LLM-ovima da samostalno usmjeravaju svoje procese i određuju kako koristiti alate i resurse za postizanje ciljeva.

[31] https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/autonomous-generative-ai-agents-still-under-development.html
[32] https://www.tensorops.ai/post/building-ai-and-llm-agents-from-the-ground-up-a-step-by-step-guide
[33] Operater OpenAI-ja: https://openai.com/index/introducing-operator/
[34] https://www.anthropic.com/research/building-effective-agents

Pregled agenata umjetne inteligencije i njihove arhitekture[35]

U sustavima koje pokreću LLM-ovi, LLM služi kao središnji "mozak" koji pruža temeljne sposobnosti razumijevanja i zaključivanja prirodnog jezika. Ova sposobnost je proširena dodatnim komponentama koje osposobljavaju agenta za planiranje, učenje i dinamičnu interakciju sa svojim okruženjem, omogućujući mu da rješava zadatke koji nadilaze samostalne LLM mogućnosti.

Arhitektura AI agenta usredotočuje se na kritične komponente koje rade zajedno kako bi omogućile sofisticirano ponašanje i prilagodljivost u stvarnim scenarijima. Arhitektura je modularna, uključuje različite komponente za percepciju, rasuđivanje, planiranje, upravljanje memorijom i djelovanje. Ova modularnost omogućuje sustavu rješavanje složenih zadataka, dinamičnu interakciju s okolinom i iterativno poboljšanje performansi.

Neki od najčešćih modula koji se trenutačno upotrebljavaju su:

1. Modul percepcije
Ovaj modul obrađuje sposobnost agenta da obrađuje ulazne podatke iz okruženja i oblikuje ih u strukturu koju LLM može razumjeti. Pretvara neobrađene ulazne podatke (npr. tekst, glasovne ili podatkovne tokove) u ugrađene ili strukturirane formate koje modul zaključivanja može obraditi.

2. Modul za rasuđivanje
Modul zaključivanja omogućuje agentu tumačenje ulaznih podataka, analizu konteksta i razgradnju složenih zadataka na manje podzadatke kojima se može upravljati. Njime se iskorištava sposobnost LLM-a da razumije i obrađuje prirodni jezik kako bi donosio odluke. Mehanizam zaključivanja omogućuje agentu da analizira korisničke unose kako bi odredio najbolji način djelovanja i iskoristio odgovarajući alat ili resurs za postizanje željenog ishoda.

3. Modul za planiranje
Modul planiranja određuje kako će agent izvršiti podzadatke utvrđene modulom zaključivanja. Organizira i sekvencira akcije za postizanje definiranog cilja.

4. Memorija i upravljanje
Kako bi održao kontekst i kontinuitet, agent prati prošle interakcije. Memorija omogućuje AI agentu pohranjivanje i dohvaćanje konteksta, kako unutar jedne interakcije tako i kroz više sesija.

- Kratkoročna memorija: Održava kontekst u trenutačnoj interakciji kako bi se osigurala dosljednost odgovora.
- Dugotrajno pamćenje: Pohranjuje korisničke postavke, prošle interakcije i naučene uvide za personalizaciju.

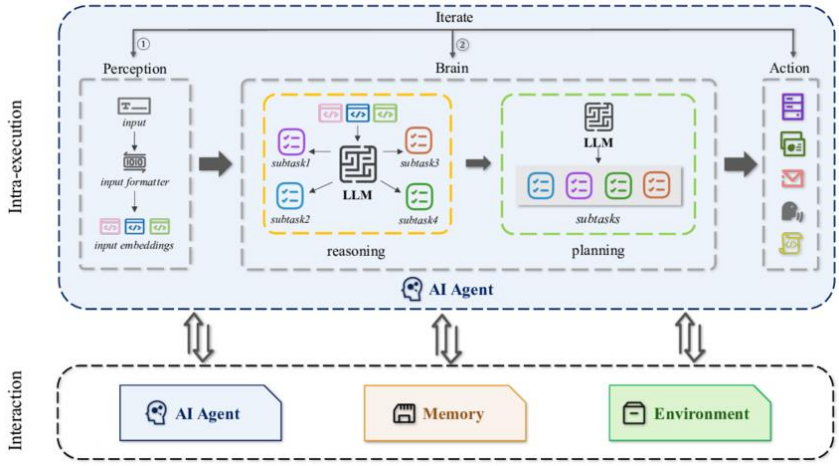
5. Modul djelovanja
Ovaj modul je odgovoran za provedbu plana i interakciju s vanjskim okruženjem. Obavlja zadatke koje su utvrdili i planirali raniji moduli. Agent mora imati pristup definiranom skupu alata, kao što su API-ji, baze podataka ili vanjski sustavi, koje može koristiti za obavljanje određenih zadataka. Na primjer, pomoćnik umjetne inteligencije može upotrebljavati API kalendara za zakazivanje ili uslugu rezervacije za rezervacije putovanja.

6. Povratne informacije i petlja iteracije
Povratna petlja omogućuje agentu da procijeni uspjeh svojih akcija i dinamički prilagodi svoje ponašanje. Uključuje korisničke ispravke, zapise sustava i mjerne podatke o izvedbi kako bi se s vremenom poboljšalo rasuđivanje, planiranje i izvršenje.

[35] idem

Interakcija između AI agenta, memorije i okoliša

Agent neprestano komunicira sa svojom memorijom i vanjskim okruženjem. Kontekst iz memorije poboljšava relevantnost i kontinuitet zadatka, dok vanjski podaci (npr. upiti korisnika, unosi senzora) pokreću donošenje odluka i izvršavanje zadatka.



Slika 6. Izvori: Agenti umjetne inteligencije pod prijetnjom: A Survey of Key Security Challenges and Future Pathways (Istraživanje o ključnim sigurnosnim izazovima i budućim putovima). https://www.researchgate.net/figure/General-workflow-of-AI-agent-Typically-an-AI-agent-consists-of-three-components_fig1_381190070

Mali jezični modeli i njihova uloga u agentima umjetne inteligencije[36]

Mali jezični modeli (SLM-ovi) su „lagani „modeli prilagođeni zadacima dizajnirani za jednostavnije ili fokusiranije zadatke u usporedbi s velikim jezičnim modelima (LLM-ovi). Dok se LLM-ovi ističu u razumijevanju i generiranju složenog jezika, SLM-ovi su optimizirani za određene aplikacije, kao što su klasifikacija teksta, prepoznavanje entiteta ili analiza sentimenta.
SLM-ovi mogu nadopuniti LLM-ove u agentskoj umjetnoj inteligenciji preuzimanjem specijaliziranih zadataka koji ne zahtijevaju opsežne računalne resurse ili općenitost LLM-ova. U agentima umjetne inteligencije, SLM-ovi mogu poboljšati učinkovitost i privatnost lokalnom obradom podataka, smanjujući oslanjanje na centralizirane LLM-ove. Ovaj modularni pristup omogućuje agentima da dodijele zadatke koji poboljšavaju ukupnu izvedbu i sigurnost.

Kako bi se u potpunosti iskoristile mogućnosti LLM-a unutar organizacija, ključno je prilagoditi modele specifičnoj bazi znanja i poslovnim procesima organizacije. Ova prilagodba, često postignuta finim podešavanjem LLM-a s podacima specifičnim za organizaciju, može rezultirati malim jezičnim modelom usmjerenim na domenu (SLM).[37]

Orkestracija modela[38]

Kako bi agentska umjetna inteligencija neprimjetno integrirala prednosti SLM-ova[39] i LLM-ova, potreban je sustav za dinamično upravljanje kojim modelom upravlja koji zadatak. Ovdje orkestracija modela igra ključnu ulogu, osiguravajući učinkovitu i sigurnu suradnju između različitih modela. U agentskoj umjetnoj inteligenciji orkestracija određuje najprikladniji model –LLM ili SLM – za određeni

[36] https://www.ibm.com/think/topics/small-language-models
[37] Biswas, Debmalya. (2024). Stateful Monitoring and Responsible Deployment of AI Agents (Nacionalno praćenje i odgovorno uvođenje agenata umjetne inteligencije).
[38] https://www.ibm.com/think/topics/llm-orchestration
[39] https://siliconangle.com/2024/09/28/llms-slms-sams-agents-redefining-ai

zadatak usmjerava ulazne podatke u skladu s tim i kombinira njihove izlazne podatke u jedinstven odgovor.

Zabrinutost u vezi s privatnošću[40]

Sve veće prihvaćanje AI agenata koje pokreću LLM-ovi donosi obećanje da će revolucionirati način na koji ljudi rade automatizacijom zadataka i poboljšanjem produktivnosti. Međutim, ti sustavi također predstavljaju znatne rizike za privatnost kojima je potrebno pažljivo upravljati:

- Kako bi učinkovito obavljali svoje zadaće, akterima umjetne inteligencije često je potreban pristup širokom rasponu korisničkih podataka, kao što su:
    - Internetska aktivnost: Pregledavanje povijesti, online pretraživanja i često posjećivane web stranice.
    - Osobne aplikacije: E-pošta, kalendari i aplikacije za razmjenu poruka za zakazivanje sastanaka i sl. ili komunikacijske zadatke.
    - Sustavi trećih strana: Financijski računi, platforme za upravljanje klijentima ili drugi organizacijski sustavi.

  Ova razina pristupa značajno povećava rizik od nezakonite obrade osobnih podataka, osobito ako su sustavi agenta ugroženi.
- Agenti umjetne inteligencije osmišljeni su tako da samostalno donose odluke, što može dovesti do pogrešaka ili izbora s kojima se korisnici možda ne slažu.
- Kao i drugi sustavi umjetne inteligencije, agenti umjetne inteligencije podložni su pristranostima koje proizlaze iz njihovih podataka za učenje, algoritama i konteksta uporabe.

***Privacy trade-offs* radi praktičnosti za korisnike:**[41] Kako agenti umjetne inteligencije postaju sposobniji, korisnici će morati razmotriti koliko su osobnih podataka voljni dijeliti u zamjenu za praktičnost. Na primjer, agent može uštedjeti vrijeme upravljanjem rezervacijama putovanja ili pregovaranjem o kupnji, ali zahtijeva pristup osjetljivim podacima kao što su podaci za plaćanje ili vjerodajnice za prijavu[42]. Uravnoteživanje tih kompromisa zahtijeva jasnu komunikaciju o politikama korištenja podataka i robusne mehanizme privola.

Odgovornost za odluke AI agenata:[43] Agenti umjetne inteligencije djeluju u složenim okruženjima i mogu se suočiti s nepredviđenim izazovima. Kada agent napravi pogrešku ili njegovi postupci uzrokuju štetu, utvrđivanje odgovornosti može biti teško. Organizacije moraju osigurati transparentnost u načinu donošenja odluka i osigurati mehanizme za korisnike da interveniraju kada se dogode pogreške.

Uobičajena uporaba LLM sustava

LLM-ovi su postali ključni u različitim industrijama, nudeći napredne sposobnosti u razumijevanju prirodnog jezika i generiranju sadržaja. Tržište nudi spektar LLM rješenja, od kojih je svako prilagođeno specifičnim aplikacijama i zahtjevima korisnika.

1. Vlasnički LLM modeli (eng. *Proprietary LLM Models*)

Vodeće tehnološke tvrtke razvile su vlasničke LLM platforme koje zadovoljavaju različite poslovne potrebe. Neke platforme nude prilagodljive LLM-ove koji se mogu osposobiti za određene skupove podataka:

[40] https://theconversation.com/what-is-an-ai-agent-a-computer-scientist-explains-the-next-wave-of-artificial-intelligence-tools-242586
[41] https://arxiv.org/abs/2309.11653
[42] Vjerodajnice za prijavu jedinstvene su informacije koje se upotrebljavaju za pristup sustavima, računima ili uslugama, a obično se sastoje od korisničkog imena i lozinke, ali mogu uključivati i dodatne metode kao što su dvofaktorska autentifikacija, biometrijski podaci ili sigurnosni PIN-ovi za dodatnu zaštitu.
[43] Zeiser, J. Odluke o vlasništvu: AI Decision-Support and the Attributability-Gap (Odluke o umjetnoj inteligenciji – potpora i zamka pripisivosti). *Sci Eng Ethics* 30, 27 (2024.). https://doi.org/10.1007/s11948-024-00485-1

- OpenAI-jeva[44] GPT serija (Generative Pre-Trained Transformer (GPT) modeli) poznata je po svojim naprednim mogućnostima obrade jezika. Ti su modeli dostupni putem API-ja, što poduzećima omogućuje da u svoje aplikacije integriraju sofisticirano razumijevanje i generiranje jezika.
- Googleovi[45] Gemini modeli dizajnirani su kako bi pomogli u raznim zadacima, pružajući korisnicima detaljne informacije i olakšavajući složene upite.
- Claudeovi antropogeni modeli razvijeni[46] su imajući na umu sigurnost i usklađenost. Claude je specijaliziran za konverzacijsku umjetnu inteligenciju s naglaskom na etičkim i sigurnim interakcijama.

Nekoliko europskih poduzeća i suradnja doprinose okružju LLM-a:

- Mistral AI, novoosnovano[47]poduzeće sa sjedištem u Parizu koje su 2023. osnovali bivši znanstvenici Google DeepMind i Meta AI nudi modele umjetne inteligencije otvorenog koda i vlasničke modele umjetne inteligencije.
- Aleph Alpha[48]ima sjedište u Heidelbergu, Njemačka, a specijalizirana je za razvoj LLM-ova osmišljenih kako bi se osigurala transparentnost u pogledu izvora koji se koriste za generiranje rezultata. Njihovi modeli namijenjeni su poduzećima i vladinim agencijama, obučenima na više europskih jezika.
- Silo AI-ov Poro[49], kroz svoju generativnu AI ruku SiloGen, razvio je Poro, obitelj višejezičnih LLM-ova otvorenog koda. Cilj je ove inicijative ojačati europski digitalni suverenitet i demokratizirati pristup LLM-ovima za sve europske jezike.
- TrustLLM[50]je koordinirani projekt Sveučilišta Linköping koji se usredotočuje na razvoj pouzdane i činjenične LLM tehnologije za Europu, naglašavajući dostupnost i pouzdanost.
- OpenEuroLLM[51] je open source obitelj izvedbenih, višejezičnih, velikih jezičnih temeljnih modela za komercijalne, industrijske i javne usluge.

2. LLM okviri i modeli otvorenog koda

Zajednica otvorenog koda znatno pridonosi okružju LLM-a. Ovdje su neke od najpoznatijih okvira i modela koji su oblikovali razvoj i implementaciju velikih jezičnih modela:

- Hugging Face's Transformers[52] je opsežna biblioteka prethodno istreniranih modela i alata, omogućujući programerima da prilagode i implementiraju LLM-ove za određene zadatke.
- Deepseek[53] je napredni jezični model koji obuhvaća 67 milijardi parametara. Obučen je od nule na ogromnom skupu podataka od 2 trilijuna tokena na engleskom i kineskom jeziku.
- Deepsetov Haystack[54] je open source okvir dizajniran za izgradnju sustava pretraživanja i aplikacija za odgovaranje na pitanja koje pokreće Large Language Models (LLM) i druge tehnike obrade prirodnog jezika (NLP).
- OLMo 32B[55] je prvi potpuno otvoreni model (svi podaci, kodovi, težine i detalji slobodno su dostupni).
- Metini modeli LLaMA usmjereni[56] su na istraživanje i praktičnu primjenu u NLP-u.

[44] https://chatgpt.com/
[45] https://gemini.google.com/
[46] https://claude.ai/
[47] https://mistral.ai/
[48] https://aleph-alpha.com/
[49] https://www.silo.ai/blog/poro-a-family-of-open-models-that-bring-european-languages-to-the-frontier
[50] https://trustllm.eu/
[51] https://openeurollm.eu/
[52] https://huggingface.co/docs/transformers/v4.17.0/en/index
[53] https://www.deepseek.com/
[54] https://haystack.deepset.ai/
[55] https://allenai.org/blog/olmo2-32B
[56] https://www.llama.com/

- BigScience[57] je razvio BLOOM kao višejezični model otvorenog koda koji može generirati tekst na više od 50 jezika, s naglaskom na pristupačnost i uključivost.
- BERT[58] je stvorio Google kako bi razumio kontekst teksta kroz dvosmjernu jezičnu reprezentaciju, izvrsnu u zadacima kao što su odgovaranje na pitanja i analiza sentimenta.
- Falcon[59] je razvijen od strane Instituta za tehnološke inovacije kao model visokih performansi optimiziran za generiranje i razumijevanje teksta, sa značajnim poboljšanjima učinkovitosti u odnosu na slične modele.
- Qwen[60] je velika obitelj jezičnih modela koju je izgradio Alibaba Cloud.
- LangChain[61] je open source okvir za izgradnju aplikacija temeljenih na velikim jezičnim modelima.

3. LLM usluge temeljene na oblaku

Glavni pružatelji usluga u oblaku nude LLM usluge koje se neprimjetno integriraju u postojeće infrastrukture pružajući pristup vlasničkim LLM-ovima i LLM-ovima otvorenog koda:

- Microsoft Azure OpenAI[62] Service surađuje s OpenAI-jem kako bi omogućio API pristup GPT modelima, omogućujući tvrtkama da uključe napredne jezične značajke u svoje aplikacije.
- Amazon Web Services (AWS) Bedrock[63] nudi niz usluga umjetne inteligencije, uključujući jezične modele koji podržavaju različite zadatke obrade prirodnog jezika.
- Google Cloud Vertex AI[64] je platforma za izgradnju, implementaciju i skaliranje modela strojnog učenja, uključujući LLM-ove. Pruža pristup modelima kao što je PaLM 2 i podržava prilagodbu za različite aplikacije, kao što su prevođenje, sažimanje i konverzacijska umjetna inteligencija.
- IBM Watson[65] pruža LLM mogućnosti koje se mogu prilagoditi prepoznavanju entiteta specifičnih za industriju, povećavajući relevantnost i točnost ekstrakcije informacija.
- Cohere[66] nudi prilagodljive LLM-ove koji se mogu prilagoditi određenim zadacima.

Primjena LLM-a

LLM-ovi se upotrebljavaju u različitim aplikacijama,[67]čime se poboljšava korisničko iskustvo i operativna učinkovitost. Ovaj popis predstavlja neke od najistaknutijih primjena LLM-a, ali nipošto nije potpun. Kontinuirano se otkrivaju novi slučajevi upotrebe LLM-ova u različitim industrijama, pokazujući njihov transformativni potencijal u različitim domenama.

- **Chatbotovi i pomoćnici za umjetnu inteligenciju:**[68] LLM-ovi pokreću virtualne asistente kao što su Siri, Alexa i Google Assistant, razumiju i obrađuju prirodni jezik, tumače namjeru korisnika i generiraju odgovore.
- **Generiranje sadržaja:**[69] LLM-ovi pomažu u stvaranju članaka, izvješća i marketinških materijala generiranjem teksta nalik ljudskom, čime se pojednostavljuju procesi stvaranja sadržaja.
- **Jezični prijevodi:**[70] Napredni LLM-ovi olakšavaju usluge prevođenja u stvarnom vremenu.

[57] https://bigscience.huggingface.co/blog/bloom
[58] https://huggingface.co/docs/transformers/model_doc/bert
[59] https://falconllm.tii.ae/
[60] https://huggingface.co/Qwen
[61] https://python.langchain.com/
[62] https://azure.microsoft.com/en-us/products/ai-services/openai-service
[63] https://aws.amazon.com/bedrock
[64] https://cloud.google.com/vertex-ai
[65] https://www.ibm.com/watson
[66] https://cohere.com/
[67] https://www.extentia.com/post/pillars-of-llm-architecture-use-cases-and-examples
[68] https://assistant.google.com/
[69] https://www.jasper.ai/
[70] https://www.deepl.com/en/translator

- **Analiza osjećaja:**[71] Tvrtke koriste LLM-ove za analizu povratnih informacija kupaca i sadržaja društvenih medija, stjecanje uvida u javno raspoloženje i informiranje o strateškim odlukama.
- **Generiranje koda i ispravljanje pogrešaka:**[72] Programeri koriste LLM-ove za generiranje isječaka koda i prepoznavanje pogrešaka, poboljšavajući učinkovitost razvoja softvera.
- Alati za potporu**obrazovanju:**[73] LLM-ovi igraju ključnu ulogu u personaliziranom učenju generiranjem obrazovnog sadržaja, objašnjenja i odgovaranjem na pitanja učenika.
- **Obrada pravnih dokumenata:**[74] LLM-ovi pomažu stručnjacima u pravnom području pregledom i sažimanjem pravnih tekstova, izdvajanjem važnih informacija i pružanjem uvida.
- **Korisnička podrška:**[75] Automatiziranje odgovora na upite kupaca i eskaliranje složenih slučajeva ljudskim agentima.
- **Autonomna vozila:**[76] Vožnja automobila s mogućnošću donošenja odluka u stvarnom vremenu.

Mjere učinkovitosti za LLMs

Ocjenjivanje performansi velikih jezičnih modela (LLM) ključno je kako bi se osiguralo da ispunjavaju svoju namjenu i željene standarde točnosti, pouzdanosti i etičke upotrebe u različitim primjenama. Kako bi se učinkovito izmjerila uspješnost velikog jezičnog modela (LLM), važno je prilagoditi pristup evaluaciji fazi životnog ciklusa LLM-a (npr. treniranje, naknadna obrada, prethodno uvođenje, produkcija) i njegovim predviđenim stvarnim primjenama. Pokazatelji učinkovitosti pomažu u utvrđivanju područja u kojima mogu biti potrebna dodatna ispitivanja ili poboljšanja prije uvođenja ili nakon što se LLM sustav počne upotrebljavati u produkcijskom okruženju.
Neki od najčešćih kriterija procjene učinkovitosti LLM-a su relevantnost odgovora, točnost, semantička sličnost, fluentnost, halucinacija, činjenična dosljednost, kontekstualna relevantnost, toksičnost, pristranost i metrika specifična za zadatak.

Uobičajeno[77] se koriste sljedeći mjerni podaci, od kojih svaki nudi različite uvide:

- **Točnost**[78]mjeri koliko se često izlaz usklađuje s točnim ili očekivanim rezultatima. U zadacima poput klasifikacije teksta ili odgovaranja na pitanja, točnost se izračunava kao omjer točnih predviđanja i ukupnog broja predviđanja. Međutim, za generativne zadatke kao što je generiranje teksta tradicionalni parametri točnosti možda neće u potpunosti obuhvatiti performanse zbog otvorene prirode mogućih točnih odgovora. U takvim slučajevima koriste se parametri kao što su BLEU (Bilingual Evaluation Understudy) i ROUGE (Recall-Oriented Understudy for Gisting Evaluation) kako bi se procijenila kvaliteta generiranog teksta usporedbom s referentnim tekstovima.
- **Preciznošću** se kvantificira omjer točno predviđenih pozitivnih ishoda i ukupnog broja pozitivnih predviđanja modela. U kontekstu LLM-ova visoka preciznost pokazuje da je model točan pri predviđanju. Međutim, u njemu se ne uzimaju u obzir relevantni slučajevi koje model ne predviđa (lažni negativni rezultati), pa se obično kombinira s opozivom (eng. *recall*) radi sveobuhvatnije evaluacije.
- **Opoziv (eng. recall),** koji se naziva i osjetljivost ili stvarna pozitivna stopa, mjeri udio stvarnih pozitivnih slučajeva koje model uspješno identificira. Visoka ocjena opoziva odražava učinkovitost modela u prikupljanju relevantnih informacija, ali ne obuhvaća nevažna predviđanja (lažno pozitivni

[71] https://surveysparrow.com/features/cognivue/
[72] https://github.com/features/copilot
[73] https://www.khanmigo.ai/
[74] https://www.luminance.com/
[75] https://www.salesforce.com/eu/
[76] https://www.tesla.com/autopilot
[77] https://www.turing.com/resources/understanding-llm-evaluation-and-benchmarks
[78] https://www.datacamp.com/blog/llm-evaluacija

rezultati). Zbog toga se opoziv obično procjenjuje zajedno s preciznošću kako bi se pružio uravnotežen pogled.

- **F1 rezultat nudi** uravnoteženu metriku kombiniranjem preciznosti i opoziva u njihovu harmonijsku sredinu. Visoka F1 ocjena pokazuje da model postiže snažnu ravnotežu između preciznosti i opoziva, što ga čini vrijednim mjerilom kada su i lažno pozitivni i lažno negativni kritični. Rezultat F1 kreće se od 0 do 1, a 1 predstavlja savršenu izvedbu na oba mjerenja.
- **Specifičnost** [79]mjeri udio stvarnih negativnih rezultata ispravno identificiranih modelom.
- **AUC** (Area Under the Curve) i **AUROC**[80](Area Under the Receiver Operating Characteristic Curve) kvantificiraju sposobnost modela da razlikuje klase. Ocjenjuje se kompromis između osjetljivosti (stvarna pozitivna stopa) i specifičnosti 1 (lažna pozitivna stopa) u različitim pragovima. Veća vrijednost AUC-a ukazuje na bolju izvedbu u klasifikacijskim zadacima.
- **AUPRC**[81](Area Under the Precision-Recall Curve) mjeri performanse modela u neuravnoteženim skupovima podataka, usredotočujući se na kompromis između preciznosti i opoziva. Visoki AUPRC ukazuje na to da model dobro funkcionira u identificiranju pozitivnih slučajeva, čak i kada su rijetki.
- **Križna entropija**[82]je mjera nesigurnosti ili slučajnosti u predviđanjima sustava. Mjeri razliku između dvije distribucije vjerojatnosti: prave oznake (stvarna distribucija podataka) i predviđene vjerojatnosti iz modela (rezultat). Niža entropija znači veće povjerenje u predviđanja, dok viša entropija ukazuje na nesigurnost.
- ***Cross Entropy***[83]proizlazi iz križne entropije i procjenjuje koliko dobro jezični model predviđa uzorak, služeći kao pokazatelj njegove sposobnosti da se nosi s nesigurnošću. Niža ocjena nesigurnosti znači bolju izvedbu, što ukazuje na to da je model sigurniji u svoja predviđanja. Neke studije upućuju na to da se *cross entropy* pokazao nepouzdanim [84] za procjenu LLM-a zbog njihovih sposobnosti dugotrajnog konteksta. Također je teško koristiti *cross entropy* kao mjerilo između modela jer njegovi rezultati ovise o čimbenicima kao što su metoda tokenizacije, skup podataka, koraci predobrade, veličina vokabulara i duljina konteksta.[85]
- **Kalibracija**[86]se odnosi na usklađivanje predviđenih vjerojatnosti modela i stvarne vjerojatnosti da su ta predviđanja točna. Dobro kalibriran model daje ocjene pouzdanosti koje točno odražavaju stvarne vjerojatnosti ishoda. Pravilna kalibracija ključna je u primjenama u kojima je važno razumijevanje sigurnosti predviđanja, kao što su medicinske dijagnoze ili analiza pravnih dokumenata.
- **MoverScore**[87]je moderna metrika razvijena za procjenu semantičke sličnosti između dva teksta.

Drugi mjerni parametri koji se upotrebljavaju za procjenu performansi i upotrebljivosti sustava koji se temelje na LLM-u, posebno u aplikacijama u stvarnom vremenu ili aplikacijama za koje postoji velika potražnja, jesu sljedeći:[88]

- **Ispunjeni zahtjevi u minuti:** Mjeri koliko zahtjeva LLM može obraditi i vratiti odgovore u jednoj minuti. Odražava učinkovitost sustava u rješavanju više upita.
- **Vrijeme do prvog tokena (TTFT):** Vrijeme od podnošenja zahtjeva do generiranja prvog tokena odgovora.

[79] https://en.wikipedia.org/wiki/Sensitivity_and_specificity
[80] https://arxiv.org/html/2406.03441v1
[81] https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc
[82] https://arxiv.org/html/2402.06216v2
[83] https://thegradient.pub/razumijevanje-evaluacija-metrics-for-language-models/
[84] https://arxiv.org/html/2410.23771v1
[85] https://www.comet.com/site/blog/perplexity-for-llm-evaluation/
[86] https://arxiv.org/abs/2211.09110
[87] https://pypi.org/project/moverscore/
[88] https://www.anyscale.com/blog/reproducible-performance-metrics-for-llm-inference

- **Latencija među tokenima (ITL):** Vremenska odgoda između generiranja uzastopnih tokena u odgovoru. Ovaj mjerni podatak procjenjuje brzinu i fluidnost generiranja teksta.
- **eng. *End to end Latency* (ETEL)** Ukupno vrijeme od trenutka podnošenja zahtjeva do dovršetka cijelog odgovora. Obuhvaća sve faze obrade, uključujući rukovanje ulazima, zaključivanje modela i generiranje izlaza.

Osim tih parametara postoje sveobuhvatni evaluacijski okviri ili mjerila kao[89]što su GLUE (Opća evaluacija razumijevanja jezika)[90], MMLU (Massive Multitask Language Understanding)[91], HELM (Holistic Evaluation of Language Models)[92], DeepEval[93] ili OpenAI Evals[94].
Mjerni podaci specifični za zadatke kao što su BLEU[95] (Bilingual Evaluation Understudy), ROUGE[96] (Recall-Oriented Understudy for Gisting Evaluation) i BLEURT[97] (Bilingual Evaluation Understudy with Representations from Transformers) naširoko se koriste za procjenu generiranja teksta, sažimanja i prevođenja.

Važno je prepoznati da kvantitativna mjerenja sama po sebi nisu dovoljna. Iako su ti parametri vrlo vrijedni za utvrđivanje rizika, posebno kada su integrirani u automatizirane kanale za evaluaciju, prvenstveno služe kao signali ranog upozoravanja, što potiče daljnju istragu kada se premaše pragovi. Mnogi kritični rizici, uključujući potencijal za zlouporabu, etička pitanja i dugoročni učinak, ne mogu se učinkovito obuhvatiti samo tim brojčanim mjerenjima.
Kako bi se osigurala cjelovitija evaluacija, organizacije bi trebale nadopuniti kvantitativne pokazatelje stručnom prosudbom, testiranjem na temelju scenarija i kvalitativnim procjenama.
Okviri otvorenog koda kao što je Inspect[98]podržavaju integrirani pristup omogućavanjem procjena prema modelu, brzim inženjeringom, praćenjem sesija i proširivim tehnikama bodovanja. Ovi alati pomažu operacionalizirati i metričke i kvalitativne procjene, nudeći bolju vidljivost i uvid u ponašanje LLM-a u stvarnim okruženjima.

Mjerenje uspješnosti agentske umjetne inteligencije

Većina trenutačnih parametara [99] za agente umjetne inteligencije usmjerena je na učinkovitost, djelotvornost i pouzdanost. To uključuje mjerne podatke o sustavu (potrošnja resursa i tehnička izvedba), dovršavanje zadataka (mjerenje postizanja ciljeva), kontrolu kvalitete (osiguravanje dosljednosti izlaznih rezultata) i interakciju alata (ocjenjivanje integracije s vanjskim alatima i API-jima).

Neki od ključnih mjernih podataka koji se koriste uključuju:

- **Točnost specifična za zadatak:**[100] Procijeniti koliko je agent ispravno obavlja određene zadatke, kao što su klasifikacija ili dohvat informacija. Najčešće se koriste metrike kao što su Exact Match (EM) i F1 Score.

[89] Referentne vrijednosti standardizirani su okviri razvijeni za procjenu dugoročnih kredita u različitim scenarijima i parametrima (vidjeti i odjeljak 10. ovog dokumenta).
[90] https://gluebenchmark.com/
[91] https://paperswithcode.com/dataset/mmlu
[92] https://crfm.stanford.edu/helm/
[93] https://github.com/confident-ai/deepeval
[94] https://github.com/openai/evals
[95] https://en.wikipedia.org/wiki/BLEU
[96] https://en.wikipedia.org/wiki/ROUGE_(metrički)
[97] https://github.com/google-research/bleurt
[98] https://inspect.aisi.org.uk/
[99] https://www.galileo.ai/blog/metrics-for-evaluating-ai-agents
[100] https://smythos.com/ai-agents/impact/ai-agent-performance-measurement/

- **Dovršetak zadatka od početka do kraja:**[101] Ocjenjuje sposobnost agenta da postigne korisničke ciljeve kroz niz akcija. Pokazatelji uključuju stopu uspjeha zadatka (TSR) i stopu dovršetka cilja (GCR).
- **Točnost na razini koraka:** Procjenjuje ispravnost pojedinačnih radnji koje je agent poduzeo unutar većeg tijeka rada. To je ključno u postupcima koji se sastoje od više koraka, kao što je rezervacija usluge ili rješavanje tehničkog problema.
- **Preciznost i *recall (opoziv):*** Mjeri koliko točno agent dohvaća relevantne informacije (preciznost) i obuhvaća li sve potrebne pojedinosti (opoziv). Ti su mjerni podaci ključni za zadatke kao što su sažimanje dokumenata ili odgovaranje na složene upite.
- **Kontekstualno razumijevanje:**[102]mjeri sposobnost agenta u održavanju i korištenju konteksta u interakcijama, što je ključno za koherentne dijaloge s više zavoja. Praćenje stanja u dijaloškom okviru relevantan[103] je mjerni podatak.
- **Zadovoljstvo korisnika:**[104]mjeri percepciju korisnika o performansama agenta, često putem ocjena povratnih informacija ili anketa i korištenjem vaga za mjerenje upotrebljivosti sustava i korisničkog iskustva.

Evaluacija agenata umjetne inteligencije s tradicionalnim mjerilima LLM-a predstavlja izazove jer često ne uspijevaju zabilježiti dinamiku u stvarnom svijetu, rasuđivanje u više koraka, upotrebu alata i prilagodljivost. Za učinkovitu procjenu potrebne su nove referentne vrijednosti kojima se mjere dugoročno planiranje, interakcija s vanjskim alatima i donošenje odluka u stvarnom vremenu. U nastavku su navedene neke od najpriznatijih referentnih vrijednosti koje se trenutačno upotrebljavaju:

- **SWE-bench:**Software[105]Engineering Benchmark skup podataka, stvoren za sustavnu procjenu sposobnosti LLM-a u rješavanju softverskih problema.
- **AgentBench:**[106][107] Namijenjen je za evaluaciju i obuku sredstava za vizualne temelje na LMM-ima.
- **MLAgentBench:**[108] Procijeniti provode li agenti koje pokreću LLM-ovi učinkovito eksperimentiranje strojnim učenjem.
- **BFCL (BerkeleyFunkcijska ljestvica):**[109] Procijeniti sposobnost različitih LLM-ova za pozivanje funkcija (koje se nazivaju i alati).
- **τ-bench:**[110] Referentna vrijednost za interakciju alata, zastupnika i korisnika u stvarnim domenama.
- **Planbench:**[111] Procijeniti LLM-ove u pogledu planiranja i zaključivanja.

Čimbenici koji mogu utjecati na točnost rezultata

Nekoliko čimbenika može utjecati na točnost izlaznih podataka koje generiraju LLM-ovi. Razumijevanje tih čimbenika ključno je za optimizaciju njihove učinkovitosti i ublažavanje rizika u praktičnim primjenama. Neka od najčešćih čimbenika su:

[101] https://aisera.com/blog/ai-agent-evaluation/
[102] https://smythos.com/artificial-intelligence/conversational-agents/conversational-agents-and-context-awareness/
[103] https://paperswithcode.com/task/dialogue-state-tracking/codeless?page=2
[104] https://essay.utwente.nl/94906/1/Bekmanis_MA_BMS.pdf
[105] https://www.swebench.com/
[106] https://github.com/THUDM/AgentBench
[107] https://paperswithcode.com/dataset/agentbench
[108] https://arxiv.org/abs/2310.03302
[109] https://huggingface.co/skupovi podataka/gorilla-llm/Berkeley-Function-Calling-Leaderboard
[110] https://github.com/sierra-research/tau-bench
[111] https://github.com/karthikv792/LLMs-Planiranje

1. Kvaliteta podataka za osposobljavanje

- **Pristranost podataka:**[112]ako podaci za učenje sadržavaju pristranosti (npr. društvene, kulturne ili jezične pristranosti), model može replicirati ili pojačati te pristranosti u svojim rezultatima.
- **Relevantnost podataka:**[113] Treniranje na zastarjelim, irelevantnim ili „šumovitim" podacima (eng. *noisy data*) može dovesti do netočnih ili kontekstualno irelevantnih odgovora.

2. Ograničenja modela

- **Razumijevanje konteksta:** [114] Unatoč naprednim arhitekturama, LLM-ovi se mogu boriti s nijansiranim kontekstima ili višestrukim razgovorima u kojima raniji dijelovi dijaloga moraju informirati kasnije odgovore.
- **Rukovanje dvosmislenostima:**[115]Dvosmisleni unos može dovesti do netočnih ili besmislenih izlaza ako model ne može zaključiti namjeravano značenje.

3. Tokenizacija i predobrada

- **Pogreške tokenizacije:**[116]Pogrešno prikazivanje ulaznog teksta zbog problema s tokenizacijom (npr. pogrešno razdvajanje riječi) može iskriviti razumijevanje modela.
- **Problemi s predobradom:**[117]Pretjerano agresivno čišćenje ili normalizacija tijekom predobrade može ukloniti važne kontekstualne informacije, smanjujući točnost.

4. Pretjerano/nedovoljno istrenirani modeli (eng. *overfitting/underfitting*)

- **Pretjerano opremanje:**[118]Treniranje sa previše iteracija na ograničenom skupu podataka može model učiniti pretjerano specijaliziranim, što dovodi do slabih performansi na nepoznatim podacima.
- **Neadekvatna**[119]**obuka** ili prejednostavni modeli možda neće uspjeti uhvatiti složenost zadatka, što rezultira općim netočnostima.

5. Brz dizajn i kvaliteta ulaza

- **Brza osjetljivost:**[120]LLM-ovi su vrlo osjetljivi na način oblikovanja ulaznih podataka. Manje varijacije u brzoj strukturi mogu dovesti do drastično različitih izlaza.
- ***Garbage in, garbage out*:**[121]Loše sročen ili nejasan unos može dovesti do netočnih ili nevažnih odgovora.

6. Ograničenja u znanju

- **Prekid znanja:** [122] LLM-ovi se obučavaju na podacima do određenog trenutka. Može im nedostajati svijest o nedavnim događajima ili novim spoznajama.
- **Činjenične pogreške:**[123]LLM-ovi mogu "halucinirati" odnosno izmisliti informacije, stvarajući vjerojatne, ali činjenično netočne odgovore zbog probabilističke prirode njihovih predviđanja.

[112] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Nesreen K. Ahmed; Pristranost i pravednost u velikim jezičnim modelima: Anketa. računalna lingvistika 2024.; 50 (3): 1097.–1179. doi.: https://doi.org/10.1162/coli_a_00524
[113] https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study
[114] https://link.springer.com/article/10.1007/s10676-024-09779-1
[115] https://arxiv.org/abs/2202.12299
[116] https://link.springer.com/article/10.1007/s10115-024-02120-8
[117] https://arxiv.org/abs/2307.06435
[118] https://www.labellerr.com/blog/evaluating-large-language-models
[119] idem
[120] https://link.springer.com/article/10.1007/s10676-024-09779-1
[121] https://arxiv.org/abs/2307.06435
[122] https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study
[123] https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

7. Nedostatak robusnosti

- **Neprijateljski ulazi (eng, adversarial inputs)** [124]LLM-ovi mogu zakazati kada se suoče s namjerno manipuliranim ulaznim podacima osmišljenim za iskorištavanje njihovih slabosti.
- **Buka i varijabilnost:** [125] Pravopisne pogreške, sleng ili nestandardni jezik mogu dovesti do pogrešnih tumačenja i niže točnosti.

8. Neodgovarajuća kalibracija

- **Prekomjerno povjerenje:**[126]Loše kalibrirani modeli mogu dodijeliti visoke ocjene pouzdanosti netočnim predviđanjima, obmanjujući korisnike. Neuspjeh u ispravnom prenošenju neizvjesnosti u predviđanjima može narušiti povjerenje u model.

3. Protok podataka i povezani rizici privatnosti u LLM sustavima

Razumijevanje protoka podataka u sustavima umjetne inteligencije koje pokreću LLM-ovi ključno je za procjenu rizika za privatnost. Taj protok može varirati ovisno o fazama rada, specifičnom sustavu koji model integrira i vrsti modela usluge koji se koristi, od kojih svaka donosi jedinstvene izazove za zaštitu podataka.

Važnost životnog ciklusa umjetne inteligencije u upravljanju rizicima u pogledu privatnosti

Životni ciklus UI sustava, kako je navedeno u normama ISO/IEC 22989 [127] i ISO/IEC 5338, [128] pruža strukturirani okvir za razumijevanje protoka podataka tijekom razvoja, uvođenja i rada UI sustava. Taj je životni ciklus ključan i za utvrđivanje i ublažavanje rizika za privatnost u svakoj fazi.

[124] https://arxiv.org/abs/2202.12299
[125] https://link.springer.com/article/10.1007/s10115-024-02120-8
[126] https://arxiv.org/abs/2402.12563
[127] ISO/IEC 22989 (Umjetna inteligencija – Koncepti i terminologija)
[128] ISO/IEC 5338:2023 Informacijska tehnologija – Umjetna inteligencija – Procesi životnog ciklusa sustava umjetne inteligencije

Slika 7. Izvor: Na temelju ISO/IEC 22989

U ovom dokumentu koristimo ovaj životni ciklus umjetne inteligencije kao referentni okvir, prepoznajući da svaka organizacija može imati vlastitu prilagođenu verziju na temelju svojih specifičnih potreba. Iako su osnovne faze životnog ciklusa općenito slične u različitim organizacijama, točne faze mogu varirati.

Svaka od faza životnog ciklusa uključuje jedinstvene rizike za privatnost koji zahtijevaju prilagođene strategije ublažavanja. Implementacija načela „privacy by design“u svaku fazu pomaže u proaktivnom rješavanju rizika, a ne retroaktivnom popravljanju istih.

Faze životnog ciklusa umjetne inteligencije i njihov utjecaj na privatnost

1. **Početak i dizajn:** U ovoj fazi donose se odluke o zahtjevima za podacima, metodama prikupljanja i strategijama obrade. Odabir izvora podataka može dovesti do rizika ako su osjetljivi ili osobni podaci uključeni bez odgovarajućih zaštitnih mjera.
2. **Priprema i predobrada podataka:** Neobrađeni podaci prikupljaju se, čiste, u nekim slučajevima anonimiziraju[129]i pripremaju za osposobljavanje ili fino podešavanje. Skupovi podataka često potječu iz različitih izvora, uključujući podatke dobivene pretraživanjem interneta, javne repozitorije, vlasničke podatke ili skupove podataka dobivene partnerstvima i suradnjom.
    - Rizici za privatnost:
        - Podaci za trening mogu nenamjerno uključivati osobne podatke, povjerljive dokumente ili druge osjetljive informacije.

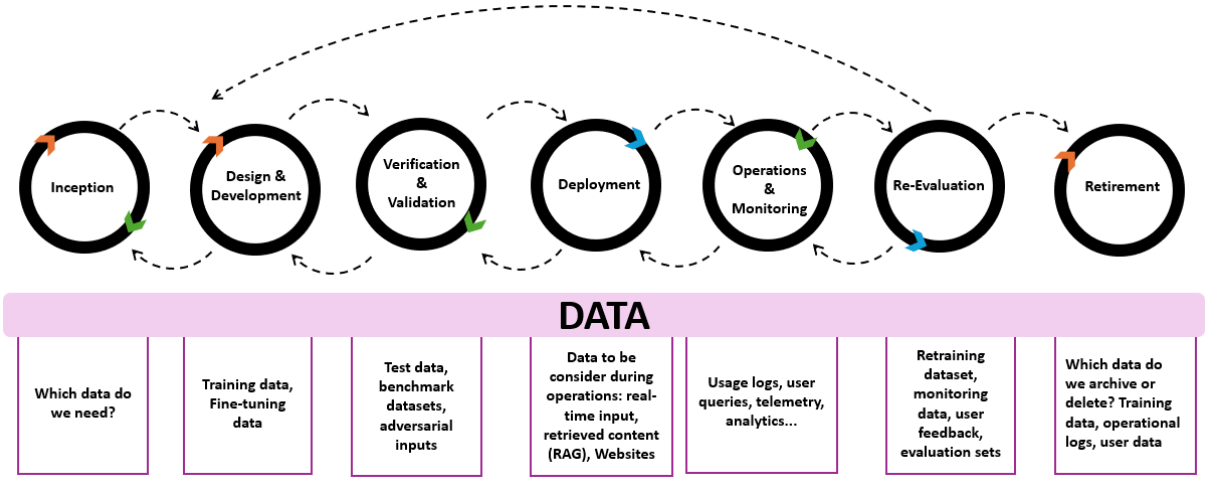
[129] Važno je uzeti u obzir mišljenje Europskog odbora za zaštitu podataka 28/2024 i odjeljak 3.2. O okolnostima u kojima bi se modeli umjetne inteligencije mogli smatrati anonimnima i povezanom demonstracijom: „..., Europski odbor za zaštitu podataka smatra da je, kako bi se model umjetne inteligencije smatrao anonimnim, razumno
znači i. vjerojatnost izravnog (uključujući vjerojatnosnog) izvlačenja osobnih podataka o pojedincima čiji su osobni podaci upotrijebljeni za osposobljavanje modela; i ii. vjerojatnost namjernog ili nenamjernog dobivanja takvih osobnih podataka iz upita trebala bi biti neznatna za bilo kojeg ispitanika.”

- Neodgovarajuća anonimizacija ili rukovanje osobnim podacima mogu dovesti do povreda osobnih podataka ili nenamjernih zaključaka tijekom kasnijih faza.
- Pristranosti prisutne u skupovima podataka mogu utjecati na predviđanja modela, što rezultira nepoštenim ili diskriminacijskim ishodima.
- Pogreške ili propusti u podacima za treniranje mogu negativno utjecati na performanse modela, smanjujući njegovu učinkovitost i pouzdanost.
- Prikupljanje i korištenje podataka za treniranje može dovesti do kršenja prava na privatnost, izostanka privole ili kršenja autorska prava i drugih zakonskih obveza.

3. **Razvoj, treniranje modela:** Pripremljeni skupovi podataka upotrebljavaju se za treniranje modela, što uključuje opsežnu obradu. Model može nenamjerno zapamtiti osobne podatke, što dovodi do potencijalnih kršenja privatnosti ako su takvi podaci izloženi u izlazima.
4. **Verifikacija & Validacija:**[130] Model se procjenjuje pomoću testnih skupova podataka, često uključujući stvarne scenarije. Podaci dobiveni ispitivanjem mogu nenamjerno otkriti osjetljive korisničke informacije, posebno ako se stvarni skupovi podataka upotrebljavaju bez anonimizacije.
5. **Uvođenje (eng. *deployment*) :** Model je u interakciji s izravnim unosima podataka korisnika, često u aplikacijama u stvarnom vremenu koje bi se mogle integrirati s drugim sustavima. Prijenosi podataka uživo mogu uključivati vrlo osjetljive informacije, što zahtijeva stroge kontrole prikupljanja, prijenosa i pohrane.
6. **Rad i praćenje:** Kontinuirani tokovi podataka ulaze u sustav radi nadzora, povratnih informacija i optimizacije performansi. Zapisi (logovi) iz sustava za nadzor mogu sadržavati osobne podatke, poput korisničkih interakcija, što stvara rizike od curenja podataka ili zlouporabe.
7. **Ponovna procjena, održavanje i ažuriranja:** Mogu se prikupiti dodatni podaci za ponovno osposobljavanje ili ažuriranje modela kako bi se poboljšala točnost ili ispunili novi zahtjevi. Korištenje korisničkih podataka uživo za ažuriranja bez odgovarajućeg pristanka ili zaštitnih mjera može kršiti načela privatnosti.
8. **Povlačenje iz upotrebe (eng. *retirement*):** Podaci povezani s modelom i njegovim operacijama arhiviraju se ili brišu. Nepravilno brisanje osobnih podataka tijekom stavljanja izvan pogona može dovesti do dugoročnih ranjivosti privatnosti.

Tijekom cijelog životnog ciklusa UI sustava važno je razmotriti kako različite vrste osobnih podataka mogu biti uključene u svaku fazu. Ovisno o fazi, osobni podaci mogu se prikupljati, obrađivati, izlagati ili transformirati na različite načine. Prepoznavanje te varijabilnosti ključno je za provedbu učinkovitih mjera za zaštitu privatnosti i podataka.

[130] Testiranje, evaluacija, validacija i verifikacija (TEVV) trajan je proces koji se odvija tijekom cijelog životnog ciklusa umjetne inteligencije kako bi se osiguralo da sustav ispunjava predviđene zahtjeve, da pouzdano funkcionira i da je usklađen sa standardima sigurnosti i usklađenosti.



Slika 8. Na slici je prikazano kako se različite vrste osobnih podataka mogu pojaviti u različitim fazama životnog ciklusa umjetne inteligencije.

Protok podataka i rizici privatnosti po modelu LLM usluge

Uobičajeno je naići na pojmove poput zatvorenih modela, otvorenih modela, zatvorenih težina i otvorenih težina u kontekstu LLM-ova. Razumijevanje ovih pojmova ključno je za procjenu rizika povezanih s različitim strategijama objave modela.

Zatvoreni modeli vlasnički su modeli koji ne omogućuju javni pristup svojim ponderima ili izvornom kodu te je interakcija s modelom ograničena, za što je obično potreban API ili pretplata, dok su **otvoreni modeli u potpunosti** javno dostupni (dostupni su ponderi, potpuni kod, podaci o treniranju i druga dokumentacija) ili **djelomično** (nije sve dostupno, obično podaci o treniranju; ili je dostupno na temelju licenci). Slično tome, **zatvorene težine (eng. *closed weights*)** označavaju zaštićene modele čiji istrenirani parametri nisu objavljeni, dok **otvorene težine (eng. *open weights*)** opisuju modele s javno dostupnim parametrima, omogućujući pregled, fino podešavanje ili integraciju u druge sustave.
Također je važno razlikovati pojam otvorenog modela od **modela otvorenog koda**. Ova klasifikacija modela kao **„otvorenog koda"** zahtijeva njegovo objavljivanje pod **licencom otvorenog koda**, koja pravno svakome daje slobodu **korištenja, proučavanja, izmjene i distribucije** modela u bilo koju svrhu[131].

Term	Rizici za privatnost
Zatvoreni modeli & zatvorene težine	Često minimalna vanjska transparentnost. Korisnici se u potpunosti oslanjaju na mjere zaštite privatnosti pružatelja usluga, zbog čega je teško neovisno provjeriti usklađenost s propisima o zaštiti podataka.
Otvoreni modeli & otvorene težine	Rizik od izloženosti osobnih podataka i povreda sigurnosti ako podaci za osposobljavanje sadržavaju osjetljiv ili štetan sadržaj. Djelomičan pristup može spriječiti potpunu kontrolu podataka za treniranje modela i ranjivosti privatnosti.
Otvoreni kod	Modeli otvorenog koda dijele iste rizike za privatnost kao otvoreni modeli i modeli s otvorenom težinom. Iako se otvorenim kodom potiču transparentnost i inovacije, njime se povećavaju i rizici jer izmjene mogu dovesti do sigurnosnih slabih točaka ili ukloniti ugrađene sigurnosne mjere.

LLM-ovi su uglavnom dostupni putem sljedećih modela usluga:

[131] Međunarodno izvješće o sigurnosti umjetne inteligencije za 2025., str. 150. https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

1. **LLM kao usluga:** Ovaj model usluge omogućuje pristup LLM-ovima putem API-ja smještenih na platformi u oblaku. Korisnici mogu slati ulazne podatke i primati izlazne podatke bez izravnog pristupa osnovnoj arhitekturi ili ponderima modela.
   Na temelju ovog modela usluga obično možemo pronaći različite dostupne varijacije LLM modela:
   - Zatvoreni modeli sa zatvorenim utezima u kojima dobavljač trenira model i zadržava kontrolu nad utezima i podacima, nudeći pristup putem API-ja. Taj pristup osigurava jednostavnost uporabe, ali zahtijeva protok korisničkih podataka kroz sustave pružatelja usluga. Primjer: OpenAI GPT-4 API[132]
   - Prilagodljivi zatvoreni utezi u kojima subjekti za uvođenje mogu prilagoditi model s pomoću vlastitih podataka u kontroliranom okruženju, iako temeljni utezi ostaju nedostupni za prilagodbu uravnoteženja sa sigurnošću. Primjer: Azure OpenAI servis[133]
   - Otvoreni ponderi u kojima neki pružatelji subjektima za uvođenje odobravaju potpun ili djelomičan pristup arhitekturi radi veće transparentnosti i fleksibilnosti putem platforme ili API-ja.[134] Primjer: Hugging Face modeli u AWS Bedrock[135]

2. LLM kao gotov proizvod (eng. *LLM 'off-the-shelf'*): U ovom modelu usluge *deployer (*onaj koji implementira sustav*)* može prilagoditi utege i fino podesiti model. To se ponekad događa putem platformi kao što su Microsoft Azure i AWS na kojima deployer može odabrati model i s njim razviti vlastito rješenje. Također se obično koristi s modelima otvorene težine, kao što su LLaMA ili BLOOM. Dok LLM kao usluga obično uključuje interakciju temeljenu na API-ju bez vlasništva modela, usluga LLM-a „izvan sustava" naglašava veću kontrolu razvojnih programera i i *deployera*. Razlika leži u ovoj razini kontrole i pristupa koji se pruža, na primjer, u Hugging Face modelima koji se mogu preuzeti lokalno.

3. Vlastito razvijeni LLM (eng. *self-developed LLM):* U ovom modelu organizacije razvijaju i implementiraju LLM-ove na vlastitu infrastrukturu, održavajući potpunu kontrolu nad interakcijom podataka i modela. Iako ova opcija može ponuditi vežu razinu zaštite osobnih podataka, ovaj model usluge zahtijeva značajne računalne resurse i stručnost.

Svaki od tri modela usluga ima poseban protok podataka. Iako postoje sličnosti među modelima, svaka faza, od korisničkog unosa do generiranja izlaznih podataka, predstavlja jedinstvene rizike koji mogu utjecati na privatnost i zaštitu podataka korisnika. U ovom odjeljku prvo ćemo ispitati protok podataka u LLM-u kao servisnom rješenju, nakon čega slijedi analiza ključnih razlika u protoku podataka kada se koristi LLM „standardni" model i samorazvijeni LLM sustav.

*Imajte na umu da se u ovom odjeljku pojmovi "dobavljač"[136] i "subjekt za uvođenje"[137] upotrebljavaju kako je definirano u Aktu o umjetnoj inteligenciji, pri čemu dobavljač upućuje na subjekt koji razvija i nudi UI sustav, a subjekt za uvođenje odnosi se na subjekt koji primjenjuje i upravlja sustavom za krajnje korisnike.

[132] https://openai.com/api/
[133] https://azure.microsoft.com/en-us/services/cognitive-services/openai-service/
[134] https://en.wikipedia.org/wiki/API
[135] https://huggingface.co/blog/bedrock-marketplace
[136] „dobavljač" znači fizička ili pravna osoba, tijelo javne vlasti, agencija ili drugo tijelo koje razvija UI sustav ili UI model opće namjene ili koje ima razvijen UI sustav ili UI model opće namjene i stavlja ga na tržište ili u uporabu pod vlastitim imenom ili žigom, uz plaćanje ili besplatno; (članak 3. stavak 3. Akta o umjetnoj inteligenciji)
[137] „subjekt koji primjenjuje sustav" znači fizička ili pravna osoba, tijelo javne vlasti, agencija ili drugo tijelo koje upotrebljava UI sustav u okviru svoje nadležnosti, osim ako se UI sustav upotrebljava u osobnoj neprofesionalnoj djelatnosti; (članak 3. stavak 4. Akta o umjetnoj inteligenciji)

1. Protok podataka u LLM-u kao usluzi (eng. LLM as a Service System)

U našem primjeru, korisnik stupa u interakciju s LLM aplikacijom koja je smještena online od strane davatelja usluga. Taj protok podataka usmjeren je isključivo na faze koje su uključene tijekom interakcije korisnika s uslugom. Priprema modela, uvođenje i integracija od strane pružatelja usluga izvan su područja primjene jer će se dodatno ispitati u primjeru vlastito razvijenog LLM sustava. Važno je napomenuti da će svaki slučaj uporabe imati poseban protok podataka ovisno o svojim jedinstvenim zahtjevima i kontekstu, a primjeri navedeni u ovom odjeljku trebali su generički prikazi.
U scenariju LLM-a kao usluge mogli bismo pronaći sljedeće opće faze protoka podataka:

- **Unos korisnika (input):**

Postupak započinje s korisnikom koji šalje unos, kao što je upit ili naredba. To se može unijeti putem internetskog sučelja, mobilne aplikacije ili drugih alata koje pruža pružatelj LLM-a.

- **Davatelj sučelja & API:**

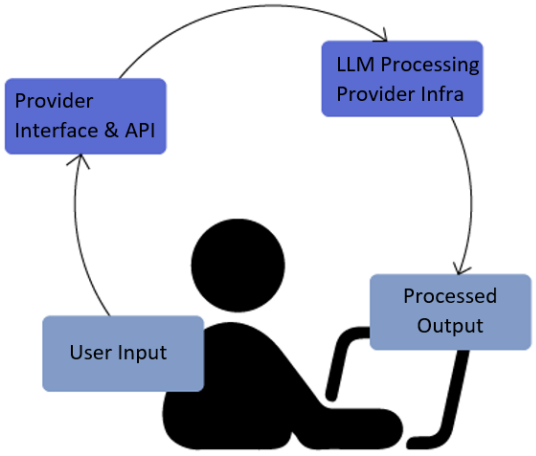
Unos se šalje putem sučelja ili aplikacije kojom upravlja pružatelj usluga (npr. internetske stranice, aplikacije ili prozora chatbota ugrađenog na internetsku stranicu). Ovo sučelje osigurava da se ulazni podaci formatiraju na odgovarajući način i sigurno prenose na infrastrukturu LLM-a.

- **Obrada LLM-a na infrastrukturi pružatelja usluga:**

API prima ulazne podatke i usmjerava ih prema LLM modelu smještenom na infrastrukturi pružatelja usluga.
LLM obrađuje ulazne podatke koristeći svoje istrenirane parametre (utege) kako bi generirao relevantan odgovor. To može uključivati korake kao što su tokenizacija, razumijevanje konteksta, rasuđivanje i generiranje teksta. Model generira odgovor.
* Evidentiranje: Ponuđač može zabilježiti korisnički unos (upit) zajedno s generiranim odgovorom kako bi analizirao interakciju i identificirao pogreške sustava ili nedostatke u kvaliteti odgovora.
Podaci bi se mogli uključiti i u skup podataka za osposobljavanje kako bi se poboljšala sposobnost modela da obrađuje slične upite u budućnosti. U tom se slučaju često primjenjuju tehnike anonimizacije i filtriranja.

- **Output (izlazni rezultat):**

Generirani izlaz vraća se korisniku putem sučelja davatelja usluga. Odgovor je obično u formatu spremnom za prikaz ili integraciju, kao što su tekst, prijedlozi ili upotrebljivi podaci.



Slika 9. Protok podataka u LLM-u kao servisnom

Pitanja zaštite osobnih podataka u ovom protoku podataka

U sljedećoj tablici istaknuti su potencijalni rizici za privatnost i zaštitu podataka te preporučene mjere ublažavanja tih rizika.

Faze	Mogući rizici & Protumjere
Unos korisnika	**Rizici:** **Osjetljivo otkrivanje podataka:** Korisnici mogu nesvjesno ili nenamjerno unijeti osjetljive osobne podatke, kao što su imena, adrese, financijske informacije ili medicinski detalji. **Neovlašteni pristup:** Ako web-sučelje, aplikacija, baze podataka ili alat za unos nemaju pouzdane kontrole pristupa, neovlaštene osobe[138] mogu dobiti pristup korisničkim računima ili sustavima, što im omogućuje pregled prethodno poslanih podataka ili upita. **Nedostatak transparentnosti:** Korisnici možda nisu u potpunosti svjesni kako će pružatelji usluga koristiti, zadržavati ili dijeliti njihove podatke. **Neprijateljski napadi:** (Zlonamjerni) korisnik može izraditi ulazne podatke osmišljene za manipuliranje ponašanjem LLM-a ili zaobilaženje njegove predviđene funkcionalnosti, kao što je ubrizgavanje neovlaštenih uputa u upite (eng. *prompt injection attack*), ili korisnici mogu pokušati zaobići sigurnosna ograničenja[139] nametnuta modelu izradom posebnih ulaznih podataka (eng. *jailbreaking attempt*).[140] **Protumjere za pružatelje:** - Provesti jasne upute za korisnike i ograničenja unosa, kao što su filtri ili upozorenja kako bi se obeshrabrilo unošenje osobnih podataka. Upotrijebite automatizirane mehanizme za otkrivanje[141][142] kako biste označili ili anonimizirali osjetljive informacije prije njihove obrade ili prijave. - Šifriranje korisničkih ulaza i izlaza tijekom prijenosa i pohrane kako bi se osjetljivi podaci zaštitili od neovlaštenog pristupa. Nedovoljna enkripcija može izložiti korisničke upite i podatke potencijalnim kršenjima, posebno tijekom faze unosa korisnika. Osigurati šifriranje u prijenosu pomoću robusnih protokola (npr. TLS) i šifriranje u mirovanju za pohranjene podatke. Osim toga, uvesti prakse odvajanja podataka kako bi se izolirali korisnički podaci, čime bi se neovlaštenim pojedincima onemogućio pristup višestrukim računima ili skupovima podataka ili ugrožavanje tih računa ili skupova podataka. Još jedno ublažavanje za sprečavanje neovlaštenog pristupa jest provedba praksi sigurne lozinke na temelju najnovijih[143] preporuka NIST-a[144] i ENISA-e: zahtijevaju minimalnu duljinu lozinke od 8 znakova, ažuriraju lozinke ako su ugrožene ili zaboravljene, provode upotrebu višestruke autentifikacije (MFA), osiguravaju da se lozinke znatno razlikuju od prethodnih, provjeravaju lozinke u odnosu na crne liste, provode pravila za zaključavanje računa, prate neuspjele pokušaje prijave, obeshrabruju upotrebu savjeta za lozinke i sigurno pohranjuju lozinke pomoću tehnika raspršivanja i soljenja s robusnim algoritmima kao što su bcrypt, Argon2 ili PBKDF2. - Jasnim i lako dostupnim pravilima o privatnosti informirajte korisnike o tome kako će se njihovi podaci upotrebljavati, zadržavati i obrađivati.

[138] https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak
[139] https://cybersecuritynews.com/new-jailbreak-techniques-expose-deepseek-llm-vulnerabilities/
[140] https://learnprompting.org/blog/injection_jailbreaking
[141] https://www.nightfall.ai/ai-security-101/data-leakage-prevention-dlp-for-llms
[142] Neki od korištenih alata su Google Cloud DLP, Microsoft Presidio, OpenAI moderatorski API, Hugging Face Fine-Tuned NER modeli i spaCy (poveznice dostupne u odjeljku 10)
[143] https://enisa.europa.eu/sites/default/files/all_files/ENISA%20guidelines%20for%20passwords.pdf
[144] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-63-3.pdf

	- Iako trenutno ne postoje mjere za[145] zaštitu od *prompt injection attack*-a, i *jailbreaking attempt*,[146][147]neke od najčešćih najboljih praksi uključuju validacija ulaznih podataka, filtriranje radi otkrivanja zlonamjernih uzoraka, praćenje LLM-ova zbog abnormalnog ponašanja ulaznih podataka, provedba ograničenja brzine, strukturiranje upita[148]i ograničavanje količine teksta koju korisnik može unijeti.[149] **Protumjere za subjekte koji primjenjuju sustav:** - Ograničite količinu osjetljivih podataka i vodite korisnike kako bi se izbjeglo dijeljenje nepotrebnih osobnih podataka s pomoću jasnih uputa, osposobljavanja i upozorenja. Surađujte s pružateljima usluga kako biste osigurali da se pridržavaju propisa o zaštiti podataka i da ne zadržavaju ili zloupotrebljavaju (osjetljive) ulazne podatke. - Zahtijevati sigurnu autentifikaciju korisnika kako bi se ograničio pristup ulaznom sučelju i zaštitili podaci o sesiji. Kako je istaknuto u mjerama ublažavanja pružatelja, subjekti za uvođenje trebali bi provoditi i prakse sigurne lozinke na temelju najnovijih preporuka NIST-a i ENISA-e,[150] poticati korisnike da usvoje upravitelje lozinki[151] i podizati svijest o sigurnim praksama i internim politikama za lozinke među korisnicima, kao što su zaposlenici. - jasno priopćiti korisnicima kako se postupa s njihovim podacima i kako se oni obrađuju u svakoj fazi protoka podataka. To se može učiniti putem (internih) pravila o privatnosti, uputa, upozorenja ili izjava o ograničenju odgovornosti u korisničkom sučelju. - Kako bi se ublažili neprijateljski napadi, može se provesti nekoliko mjera kao što su dodavanje sloja za dezinfekciju i filtriranje ulaznih podataka, praćenje i bilježenje korisničkih upita radi otkrivanja neuobičajenih obrazaca te uključivanje slojeva za naknadnu obradu radi provjere valjanosti izlaznih podataka. Osim toga, educiranje korisnika o pravilnoj uporabi može pomoći u smanjenju vjerojatnosti nenamjernih unosa koji mogu dovesti do štetnih ishoda.
Davatelj sučelje & API	**Rizici:** **Presretanje podataka:** Nedovoljna enkripcija tijekom prijenosa podataka na poslužitelje pružatelja usluga može izložiti unos presretanju koje provode treće strane. **Zlouporaba API-ja:** Ako pristup API-ju nije ograničen i zaštićen, napadači bi mogli iskoristiti API za presretanje podataka ili manipuliranje njima. Napadači bi također mogli preopteretiti API prekomjernim prometom kako bi poremetili njegovu dostupnost (napadi na uskraćivanje usluge). **Slabosti sučelja:** Slabosti sučelja odnose se na slabosti u korisničkom sučelju pružatelja usluga koje mogu izložiti korisničke podatke zlonamjernim akterima. Te ranjivosti mogu proizlaziti iz tehničkih nedostataka (npr. nedovoljna provjera valjanosti ulaznih podataka, pogrešne konfiguracije krajnjih točaka API-ja) ili taktika socijalnog inženjeringa kao što je phishing. Na primjer, napadači bi mogli izraditi lažne verzije sučelja chatbota (npr. repliciranje dizajna i brendiranja) kako bi prevarili korisnike da unesu osjetljive informacije kao što su vjerodajnice, podaci o plaćanju ili osobni podaci. Zlonamjerni akteri mogli bi razviti obmanjujuće aplikacije koje tvrde da su legitimne integracije s API-jem pružatelja usluga, čime bi krajnje korisnike prevarili u dijeljenju osjetljivih podataka. **Protumjere za pružatelje:**

[145] https://www.ibm.com/think/insights/prevent-prompt-injection
[146] https://arxiv.org/abs/2411.07494
[147] https://arxiv.org/abs/2410.15236
[148] https://arxiv.org/abs/2402.06363
[149] https://platform.openai.com/docs/guides/safety-best-practices#constrain-user-input-and-limit-output-tokens
[150] https://community.trustcloud.ai/article/nist-password-guidelines-2025-15-pravila za praćenje/
[151] https://en.wikipedia.org/wiki/Password_manager

	- Implementirati *end-to-end* enkripciju na kraj za sve prijenose podataka, redovito ažurirati protokole šifriranja i primjenjivati prakse sigurnog upravljanja ključevima. - Provedba pouzdane autentifikacije (npr. ključevi API-ja, OAuth), provedba ograničenja stope,[152]praćenje sumnjivih aktivnosti (otkrivanje nepravilnosti) i redovita revizija sigurnosti API-ja. - Obavljajte redovita sigurnosna testiranja, primijenite provjeru valjanosti ulaznih podataka kako biste spriječili napade i primijenite robusne kontrole upravljanja sesijama. Kad je riječ o phishingu, i pružatelj i subjekt za uvođenje imaju ulogu u rješavanju tog rizika. Pružatelj bi trebao uvesti zaštite na razini platforme, kao što su zaštita autentičnosti svojeg sučelja (npr. mjere protiv krađe identiteta, zaštita robne marke, sigurni API-ji), praćenje sumnjivih aktivnosti i osiguravanje alata koji će subjektima za uvođenje pomoći u otkrivanju i sprečavanju zlouporabe. **Protumjere za subjekte koji primjenjuju sustav:** - Ako subjekt za uvođenje upotrebljava samo dobavljačevo sučelje, njegova je odgovornost ograničena na sigurno upravljanje vjerodajnicama za pristup i poštovanje politika postupanja s podacima; međutim, ako subjekt za uvođenje integrira dobavljačev API u vlastite sustave, dodatno je odgovoran za osiguravanje integracije, uključujući šifriranje, praćenje i zaštitu podataka u prijenosu. I pružatelj usluga i subjekt za uvođenje trebali bi osmisliti, razviti, uvesti i testirati aplikacije i API-je u skladu s vodećim industrijskim standardima (npr. OWASP za internetske aplikacije)[153]i poštovati primjenjive pravne, zakonske ili regulatorne obveze usklađenosti. - Subjekt za uvođenje trebao bi educirati zaposlenike i krajnje korisnike o novim tehnikama phishinga, kao što su lažna sučelja, obmanjujuća elektronička pošta ili lažne integracije, koje bi pojedince mogle navesti na otkrivanje osjetljivih informacija. Edukacija bi se trebala usredotočiti na prepoznavanje sumnjivih ponašanja i provjeru legitimnosti komunikacija i sučelja.
Obrada LLM-a na infrastrukturi pružatelja usluga	**Rizici:** **Rizici za izvođenje zaključaka modela (eng. *model inference risks*):** Tijekom obrade model može nenamjerno izvoditi zaključke o osjetljivim ili neprimjerenim rezultatima na temelju podataka za osposobljavanje ili pruženih ulaznih podataka. **(Ne)planirano bilježenje podataka:** Pružatelji usluga mogu bilježiti korisničke ulazne upite i izlaze za ispravljanje pogrešaka ili poboljšanje modela, potencijalno pohranjujući osjetljive podatke[154] bez izričitog pristanka korisnika. Ako su upiti prijavljenih korisnika uključeni u podatke o osposobljavanju, u slučaju kontradiktornog napada napadači mogu uvesti zlonamjerni ili obmanjujući sadržaj kako bi manipulirali budućim rezultatima modela (eng. *data poisoning attack*).[155] **Neuspjesi anonimizacije:** Neodgovarajuće tehnike anonimizacije ili filtriranja mogle bi dovesti do uključivanja korisničkih podataka koji se mogu identificirati u skupove podataka za učenje modela, što izaziva zabrinutost u pogledu privatnosti. **Neovlašteni pristup zapisnicima:** Evidencijama koje sadrže korisničke unose i izlaze može pristupiti neovlašteno osoblje ili se mogu iskoristiti u slučaju povrede podataka. Rizici agregiranja**podataka:** Ako se zapisi agregiraju tijekom vremena, mogli bi formirati sveobuhvatan skup podataka koji može otkriti obrasce o pojedincima, organizacijama ili drugim osjetljivim aktivnostima. **Izloženost prema trećim osobama:** Ako se pružatelj usluga oslanja na vanjsku infrastrukturu u oblaku ili alate trećih strana za obradu LLM-a, postoji dodatni rizik od

[152] https://llm-engine.scale.com/guides/rate_limits/
[153] https://owasp.org/www-project-top-ten/
[154] Pohranjeni podaci mogli bi biti osjetljivi podaci kao što su brojevi kreditnih kartica ili posebna kategorija podataka kao što su zdravstveni podaci (članak 9. Opće uredbe o zaštiti podataka).
[155] https://www.riskinsight-wavestone.com/en/2024/10/data-poisoning-a-threat-to-llms-integrity-and-security/

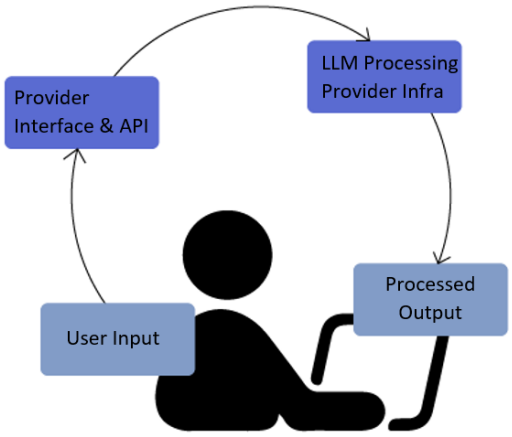
	izloženosti podacima zbog tih ovisnosti. Te ovisnosti uključuju vanjske sustave, koji mogu imati vlastite ranjivosti. **Nedostatak politika zadržavanja podataka:** Pružatelj usluga mogao bi pohraniti podatke na neodređeno vrijeme bez postojanja pravila o zadržavanju. **Protumjere za pružatelje:** - Provesti stroge mehanizme filtriranja sadržaja i postupke ljudskog pregleda kako bi se označili osjetljivi ili neprimjereni rezultati. - Smanjite bilježenje podataka, prikupite samo potrebne informacije i osigurajte da imate odgovarajuću pravnu osnovu za sve obrađene podatke. Upotrijebite pouzdane izvore za podatke za učenje i potvrdite njihovu kvalitetu. Sanitizirajte i unaprijed obradite podatke o treniranju kako biste uklonili ranjivosti ili pristranosti. Redovito pregledavati i revidirati podatke o osposobljavanju i prilagođavati postupke za probleme ili manipulacije. Uvesti sustave praćenja i uzbunjivanja kako bi se otkrilo neobično ponašanje ili potencijalno *data poisoning attack*[156]. - Primijeniti robusne tehnike anonimizacije, redovito ih testirati na učinkovitost i koristiti automatizirane alate za identifikaciju i uklanjanje podataka koji se mogu identificirati prije uporabe u treningu. - Provoditi stroge kontrole pristupa, šifrirati podatke zapisnika i pratiti zapise o pristupu za sumnjive aktivnosti kako bi se spriječile povrede. - Pružatelji usluga moraju provoditi robustan program upravljanja rizicima treće strane, pridržavajući se najpoznatijih okvira[157] kako bi osigurali sigurno okruženje. Ključne mjere uključuju provedbu temeljitih procjena dobavljača, osiguravanje usklađenosti sa sigurnosnim standardima, zahtijevanje snažnog šifriranja podataka tijekom prijenosa i pohrane, provođenje sigurnosnih revizija, provedbu praćenja u stvarnom vremenu i planova odgovora na incidente prilagođenih ovisnostima trećih strana. Pružatelji bi također trebali uvesti zaštitu od prijetnji kao što su napadi u skladu s DoS-om/DDoS-om, koji mogu poremetiti poslovanje i izložiti sustave daljnjim rizicima. - Jasno definirati politike zadržavanja, uskladiti ih s pravnim zahtjevima i, ako je moguće, pružiti korisnicima mogućnosti brisanja njihovih podataka.
eng. Processed Output	**Rizici:** **Netočni ili osjetljivi odgovori:** Model može generirati izlazne rezultate kojima se otkrivaju nenamjerne osjetljive informacije ili pružaju netočne ili obmanjujuće informacije (halucinacije),[158]što dovodi do štete ili pogrešnih informacija. **Rizici ponovne identifikacije:** Rezultati bi mogli nenamjerno otkriti informacije o upitu ili kontekstu korisnika koje se mogu povezati s njima. **Zlouporaba**[159]**proizvodnje:** Korisnici ili treće strane mogu zloupotrijebiti generirani rezultat. **Protumjere za pružatelje:** - Implementirajte filtre za naknadnu obradu kako biste otkrili i uklonili osjetljiv ili netočan sadržaj te redovito prekvalificirali model pomoću ažuriranih i provjerenih skupova podataka kako biste poboljšali točnost odgovora. Uvođenje izjava o ograničenju odgovornosti kako bi se istaknula potencijalna ograničenja odgovora stvorenih umjetnom inteligencijom.

[156] https://owasp.org/www-project-top-10-for-large-language-model-applications/Archive/0_1_vulns/Training_Data_Poisoning.html
[157] https://www.cisecurity.org/controls/cis-controls-list
[158] https://en.wikipedia.org/wiki/Hallucination_(umjetna_inteligencija)
[159] https://genai.owasp.org/llmrisk/llm052025-improper-output-handling/

	- Primijenite tehnike očuvanja privatnosti kako biste lakše uklonili osjetljive identifikatore u izlazima i smanjili uključivanje nepotrebnih kontekstualnih detalja u generirane odgovore. - Osigurati jasne politike upotrebe, educirati korisnike o etičkoj upotrebi izlaznih proizvoda i provesti mehanizme za otkrivanje i sprečavanje zlouporabe generiranog sadržaja ako je to izvedivo. **Protumjere za subjekte koji primjenjuju sustav:** - Kad je riječ o kritičnim primjenama, osigurati da ljudi pregledaju generirane rezultate prije provedbe ili širenja. - Educirati krajnje korisnike o etičkoj i odgovarajućoj upotrebi rezultata, uključujući izbjegavanje prekomjernog oslanjanja na model za donošenje ključnih odluka ili odluka s velikim ulogom bez provjere. - Sigurno pohranjivati izlaze i ograničiti pristup samo ovlaštenom osoblju ili sustavima.

2. Protok podataka u 'off-the-shelf' LLM sustavu

Najčešći primjer upotrebe ovog modela usluga uključuje organizacije koje koriste prethodno uvježbani model s platforme za razvoj i uvođenje vlastitog sustava umjetne inteligencije. Nakon što sustav umjetne inteligencije postane operativan, protok podataka vrlo je sličan protoku LLM-a kao usluge, posebno tijekom interakcije korisnika i generiranja izlaznih podataka.



Slika 10. Protok podataka u lako dostupnom LLM

Međutim, nekoliko ključnih razlika i ograničenja razlikuje ove modele:

- **Uloge i odgovornosti:**

Organizacije koje razvijaju LLM sustav primjenom „standardnog" modela mogu se smatrati pružateljima usluga,[160] posebno ako namjeravaju staviti sustav na tržište kako bi ga upotrebljavali drugi (subjekti koji primjenjuju njihov sustav i krajnji korisnici). Time se uvodi dodatna razina odgovornosti za postupanje s podacima, sigurnost i usklađenost s propisima o privatnosti. Organizacija također može razviti UI sustav za vlastitu internu uporabu.

- **Hosting i obrada:**

U 'off-the-shelf' LLM sustavu pružatelj smješta model na svoju infrastrukturu ili okruženje u oblaku treće strane po vlastitom izboru. To je u suprotnosti s modelom LLM kao usluge, gdje hostingom i obradom u

[160] U skladu s člankom 25. Akta o umjetnoj inteligenciji subjekt koji primjenjuje visokorizični UI sustav postaje dobavljač ako znatno izmijeni postojeći UI sustav, među ostalim prilagodbom prethodno poučenog modela za nove primjene. U takvim slučajevima subjekt za uvođenje preuzima odgovornosti dobavljača u skladu s Aktom o umjetnoj inteligenciji.

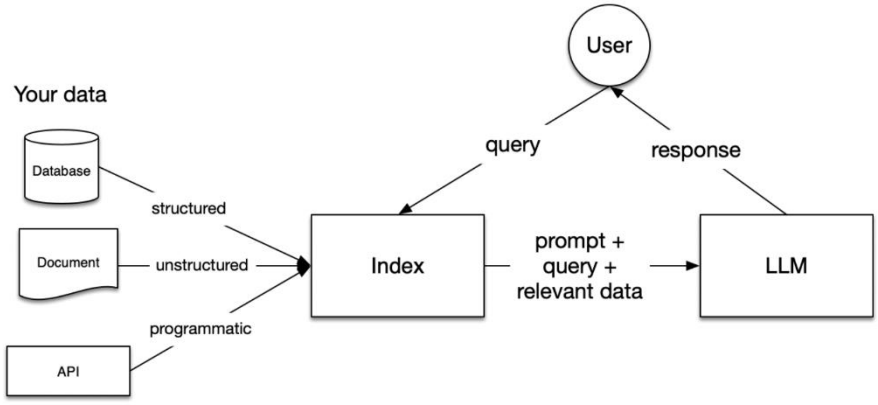
potpunosti upravlja izvorni davatelj modela. Novi pružatelj usluga sada je odgovoran za sve aspekte integracije, održavanja i sigurnosti sustava.

- **Prilagodba i trening:**

Značajna je razlika u tome što je početni trening i prilagodbu modela proveo izvorni pružatelj usluga, što može dovesti do rizika i ograničenja:

- Novi pružatelj često nema nadzor ili znanje o sadržaju skupa podataka koji se upotrebljava tijekom početnog treninga modela, što može dovesti do pristranosti, netočnosti ili nepoznatih rizika za privatnost i zaštitu osobnih podataka[161]
- Novi pružatelj usluga i dalje ovisi o izvornom pružatelju usluga za ažuriranja ili ispravke grešaka u arhitekturi modela, što može odgoditi ključna poboljšanja ili ispravke.
- Fino podešavanje (eng. *fine tuning*) može biti ograničeno mogućnostima standardnog modela. Novi pružatelji usluga mogli bi samo prilagoditi određene parametre ili dodati nove slojeve umjesto da u potpunosti prekvalificiraju model, čime bi se ograničila njegova prilagodljivost za vrlo specifične slučajeve upotrebe.
- U takvim slučajevima, retrieval-augmented generation (RAG) je najčešće korištena alternativa. Umjesto ugrađivanja znanja specifičnog za određeno područje u sam model, RAG povezuje model s vanjskom bazom znanja i dohvaća relevantne dokumente tijekom izvođenja kako bi utemeljio svoje odgovore. To omogućuje dinamične, točne i ažurirane odgovore bez izmjene osnovnog modela, što je ključna prednost za područja s promjenjivim informacijama ili regulatornim zahtjevima.
  Sličan pristup je cache-augmented generation (CAG) [162] koji može smanjiti latenciju, smanjiti troškove računanja i osigurati dosljednost u odgovorima u ponovljenim interakcijama, ali to je manje praktično za velike skupove podataka koji se često ažuriraju.

  Na slici u nastavku prikazano je kako RAG[163] funkcionira: upit korisnika prvo se poboljšava relevantnim informacijama koje se dohvaćaju iz vanjske baze podataka, a taj se obogaćeni unos zatim šalje jezičnom modelu kako bi se dobio točniji i utemeljeniji odgovor.



Slika 11. Dijagram RAG-a – Otvorena knjižica o

Neki od uobičajenih rizika za privatnost pri korištenju RAG-a su:

[161] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.
[162] https://developer.ibm.com/articles/awb-cache-rag-efficiency-speed-ai/
[163] https://cookbook.openai.com/examples/evaluation/evaluate_rag_with_llamaindex

- Nesigurno bilježenje ili predmemoriranje: Upiti korisnika i preuzeti dokumenti mogu se pohraniti nesigurno, povećavajući rizik od neovlaštenog pristupa ili curenja podataka.
- Obrada podataka koju provodi treća strana: Ako sustav za dohvat upotrebljava vanjske API-je ili usluge, upiti korisnika mogu se slati trećim stranama, gdje se mogu evidentirati, pratiti ili pohranjivati bez pristanka korisnika.
- Izloženost osjetljivim podacima: Model može dohvatiti osobne podatke ili povjerljive informacije ako su pohranjene u bazi znanja.

3. Protok podataka u vlasničkim LLM sustavima (eng. *Self-developed LLM System*)

U samostalno razvijenom LLM sustavu, organizacija preuzima punu odgovornost za projektiranje, obuku, a u nekim slučajevima i implementaciju modela. Ovaj pristup pruža maksimalnu kontrolu nad LLM modelom, ali također uvodi jedinstvene izazove u protoku podataka.

Budući da smo u prethodnom odjeljku već istražili primjer rizika za privatnost u protoku podataka životnog ciklusa umjetne inteligencije, ovdje ćemo zauzeti općenitiji pristup, usredotočujući se na neke od važnih faza tog modela usluge. Opće faze protoka podataka mogu biti sljedeće:

- **Prikupljanje i priprema skupova podataka:**

Organizacija prikuplja i čuva[164] velike skupove podataka za obuku LLM-a.

- **Osposobljavanje prema modelu:**

Faza osposobljavanja uključuje upotrebu prikupljenog skupa podataka za razvoj LLM-a. To obično zahtijeva značajne računalne resurse i specijaliziranu infrastrukturu, kao što su GPU-ovi visokih performansi ili distribuirani računalni sustavi. Prije implementacije, model prolazi strogu evaluaciju i testiranje pomoću zasebnih validacijskih i testnih skupova podataka kako bi se osigurala njegova točnost, pouzdanost i usklađenost s predviđenim slučajevima uporabe.

- **Fino podešavanje:**

Nakon početnog osposobljavanja model se može prilagoditi upotrebom dodatnih skupova podataka kako bi se njegove sposobnosti specijalizirale za određene zadatke ili domene.

- ***Deployment:***

Osposobljeni i fino podešeni LLM integriran je u infrastrukturu organizacije, čime je sučelje dostupno krajnjim korisnicima.

- **Unos korisnika (input):**

Krajnji korisnici stupaju u interakciju s primijenjenim UI sustavom slanjem ulaznih podataka putem sučelja kao što su aplikacija, chatbot ili prilagođeni API.

- **Davatelj sučelje & API:**

Unos se šalje putem sučelja ili aplikacije. Ovo sučelje osigurava da se ulazni podaci formatiraju na odgovarajući način i sigurno prenose na infrastrukturu LLM-a.

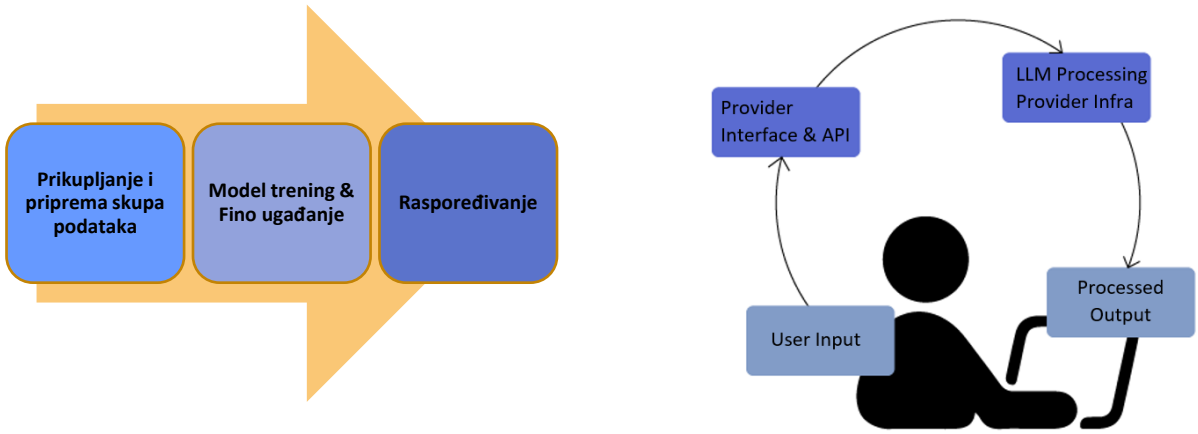
- **Obrada modela:**

Samorazvijeni LLM obrađuje korisničke unose lokalno ili na infrastrukturi organizacije (u oblaku), generirajući kontekstualno relevantne odgovore koristeći svoje istrenirane parametre.

- **Prerađena isporuka outputa (eng. processed output delivery):**

Obrađeni izlazi isporučuju se krajnjim korisnicima ili se integriraju u *downstream* sustave za upotrebljivu uporabu. Rezultati mogu uključivati tekstualne odgovore, uvide ili preporuke.

[164] https://atlan.com/data-curation-in-machine-learning/



Slika 12. Protok podataka u samorazvijenom LLM

Pitanja privatnosti u u vlasničkim LLM sustavima (eng. *Self-developed LLM System*)

Samostalni razvoj LLM sustava pruža značajnu kontrolu, ali također uvodi rizike privatnosti i zaštite podataka u svakoj fazi protoka podataka. U nastavku su navedeni neki od ključnih rizika i predložene mjere ublažavanja:

Faze	Mogući rizici & Protumjere
Prikupljanje i priprema skupova podataka	**Rizici:** **Osjetljivo uključivanje podataka:** Prikupljeni skupovi podataka mogu (nenamjerno) uključivati osobne ili osjetljive informacije. **Pravna neusklađenost:** Podaci bi se mogli prikupljati nezakonito kršeći propise o zaštiti podataka kao što je GDPR. **Pristranost i diskriminacija:** Skupovi podataka mogli bi odražavati društvene ili povijesne pristranosti, što bi dovelo do diskriminirajućih rezultata. ***Data poisoning*:** Zlonamjerni akteri mogu namjerno manipulirati skupovima podataka tijekom prikupljanja ili pripreme, uvodeći oštećene ili kontradiktorne podatke kako bi zavarali model tijekom treninga. **Protumjere za pružatelje:** - Primijeniti tehnike anonimizacije i pseudonimizacije[165] kako bi se rizici za privatnost sveli na najmanju moguću mjeru. - U određenim slučajevima upotrebe i nakon pažljivog odvagivanja mogućih prednosti i nedostataka,[166][167] stvaranje sintetičkih podataka upotrebom LLM-ova [168] moglo bi biti alternativa upotrebi stvarnih osobnih podataka. Međutim, sintetički podaci možda nisu uvijek prikladni[169]jer njihova kvaliteta i korisnost ovise o posebnim zahtjevima i kontekstu zahtjeva. - Osigurati da je prikupljanje podataka u skladu s propisima.

[165] ENISA, Smjernice o oblikovanju tehnologije u skladu s odredbama Opće uredbe o zaštiti podataka: https://www.enisa.europa.eu/sites/default/files/publications/Guidelines%20on%20shaping%20technology%20according%20to%20GDPR%20provisions.pdf
[166] https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-When-training-ai-systems
[167] https://proceedings.mlr.press/v202/van-breugel23a/van-breugel23a.pdf
[168] https://www.confident-ai.com/blog/the-definitive-guide-to-synthetic-data-generation-using-llms
[169] https://www.tmlt.io/resources/fundamental-trilemma-synthetic-data-generation

	- Redovito pregledavajte skupove podataka zbog pristranosti i osjetljivog sadržaja, uklanjajući sve problematične unose. - Provesti pouzdanu validaciju i praćenje podataka kako bi se otkrili i spriječili zlonamjerni ili oštećeni podaci. Upotrijebite pouzdane izvore podataka, primijenite automatizirane provjere za nepravilnosti i unakrsno potvrdite podatke iz više izvora.
Trening modela	**Rizici:** **Nezaštićeno okruženje za osposobljavanje:** Infrastruktura za osposobljavanje može biti izložena neovlaštenom pristupu, što bi moglo izložiti osjetljive podatke ili omogućiti zlonamjernim akterima da ugroze postupak osposobljavanja. **Prekomjerno dopunjavanje podataka:** Model može nenamjerno zapamtiti osjetljive informacije umjesto generaliziranja uzoraka. **Protumjere za pružatelje:** - Kibernetička sigurnost bi trebala slijediti slojevit pristup, provodeći višestruku obranu kako bi se spriječio neovlašteni pristup i ublažio njegov učinak. Protumjere koje bi se mogle provesti uključuju: upotreba sigurnih računalnih okruženja sa snažnim kontrolama pristupa tijekom osposobljavanja (npr. višefaktorska autentifikacija (MFA), upravljanje povlaštenim pristupom (PAM)[170]ili kontrole pristupa na temelju uloga (RBAC));[171] primjenom segmentacije mreže kako bi se izolirala infrastruktura za osposobljavanje od drugih sustava i smanjila površina za napad; praćenje i bilježenje pristupa kako bi se odmah otkrile neovlaštene aktivnosti i odgovorilo na njih; i upotreba šifriranja za podatke u mirovanju i u prijenosu kako bi se osigurali osjetljivi podaci za učenje. Još jedna mjera ublažavanja za smanjenje izloženosti podacima jest integracija tehnika diferencijalne privatnosti, kojima se dodaje buka podacima za učenje kako bi se spriječila ponovna identifikacija pojedinačnih točaka podataka, čak i ako je model ugrožen. Važno je procijeniti konkretan slučaj upotrebe kako bi se utvrdilo je li diferencijalna privatnost prikladna jer može utjecati na točnost modela u određenim scenarijima. - Procijenite model za preopterećenje i osigurajte da osjetljivi podaci nisu izloženi u izlazima.
Fino podešavanje (eng. *fine tuning*)	**Rizici:** **Izloženost vlasničkim ili osjetljivim podacima:** Fino podešavanje podataka može uključivati osjetljive ili vlasničke informacije, riskirajući curenje. **Rizici trećih strana:** Ako se za fino podešavanje upotrebljavaju vanjske platforme, osjetljivi podaci mogu biti izloženi dodatnim rizicima. **Protumjere za pružatelje:** - Šifriranje preciznih skupova podataka i ograničavanje pristupa ovlaštenom osoblju. - Koristite pouzdane platforme s robusnim jamstvima privatnosti za fino podešavanje. - Uključiti samo podatke koji su nužno potrebni za precizno podešavanje zadataka.
Deployment	**Rizici:** **Neovlašteni pristup:** Slabe kontrole pristupa mogle bi neovlaštenim stranama omogućiti interakciju s modelom ili pristup osnovnim sustavima. **Nesiguran hosting:** Hostiranje modela na nezaštićenom poslužitelju ili okruženju u oblaku moglo bi izložiti osjetljive podatke.

[170] https://en.wikipedia.org/wiki/Privileged_access_management
[171] https://en.wikipedia.org/wiki/Role-based_access_control

	Protumjere za pružatelje: - Provesti pouzdanu autentifikaciju i kontrole pristupa koje se temelje na ulogama za pristup modelu i sustavu. - Koristite okruženja u oblaku s jakim šifriranjem i praćenjem. - Periodično provjeravajte postoje li ranjivosti u konfiguracijama implementacije.
Unos korisnika	Vidjeti prethodnu tablicu o rizicima u protoku podataka LLM-a kao usluge, u kojoj je ta faza detaljno opisana. Primjenjive mjere ublažavanja prvenstveno se odnose na dobavljače, ali se proširuju i na subjekte za uvođenje u scenarijima u kojima razvojni programeri uvode i upotrebljavaju vlastite UI sustave.
Davatelj sučelje & API	Idem
Obrada LLM-a na infrastrukturi pružatelja usluga	Idem
Obrađena proizvodnja	Idem

4. Protok podataka u agentskim sustavima koji se temelje na LLM-u

Agenti umjetne inteligencije obuhvaćaju i faze protoka podataka o kojima se dosad raspravljalo, ali uvode dodatnu složenost zbog brojnih interakcija s drugim sustavima i aplikacijama. Ti agenti ne samo da obrađuju korisničke unose, već i surađuju s vanjskim aplikacijama kako bi dohvatili informacije, izvršili naredbe ili izvršili radnje. To se često čini putem funkcijskih poziva, pri čemu agent upotrebljava strukturirana sučelja (npr. API-je) za interakciju s vanjskim alatima. U nekim se slučajevima protokol za upravljanje kontekstom (CMP)[172] upotrebljava za upravljanje tim interakcijama i njihovo praćenje, čime se agentu osigurava pristup relevantnom kontekstu u više koraka ili alata. U ovom primjeru istražujemo protok podataka agenta umjetne inteligencije koji je u interakciji s dvije vanjske aplikacije, naglašavajući najčešće izazove u pogledu privatnosti i zaštite podataka koje te interakcije donose.

Pojednostavnjeni pregled najčešćih faza uključenih u taj protok podataka uključuje:

- **Unos korisnika:**

Postupak započinje tako da korisnik agentu umjetne inteligencije dostavi podatke kao što su upit, naredba ili opis zadatka (npr. „Rezerviraj mi let i hotel za moje putovanje u Amsterdam").
    - Akcije:
        - Ulaz se prikuplja putem korisničkog sučelja, kao što je chatbot ili glasovni asistent.
        - Predobrada se može provoditi lokalno radi čišćenja i standardizacije ulaznih podataka.
- **AI agent:**

Agent umjetne inteligencije upotrebljava integrirani LLM za razumijevanje i obradu korisničkih unosa. To uključuje raščlanjivanje zahtjeva i identificiranje radnji potrebnih za ispunjavanje zadatka.
    - Akcije:
        - Ulazni podatak je tokeniziran i interpretiran od strane LLM-a.
        - Agent odlučuje koje vanjske aplikacije kontaktirati i formulira upite ili naredbe za njih.
- **Interakcija s aplikacijom 1 (npr. sustav rezervacija letova):**

Agent šalje upit ili naredbu prvoj vanjskoj aplikaciji za dohvaćanje ili obradu informacija. Na primjer, može zatražiti dostupne letove na temelju korisnikovih preferencija u pogledu putovanja.
    - Akcije:
        - Podaci (npr. korisničke postavke) prenose se u vanjsku aplikaciju.

[172] https://www.anthropic.com/news/model-context-protocol

- Aplikacija obrađuje upit i vraća odgovor, kao što je popis dostupnih letova.
- Agent prima i obrađuje odgovor za integraciju u cjelokupni tijek rada.

➢ **Interakcija s aplikacijom 2 (npr. sustav rezervacija hotela):**

Agent se uključuje u drugu vanjsku aplikaciju kako bi dovršio drugi dio zadatka. Na primjer, može zatražiti opcije hotela na temelju odredišta i datuma putovanja.

- Akcije:
  - Podaci (npr. datumi putovanja) prenose se u aplikaciju.
  - Aplikacija daje odgovor, kao što su dostupni hoteli, koji obrađuje agent.

➢ **Zbrajanje odgovora:**

Agent umjetne inteligencije integrira odgovore iz obje aplikacije kako bi stvorio kohezivan rezultat. Na primjer, objedinjuje opcije leta i hotela u jedan izlaz za korisnika.

- Akcije:
  - Odgovori se potvrđuju i formatiraju radi jasnoće i relevantnosti.
  - Rješavaju se moguće pogreške ili sukobi (npr. preklapajući rasporedi).

➢ **Proizvodnja outputa:**

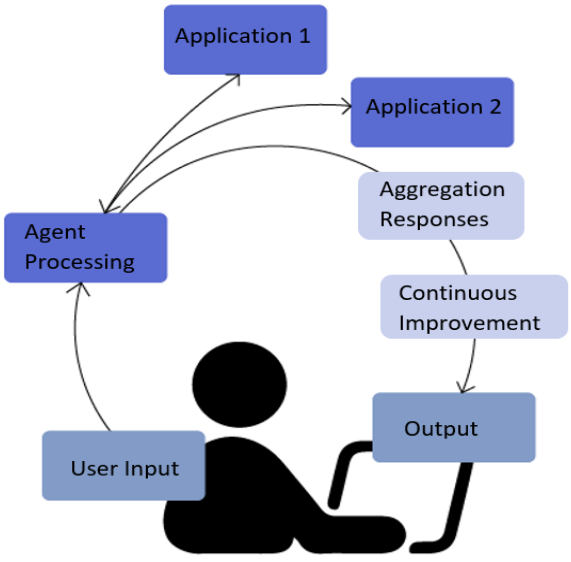
Agent korisniku dostavlja zbirni rezultat u formatu prilagođenom korisniku, kao što je sažetak opcija rezervacije ili provedivih preporuka.

- Akcije:
  - Izlaz se prikazuje putem korisničkog sučelja ili se prenosi u drugi sustav radi daljnjeg djelovanja.
  - Ako je potrebno, agent daje upute za daljnje postupanje kako bi poboljšao preferencije ili odabire korisnika.

➢ **Logiranje i stalno poboljšanje:**

Zapisi o interakciji mogu se privremeno pohraniti radi uklanjanja pogrešaka, poboljšanja sustava ili ponovnog osposobljavanja, ovisno o pravilima organizacije.

- Akcije:
  - Zapisi se analiziraju kako bi se optimizirala uspješnost agenta i poboljšalo korisničko iskustvo.



Slika 13. Protok podataka u agentskim sustavima

Pregled mogućih rizika i mjera ublažavanja

Agenti umjetne inteligencije predstavljaju znatan napredak u iskorištavanju dugotrajno nezaposlenih osoba za složene zadatke s više primjena. Njihov protok podataka uključuje tradicionalne faze koje se

vide u LLM sustavima, ali njihova integracija s drugim sustavima i njihova modularna agentska arhitektura, koja obuhvaća percepciju, rasuđivanje, planiranje, pamćenje, djelovanje i petlje povratnih informacija, donosi jedinstvene izazove u pogledu privatnosti. U nastavku je pregled nekih ključnih rizika za privatnost i odgovarajućih ublažavanja[173] za svaku fazu, s naglaskom na tome kako se te faze usklađuju s agentskom arhitekturom.

Faze	Mogući rizici & Protumjere
Percepcija (prikupljanje korisničkih unosa i predobrada)	**Rizici:** - Mogu se prikupljati i izlagati osjetljivi unosi korisnika (npr. osobni, financijski) ili posebne kategorije podataka (npr. medicinski podaci). - Nedostatak odgovarajuće predobrade može zadržati identifikacijske podatke. - Ulazni podaci uneseni u osjetljiva sučelja mogu se presresti ili zloupotrijebiti. - Nedostatak transparentnosti: Korisnici možda nisu u potpunosti upoznati s načinom na koji će se njihovi podaci upotrebljavati, zadržavati ili dijeliti s različitim aplikacijama koje će sustav integrirati i komunicirati. **Protumjere za pružatelje:** - Ograničite prikupljanje podataka samo na ono što je nužno potrebno. - Anonimizirajte i unaprijed obradite ulazne podatke kako biste uklonili osjetljive elemente. - Pružanje alata, sučelja ili API-ja koji subjektima za uvođenje omogućuju prikupljanje pristanka korisnika kada je to potrebno - Sigurna korisnička sučelja s mehanizmima šifriranja i autentifikacije za zaštitu ulaznih podataka. - Jasnim i lako dostupnim pravilima o privatnosti informirajte korisnike o tome kako će se njihovi podaci upotrebljavati, zadržavati i obrađivati. **Protumjere za subjekte koji primjenjuju sustav:** - Konfigurirajte integracije kako biste prikupljali i pohranjivali samo potrebne ulazne podatke. Izbjegavajte traženje ili pohranjivanje osobnih podataka, osim ako to nije izričito potrebno za zadatak. - Uvesti *preprocessing pipelines* kako bi se uklonile osjetljive informacije prije slanja ulaznih podataka u sustav pružatelja usluga. - Kada je to potrebno, implementirajte obrasce za privolu ili sučelja za korisnike prilikom prikupljanja podataka za upotrebu s LLM-om. - Sigurna korisnička sučelja s mehanizmima šifriranja i autentifikacije kako bi se zaštitili ulazni podaci i osiguralo pravilno postupanje s podacima na razini uvođenja putem šifrirane lokalne pohrane i sigurnih API veza. - Jasno priopćite korisnicima kako se postupa s njihovim podacima i kako se oni obrađuju. To se može učiniti putem pravila o privatnosti, upozorenja, uputa ili izjava o ograničenju odgovornosti u korisničkom sučelju.
Obrazloženje (agent processing)	**Rizici:** - Osjetljivi podaci koji se upotrebljavaju u zadacima zaključivanja mogu biti zloupotrijebljeni ili izloženi. - Nepravilno postupanje s korisničkim podacima tijekom raščlanjivanja zadatka može dovesti do nezakonite obrade osobnih podataka i širenja osjetljivih informacija. - Zaključci (eng. *inferences*) doneseni tijekom rasuđivanja mogli bi nenamjerno otkriti osobne uvide. - Ograničena objašnjivost mogla bi dovesti do netočnih ili neoptimalnih rezultata, čime bi se smanjilo povjerenje i pouzdanost, posebno u složenim zadaćama donošenja

[173] OWASP https://genaisecurityproject.com/resource/agentic-ai-prijetnje-i-ublažavanja/

	odluka. Bez jasnih lanaca zaključivanja korisnici mogu imati poteškoća u razumijevanju ili provjeri načina na koji se donose zaključci, čime se povećava vjerojatnost pogrešaka i nenamjernih ishoda. **Protumjere za pružatelje:** - Uvesti mehanizme za anonimizaciju ili prethodnu obradu osjetljivih ulaznih podataka kako bi se smanjio rizik od upotrebe osjetljivih informacija tijekom zadataka zaključivanja. - Osigurajte robusne značajke kontrole pristupa koje ograničavaju subjekte koji mogu obrađivati osjetljive podatke, čak i tijekom složenih zadataka zaključivanja. - Osigurati da se svi rezultati zaključivanja sigurno evidentiraju i da ih subjekti za uvođenje mogu revidirati, čime se omogućuje praćenje zaključaka o curenju osjetljivih informacija. - Integrirati mehanizme za stvaranje lanaca zaključivanja (npr. lanac mišljenja) kako bi postupak donošenja odluka LLM-a bio transparentniji i kako bi ga se moglo revidirati. **Protumjere za subjekte koji primjenjuju sustav:** - Smanjite upotrebu osjetljivih podataka tijekom obrade i zaključivanja. - Osigurati da se svi zaključci i posredni podaci mogu provjeriti i sigurno pohraniti. - Primjenjuju stroge kontrole pristupa i bilježenja kako bi nadzirali i ograničili zlouporabu. - Provesti[174] okvir lanca mišljenja (CoT) kako bi se poboljšale sposobnosti zaključivanja LLM-ova. To se može postići poticanjem korisnika, pri čemu se logika za rješavanje problema pruža ručno, ili automatiziranim CoT-om (Auto-CoT), koji grupira pitanja i stvara lance zaključivanja bez ljudske intervencije. Auto-CoT je posebno učinkovit za LLM-ove s oko 100B parametara, ali može biti manje točan za manje modele.[175]
Planiranje (organizacija zadataka i vanjske interakcije)	**Rizici:** - Osjetljivi podaci mogli bi se prenositi vanjskim aplikacijama bez odgovarajućih zaštitnih mjera. Funkcijski pozivi mogu prenositi prekomjerne ili nepotrebne korisničke podatke vanjskim aplikacijama, posebno ako filtriranje parametara nije pravilno provedeno. Sustavi trećih strana možda se ne pridržavaju istih standarda privatnosti i sigurnosti. **Protumjere za pružatelje i subjekte za uvođenje:**[176] - Koristite anonimizaciju i enkripciju prilikom prijenosa podataka vanjskim aplikacijama. - Pratite aplikacije trećih strana kako biste osigurali da se vanjski sustavi koji su u interakciji s agentima umjetne inteligencije pridržavaju istih standarda privatnosti, sigurnosti i regulatornih standarda kao i vaš vlastiti sustav: provoditi procjene dobavljača, provjeravati imaju li pružatelji usluga treće strane ažurirane sigurnosne certifikate, kao što su ISO 27001, SOC 2 ili slično, kako bi potvrdili da ispunjavaju priznate sigurnosne standarde, sklapati ugovore ili sporazume o obradi podataka koji uključuju jasna očekivanja u pogledu usklađenosti s pravilima o privatnosti, upotrebljavati alate za praćenje aplikacija trećih strana u stvarnom vremenu za sumnjive aktivnosti i osigurati da pružatelji usluga treće strane imaju pouzdane planove za odgovor na incidente kako bi pravodobno riješili moguće povrede i obavijestili dionike.

[174] https://www.ibm.com/think/topics/chain-of-thoughts

[175] https://www.linkedin.com/pulse/stateful-responsible-aiagents-debmalya-biswas-runze/

[176] Pružatelji usluga odgovorni su za osiguravanje da temeljni model i platforma budu sigurni, usklađeni s privatnošću i opremljeni značajkama za sigurnu implementaciju koje subjekti za implementaciju mogu konfigurirati. Subjekti za uvođenje odgovorni su za sigurnu integraciju, konfiguriranje i korištenje LLM-a u svojem specifičnom kontekstu. Podjela odgovornosti ovisi o razini prilagodbe koju zahtijeva subjekt za uvođenje i kontekstu uvođenja, pri čemu obje strane dijele odgovornost za provedbu potrebnih mjera ublažavanja.

	- Provedba učinkovitog upravljanja identitetom i pristupom:[177] uvesti sigurnosni model *zero trust* ckojim se kontinuirano provjeravaju identitet i uređaj, uz pretpostavku da ne postoji implicitno povjerenje; implement dinamičke kontrole pristupa koje su svjesne konteksta prilagođavaju dozvole na temelju čimbenika u stvarnom vremenu kao što su lokacija, status uređaja ili ponašanje kako bi se smanjili rizici; odobriti pristup „točno na vrijeme", čime se osigurava da agenti umjetne inteligencije imaju dozvole samo za vrijeme trajanja svojih zadaća, čime se smanjuje puzanje privilegija; preispitivanje i ažuriranje kontrola pristupa tijekom životnog ciklusa agenta umjetne inteligencije; automatizirati rotaciju vjerodajnica, upravljanje ključevima i ažuriranja certifikata kako bi se održala sigurnost i smanjila ljudska pogreška. - Pribavite, prema potrebi, privolu korisnika za vanjske interakcije i transparentno osigurajte politike. - Implementirajte parametre filtra u funkcijske pozive i izbjegavajte nepotrebno slanje i zadržavanje osjetljivih prijelaznih podataka.
Memorija (pohrana i zadržavanje podataka)	**Rizici:** - Dugotrajna pohrana korisničkih podataka povećava rizik od neovlaštenog pristupa ili zlouporabe. - Zadržavanje osobnih podataka u svim interakcijama može kršiti propise o zaštiti osobnih podataka. CMP-ovi održavaju kontekst u višestrukim interakcijama, što može rezultirati dugotrajnim sesijama u kojima se akumuliraju osobnipodaci. **Protumjere za pružatelje i subjekte za uvođenje:**[178] - Omogućite korisnicima upravljanje pohranjenim podacima (npr. brisanje, uređivanje). - Primijenite sigurna rješenja za pohranu s robusnim kontrolama pristupa i šifriranjem. - Ograničite razdoblja čuvanja i implementirajte politike automatiziranog brisanja za osjetljive podatke.
Aktivnost (proizvodnja i isporuka)	**Rizici:** - Generirani izlazi mogu nenamjerno uključivati osobne podatke. - Rezultati podijeljeni s vanjskim sustavima mogli bi se presresti ili zloupotrijebiti. - Prilikom orkestriranja više AI agenata, vjerojatnost halucinacija povećava se kako broj uključenih agenata raste. **Protumjere za pružatelje i subjekte za uvođenje:**[179] - Potvrdite i filtrirajte izlaze kako biste bili sigurni da ne otkrivaju osjetljive informacije. - Sigurni kanali za isporuku izlaza s mehanizmima za šifriranje i autentifikaciju. - Pratite interakcije s vanjskim sustavima radi pridržavanja standarda privatnosti. - Kako bi se smanjile halucinacije, iako izrađeni upiti mogu biti korisni, nude ograničenu učinkovitost. LLM-ove bi trebalo prilagoditi uređenim, visokokvalitetnim podacima i konfigurirati tako da se prostor za pretraživanje odgovora ograniči na relevantne i ažurirane informacije.[180]
Povratne informacije i iteracija Loop (učenje i poboljšanje)	**Rizici:** - Povratne informacije korisnika mogu se pohraniti ili koristiti za prekvalifikaciju modela bez privole korisnika. - Osjetljive povratne informacije mogu nenamjerno ostati u zapisnicima ili skupovima podataka.

[177] https://www.accenture.com/us-en/blogs/security/strengthening-ai-agent-security-identity-management
[178] Vidjeti bilješku 176.
[179] idem
[180] Vidjeti bilješku 175.

	Protumjere za pružatelje: - Osigurati da platforma uključuje jasne mehanizme prilagođene korisnicima kako bi se korisnici mogli uključiti ili isključiti da se njihovi podaci o povratnim informacijama upotrebljavaju u ponovnom treningu modela. - Nudite ugrađene alate ili značajke za automatsku anonimizaciju ili pseudonimizaciju podataka o povratnim informacijama prije pohrane ili obrade u svrhu prekvalifikacije. - Ograničite razdoblja čuvanja zapisa prema zadanim postavkama i subjektima za uvođenje pružite mogućnosti koje se mogu konfigurirati kako biste osigurali usklađenost s propisima o privatnosti. **Protumjere za subjekte koji primjenjuju sustav:** - jasno priopćiti korisnicima kako će se upotrebljavati njihovi podaci o povratnim informacijama i osigurati pouzdano praćenje mehanizama pristanka korisnika (opt-in/opt-out) koje pruža platforma. - Koristite alate za anonimizaciju i pseudonimizaciju za sigurno rukovanje povratnim podacima. - Ograničite razdoblja čuvanja zapisnika i osigurajte usklađenost s propisima o privatnosti**.**

Razmatranja o vanjskim integracijama u protoku podataka LLM-a

LLM-ovi mogu biti u interakciji s vanjskim sustavima u svim uslužnim modelima, neovisno o tome temelje li se na SaaS-u, standardnoj, vlasničkim LLM sustavima ili agentskoj umjetnoj inteligenciji, čime se povećava složenost protoka podataka. Te interakcije mogu uključivati dohvaćanje informacija iz vanjskih baza znanja, pristup internetu za podatke u stvarnom vremenu ili integraciju s drugim aplikacijama putem API-ja ili dodataka. Takve integracije uvode dodatne slojeve protoka podataka koji se moraju pažljivo mapirati i razumjeti. Pri dizajniranju sustava temeljenog na LLM-u ključno je uzeti u obzir kako se pristupa vanjskim izvorima podataka i sustavima, kako se podaci prenose i obrađuju te koje su zaštitne mjere uspostavljene kako bi se zaštitila privatnost korisnika i sigurnost podataka.

Filtri kao zaštita i kao dodatni sloj za ulaz i izlaz

Većina LLM sustava uključuje ulazne i izlazne filtre kao zaštitne mjere, uvodeći dodatni sloj u protok podataka. Ti filtri djeluju kao kontrolni mehanizmi za predobradu ulaznih podataka ili poboljšanje generiranih izlaza, pomažući u provedbi standarda privatnosti, sigurnosti i sadržaja. Mogu biti u različitim oblicima, uključujući Pythonove skripte, promptne predloške ili čak druge LLM-ove.[181]

Ulazni filtri funkcioniraju kao nadzornici pristupa, pregledavajući podatke prije nego što dođu do osnovnog modela. Na primjer, mogu blokirati osobne podatke, otkriti štetne upute ili dezinficirati neprikladan jezik.

Izlazni filtri osiguravaju da odgovori sustava zadovoljavaju zahtjeve privatnosti, etičke ili kontekstualne zahtjeve prije nego što se pokažu korisnicima ili proslijede nizvodnim sustavima. Izlazni filtar može, na primjer, ukloniti osjetljivi sadržaj ili preformulirati odgovor kako bi se uskladio s organizacijskim pravilima.

Dodavanjem filtara u arhitekturu sustava uvodi se složenost: Filtri mogu dodati vrijeme latencije/obrade, što utječe na vrijeme odziva u sustavima u stvarnom vremenu. Moraju biti sigurne jer bi ranjivosti mogle razotkriti osjetljive podatke ili omogućiti zlonamjernim ulaznim podacima da zaobiđu kontrolu. Potrebno ih je pratiti i redovito ažurirati kako bi se prilagodili novim rizicima, promjenama propisa ili zahtjevima sustava koji se mijenjaju.

[181] https://ai.meta.com/istraživanje/publikacije/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/

Uloge u modelima pružanja usluga velikih jezičnih modela prema Aktu o umjetnoj inteligenciji i Općoj uredbi o zaštiti podataka

Uloge dobavljača i subjekta za uvođenje u skladu s Aktom o umjetnoj inteligenciji, kao i voditelja obrade i izvršitelja obrade u skladu s Općom uredbom o zaštiti podataka, mogu se razlikovati ovisno o različitim modelima usluga. U nastavku je objašnjenje kako se te uloge mogu primijeniti i obrazloženje njihova dodjeljivanja, kategorizirano prema svakom modelu usluge. Imajte na umu da bi se kvalifikacija organizacija kao voditelja obrade ili izvršitelja obrade trebala procijeniti na temelju okolnosti svakog slučaja, a ovdje navedeno objašnjenje namijenjeno je samo u referentne svrhe i ne podrazumijeva da će se uvijek primjenjivati na isti način.

1. LLM kao usluga (eng. *LLM as a Service Model*)

U tom se modelu usluge sva obrada odvija na infrastrukturi pružatelja, a korisnik komunicira s LLM-om putem API-ja ili platformi u oblaku. Pružatelji usluga također mogu koristiti prikupljene podatke za poboljšanje modela, prilagodbu ili istraživanje.

Primjer alata: OpenAI-jev GPT API (npr. ChatGPT)
Slučaj primjene: Poduzeća upotrebljavaju OpenAI-jev API za integraciju ChatGPT-a u svoje aplikacije za korisničku podršku ili generiranje sadržaja. Obrada se u cijelosti obavlja na infrastrukturi OpenAI-ja.
Pružatelj: OpenAI.
Subjekt koji primjenjuje sustav: Tvrtka koja integrira ChatGPT API u svoj tijek rada.

Uloge iz Akta o umjetnoj inteligenciji

- Pružatelj (provider): Organizacija koja razvija i nudi LLM kao uslugu. Dobavljači su odgovorni za osiguravanje usklađenosti s Aktom o umjetnoj inteligenciji, uključujući upravljanje rizicima, transparentnost i tehničku stabilnost (npr. OpenAI pruža GPT modele putem API-ja).
- Subjekt koji primjenjuje sustav (deployer): Organizacija koja koristi LLM (npr. tvrtka koja koristi predviđeno sučelje za bilo koji određeni zadatak).
- Novi pružatelj (provider): Organizacija koja integrira API LLM-a kao uslugu u svoj komercijalni UI sustav (npr. chatbot) također bi se mogla smatrati dobavljačem na temelju Akta o umjetnoj inteligenciji ako se njezin sustav smatra visokorizičnim i obuhvaćen je područjem primjene članka 25. Akta o umjetnoj inteligenciji.

Uloge u Općoj uredbi o zaštiti podataka

- Subjekt koji primjenjuje sustav (deployer): kao voditelj obrade: Subjekt za uvođenje koji upotrebljava LLM kao uslugu obično djeluje kao voditelj obrade podataka jer određuje svrhe i sredstva obrade podataka (npr. prikupljanje upita korisnika radi poboljšanja usluga ili upotreba alata LLM u svrhu sažetka).
- Pružatelj usluga (provider) kao voditelj obrade: Kada pružatelji usluga prikupljaju ili zadržavaju podatke za vlastite potrebe (npr. fino podešavanje modela ili poboljšanje značajki), oni također preuzimaju ulogu voditelja obrade. To je slučaj u većini LLM-a kao uslužnih rješenja u kojima pružatelji usluga imaju vlasništvo nad modelom i podacima o osposobljavanju. U tom bi scenariju zajedničko vođenje obrade moglo biti prikladnija opcija.
- Izvršitelj obrade: Pružatelj usluga (provider) djeluje kao izvršitelj obrade pri rukovanju podacima strogo u skladu s uputama subjekta za uvođenje za određene zadatke, kao što je generiranje odgovora. To bi moglo biti teško u ovom modelu usluga zbog vlasništva modela pružatelja.

U scenariju LLM-a kao modela usluge često govorimo o konceptu **zajedničke odgovornosti,** u kojem i pružatelj i subjekt za uvođenje imaju različite, ali komplementarne uloge u osiguravanju privatnosti, sigurnosti i usklađenosti. Pružatelj je odgovoran za infrastrukturu, osposobljavanje i održavanje modela, a subjekt za uvođenje mora osigurati sigurnu upotrebu, pravilnu integraciju i poštovanje primjenjivih propisa u svojem specifičnom kontekstu uvođenja. Ta podjela odgovornosti zahtijeva jasne sporazume i snažnu suradnju kako bi se učinkovito upravljalo rizicima.

2. LLM „samostalni" servisni model (eng. *LLM 'Off-the-Self' Service Model)*

U modelu usluga „izvan sustava" uloge pružatelja i subjekta za uvođenje određuju se operativnim ustrojem i posebnom upotrebom sustava LLM te načinom na koji platforma i subjekt za uvođenje stupaju u interakciju s modelom. Općenito, sudjelovanje izvornog pružatelja usluga vrlo je ograničeno. Često nemaju pristup podacima subjekta za uvođenje ili zadaćama koje se izvršavaju, a njihovo je sudjelovanje ograničeno na osiguravanje funkcionalnosti, pouzdanosti i usklađenosti modela s regulatornim standardima na mjestu isporuke. Pružatelj može nastaviti nuditi ažuriranja, ispravke grešaka ili poboljšanja osnovne funkcionalnosti modela, ali to obično ne uključuje pristup podacima subjekta za uvođenje ili njihovu obradu.
Subjekt za uvođenje djeluje neovisno, koristeći model koji se pruža putem platforme za izgradnju vlastitog LLM sustava.

Primjer alata: Hugging Face's Transformers Library
Slučaj primjene: Subjekt za uvođenje preuzima prethodno osposobljeni jezični model (npr. BERT) iz repozitorija aplikacije Hugging Face i prilagođava ga ili integrira u svoj sustav.
Pružatelj (provider): Hugging Face (kao platforma) i Google (razvojni inženjer BERT-a) odgovorni su za robusnost i usklađenost originalnog modela.
Subjekt koji primjenjuje sustav (deployer): Organizacija koja koristi model u prilagođene svrhe, kao što je stvaranje chatbota.

Uloge iz Akta o umjetnoj inteligenciji

- Pružatelj (provider): Organizacija koja razvija, stavlja na tržište ili u uporabu gotov LLM model. Dobavljači su dužni osigurati da je model u skladu sa zahtjevima Akta o umjetnoj[182]inteligenciji. U slučaju LLM-ova izdanih na temelju besplatnih licencija otvorenog koda, trebalo bi razmotriti mogućnost osiguravanja visoke razine transparentnosti i otvorenosti ako su njihovi parametri, uključujući pondere, informacije o arhitekturi modela i informacije o upotrebi modela javno dostupni.[183]
    - Ako pružatelj platforme razvije, istrenira ili znatno prilagodi LLM i stavi ga na raspolaganje subjektima za uvođenje, djelovao bi kao *provider* u skladu s Aktom o umjetnoj inteligenciji.

[182] Napomena na temelju uvodne izjave 104. Akta o umjetnoj inteligenciji: Dobavljači modela umjetne inteligencije opće namjene izdanih na temelju besplatne licencije otvorenog koda s javno dostupnim parametrima (uključujući pondere, pojedinosti o arhitekturi i informacije o uporabi) trebali bi podlijegati iznimkama u pogledu zahtjeva povezanih s transparentnošću na temelju Akta o umjetnoj inteligenciji. Međutim, iznimke se ne bi trebale primjenjivati ako takvi modeli predstavljaju sistemski rizik. U takvim slučajevima transparentnost i licenca otvorenog koda same po sebi ne bi trebale biti dovoljne za izuzimanje pružatelja od ispunjavanja obveza iz Uredbe. Nadalje, objava modela otvorenog koda sama po sebi ne jamči znatno otkrivanje skupova podataka koji se upotrebljavaju za osposobljavanje ili prilagodbu niti osigurava usklađenost sa zakonodavstvom o autorskom pravu. Stoga bi od pružatelja usluga i dalje trebalo zahtijevati da izrade sažetak sadržaja koji se upotrebljava za ogledno osposobljavanje i provedu politiku usklađivanja s pravom Unije o autorskom pravu, uključujući utvrđivanje i poštovanje rezervi prava kako je navedeno u članku 4. stavku 3. Direktive (EU) 2019/790.
[183] Uvodna izjava 102. Akta o umjetnoj inteligenciji

  - Platforma bi također mogla imati samo ulogu infrastrukturnog pokretača i ne smatrati se providerom, već distributerom.
- Subjekt koji uvodi sustav (deployer): Organizacija koja koristi standardni model za izgradnju ili poboljšanje vlastitih usluga preuzima ulogu deployera. Međutim, u slučaju visokorizičnih UI sustava subjekt za uvođenje može preuzeti i ulogu dobavljača ako znatno izmijeni ili doradi model ili ga stavi na raspolaganje drugima u okviru vlastitih usluga. Ta je dvostruka uloga obuhvaćena člankom 25. Akta o umjetnoj inteligenciji.

Uloge u Općoj uredbi o zaštiti podataka

- *Deployer* kao voditelj obrade: Subjekt za uvođenje obično djeluje kao voditelj obrade jer određuje svrhu i sredstva obrade osobnih podataka tijekom upotrebe LLM-a.
- *Provider* kao voditelj obrade: Izvorni pružatelj modela može djelovati kao voditelj obrade u ograničenim scenarijima u kojima obrađuje podatke za vlastite potrebe. Ako pružatelj platforme bilježi, analizira ili zadržava podatke korisnika ili subjekta za uvođenje u svrhe kao što su poboljšanje usluga platforme, uklanjanje pogrešaka ili praćenje rada sustava, mogao bi preuzeti ulogu voditelja obrade za tu konkretnu obradu podataka.
- Izvršitelj obrade: Ta bi uloga mogla biti izvršavanje zadaća u oblaku koje je subjekt za uvođenje izričito naložio. Na primjer, tijekom zadaća zaključivanja podataka podaci se mogu obrađivati u skladu s uputama subjekta koji primjenjuje sustav. U tom bi slučaju platforma koja pruža model mogla djelovati kao izvršitelj obradeu skladu s Općom uredbom o zaštiti podataka.

Pružatelj (provider) je i dalje odgovoran za usklađenost i funkcionalnost temeljnog modela. Subjekt za uvođenje (deployer) odgovoran je za način na koji se model provodi, prilagođava i upravlja u njihovu specifičnom kontekstu, posebno u scenarijima u kojima se podaci obrađuju lokalno ili zadatke u oblaku vodi subjekt za uvođenje. Ta dvoslojna odgovornost naglašava potrebu za jasnim ugovornim sporazumima i čvrstim mehanizmima upravljanja.

3. Vlasnički LLM sustavi (eng. Self-developed LLMs)

Sve operacije, od razvoja modela, infrastrukture, prikupljanja ulaznih podataka do obrade modela, provode se pod odgovornošću pružatelja usluga koji često primjenjuje model za vlastitu upotrebu.

Primjeri alata: PyTorch, DeepSpeed, TensorRT-LLM
Slučaj primjene: Tvrtka koristi zbirku različitih alata za razvoj i treniranje prilagođenog LLM-a u potpunosti na svojoj infrastrukturi, uključujući pripremu podataka, obuku i implementaciju.
Pružatelj: Organizacija koja razvija LLM.
Subjekt koji primjenjuje sustav: Ista organizacija (ako također implementira model) ili klijent treće strane.

Uloge iz Akta o umjetnoj inteligenciji

- Pružatelj (provider): Subjekt koji razvija LLM.
- Subjekt koji primjenjuje sustav (deployer): Organizacija koja implementira rješenje i preuzima većinu operativnih odgovornosti, uključujući praćenje, upravljanje rizikom i transparentnost.

U ovom specifičnom modelu usluge, organizacija koja razvija LLM sustav mogla bi biti ista organizacija koja stavlja sustav u vlastitu uporabu. U tom bi se scenariju ista organizacija smatrala dobavljačem (providerom) i subjektom za uvođenje (deployerom) u skladu s Aktom o umjetnoj inteligenciji.

Uloge u Općoj uredbi o zaštiti podataka

- Pružatelj usluga (provider) kao voditelj obrade: Razvojni programer LLM sustava, budući da kontrolira i izvršava sve aktivnosti obrade podataka unutar svoje lokalne infrastrukture tijekom razvoja.
- Subjetk koji uvodi sustav (deployer) kao voditelj obrade: subjekt za uvođenje jer određuje svrhu i načine obrade osobnih podataka tijekom upotrebe LLM-a.
- Izvršitelj obrade: Tu ulogu može preuzeti bilo koja treća strana koja obrađuje podatke u ime voditelja obrade.

Zahvaljujući potpunoj kontroli nad infrastrukturom i podacima voditelj obrade odgovoran je za usklađenost sa zahtjevima Opće uredbe o zaštiti podataka i Akta o umjetnoj inteligenciji.
Uloga izvršitelja obrade ograničena je na bilo koji alat ili komponentu treće strane kojima bi se voditelj obrade mogao koristiti u tom postupku.

4. Agentski sustavi umjetne inteligencije

Agentski UI sustavi uvode jedinstvenu dinamiku protoka podataka i dodjele uloga u skladu s Općom uredbom o zaštiti podataka i Aktom o umjetnoj inteligenciji zbog svojeg autonomnog i dinamičnog ponašanja.

Primjer alata: Auto-GPT
Slučaj primjene: Auto-GPT samostalno obavlja zadatke kao što su prikupljanje podataka i planiranje na temelju uputa visoke razine. Subjekt za implementaciju može prilagoditi agenta ili ga integrirati s određenim sustavima.
Pružatelj (provider): Razvojni programeri koji stvaraju temeljne Auto-GPT arhitekture.
Subjekt koji uvodi sustav (deployer): Organizacije koje koriste ili mijenjaju Auto-GPT za određene tijekove rada (npr. automatizacija poslovnih operacija).

Uloge iz Akta o umjetnoj inteligenciji

- Pružatelj (provider): Subjekt koji razvija i isporučuje LLM ili osnovnu agentsku arhitekturu.
- Subjekt koji uvodi sustav (deployer): Organizacija koja primjenjuje agentski UI sustav za vlastitu upotrebu ili upotrebu treće strane. U visokorizičnim UI sustavima, ako subjekt za uvođenje prilagodi agenta, integrira ga s određenim sustavima ili znatno izmijeni njegovu arhitekturu, može preuzeti i ulogu dobavljača u skladu s člankom 25. Akta o umjetnoj inteligenciji, odgovornog za usklađenost izmijenjenog sustava.

U ovom modelu usluge subjekt koji primjenjuje sustav često je istodobno i pružatelj i primjenitelj, ovisno o razini prilagodbe, dodatnog podešavanja ili daljnje implementacije agentskog sustava umjetne inteligencije.Uloge u Općoj uredbi o zaštiti podataka

- Deployer kao voditelj obrade: Subjekt za uvođenje obično preuzima ulogu voditelja obrade jer određuje svrhu i sredstva obrade osobnih podataka. To uključuje ulaze, izlaze, upravljanje memorijom i interakcije s vanjskim sustavima.

- Izvršitelj obrade: Ako subjekt za uvođenje upotrebljava alate trećih strana, vanjske API-je ili usluge računalstva u oblaku kao dio operacija agentske umjetne inteligencije, ti treći dobavljači mogli bi djelovati kao izvršitelji obrade. Na primjer, ako vanjski API ili usluga olakšavaju dohvaćanje podataka u stvarnom vremenu ili poboljšavaju funkcionalnost, preuzimaju ulogu obrade u skladu s uputama subjekta za uvođenje. U nekim slučajevima treće strane mogu djelovati kao zajednički voditelji obrade.

Dijeljenje odgovornosti:
Subjekt za uvođenje snosi znatnu odgovornost za upravljanje izlaznim rezultatima i interakcijama agenta umjetne inteligencije. Međutim, pružatelji temeljnih LLM-ova ili modula također bi mogli dijeliti odgovornost za usklađenost prije uvođenja.

U ovoj tablici prikazan je pregled mogućih uloga po modelu usluge, uvijek podložno procjeni postojećih okolnosti:

Model	Deployer kao voditelj obrade	Provider kao voditelj obrade	Izvršitelji obrade
LLM kao usluga	Kada subjekt za uvođenje (deployer) upotrebljava LLM u svrhe specifične za aplikaciju, definiranje ciljeva i metoda obrade podataka (npr. obrada korisničkih upita).	Ako provider provodi fino podešavanje, osposobljavanje ili analitiku izvan uputa subjekta za uvođenje (npr. zadržavanje podataka za potrebe ponovnog treniranja ili praćenja).	Provider: Za obradu podataka prema uputama subjekta za uvođenje (deployer), kao što je rukovanje ulaznim/izlaznim podacima tijekom zadaća zaključivanja.
LLM 'Off-theSelf'	Za određivanje upotrebe LLM sustava, kontrolu podataka tijekom predobrade, rukovanje izlazom i prilagodbu LLM-a za određene tijekove rada.	Kada izvorni pružatelj modela zadrži ili ponovno upotrijebi podatke u svoje svrhe (npr. ispravljanje pogrešaka ili praćenje učinkovitosti).	Platforma: Za zadatke obrade u oblaku koji se obavljaju prema uputama subjekta za uvođenje (deployera) npr. hosting modela za izvođenje zaključaka (eng. inference)
(eng. Self-developed LLMs) Samorazvijeni (vlasnički) LLM	Potpuno primjenjivo kada je razvojni programer također deployer, jer organizacija izravno definira ciljeve obrade podataka.	Potpuno primjenjivo, kako organizacija razvija i kontrolira model.	Nije primjenjivo jer nijedna treća strana koja pruža modele nije uključena u obradu. Mogu postojati i druge treće strane koje djeluju kao izvršitelji obrade.
Agentska umjetna inteligencija	Deployer djeluje kao primarni voditelj obrade, upravlja ulazima podataka, zadacima, pohranom memorije i vanjskim interakcijama sustava, dok nadzire autonomiju agenta.	Kada provider temeljnog modela ili provider modula zadržava podatke iz interakcija za vlastite potrebe, kao što je poboljšanje komponenti ili alata za rasuđivanje. Kada provider provodi osposobljavanje izvan uputa subjekta za uvođenje.	Davatelj modela i alata (provideri)/ ostale treće strane: Kada se vanjski API-ji, alati ili platforme upotrebljavaju za posebne funkcije agenata u skladu s uputama subjekta za uvođenje (deployera).

4. Procjena rizika u području zaštite podataka i privatnosti: Identifikacija rizika

Procjena rizika općenito se smatra prvom fazom upravljanja rizikom. Obuhvaća analizu rizika, koja uključuje utvrđivanje, procjenu i procjenu potencijalnih rizika. Za početak, analiza rizika zahtijeva pažljivo utvrđivanje rizika koji se mogu pojaviti u određenom kontekstu. U ovom se odjeljku razmatra kako pristupiti utvrđivanju rizika za privatnost i zaštitu podataka u LLM sustavima.

Kriteriji koje treba uzeti u obzir pri utvrđivanju rizika

Čimbenici rizika

Kako bismo lakše identificirali rizike povezane s korištenjem LLM-ova, možemo iskoristiti različite čimbenike rizika.
Čimbenici rizika su uvjeti povezani s većom vjerojatnošću neželjenih ishoda. Mogu pomoći u prepoznavanju, procjeni i određivanju prioriteta potencijalnih rizika. Na primjer, obrada osjetljivih podataka i velike količine podataka dva su čimbenika rizika s visokom razinom rizika. Njihovo priznavanje u slučaju vlastite uporabe može vam pomoći da utvrdite povezane potencijalne rizike i njihovu ozbiljnost.

Čimbenici rizika prikazani u nastavku rezultat su analize sadržaja pravnih instrumenata kao što su Opća uredba o zaštiti podataka[184], Uredba EU-a o zaštiti podataka[185], Povelja EU-a i[186] druge primjenjive smjernice povezane s privatnošću i zaštitom podataka.[187]

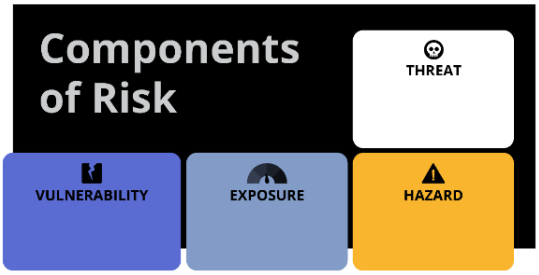
Visok rizik / Važni razlozi za zabrinutost	**Primjeri primjenjivosti**
Osjetljiva & utjecajna svrha obrade Upotreba LLMS-a za odlučivanje o ostvarivanju ili sprečavanje ostvarivanja temeljnih prava pojedinaca ili o njihovu pristupu usluzi, izvršenju ili izvršenju ugovora ili pristupu financijskim uslugama razlog je za zabrinutost, posebno ako će te odluke biti automatizirane bez ljudske intervencije. Pogrešne odluke mogle bi negativno utjecati na pojedince.	Uvođenje LLM-a za utvrđivanje kreditne sposobnosti ili odobrenja kredita bez ljudskog nadzora ili za automatizaciju odluka o zapošljavanju, promaknućima ili otkazima radnih mjesta bez odgovarajućih zaštitnih mjera moglo bi negativno utjecati na pojedince.
Obrada osjetljivih podataka Kada LLM obrađuje osjetljive podatke kao što su posebne kategorije podataka, osobni podaci povezani s osuđujućim presudama i kaznenim djelima, financijski podaci, podaci o ponašanju, jedinstveni identifikatori, podaci o lokaciji itd. To je razlog za zabrinutost jer bi neprimjerena obrada tih osobnih podataka mogla negativno utjecati na pojedince.	Korištenje sustava temeljenog na LLM-u za analizu kartona pacijenata, dijagnoza ili planova liječenja ili podataka povezanih s kaznenim osudama, sudskim zapisima ili istražnim izvješćima.
Obrada velikih razmjera Obrada velikih količina osobnih podataka razlog je za zabrinutost, posebno ako su ti osobni podaci osjetljivi. Što je veća količina, to je veći učinak u slučaju povrede podataka ili bilo koje druge situacije koja ugrožava pojedince.	LLM koji se primjenjuje na velikoj platformi za e-trgovinu koja obrađuje velike količine korisničkih podataka ili LLM koji se upotrebljava na platformi društvenih medija.

[184] Opća uredba o zaštiti podataka (2016/679)
[185] Uredba Europske unije o zaštiti podataka (Uredba 2018/1725)
[186] Povelja Europske unije o temeljnim pravima (2012/C 326/02)
[187] Pag. 79, AEPD, „Risk Management and Impact Assessment in Processing of Personal Data" (Upravljanje rizicima i procjena učinka u obradi osobnih podataka), 2021.

Obrada podataka ranjivih pojedinaca To je razlog za zabrinutost jer je ranjivim pojedincima često potrebna posebna zaštita. Obrada njihovih osobnih podataka bez odgovarajućih zaštitnih mjera može dovesti do kršenja njihovih temeljnih prava. Neki primjeri ranjivih pojedinaca su djeca, starije osobe, osobe s mentalnim bolestima, osobe s invaliditetom, pacijenti, osobe kojima prijeti socijalna isključenost, tražitelji azila, osobe koje pristupaju socijalnim uslugama, zaposlenici itd.	To bi mogao biti slučaj kada se LLM sustavi koriste u zdravstvenom sektoru, u školama, organizacijama socijalnih službi, vladinim institucijama, poslodavcima itd. Na primjer, platforma temeljena na LLM-u koja se koristi u školama za procjenu uspješnosti učenika i pružanje personaliziranih preporuka za učenje obrađuje podatke o djeci.
Niska kvaliteta podataka Niska kvaliteta ulaznih podataka i/ili podataka za osposobljavanje razlog je za zabrinutost jer može dovesti do rizika od netočnosti u generiranom rezultatu, što bi moglo uzrokovati pogrešnu identifikaciju znakova i imati druge negativne učinke, ovisno o slučaju upotrebe.	LLM-ovi se u velikoj mjeri oslanjaju na kvalitetu ulaznih podataka koje dostavljaju korisnici i podataka koji se upotrebljavaju za osposobljavanje modela. Bilo kakve netočnosti, pristranosti ili nepotpunosti u podacima mogu imati dalekosežne posljedice jer LLM-ovi generiraju izlazne rezultate na temelju obrazaca koje otkriju u svojim vježbama i ulaznim podacima. Stupanj rizika koji predstavlja niska kvaliteta podataka uvelike ovisi o primjeni. U manje kritičnim slučajevima upotrebe, kao što je generiranje sadržaja, netočnosti mogu imati manji učinak. Međutim, u scenarijima s velikim ulozima, kao što su zdravstvena skrb, financije ili javna politika, čak i manje netočnosti mogu imati znatne negativne posljedice.
Nedostatne sigurnosne mjere Nedostatak dostatnih zaštitnih mjera mogao bi biti uzrok povrede podataka. Podaci se također mogu prenositi državama ili organizacijama u drugim zemljama bez odgovarajuće razine zaštite.	To bi mogao biti slučaj ako nisu provedene dostatne zaštitne mjere za zaštitu ulaznih podataka i rezultata obrade. To bi se moglo primijeniti na svaki slučaj upotrebe. LLM-ovi koji se nude kao SaaS rješenja u nekim slučajevima uključuju slanje podataka na obradu poslužiteljima u zemljama bez odgovarajućih zakona o zaštiti podataka, čime se povećava izloženost rizicima za privatnost.

Ostale sastavnice rizika povezanog s umjetnom inteligencijom

Aktom o umjetnoj inteligenciji uvode se ključni sigurnosni koncepti u postupak upravljanja rizicima UI sustava, što odražava njegovu prirodu propisa o sigurnosti proizvoda. Razumijevanje tih koncepata ključno je pri pokretanju procjene rizika povezanih sa sustavom koji se temelji na LLM-u.
Ključni pojmovi kao[188] što su opasnost, izloženost opasnosti, sigurnost, prijetnje, ranjivosti i njihovo međudjelovanje s temeljnim pravima sastavnice su rizika koje pružaju temeljni okvir za evaluaciju rizika povezanih s umjetnom inteligencijom.



Slika 14. Ostale sastavnice rizika

Opasnost se odnosi na potencijalni izvor štete, dok **izloženost opasnosti** opisuje uvjete ili opseg u kojem su pojedinci ili sustavi izloženi toj šteti u opasnoj situaciji. **Sigurnost** predstavlja mjere koje se provode kako bi se smanjila ili ublažila šteta, osiguravajući da sustav radi kako je predviđeno bez izazivanja nepotrebnog rizika. **Prijetnje su** vanjski čimbenici koji mogu iskoristiti **ranjivosti** unutar sustava

[188] Ti su pojmovi ovdje objašnjeni na pristupačan način kako bi se olakšalo razumijevanje, ali nisu službene definicije. Europska usklađena norma za upravljanje rizicima, koju trenutačno razvija CEN/CENELEC na zahtjev Europske komisije, sadržavat će standardizirane definicije i pružiti formalizirane smjernice o tim pojmovima.

temeljenog na LLM-u, a to su slabosti koje bi se mogle iskoristiti za ugrožavanje funkcionalnosti, sigurnosti ili zaštite podataka. Aktom o umjetnoj inteligenciji naglašava se zaštita temeljnih prava, uključujući privatnost, kako bi se osiguralo da sustavi umjetne inteligencije ne utječu negativno na pojedince.

Pri utvrđivanju rizika sustava koji se temelji na LLM-u važno je uzeti u obzir sve te komponente rizika[189] koje bi mogle utjecati na privatnost i zaštitu podataka. Rizici za privatnost često proizlaze iz opasnosti ili ranjivosti unutar sustava koje bi mogle biti iskorištene vanjskim ili unutarnjim prijetnjama. Izloženost opasnosti u tom kontekstu mogla bi se odnositi na način na koji su osobni podaci pojedinaca izloženi tim rizicima upotrebom sustava koji se temelji na LLM-u, na primjer tijekom pretraživanja ulaznih podataka. Razumijevanje tih međusobno povezanih koncepata olakšava postupak upravljanja rizicima UI sustava koji moraju biti u skladu s Općom uredbom o zaštiti podataka i Aktom o umjetnoj inteligenciji, a krajnji je cilj zaštita pojedinaca.

Važnost namjere i konteksta u identifikaciji rizika

U uvodnoj izjavi 90. Opće uredbe o zaštiti podataka naglašava se važnost uspostave konteksta: „uzimajući u obzir prirodu, opseg, kontekst i svrhe obrade te izvore rizika".
To je ključno načelo pri provedbi procjene rizika za privatnost jer se njime osigurava da se rizici za prava i slobode fizičkih osoba procjenjuju u njihovim posebnim operativnim i kontekstualnim okruženjima. To je usko usklađeno s konceptom **„namjene"** iz Akta o umjetnoj inteligenciji, u kojem se naglašava potreba za definiranjem i procjenom očekivanog načina rada UI sustava.
Koncepti **namjene**[190]**i konteksta** temelj su za utvrđivanje rizika u UI sustavima i upravljanje njima.[191] Namjena se odnosi na posebne svrhe i scenarije za koje je UI sustav osmišljen, dok kontekst obuhvaća okruženje, korisničku bazu i operativne postavke u kojima sustav funkcionira. Razumijevanje tih dimenzija ključno je jer se rizici često javljaju kada se sustavi upotrebljavaju na nenamjerne načine ili u kontekstima koji dovode do nepredviđenih ranjivosti.

Jasnim definiranjem predviđene svrhe možete procijeniti jesu li dizajn i funkcionalnosti u skladu s predviđenom upotrebom aplikacije. To je korisno i za utvrđivanje moguće zlouporabe sustava i štete za određene skupine korisnika. Slično tome, razumijevanje šireg konteksta, uključujući demografiju korisnika, jezik, kulturne čimbenike i poslovne modele, omogućuje vam da procijenite interakciju sustava i njegova okruženja kako biste predvidjeli moguće probleme.

Uloga modeliranja prijetnji u identifikaciji rizika privatnosti

S obzirom na širok spektar rizika povezanih s umjetnom inteligencijom, metodologije kao što je modeliranje prijetnji [192] imaju ključnu ulogu u sustavnom utvrđivanju rizika za privatnost. Te metodologije često koriste biblioteke specifičnih prijetnji, opasnosti i ranjivosti umjetne inteligencije, pružajući strukturiranu evaluaciju rizika tijekom cijelog životnog ciklusa UI sustava, uključujući one koji proizlaze iz namjerne i nenamjerne uporabe sustava.

[189] Novelli, C., Casolari, F., Rotolo, A. *et al.* Procjena rizika umjetne inteligencije: Proporcionalna metodologija Akta o umjetnoj inteligenciji koja se temelji na scenarijima. *DISO* **3**, 13. (2024.). https://doi.org/10.1007/s44206-024-00095-1
[190] U članku 9. stavku 2. točki (a) Akta o umjetnoj inteligenciji za visokorizične UI sustave za upravljanje rizicima zahtijeva se „utvrđivanje i analiza poznatih i razumno predvidljivih rizika koje visokorizični UI sustav može predstavljati za zdravlje, sigurnost ili temeljna prava ako se visokorizični UI sustav upotrebljava u skladu sa svojom namjenom;
[191] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[192] https://www.threatmodelingmanifesto.org/

Modeliranje prijetnji može pomoći u utvrđivanju potencijalnih površina napada,[193]slučajeva zlouporabe i ranjivosti, čime se omogućuje proaktivan pristup utvrđivanju i ublažavanju rizika. Na primjer, utvrđivanjem protoka podataka i ovisnosti o sustavu modeliranje prijetnji može otkriti rizike kao što je neovlašteni pristup podacima koji možda nisu odmah očiti. Utvrđene prijetnje koje proizlaze iz sesije modeliranja prijetnji mogu se integrirati u evaluacije LLM-a, u kojima se modeli sustavno testiraju protiv prijetnji putem kontradiktornog testiranja, crvenog timiranja ili procjena na temelju scenarija.

Važnost praćenja i prikupljanja dokaza

Kako biste učinkovito upravljali rizicima u LLM sustavima, ključno je svoju procjenu temeljiti i na čvrstim dokazima[194]. To uključuje prikupljanje podataka iz više izvora kako bi se osiguralo da analiza točno odražava potencijalne štete i ranjivosti. Nakon uvođenja, podaci o praćenju kao što su zapisi i obrasci upotrebe mogu pružiti uvid u to kako se sustav upotrebljava u praksi i je li usklađen sa svojom namjenom. Tijekom cijelog životnog ciklusa rezultati evaluacije dobiveni mjerenjem, testiranjem, vježbama crvenih timova [195] i vanjskim revizijama mogu ukazati na nedostatke u funkcionalnosti, pristranosti ili probleme s uspješnošću. Osim toga, povratne informacije korisnika ili zviždača,[196]koje sadržavaju pritužbe, prijave ili obrasce ponašanja, pružaju vrijedne perspektive o stvarnim rizicima i mogućim područjima za poboljšanje.

Uključivanjem dokaza[197]iz osnovnih i poboljšanih izvora osigurava se sveobuhvatno razumijevanje rizika. Temeljni dokazi uključuju postojeće podatke o značajkama sustava i interakcijama korisnika, dok poboljšani dokazi mogu uključivati savjetovanje sa stručnjacima, provođenje ciljanih istraživanja ili upotrebu rezultata moderiranja sadržaja ili sustava tehničke evaluacije. Taj višedimenzionalni pristup ne samo da pomaže u utvrđivanju rizika, već pruža i dokumentiranu osnovu za odluke o upravljanju rizicima.

Primjeri rizika privatnosti u LLM sustavima

LLM-ovi mogu predstavljati širok raspon rizika za privatnost i zaštitu podataka. Ti rizici proizlaze iz različitih čimbenika, uključujući poseban slučaj uporabe, kontekst primjene te čimbenike rizika i dokaze utvrđene tijekom postupka procjene. Prepoznavanje i rješavanje tih rizika ključno je za organizacije koje nastoje odgovorno nabaviti, razviti ili implementirati sustave temeljene na LLM-u.

U tablici u nastavku kategoriziraju se zajednički rizici za privatnost LLM sustava na temelju njihove primjenjivosti na uloge pružatelja usluga i subjekata za uvođenje. I pružatelji i subjekti za uvođenje odgovorni su za sve rizike na stolu, ali stupanj odgovornosti ovisi o razini kontrole nad sustavom (npr. pružatelji infrastrukture, subjekti za uvođenje za upotrebu i konfiguraciju).
Važno je razmotriti kako se rizici razlikuju ovisno o perspektivi. Na primjer, iako se pružatelj može suočiti s regulatornim obvezama kako bi se pohrana podataka svela na najmanju moguću mjeru, subjekt za

[193] Meta, Frontier AI Framework, https://ai.meta.com/static-resource/meta-frontier-ai-framework/?utm_source=newsroom&utm_medium=web&utm_content=Frontier_AI_Framework_PDF&utm_campaign=Our_Approach_to_Frontier_AI_blog
[194] https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=387549
[195] https://research.ibm.com/blog/what-is-red-teaming-gen-AI
[196] Uvodna izjava 172. Akta o umjetnoj inteligenciji: „Osobe koje djeluju kao zviždači u slučaju kršenja ove Uredbe trebale bi biti zaštićene pravom Unije. Direktiva (EU) 2019/1937 Europskog parlamenta i Vijeća (54) trebala bi se stoga primjenjivati na prijave povreda ove Uredbe i zaštitu osoba koje prijavljuju takve povrede."
[197] https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/270826-consultation-protecting-people-from-illegal-content-online/associated-documents/annex-5-draft-service-risk-assessment-guidance?v=330403

uvođenje mora procijeniti rizike od povjeravanja osjetljivih informacija pružatelju. Te uloge imaju različite odgovornosti koje zahtijevaju posebne strategije upravljanja rizicima.

Pružatelji, subjekti za uvođenje i timovi **za nabavu** moraju zajednički rješavati te rizike. Nabava posebno ima ključnu ulogu u premošćivanju odgovornosti pružatelja usluga i subjekata za uvođenje tako što osigurava da odabrani sustavi ispunjavaju regulatorne standarde i organizacijske zahtjeve u pogledu privatnosti. Ključna razmatranja tijekom javne nabave uključuju procjenu politika pružatelja usluga, osiguravanje usklađenosti s relevantnim propisima i ugradnju klauzula kojima se ograničava zlouporaba podataka i podupiru prava ispitanika.
Subjekti koji **primjenjuju** LLM-ove moraju uzeti u obzir rizike povezane s njihovim posebnim slučajevima upotrebe i kontekstom. Primjenom čimbenika rizika ili evaluacijskih kriterija može se olakšati utvrđivanje tih rizika. Na primjer, kriteriji „niske kvalitete podataka" već mogu potaknuti utvrđivanje rizičnih aktivnosti obrade koje bi mogle prouzročiti štetu.
Pružatelji i **razvojni programeri** LLM-ova moraju provoditi upravljanje rizicima kao iterativni proces za utvrđivanje i rješavanje rizika, prepoznajući da se ti rizici mogu pojaviti u različitim fazama razvojnog ciklusa, kako je objašnjeno u prethodnim odjeljcima.

Pregled iz tablice u nastavku može poslužiti kao praktična polazna točka za utvrđivanje i analizu rizika za privatnost i zaštitu podataka tijekom životnog ciklusa sustava koji se temelje na LLM-u. U tablici je prikazan konsolidirani sažetak rizika za privatnost kojim se dopunjuju pojedinosti koje su već navedene u odjeljku 3. pod naslovom Tok podataka i povezani rizici za privatnost u LLM sustavima.

				Primjenjivost rizika		
Zaštita podataka i rizici za privatnost	**Opis rizika**	**Potencijalni utjecaj GDPR-a**	**Primjeri**	**Servisni model**	**Davatelj**	**Subjekt koji primjenjuje sustav**
1. Nedovoljna zaštita osobnih podataka koji u konačnici mogu biti uzrok povrede podataka.	Zaštitne mjere za zaštitu osobnih podataka ne provode se ili su nedostatne.	Povreda: Članak 32. Sigurnost obrade, članak 5. stavak 1. točka (f) Cjelovitost i povjerljivost te članak 9. Obrada posebnih kategorija osobnih podataka	Osjetljivo otkrivanje podataka u unosima korisnika ili tijekom osposobljavanja, zaključivanja i izlaza. Neovlašteni pristup, nedovoljna enkripcija tijekom prijenosa podataka, zlouporaba API-ja, ranjivosti sučelja, neodgovarajuće tehnike anonimizacije ili filtriranja, izloženost treće strane.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
2. Pogrešno klasificiranje podataka za obuku kao anonimnih od strane voditelja obrade kada sadrže podatke koji se mogu identificirati.	Voditelji obrade mogu pogrešno pretpostaviti da su podaci za osposobljavanje anonimni jer ne provode potrebne zaštitne mjere za zaštitu osobnih podataka.	Povreda: Članak 5. stavak 1. točka (a) (Zakonitost, pravednost i transparentnost), članak 5. stavak 1. točka (b) (Ograničenje svrhe), članak 25. (Integrirana i zadana zaštita podataka)	LLM obučen za nepropisno anonimizirane korisničke dnevnike otkriva prepoznatljive korisničke podatke putem napada zaključivanja modela. Subjekt za uvođenje otkriva da je LLM treće strane koji upotrebljava osposobljen za neanonimizirane osobne podatke, a prodavatelj ne provodi odgovarajuće zaštitne mjere, zbog čega je subjekt za uvođenje izložen rizicima u pogledu usklađenosti.	✓ LLM kao usluga[198] ✓ LLM „lako dostupni" [199] proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
3. Nezakonita obrada osobnih podataka u skupovima za osposobljavanje.	Osobni podaci uključeni su u skupove podataka za obuku bez odgovarajuće pravne osnove, zaštitnih mjera ili pristanka korisnika.	Povreda: Članak 5. stavak 1. točka (a) (Zakonitost, pravednost i transparentnost) Članak 6. stavak 1. (Zakonitost obrade), članak 7. (Pristanak), članak 5. stavak 1. točka (c) (Smanjenje količine podataka)	Platforma za e-trgovinu koristi povijest kupnje kupaca za obuku LLM-a bez informiranja kupaca ili dobivanja njihovog pristanka.	✓ LLM kao usluga[200] ✓ LLM „lako dostupni" [201] proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
4. Nezakonita obrada posebnih kategorija osobnih podataka i podataka povezanih s kaznenim osudama i	Skupovi podataka za treniranje uključuju osjetljive podatke, kao što su zdravstvena ili kaznena evidencija,	Povreda: Članak 9. stavci 1. i 2. (Posebne kategorije podataka), članak 10. (Kaznene presude i kaznena djela).	Medicinski zapisi prikupljeni iz neosiguranih internetskih izvora upotrebljavaju se za obuku zdravstvenog chatbota bez primjene zaštitnih mjera usklađenih s Općom uredbom o zaštiti podataka.	✓ LLM kao usluga[202]	✓	✓

[198] Taj se rizik prvenstveno odnosi na pružatelja; međutim, subjekt za uvođenje dijeli odgovornost tako što osigurava da surađuje sa zakonitim dobavljačima. Uloga subjekta za uvođenje uključuje provođenje dubinske analize kako bi se provjerilo poštuje li dobavljač pravne obveze i djeluje li u okviru primjenjivih propisa.
[199] Idem
[200] Idem
[201] Idem
[202] Idem

kažnjivim djelima u podacima za osposobljavanje.	bez poštovanja iznimaka iz Opće uredbe o zaštiti podataka za zakonitu obradu.			✓ LLM „lako dostupni" [203] proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM		
5. Mogući negativan učinak na ispitanike koji bi mogao negativno utjecati na temeljna prava.	Izlaz sustava mogao bi negativno utjecati na pojedinca.	Povreda: Članak 5. stavak 1. točka (d) Točnost, članak 5. stavak 1. točka (a) Pravednost, članak 22. Automatizirano pojedinačno donošenje odluka, uključujući izradu profila, članak 25. Tehnička i integrirana zaštita podataka	Sustav koji pruža izlazne podatke koji nisu točni ili sadržavaju pristranost i ne pruža mehanizme za izmjenu pogrešaka. Rezultat LLM-a mogao bi se upotrijebiti za donošenje automatskih odluka koje proizvode pravne učinke ili slično značajne učinke na ispitanike.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
6. Nemogućnost ljudske intervencije za obradu koja može imati pravni ili važan učinak na ispitanika.	Automatizirane odluke koje znatno utječu na pojedince donose se bez ljudskog preispitivanja, čime se krše zahtjevi Opće uredbe o zaštiti podataka za ljudski nadzor ili se temelje na neprimjerenoj osnovi.[204]	Povreda: Članak 22. stavci 1. i 3. (Automatizirano donošenje odluka), članak 12. (Transparentna komunikacija).	Chatbot automatizira odobrenja kredita na temelju korisničkih podataka, odbijajući zahtjeve bez uključivanja ljudskog recenzenta.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
7. neodobravanje prava ispitanicima.	Prava ispitanika ne mogu se u potpunosti ili djelomično dodijeliti.	Povreda: Članci od 12. do 14.: Informacije koje treba pružiti pri prikupljanju osobnih podataka Članak 16. i članak 17.: Pravo na ispravak i pravo na brisanje Članak 18. Pravo na ograničenje obrade i članak 21. Pravo na prigovor	Zahtjevi ispitanika za ispravak ili brisanje osobnih podataka ne mogu se ispuniti. Korisnici nisu svjesni kako će pružatelji usluga koristiti, zadržavati ili dijeliti njihove podatke.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓

[203] Idem

[204] U skladu s iznimkama navedenima u članku 22. stavku 2. Opće uredbe o zaštiti podataka automatizirano pojedinačno donošenje odluka dopušteno je samo ako se temelji na ugovornoj nužnosti, izričitoj privoli ili ako je odobreno pravom EU-a ili pravom države članice.

8. Nezakonita prenamjena osobnih podataka.	Osobni podaci koriste se u drugu svrhu.	Povreda: Članak 5. stavak 1. točka (b) Ograničenje svrhe, članak 5. stavak 1. točka (a) Zakonitost, pravednost i transparentnost, Članak 28. stavak 3. točka (a)[205] i članak 29. Obrada pod nadzorom voditelja obrade ili izvršitelja obrade	To bi mogao biti slučaj ako pružatelj upotrebljava ulazne i/ili izlazne podatke za osposobljavanje LLM-a, a da to prethodno nije službeno dogovoreno.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
9. Nezakonito neograničeno pohranjivanje osobnih podataka.	Ulazni i/ili izlazni podaci pohranjuju se dulje nego što je potrebno.	Povreda: Članak 5. stavak 1. točka (e) Ograničenje pohrane i članak 25. Tehnička i integrirana zaštita podataka	Sustav bi mogao nepotrebno pohranjivati ulazne podatke koji nisu izravno relevantni za postupak LLM-a. U nekim slučajevima subjekt za uvođenje može pohraniti izlazne podatke dulje nego što je potrebno. Pružatelji usluga također mogu bilježiti korisničke unose i izlaze za ispravljanje pogrešaka ili poboljšanje modela.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓
10. Nezakonit prijenos osobnih podataka.	Podaci se obrađuju u zemljama bez odgovarajuće razine zaštite.	Povreda: Članak 44. Opće načelo za prijenose, članak 45. Prijenosi na temelju odluke o primjerenosti, članak 46. Prijenosi koji podliježu odgovarajućim zaštitnim mjerama	Pružatelji dugotrajne skrbi mogli bi obrađivati podatke u zemljama koje ne nude dovoljno zaštitnih mjera.[206]	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Agentski LLM	✓	✓
11. Kršenje načela smanjenja količine podataka.	Opsežna obrada osobnih podataka za osposobljavanje modela.	Povreda: Članak 5. stavak 1. točka (c) Smanjenje količine podataka, članak 6. u mjeri u kojoj smanjenje količine podataka utječe na najprikladniju zakonitu osnovu (npr. legitimni interes u skladu s člankom 6. stavkom 1. točkom (f)) i članak	LLM-ovi zahtijevaju znatne količine podataka za osposobljavanje. Slično tome, subjekti za uvođenje mogu upotrebljavati skupove podataka kako bi prilagodili sustav koji se temelji na LLM-u za svoje posebne slučajeve upotrebe.	✓ LLM kao usluga ✓ LLM „lako dostupni" proizvodi ✓ Samorazvijeni LLM ✓ Agentski LLM	✓	✓

[205] „obrađuje osobne podatke samo prema dokumentiranim uputama voditelja obrade, među ostalim s obzirom na prijenose osobnih podataka trećoj zemlji ili međunarodnoj organizaciji, osim ako se to zahtijeva pravom Unije ili pravom države članice kojem podliježe izvršitelj obrade; u tom slučaju izvršitelj obrade obavješćuje voditelja obrade o tom pravnom zahtjevu prije obrade, osim ako se tim pravom zabranjuje takvo obavješćivanje zbog važnih razloga od javnog interesa;"

[206] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10097450?mkt_tok=MTM4LUVaTS0wNDIAAAGYXIH0PW4qTzz-TKclqJPRoyU5yZoUVox1JLxNIcVP7RTnC_bvlu_rRyXg8hy6RdOqFw9BgFYU8wXP1XmPVVBTU7DCNt1660jK9umFkCSnLY4e#english

		25. Tehnička i integrirana zaštita podataka				

Pri procjeni rizika povezanih s LLM-ovima ključno je razmotriti šira pitanja povezana s načelima Opće uredbe o zaštiti podataka kao što su zakonitost, pravednost, transparentnost i odgovornost. Osim pitanja privatnosti, potrebno je riješiti i pitanja povezana s autorskim pravima, prekomjernim oslanjanjem i manipulacijom.

Zakonitost, transparentnost i pravednost
Transparentnost osigurava da pojedinci razumiju kako se njihovi podaci obrađuju, dok pravednost zahtijeva da se s podacima postupa pravedno i bez obmane, izbjegavajući metode koje su štetne, nezakonito diskriminirajuće ili obmanjujuće. Međutim, ta se načela često dovode u pitanje zbog netransparentnosti koja je svojstvena dugim svjetlosnim ciklusima. Na primjer, korisnici mogu imati poteškoća s razumijevanjem načina na koji LLM-ovi generiraju odgovore, određuju prioritete izlaznih proizvoda ili donose odluke, što im otežava procjenu ili osporavanje rezultata.
Načelo pravednosti, usko povezano s transparentnošću, naglašava važnost pružanja jasnih, pristupačnih i smislenih informacija o aktivnostima obrade podataka. Usklađenost s obvezama transparentnosti, kako je navedeno u člancima od 12. do 14. Opće uredbe o zaštiti podataka, uključuje detaljno navođenje logike, značaja i mogućih posljedica automatiziranog donošenja odluka, uključujući izradu profila. To je posebno važno za LLM-ove s obzirom na njihovu složenost i opsežnu upotrebu podataka tijekom razvoja. Oslanjanje na iznimke kao što je članak 14. stavak 5. točka (b) Opće uredbe o zaštiti podataka strogo je ograničeno na scenarije u kojima su svi pravni zahtjevi u potpunosti ispunjeni.[207]

Kako bi se uskladili s tim načelima, razvojni programeri LLM-a moraju aktivno pratiti rezultate, rješavati potencijalne pristranosti[208]upotrebom visokokvalitetnih i nepristranih podataka za osposobljavanje te pružati razumljive informacije prilagođene korisnicima o postupcima donošenja odluka u sustavu. Tim se koracima osigurava usklađenost s Općom uredbom o zaštiti podataka, ali i poštovanje pravednosti i transparentnosti, poticanje povjerenja u tehnologije umjetne inteligencije i zaštita prava pojedinaca.

Autorska prava[209]
LLM-ovi osposobljeni za web-scraped ili javno dostupne podatke često uključuju materijale zaštićene autorskim pravima, što izaziva zabrinutost zbog kršenja intelektualnog vlasništva. Rezultati dobiveni takvim modelima mogu nenamjerno replicirati zaštićeni sadržaj, stvarajući pravne rizike i za pružatelje i za subjekte za uvođenje. Ta pitanja naglašavaju važnost osiguravanja da se podaci koji se upotrebljavaju za osposobljavanje LLM-ova prikupljaju i obrađuju zakonito i u skladu sa zakonima o autorskim pravima.

Prekomjerno oslanjanje & Manipulacija
Prekomjerno oslanjanje[210] može ugroziti autonomiju i odgovornost korisnika. LLM-ovi mogu nenamjerno utjecati na ponašanje korisnika putem prilagođenih preporuka ili izvedenih preferencija, smanjujući autonomiju. Korisnici mogu imati povjerenja u rezultate generirane LLM-om u kritičnim područjima, kao što su financijski savjeti ili odluke o zdravstvenoj skrbi, bez dovoljnog razumijevanja ili nadzora. To prekomjerno oslanjanje može prikriti pogreške ili pristranosti u sustavu, što dovodi do odluka koje nisu ni poštene ni transparentne. LLM-ovi također mogu stvoriti realističan lažni sadržaj, kao što su deepfake, širenje pogrešnih informacija ili manipuliranje mišljenjima.

[207] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.
[208] https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/large-language-models-llm_en
[209] https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_hr
[210] https://www.aporia.com/learn/risks-of-overreliance-on-llms/

Neke strategije ublažavanja uključuju zahtijevanje ljudskog nadzora za ključne odluke, promicanje digitalne pismenosti i osiguravanje da su rezultati jasno označeni kao preporuke stvorene umjetnom inteligencijom.

5. Procjena rizika u području zaštite podataka i privatnosti: Procjena rizika & Evaluacija

Od identifikacije rizika do procjene rizika

Nakon što se utvrde rizici, sljedeći ključni koraci u fazi analize rizika jesu procjena i evaluacija rizika. To uključuje klasifikaciju i određivanje prioriteta rizika na temelju njihove vjerojatnosti[211]i ozbiljnosti ili potencijalnog učinka. Stvarna razina rizika ili klasifikacija uvelike će ovisiti o konkretnom slučaju upotrebe, operativnom kontekstu, praćenju sustava, rezultatima evaluacije modela i pogođenim dionicima.
Tijekom ove faze rizici se analiziraju kako bi se detaljnije razumjele njihove implikacije. Taj proces uključuje procjenu čimbenika kao što su vjerojatnost nastanka rizika, potencijalna šteta koju bi on mogao prouzročiti i ranjivosti koje to omogućuju.
Suradnja dionika[212] ima ključnu ulogu u tom procesu, posebno s obzirom na multidisciplinarnu prirodu umjetne inteligencije, u kojoj su doprinosi iz tehničkih, pravnih, etičkih, sigurnosnih i operativnih perspektiva ključni za sveobuhvatno upravljanje rizicima.
Etička matrica[213] može biti vrijedan alat za utvrđivanje na koje bi dionike sustav koji se temelji na dugotrajnom radu mogao izravno ili neizravno utjecati. Mapiranje dionika na temelju njihove razine uključenosti i potencijalnog učinka na njih omogućuje organizacijama da uključe perspektive i zabrinutosti pogođenih osoba u postupak klasifikacije i ublažavanja rizika. Time se osigurava rješavanje etičkih i praktičnih pitanja uvođenja dugoročnih jamstava, čime se provedba sustava usklađuje s potrebama i pravima svih uključenih strana.

Nakon utvrđivanja rizika sljedeći su koraci:[214]

- Procjena vjerojatnosti i ozbiljnosti utvrđenih rizika.
- Procijeniti je li potrebno postupati s rizicima kako bi se osigurala zaštita osobnih podataka i dokazala usklađenost s Općom uredbom o zaštiti podataka i Uredbom EU-a o zaštiti podataka.

Za klasifikaciju i procjenu rizika dostupne su različite metodologije upravljanja rizicima. Cilj ovog dokumenta nije propisati ili definirati specifičnu metodologiju jer bi odabir trebala odrediti svaka organizacija. Međutim, za potrebe ovog dokumenta uputit ćemo na međunarodne norme koje su prethodno istaknute u[215] smjernicama Radne skupine iz članka 29.[216] i AEPD-a, kao i na rad koji se trenutačno obavlja u području europske normizacije umjetne inteligencije.

[211] Napomena: U ovom dokumentu upotrebljavamo pojam „vjerojatnost" umjesto „vjerojatnost" kako bismo se uskladili s terminologijom iz definicija kao što je definicija rizika iz Akta o umjetnoj inteligenciji. Dok u upravljanju rizicima „vjerojatnost" obično upućuje na kvalitativni pristup upravljanju rizicima, „vjerojatnost" podrazumijeva kvantitativnu metodu procjene rizika.
[212] https://ecnl.org/publications/framework-meiningful-engagement-human-rights-impact-assessments-ai
[213] https://blog.dataiku.com/algorithmic-stakeholders-an-ethical-matrix-for-ai
[214] Smjernice o procjeni učinka na zaštitu podataka i utvrđivanje može li obrada „vjerojatno prouzročiti visok rizik" za potrebe Uredbe 2016/679, Radna skupina za zaštitu podataka iz članka 29., posljednja revizija 2017.
[215] ISO 31010:2019, Upravljanje rizicima – Tehnike procjene rizika, Međunarodna organizacija za normizaciju (ISO)
[216] ISO 31000:2009, Upravljanje rizicima – Načela i smjernice, Međunarodna organizacija za normizaciju (ISO); ISO/IEC 29134, Informacijska tehnologija – Sigurnosne tehnike – Procjena učinka na privatnost – Smjernice, Međunarodna organizacija za normizaciju (ISO).

Općenito govoreći, rizik se može izraziti kao:

Rizik = vjerojatnost x ozbiljnost

Ova jednadžba naglašava da je rizik određen vjerojatnošću nastanka događaja, u kombinaciji s potencijalnim utjecajem ili ozbiljnošću nastale štete.
Rizik je u Općoj uredbi o zaštiti podataka (uvodna izjava 75.) definiran kao potencijalna šteta za prava i slobode pojedinaca, različite vjerojatnosti i težine, koja proizlazi iz obrade osobnih podataka. Slično tome, u Aktu o umjetnoj inteligenciji (članak 3.) rizik se definira kao „kombinacija vjerojatnosti nastanka štete i ozbiljnosti te štete;".
Kako bi se procijenila razina rizika u pogledu zaštite podataka i privatnosti pri nabavi, razvoju ili upotrebi LLM-ova, ključno je procijeniti i vjerojatnost i ozbiljnost ostvarenja utvrđenih rizika.

Kriteriji za utvrđivanje vjerojatnosti rizika u LLM sustavima

Kako procijeniti vjerojatnost

Za određivanje vjerojatnosti rizika LLM-a koristit ćemo sljedeću matricu klasifikacije rizika na četiri razine:

Razina posljedica	**Definicija vjerojatnosti**
vrlo visoka	Velika vjerojatnost da će se neki događaj dogoditi
visoka	Znatna vjerojatnost nastanka događaja
Niska	Mala vjerojatnost nastanka događaja
Malo vjerojatno	Ni u kojem slučaju ne postoje dokazi o ostvarenju takvog rizika.

Utvrđivanje vjerojatnosti mora biti prilagođeno posebnim rizicima i slučajevima upotrebe koji se procjenjuju. Iako ta opća matrica pruža strukturirani pristup, primjenom detaljnijih kriterija može se povećati točnost procjene vjerojatnosti.

U tablici u nastavku nalazi se primjer kriterija[217] koji mogu poslužiti kao smjernice za taj postupak i pomoći u poboljšanju procjene vjerojatnosti za određene scenarije. Imajte na umu da se neki kriteriji odnose na atribute na razini sustava, dok su drugi specifični za kontekst.

		Razine vjerojatnosti			
Kriteriji	**Opis**	**Razina 1 (vjerojatno)**	**Razina 2 (niska razina)**	**Razina 3 (Visoki)**	**Razina 4 (Vrlo visoka)**
1. Učestalost korištenja	Koliko se često UI sustav upotrebljava, čime se povećava izloženost potencijalnom riziku koji utječe na pouzdanost (očekivano vrijeme prije kvara)	Sustav se rijetko koristi ili ima rijetke interakcije (npr. godišnje ili manje).	Sustav se povremeno upotrebljava, ali ne u kritičnim operacijama (npr. mjesečno).	Sustav se često upotrebljava i integrira u važne operacije (npr. tjedno).	Sustav se upotrebljava kontinuirano ili u kritičnim operacijama u stvarnom vremenu (npr. svakodnevno).

[217] Barberá, I. "FRASP, A Structured Framework for Assessing the Severity & Probability of Fundamental Rights Interferences in AI Systems" (FRASP, Strukturirani okvir za procjenu ozbiljnosti i prednosti; Vjerojatnost uplitanja u temeljna prava u sustavima umjetne inteligencije) (2025.)

2. Izloženost visokorizičnim scenarijima	Mjera u kojoj UI sustav radi u osjetljivim okruženjima ili okruženjima s velikim brojem sudionika.	Sustav se ne koristi u osjetljivim scenarijima ili scenarijima s velikim ulozima.	Sustav radi u umjereno osjetljivim okruženjima s minimalnim ulogom.	Sustav se koristi u okruženjima s visokim ulozima s potencijalom za značajan utjecaj.	Sustav djeluje u vrlo osjetljivim ili kritičnim okruženjima (npr. zdravstvena skrb, sigurnost).
3. Povijesni presedani	Prethodni slučajevi sličnih rizika ili kvarova u istim ili usporedivim UI sustavima.	U usporedivim sustavima nisu zabilježeni slični rizici ili kvarovi.	U usporedivim sustavima došlo je do malog broja sličnih rizika ili kvarova.	Slični rizici ili kvarovi često su se javljali u usporedivim sustavima.	U usporedivim sustavima zabilježeni su česti i znatni rizici ili kvarovi.
4. Okolišni čimbenici	Vanjski, nekontrolirani uvjeti koji utječu na performanse ili pouzdanost sustava (npr. politička nestabilnost, regulatorni nedostaci, financijska ograničenja).	Vanjski uvjeti su stabilni i ne utječu na performanse sustava.	Vanjski uvjeti povremeno utječu na performanse sustava, ali se njima može upravljati.	Vanjski uvjeti često utječu na performanse sustava, stvarajući ranjivosti.	Vanjski uvjeti ozbiljno utječu na performanse sustava, stvarajući stalne rizike.
5. Robusnost sustava	Stupanj u kojem je UI sustav otporan na kvarove ili nenamjerno ponašanje.	Sustav je vrlo pouzdan s višestrukim viškovima i zaštitnim mjerama.	Sustav je umjereno stabilan, uz određena otpuštanja, ali povremene slabosti.	Sustav je donekle otporan, ali sadržava znatne slabosti ili slabe zaštitne mjere.	Sustavu nedostaje robusnost, zaštitne mjere ili je sklon čestim kvarovima.
6. Kvaliteta i cjelovitost podataka	Mjera u kojoj se UI sustav oslanja na točne, nepristrane i potpune podatke. Promjenjivo boljim prilagođavanjem ili validacijom skupa podataka.	Podaci su vrlo točni, nepristrani i potpuni s minimalnim rizikom od pogrešaka.	Podaci su uglavnom točni i potpuni, ali imaju povremene manje pristranosti ili pogreške.	Podaci su djelomično točni ili potpuni, sa značajnim pristranostima ili nedosljednostima.	Podaci su znatno netočni, pristrani ili nepotpuni, što dovodi do visokog rizika.
7. Ljudski nadzor i stručnost	Kako vještine ljudskih operatera i donošenje odluka utječu na pouzdanost sustava i vjerojatnost rizika. Može se izmijeniti osposobljavanjem ili poboljšanjima nadzora.	Operateri su visoko obučeni, iskusni i dosljedno učinkoviti u donošenju odluka.	Operateri su umjereno obučeni i učinkoviti, ali se događaju povremene pogreške.	Operateri su nedovoljno osposobljeni ili nedosljedni, što dovodi do redovitih pogrešaka u donošenju odluka.	Operateri su neobučeni ili neučinkoviti, što uzrokuje česte i ozbiljne pogreške.

Da biste koristili kriterije za određivanje vjerojatnosti rizika, možete učiniti sljedeće:

Korak 1.: Ukupni rezultati

- Procijenite svaki kriterij i dodijelite mu ocjenu na temelju unaprijed definiranih razina vjerojatnosti od 1 do 4.
- Dodajte ocjene svih čimbenika i podijelite ukupni zbroj s brojem čimbenika za izračun zbirne ocjene vjerojatnosti. To se može učiniti na jedan od sljedećih načina:

    - Ponderirani prosjek: Dodijelite veću važnost određenim čimbenicima ponderiranjem prije uprosječivanja.
    - Jednostavan prosjek: Postupajte sa svim faktorima jednako i izračunajte srednju vrijednost.

Formula: **Ukupna ocjena vjerojatnosti = zbroj svih ocjena / broj čimbenika**

Drugi korak: Karta zbirni rezultat na razinu vjerojatnosti

Nakon što se izračuna zbirna ocjena, rasporedite je na jednu od unaprijed definiranih razina vjerojatnosti iz matrice na temelju sljedećih raspona:

1.0 - 1.5: Malo vjerojatno
1.6 - 2.5: Niska
2.6 - 3.5: visoka
3.6 - 4.0: vrlo visoka

Ovo mapiranje pruža jasnu, kategoriziranu razinu vjerojatnosti za svaki rizik, što pomaže u određivanju prioriteta rizika na temelju njihove potencijalne pojave. Kasnije ćemo u ovom dokumentu istražiti kako se taj okvir može primijeniti u praksi u jednom konkretnom slučaju upotrebe.

Kriteriji za utvrđivanje ozbiljnosti rizika u LLM sustavima

Kako procijeniti ozbiljnost

Za određivanje ozbiljnosti rizika LLM-ova upotrijebit ćemo matricu klasifikacije rizika na četiri razine:[218]

Razina ozbiljnosti	**Definicija ozbiljnosti**
vrlo značajan Katastrofalna šteta	Ono utječe na ostvarivanje temeljnih prava i javnih sloboda, a njegove su posljedice nepovratne i/ili su posljedice povezane s posebnim kategorijama podataka ili kaznenim djelima i nepovratne i/ili uzrokuje znatnu društvenu štetu, kao što je diskriminacija, te je nepovratno i/ili na nepovratan način utječe na posebno ranjive ispitanike, posebno djecu, i/ili uzrokuje znatne i nepovratne moralne ili materijalne gubitke.
Značajno Kritična šteta	Prethodno navedeni slučajevi u kojima su učinci reverzibilni i/ili postoji gubitak kontrole ispitanika nad njegovim osobnim podacima, u kojima je opseg podataka visok u odnosu na kategorije podataka ili broj ispitanika i/ili dolazi ili može doći do krađe identiteta ispitanika i/ili može doći do znatnih financijskih gubitaka ispitanika i/ili gubitka povjerljivosti ispitanika ili kršenja obveze povjerljivosti i/ili postoji društvena šteta za ispitanike ili određene skupine ispitanika
Ograničeno Ozbiljna šteta	Vrlo ograničen gubitak kontrole nad nekim osobnim podacima i određenim ispitanicima, osim nad posebnom kategorijom ili nepovratnim kaznenim djelima ili osudama i/ili zanemarivi i nepovratni financijski gubici i/ili gubitak povjerljivosti podataka koji podliježu obvezi čuvanja poslovne tajne, ali ne i posebne kategorije ili sankcije zbog kršenja
vrlo ograničeno Umjerena ili manja šteta	U gornjem slučaju (ograničeno) kada su svi učinci reverzibilni

Imajte na umu da metodologija upravljanja[219] rizikom HUDERIA, koju je razvio Odbor za umjetnu inteligenciju Vijeća Europe, također koristi matricu ozbiljnosti na četiri razine. Međutim, koristi malo drugačiju terminologiju, kao što je prikazano u ovoj matrici (italizirano): *Katastrofalna šteta, kritična šteta, ozbiljna šteta i umjerena ili manja šteta.*

Slično procjeni vjerojatnosti, procjena ozbiljnosti može imati koristi i od primjene različitih kriterija ozbiljnosti[220] kako bi se smanjila subjektivnost u procesu. Kriteriji ozbiljnosti povezani su s gubitkom privatnosti s kojim se suočava ispitanik, ali koji može imati daljnje povezane posljedice koje utječu na druge pojedince i/ili društvo.

[218] Pag. 77, AEPD, „Risk Management and Impact Assessment in Processing of Personal Data" (Upravljanje rizicima i procjena učinka u obradi osobnih podataka), 2021.
[219] https://rm.coe.int/cai-2024-16rev2-metodology-for-the-risk-and-impact-assessment-of-arti/1680b2a09f
[220] (...) 7/ Rizike povezane s mogućim negativnim učinkom na prava, slobode i interese ispitanika trebalo bi utvrditi uzimajući u obzir posebne objektivne kriterije kao što su priroda osobnih podataka (npr. osjetljivi ili neosjetljivi), kategorija ispitanika (npr. maloljetnik ili ne), broj ispitanika na koje to utječe i svrha obrade. Ozbiljnost i vjerojatnost učinaka na prava i slobode ispitanika elementi su koje treba uzeti u obzir pri procjeni rizika za privatnost pojedinca", str. 4., Radna skupina iz članka 29. (WP 208) „Izjava o ulozi pristupa temeljenog na riziku u pravnim okvirima za zaštitu podataka", donesena 30. svibnja 2014., https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf

U tablici u nastavku navedeni su različiti[221] kriteriji ozbiljnosti. Izračun ozbiljnosti može slijediti iste korake kao i oni koji se koriste za određivanje vjerojatnosti, uključujući zbrajanje ocjena i njihovo raspoređivanje na razine ozbiljnosti. Međutim, za ozbiljnost, određeni kriteriji (brojevi od 1 do 5 i 7 & 8) djeluju kao "zaustavljači" To znači da će krajnji rezultat uvijek biti najviši od tih kriterija bez obzira na rezultat agregiranja. Na primjer, ako se bilo koji od tih kriterija ocijeni na najvišoj razini (4), ukupna ocjena ozbiljnosti odmah se dodjeljuje 4. razini. Tim se pristupom osigurava da se kritičnim štetama, kao što su one koje uključuju nepopravljivu štetu, da odgovarajući prioritet i da ih se označi za hitne i sveobuhvatne mjere ublažavanja.

[221] Vidjeti bilješku 217.

		RAZINE SVOJSTVA			
Kriteriji	**Opis**	**Razina 1 (vrlo ograničena)**	**Razina 2 (ograničeno)**	**Razina 3 (značajna)**	**Razina 4 (vrlo značajna)**
		Umjerena ili manja šteta	**Ozbiljna šteta**	**Kritična šteta**	**Katastrofalna šteta**
		Umjerene ili manje predrasude ili oštećenja pri ostvarivanju temeljnih prava i sloboda koji ne dovode ni do kakvog znatnog, trajnog ili privremenog narušavanja ljudskog dostojanstva, autonomije, fizičkog, psihičkog ili moralnog integriteta ili integriteta zajedničkog života, demokratskog društva ili pravednog pravnog poretka.	Ozbiljne predrasude ili oštećenja u ostvarivanju temeljnih prava i sloboda koje dovode do privremenog narušavanja ljudskog dostojanstva, autonomije, fizičkog, psihičkog ili moralnog integriteta ili integriteta zajedničkog života, demokratskog društva ili pravednog pravnog poretka ili štete informacijskom i komunikacijskom okruženju.	Kritične predrasude ili oštećenja u ostvarivanju temeljnih prava i sloboda koje dovode do značajnog i trajnog narušavanja ljudskog dostojanstva, autonomije, fizičkog, psihičkog ili moralnog integriteta ili integriteta zajedničkog života, demokratskog društva ili pravednog pravnog poretka.	katastrofalne predrasude ili oštećenja u ostvarivanju temeljnih prava i sloboda koja dovode do oduzimanja prava na život; nepopravljivu ozljedu tjelesnog, psihičkog ili moralnog integriteta; uskraćivanje dobrobiti cijelim skupinama ili zajednicama; katastrofalna šteta za demokratsko društvo, vladavinu prava ili preduvjete demokratskog načina života i pravednog pravnog poretka; lišavanje osobne slobode i prava na slobodu i sigurnost; Oštećenje biosfere.
1. Priroda temeljnog prava i usklađivanje sa zakonskim ograničenjem	Ovim se kriterijem ocjenjuje priroda zahvaćenog temeljnog prava, bez obzira na to je li apsolutno ili podložno ograničenjima, te se procjenjuje u kojoj je mjeri slučaj uporabe UI sustava usklađen sa zakonitim i razmjernim ograničenjima. Apsolutna prava nisu odvojiva i ne mogu se ograničiti ni pod kojim okolnostima, dok se druga prava mogu ograničiti samo ako zadiranje ispunjava stroge pravne zahtjeve, zahtjeve proporcionalnosti i nužnosti. Ovaj kriterij pomaže u određivanju ozbiljnosti utjecaja na temelju stupnja neusklađenosti ili kršenja zaštite prava.	Predmetno temeljno pravo vrlo je ograničeno u pogledu područja primjene i primjenjivosti, što znači da često podliježe zakonitim ograničenjima s minimalnim zahtjevima za opravdanje zadiranja. Slučaj upotrebe jasno je i u potpunosti usklađen s dopuštenim pravnim ograničenjima, a uplitanje je rutinsko i široko prihvaćeno, bez uzrokovanja znatnih kršenja pravnih ili normativnih okvira.	Zahvaćeno temeljno pravo umjereno je ograničeno, što znači da su ograničenja zakonita, ali podliježu strožim zahtjevima za opravdanje i konkretnijim uvjetima. Slučaj primjene usklađen je s pravnim ograničenjima, ali zadiranje zahtijeva umjerenu razinu opravdanja, kao što je dokazivanje proporcionalnosti i nužnosti, jer u protivnom može dovesti do manjih kršenja pravnih ili normativnih okvira.	Zahvaćeno temeljno pravo minimalno je ograničeno, što znači da su ograničenja zakonita samo u iznimnim i strogo kontroliranim okolnostima. Slučaj upotrebe djelomično je usklađen sa zakonitim iznimkama, ali postoje nesigurnosti u pogledu proporcionalnosti, nužnosti ili legitimnosti uplitanja, što uzrokuje moguća velika kršenja pravnih ili normativnih okvira.	Zahvaćeno temeljno pravo apsolutno je i ne može se odstupiti, što znači da ni u kojim okolnostima nije dopušteno nikakvo zakonito ograničenje. Alternativno, slučaj upotrebe nije u skladu sa zakonitim i razmjernim ograničenjima, čak i ako pravo nije apsolutno, što uzrokuje teška kršenja pravnih ili normativnih okvira. * Povelja ne navodi izričito prava koja su apsolutna. Na temelju objašnjenja Povelje, EKLJP-a i sudske prakse europskih sudova tvrdi se da ljudsko dostojanstvo (članak 1. Povelje), zabrana mučenja i nečovječnog ili ponižavajućeg postupanja ili kažnjavanja (članak 4. Povelja), zabrana ropstva i prisilnog rada (članak 5. stavci 1. i 2. Povelje), unutarnja sloboda mišljenja, savjesti i vjeroispovijedi (članak 10. stavak 1.

					Povelje), pretpostavka nedužnosti i pravo na obranu (članak 48. Povelje), načelo zakonitosti (članak 49. stavak 1. Povelje) i pravo da se ne bude dva puta suđen ili kažnjen za isto kazneno djelo (članak 50. Povelje) mogu se smatrati apsolutnim pravima.
2. Priroda osobnih podataka	Ovim se kriterijem procjenjuje osjetljivost osobnih podataka koji se obrađuju, uzimajući u obzir njihov potencijal da prouzroče štetu u slučaju zlouporabe. Posebna kategorija podataka (npr. zdravstveni, biometrijski ili genetski podaci) predstavlja veći rizik za temeljna prava kao što su privatnost i autonomija.	Neosjetljivi, javno dostupni podaci (npr. anonimizirani podaci, javne evidencije).	Umjereno osjetljivi podaci (npr. financijski podaci, povijest pretraživanja).	Vrlo osjetljivi podaci	Najosjetljiviji podaci & posebna kategorija podataka, npr. genetski podaci, psihološki profili, biometrija ili podaci koji otkrivaju kaznenu povijest.
3. Kategorija ispitanika (npr. maloljetnik ili ne)	Ovim se kriterijem procjenjuje ranjivost pojedinaca čiji se podaci obrađuju. Ranjive skupine (npr. maloljetnici, marginalizirane zajednice) izložene su većem riziku od zlouporabe podataka.	Ispitanici nisu ranjivi (npr. odrasle osobe u rutinskim, neosjetljivim kontekstima).	Ispitanici uključuju neke pojedince iz potencijalno ranjivih skupina (npr. zaposlenike, klijente).	Ispitanici uključuju pojedince u osjetljivim ili visokorizičnim ulogama (npr. novinari, aktivisti).	Ispitanici su vrlo ranjivi (npr. maloljetnici, osobe s invaliditetom ili progonjene skupine).
4. Svrha obrade	Tim se kriterijem ocjenjuje legitimnost, nužnost i proporcionalnost svrhe u koju se osobni podaci obrađuju. Nezakonite ili nerazmjerne svrhe povećavaju ozbiljnost.	Jasno legitimne i razmjerne svrhe s minimalnim rizicima (npr. operativne svrhe).	Legitimne svrhe s umjerenim rizicima ili neizravnim učincima (npr. ciljani marketing).	Legitimne svrhe, ali uz upitnu proporcionalnost ili nužnost (npr. izrada profila za ocjenjivanje kreditne sposobnosti).	Nezakonite, nejasne ili neproporcionalne svrhe sa znatnim rizicima (npr. nadzor, diskriminirajuće profiliranje).
5. Ljestvica učinka (društveni, skupni, pojedinačni) & broj pogođenih ispitanika	Opseg kršenja na društvenoj, grupnoj i pojedinačnoj razini. Ovim se kriterijem uzima u obzir razmjer učinka na temelju broja pojedinaca čiji su podaci pogođeni.	Utjecaj je ograničen na malu, lokaliziranu grupu ili pojedinca. Pogođeno je manje od 100 osoba.	Učinak je ograničen na određene skupine ili mali društveni segment. Pogođeno je između 100 i 1000 osoba.	Učinak obuhvaća više skupina ili društvenih domena. Pogođeno je između 1000 i 100.000 osoba.	Učinak je raširen i utječe na društvenu, grupnu i pojedinačnu razinu. Ugroženo je više od 100 tisuća ljudi.

6. Kontekstualna osjetljivost i osjetljivost na domenu	Kako specifični kontekstualni čimbenici ili domene pojačavaju ozbiljnost interferencije. Uključuje rizike na osnovi indicija kao što su društveno-politička nestabilnost i ako su pogođena djeca i druge ranjive skupine.	Kontekst ili domena ne pojačavaju ozbiljnost miješanja u temeljno pravo.	Kontekst ili domena umjereno pojačava ozbiljnost miješanja u temeljno pravo.	Kontekst ili domena znatno pojačava ozbiljnost zadiranja u temeljno pravo.	Kontekst ili domena znatno pojačava ozbiljnost miješanja u temeljno pravo.
7. Obnovljivost, oporavak, stupanj mogućnosti popravka	Poteškoće ili izvedivost preokretanja štete i vrijeme potrebno za oporavak. Uključuje previsoke rizike ako je šteta nepovratna.	Šteta je potpuno reverzibilna u kratkom razdoblju uz minimalan napor.	Šteta je reverzibilna uz umjeren napor tijekom razumnog vremenskog okvira.	Štetu je teško preokrenuti, što zahtijeva znatan napor ili resurse.	Šteta je nepovratna i nema izvedivih načina oporavka.
8. Trajanje i postojanost štete	Duljina vremena i postojanost štetnih učinaka uzrokovanih smetnjama.	Štetni učinci su minimalni i ne traju tijekom vremena.	Štetni učinci traju kratko, ali ne dovode do dugoročnih posljedica.	Štetni učinci traju dulje vrijeme i mogu utjecati na više skupina.	Štetni učinci su trajni ili traju neograničeno.
9. Brzina za materijalizaciju	Brzina kojom se rizik materijalizira: Postupno, iznenadno, stalno se mijenja.	Rizik se ostvaruje postupno, osiguravajući dovoljno vremena za intervenciju.	Rizik se ostvaruje umjerenim tempom, što omogućuje korektivne mjere.	Rizik se iznenada materijalizira, ostavljajući ograničeno vrijeme za intervenciju.	Rizik se brzo materijalizira, bez mogućnosti intervencije.
10. Transparentnost i mehanizmi odgovornosti	Stupanj transparentnosti sustava i mehanizmi odgovornosti.	Sustav je vrlo transparentan s jasnim i učinkovitim mehanizmima odgovornosti.	Sustavu nedostaje određena transparentnost, ali ima osnovne mehanizme odgovornosti.	Sustavu nedostaje transparentnosti i ima slabe mehanizme odgovornosti.	Sustav je potpuno netransparentan, bez mehanizama odgovornosti.
11. Ripple i kaskadni učinci	Mjera u kojoj smetnje uzrokuju dodatne štete u sustavima ili domenama.	nema kaskadnih učinaka; rizik je izoliran i obuzdan.	Minimalni kaskadni učinci; Učinci su uglavnom ograničeni.	primjetni kaskadni učinci; učinci se protežu u svim područjima.	teški kaskadni učinci; Učinci se šire u velikoj mjeri.

Procjena rizika: Klasifikacija rizika

Procjena vjerojatnosti i ozbiljnosti temelj je za utvrđivanje ukupne razine rizika utvrđenih rizika za privatnost i zaštitu podataka. Koristeći klasifikacijsku matricu na četiri razine za vjerojatnost i ozbiljnost, rizici se mogu kategorizirati u konačne klasifikacije vrlo visoke, visoke, srednje ili niske.

Matrica, kako je prikazano u nastavku, služi kao praktičan alat za dobivanje tih klasifikacija, nudeći jasno i strukturirano rangiranje za određivanje prioriteta rizika i usmjeravanje odgovarajućih strategija ublažavanja. Ova je klasifikacija ključan korak u sljedećem procesu obrade rizika jer osigurava da su resursi usmjereni na učinkovito rješavanje najhitnijih rizika.

Vjerojatnost	**vrlo visoka**	Srednja vrijednost	visoka	vrlo visoka	vrlo visoka
	visoka	Niska	visoka	vrlo visoka	vrlo visoka
	Niska	Niska	Srednja vrijednost	visoka	vrlo visoka
	Malo vjerojatno	Niska	Niska	Srednja vrijednost	vrlo visoka
		vrlo ograničeno	**Ograničeno**	**Značajno**	**vrlo značajan**
		Ozbiljnost			

Slika 15. Matrica procjene rizika

Najbolje prakse u upravljanju rizicima upućuju na to da bi ublažavanje vrlo visokih i visokih rizika trebalo biti prioritet.[222] Nakon što se utvrde ti kritični rizici, sljedeći je ključni korak izrada i provedba plana postupanja s rizicima.

Kriteriji za prihvaćanje rizika

U fazi procjene rizika primjenjuju se kriteriji rizika kako bi se utvrdilo je li rizik prihvatljiv ili ga je potrebno tretirati. Ti kriteriji odražavaju spremnost i sposobnost organizacije da podnese rizike unutar zakonskih i operativnih ograničenja te se moraju uskladiti s primjenjivim zakonima i propisima, kao što su zahtjevi iz Opće uredbe o zaštiti podataka i Zakona o umjetnoj inteligenciji, kako bi se osigurala usklađenost i zaštita prava pojedinaca.
Različiti okviri i najbolje prakse[223] mogu pomoći organizacijama u definiranju tih kriterija. Oni se mogu utvrditi uzimajući u obzir čimbenike kao što su društvene norme, očekivane koristi, potencijalne štete, utvrđeni pragovi parametara, rezultati evaluacije i usporedivi slučajevi upotrebe.[224] Opravdanje tih odluka od ključne je važnosti jer su organizacije odgovorne za dokazivanje kako se upravlja rizicima i kako ih se ublažava. To je u skladu s načelom odgovornosti iz Opće uredbe o zaštiti podataka,[225]kojim se od organizacija zahtijeva da dokumentiraju i obrazlože svoje odluke o ublažavanju rizika i prihvaćanju.

[222] https://www.pmi.org/learning/library/high-risk-critical-path-projects-7675
[223] https://risk-engineering.org/static/PDF/slides-risk-acceptability.pdf
[224] https://www.sciencedirect.com/topics/engineering/residual-risk
[225] Članak 5. stavak 2., uvodna izjava 74. Opće uredbe o zaštiti podataka

6. Zaštita podataka i kontrola rizika privatnosti

Kriteriji za postupanje s rizikom

tj. ublažiti, prenijeti, izbjeći ili prihvatiti rizik.

Postupanje s rizicima uključuje razvoj strategija za ublažavanje utvrđenih rizika i izradu provedivih planova provedbe. Odabir odgovarajuće opcije obrade trebao bi biti prilagođen kontekstu, vođen analizom izvedivosti kao[226] što je sljedeće:

- Procijeniti vrstu rizika i dostupne mjere ublažavanja koje se mogu provesti.
- Usporedite moguće koristi ostvarene provedbom ublažavanja s uključenim troškovima i naporima te mogućim učinkom.
- Procijeniti utjecaj na predviđenu svrhu implementacije LLM sustava.
- Razmotrite razumna očekivanja pojedinaca na koje sustav utječe.
- Provesti analizu kompromisa kako bi se ocijenio učinak mogućih ublažavanja na aspekte kao što su uspješnost, transparentnost i pravednost, osiguravajući da obrada ostane etična i usklađena na temelju konkretnog slučaja upotrebe.

Analiza tih kriterija ključna je za učinkovito ublažavanje rizika i planiranje upravljanja rizicima, čime se pruža jasnoća o tome jesu li posebni napori za ublažavanje opravdani. U svim slučajevima odabrana mogućnost liječenja trebala bi biti jasno obrazložena i detaljno dokumentirana kako bi se osigurala odgovornost i usklađenost.

Najčešći kriteriji za postupanje s rizikom su: **Ublažavanje, prijenos, izbjegavanje** i **prihvaćanje**.
Za svaki utvrđeni rizik odabrat će se jedna od opcija kriterija:

- ✓ Ublažavanje – provedba mjera za smanjenje vjerojatnosti ili ozbiljnosti rizika.
- ✓ Prijenos – prijenos odgovornosti za rizik na drugu stranu (npr. osiguranjem ili eksternalizacijom).
- ✓ Izbjegavanje – potpuno uklanjanje rizika rješavanjem njegova temeljnog uzroka.
- ✓ Prihvati – odluči ne poduzimati nikakve radnje, prihvaćajući rizik jer je unutar prihvatljivih granica definiranih u kriterijima rizika.

Odlučivanje o tome može li se rizik ublažiti uključuje procjenu njegove prirode, potencijalnog učinka i dostupnih mjera ublažavanja kao što su provedba kontrola, donošenje najboljih praksi, izmjena postupaka ili upotreba alata za smanjenje vjerojatnosti ili ozbiljnosti rizika.
Ne mogu se svi rizici u potpunosti ublažiti. Neki rizici mogu biti inherentni i ne mogu se u potpunosti izbjeći. U takvim je slučajevima cilj smanjiti rizik na prihvatljivu razinu ili provesti mjere ublažavanja i kontrole rizika kojima se učinkovito upravlja njegovim učinkom.

Također je važno voditi dinamičan registar rizika koji sadržava trajne, lako dostupne i jasne evidencije o rizicima koje se dosljedno ažuriraju kako bi se osigurala točnost i relevantnost.[227]

[226] Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR (Rizik, visoki rizik, procjene rizika i procjene učinka na zaštitu podataka u okviru Opće uredbe o zaštiti podataka), CIPL GDPR Interpretation and Implementation Project (Projekt tumačenja i provedbe Opće uredbe o zaštiti podataka), 2016.
[227] https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/volume-1-governance-and-risks-management.pdf?v=387545

Rizicima bi također trebalo dodijeliti jasno vlasništvo te bi trebalo provoditi redovita preispitivanja kako bi se osiguralo da prakse upravljanja rizicima ostanu proaktivne.

Primjer mjera ublažavanja povezanih s rizicima LLM sustava

Odabir odgovarajućih mjera ublažavanja trebao bi se provoditi na pojedinačnoj osnovi. U tablici u nastavku navedene su neke od mogućih mjera ublažavanja koje bi se mogle provesti kako bi se uklonili rizici povezani s privatnošću i zaštitom podataka u dugotrajno nezaposlenim osobama. Te su mjere općenite i nisu povezane ni s jednim konkretnim slučajem upotrebe ni s njegovom namjenom. Preporuke za ublažavanje mogu se primjenjivati i na pružatelje i na subjekte za uvođenje. U nekim se slučajevima ublažavanja primjenjuju i na subjekta za uvođenje ako izmijeni model.

Zaštita podataka i rizici za privatnost	Preporučene mjere ublažavanja	Davatelj	Subjekt koji primjenjuje sustav
1. Nedovoljna zaštita osobnih podataka koji u konačnici mogu biti uzrok povrede podataka Specifično za Smjernice za regionalne potpore: Nedovoljna zaštita u sustavima RAG-ova koja može dovesti do povreda podataka, neizravnog brzog unosa i dohvaćanja zastarjelih ili netočnih informacija, što dovodi do lošeg donošenja odluka i potencijalne štete za pojedince. Posebni upiti također mogu nenamjerno otkriti osobne podatke uključene u obuku, kršeći propise o privatnosti.[228]	Kao pružatelj i subjekt koji primjenjuje sustav važno je provjeriti[229] sljedeće: ▪ API-ji su sigurno implementirani (upotrijebite API pristupnike s mogućnostima ograničavanja brzine i praćenja za kontrolu i praćenje pristupa) ▪ Prijenos podataka zaštićen je odgovarajućim protokolima šifriranja, podaci u mirovanju su šifrirani. ▪ Uveden je odgovarajući mehanizam kontrole pristupa. ▪ Provode se mjere za anonimizaciju i pseudonimizaciju osobnih podataka ili za prikrivanje podataka ili upotrebu sintetičkih podataka. ▪ Provode se dodatne tehničke i organizacijske mjere kako bi se dodatno poboljšala sigurnost. Te mjere mogu uključivati, među ostalim, osiguravanje sigurnog okruženja segregacijom podataka, sigurnosnim kopijama, redovitim revizijama fizičke i digitalne sigurnosti, mehanizmima za odgovor na incidente, informiranjem i osposobljavanjem. ▪ [230]Pristup obrane u dubini može se provesti raščlanjivanjem višestrukih mjera za smanjenje rizika kako bi se spriječile pojedinačne točke neuspjeha. To može uključivati zaštitu na razini modela, mrežnu sigurnost, autentifikaciju korisnika, šifriranje, PET-ove,[231]kontrolu pristupa i stalno praćenje radi otkrivanja zlouporabe, pristranosti ili ranjivosti u stvarnom vremenu. ▪ Kako biste ublažili rizik od pamćenja, primijenite diferencijalne tehnike privatnosti kako biste spriječili kodiranje osjetljivih podataka i redovito testirali regurgitaciju podataka. Razmotrite i upotrebu manjih modela kako biste izbjegli[232] učinak pamćenja od modela s prekomjernom parametrijom. ▪ Isto tako, mjere za zaštitu i identifikaciju unutarnjih prijetnji, mjere za ublažavanje napada u lancu opskrbe kojima bi se mogao omogućiti pristup podacima za učenje i/ili ključevima za pohranu i šifriranje podataka, mjere koje se	✓	✓

[228] https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_hr
[229] To se može učiniti izvođenjem pentesta i/ili traženjem rezultata pentesta od dobavljača.
[230] Međunarodno izvješće o sigurnosti umjetne inteligencije, 2025., stranica 167. https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf
[231] Feretzakis, G.; Papaspyridis, K.; Gkoulalas-Divanis, A.; Verykios, V.S. Tehnike očuvanja privatnosti u generativnoj umjetnoj inteligenciji i velikim jezičnim modelima: Narativni osvrt. Informacije 2024., 15., 697. https://doi.org/10.3390/info15110697
[232] https://blog.kjamistan.com/category/ml-memorization.html

	provode kako bi se spriječili rizici povezani s različitim sigurnosnim prijetnjama u vezi s dugotrajnim radom,[233]kao[234][235] što su zaključci o članstvu, napadi inverzijom modela[236] i trovanjem.[237] ▪ Uspostavljeni su i zapisnici o pristupu i promjenama kako bi se pristupilo dokumentima i promjenama u digitaliziranim zapisima. ▪ Zaposlenici i korisnici obučeni su o najboljim sigurnosnim praksama. ▪ Učinkoviti RAG sustavi zahtijevaju pažljivo usklađivanje modela kako bi se spriječio neovlašteni pristup i osjetljiva izloženost podacima. Integracija s više izvora podataka zahtijeva snažne sigurnosne mjere kako bi se osigurala povjerljivost i cjelovitost podataka, uz pridržavanje načela zaštite podataka kao što su nužnost i proporcionalnost. Za eksternalizirane modele RAG-a koji uključuju prijenos osobnih podataka, usklađenost s pravilima GDPR-a o prijenosu podataka ključna je za održavanje povjerljivosti i pravnih obveza.[238]		
2. Pogrešna klasifikacija podataka za osposobljavanje kao anonimnih od strane voditelja obrade ako sadržavaju informacije koje se mogu identificirati, što dovodi do neprovođenja odgovarajućih zaštitnih mjera za zaštitu podataka. (djelomično povezano s rizikom 3.) *Kad god se informacije koje se odnose na identificirane pojedince ili pojedince koje se može identificirati, a čiji su osobni podaci upotrijebljeni za učenje modela, mogu dobiti iz modela umjetne inteligencije sredstvima za koja je razumno vjerojatno da će se upotrebljavati,* [239]*može se zaključiti da takav model nije anoniman.* [240]	▪ Provesti pouzdane postupke testiranja i validacije kako bi se osiguralo i. da se osobni podaci povezani s podacima za osposobljavanje ne mogu izvući iz modela upotrebom razumnih sredstava i ii. da se rezultati dobiveni modelom ne povezuju s ispitanicima čiji su se osobni podaci upotrebljavali tijekom osposobljavanja niti ih identificiraju. ▪ Ta bi se procjena trebala provesti uzimajući u obzir „sva sredstva za koja je vjerojatno da će se upotrijebiti”, uzimajući u obzir objektivne čimbenike kao što su: [241] o Značajke podataka za učenje, modela umjetne inteligencije i postupka učenja. o kontekst u kojem se model umjetne inteligencije objavljuje ili obrađuje. o dostupnost dodatnih informacija koje bi mogle omogućiti identifikaciju. o troškove i vrijeme potrebno za pristup takvim dodatnim informacijama, ako one nisu lako dostupne. o Trenutačne tehnološke mogućnosti i mogući budući napredak. ▪ Provesti alternativne pristupe anonimizaciji ako pružaju jednaku razinu zaštite, osiguravajući njihovu usklađenost s najnovijim dostignućima. ▪ Provodim strukturirano testiranje protiv najmodernijih napada poput zaključivanja atributa i članstva, izvlačenja, regurgitacije inverzije modela podataka za obuku ili rekonstrukcijskih napada. ▪ Dokumentirati i čuvati dokaze kojima se dokazuje usklađenost s tim zaštitnim mjerama u skladu s obvezama odgovornosti iz članka 5. stavka 2. Opće uredbe o zaštiti podataka. Dokumentacija bi trebala uključivati: o Pojedinosti procjena učinka na zaštitu podataka, uključujući procjene i odluke o njihovoj nužnosti. o Savjeti ili povratne informacije službenika za zaštitu podataka (ako je primjenjivo). o Informacije o tehničkim i organizacijskim mjerama za smanjenje rizika identifikacije tijekom dizajna modela, uključujući modele prijetnji i procjene rizika za skupove podataka za učenje (npr. izvorni URL-ovi i zaštitne mjere).	✓	? [242]

[233] Shokri i dr., „Membership Inference Attacks Against Machine Learning Models”, 2017.
[234] https://genai.owasp.org/llm-top-10/
[235] https://arxiv.org/html/2310.01424v2
[236] Zhang i dr., „Generative Model-Inversion Attacks Against Deep Neural Networks” (Generativni napadi inverzijom modela na duboke neuronske mreže), 2020.
[237] Junfeng Guo, Cong Liu, „Pactical Poisoning Attacks on Neural Networks” (Praktični napadi trovanjem na neuronske mreže), 2020.
[238] https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_hr
[239] Napadi iz zaključivanja članstva i napadi inverzijom modela.
[240] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.
[241] idem
[242] Subjekti za uvođenje trebali bi provjeriti je li pružatelj djelotvorno otklonio taj rizik. Ova je preporuka jednako relevantna u slučajevima u kojima subjekti za uvođenje sudjeluju u modelima fine prilagodbe ili ponovnog osposobljavanja.

	○ Mjere poduzete tijekom cijelog životnog ciklusa modela umjetne inteligencije kako bi se spriječilo ili provjerilo nepostojanje osobnih podataka u modelu. ○ Dokaz teoretske otpornosti modela na tehnike ponovne identifikacije, uključujući metrike, izvješća o testiranju i analizu otpornosti na napad (npr. regurgitacija, zaključak o članstvu). ○ Dokumentacija koja se dostavlja voditeljima obrade i ispitanicima s pojedinostima o mjerama za smanjenje rizika za identifikaciju i rješavanje potencijalnih preostalih rizika.		
3. Nezakonita obrada osobnih podataka u skupovima za osposobljavanje	▪ Dokumentirati sve izvore podataka o osposobljavanju (npr. baze podataka knjiga, internetske stranice) kako bi se osigurala odgovornost u skladu s člankom 5. stavkom 2. Opće uredbe o zaštiti podataka. ▪ Provjerite podatke o osposobljavanju za statističke poremećaje ili pristranosti i izvršite potrebne prilagodbe. ▪ Isključite podatke za vježbanje koji uključuju neovlašteni sadržaj, kao što su lažne vijesti, govor mržnje ili teorije zavjere. ▪ Isključiti sadržaj iz publikacija koje mogu sadržavati osobne podatke koji predstavljaju rizik za pojedince ili skupine, kao što su osobe podložne zlouporabi, predrasudama ili šteti. ▪ Uklonite nepotrebne osobne podatke (npr. brojeve kreditnih kartica, adrese e-pošte, imena) iz skupa podataka za učenje. [243] ▪ Koristiti metodološke izbore koji značajno smanjuju ili eliminiraju prepoznatljivost, kao što je korištenje regularizacijskih metoda za poboljšanje generalizacije modela i minimiziranje preklapanja. ▪ Implementirati robusne tehnike očuvanja privatnosti, kao što je diferencijalna privatnost.[244] ▪ Pri upotrebi internetskog struganja kao metode prikupljanja podataka osigurati usklađenost s člankom 6. stavkom 1. točkom (f) Opće uredbe o zaštiti podataka provođenjem temeljite pravne procjene. To uključuje ocjenjivanje: ○ postojanje legitimnog interesa za obradu podataka. Interes bi trebao biti zakonit, jasno artikuliran i stvaran, a ne spekulativan. ○ ii. nužnost obrade kojom se osigurava da su prikupljeni osobni podaci primjereni, relevantni i ograničeni na ono što je nužno za navedenu svrhu[245]i ○ nije ih bilo moguće u razumnoj mjeri ispuniti drugim sredstvima." ○ pažljivog odvagivanja interesa, pri čemu se temeljna prava i slobode ispitanika odvažu u odnosu na legitimne interese voditelja obrade podataka. Također bi trebalo razmotriti razumna očekivanja ispitanika u pogledu uporabe njihovih podataka.[246]	✓	? [249]

[243] Kontrolni popis za umjetnu inteligenciju usklađen sa zaštitom podataka s kriterijima ispitivanja u skladu s Općom uredbom o zaštiti podataka. Bavarski državni ured za nadzor zaštite podataka

[244] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.

[245] U uvodnoj izjavi 39. Opće uredbe o zaštiti podataka pojašnjava se da bi se „osobni podaci trebali obrađivati samo ako se svrha obrade opravdano ne bi mogla postići drugim sredstvima".

[246] Europski odbor za zaštitu podataka, Izvješće o radu radne skupine ChatGPT

[249] Subjekti za uvođenje trebali bi provjeriti je li pružatelj djelotvorno otklonio taj rizik. Ova je preporuka jednako relevantna u slučajevima u kojima subjekti za uvođenje sudjeluju u modelima fine prilagodbe ili ponovnog osposobljavanja.

	▪ Uključiti službenika za zaštitu podataka u test ravnoteže, ako je primjenjivo.[247] ▪ Kad je riječ o izvlačenju podataka s interneta, procijeniti primjenjuje li se izuzeće iz članka 14. stavka 5. točke (b) te osigurati da su ispunjeni svi kriteriji kako bi se opravdalo neobavješćivanje svakog ispitanika pojedinačno. Transparentnost: ▪ Pružanje javnih i lako dostupnih informacija koje nadilaze zahtjeve iz članaka 13. i 14. Opće uredbe o zaštiti podataka, uključujući pojedinosti o kriterijima prikupljanja i korištenim skupovima podataka, s posebnim naglaskom na zaštiti djece i ranjivih pojedinaca. ▪ Upotrijebite inovativne pristupe za informiranje ispitanika, kao što su medijske kampanje, obavijesti e-poštom, grafičke vizualizacije, često postavljana pitanja, oznake transparentnosti, modeli kartica i dobrovoljna godišnja izvješća o transparentnosti.[248] ▪ Provesti popis izuzeća kojim upravlja voditelj obrade, čime se ispitanicima omogućuje da se usprotive prikupljanju svojih podataka s određenih internetskih stranica ili platformi pružanjem identifikacijskih informacija prije početka prikupljanja podataka.		
4. Nezakonita obrada posebnih kategorija osobnih podataka i podataka povezanih s kaznenim osudama i kažnjivim djelima u podacima za osposobljavanje.	▪ Za zakonitu obradu posebnih kategorija osobnih podataka osigurati primjenu iznimke iz članka 9. stavka 2. Opće uredbe o zaštiti podataka.[250] Pri pozivanju na članak 9. stavak 2. točku (e) potvrditi da je ispitanik jasnom potvrdnom radnjom izričito i namjerno stavio podatke na raspolaganje javnosti. Sama činjenica da su osobni podaci javno dostupni ne znači da je ispitanik očito objavio takve podatke.[251] ▪ S obzirom na izazove procjene od slučaja do slučaja u masovnom izvlačenju podataka s interneta, uvesti zaštitne mjere kao što je filtriranje kako bi se isključili podaci obuhvaćeni člankom 9. stavkom 1. Opće uredbe o zaštiti podataka tijekom i neposredno nakon prikupljanja podataka. ▪ Održavati pouzdanu dokumentaciju i dokaze o tim mjerama kako bi se ispunili zahtjevi u pogledu odgovornosti iz članka 5. stavka 2. i članka 24. Opće uredbe o zaštiti podataka.[252]		? [253]

[247] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.
[248] Idem
[250] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.: **„Europski**odbor za zaštitu podataka podsjeća na zabranu iz članka 9. stavka 1. Opće uredbe o zaštiti podataka u pogledu obrade posebnih kategorija podataka i na ograničene iznimke iz članka 9. stavka 2. Opće uredbe o zaštiti podataka. U tom je pogledu Sud Europske unije dodatno pojasnio da „ako se skup podataka koji sadržava i osjetljive i neosjetljive podatke [...] prikuplja u paketu, a da pritom nije moguće razdvojiti podatke u trenutku prikupljanja, obrada tog skupa podataka mora se smatrati zabranjenom u smislu članka 9. stavka 1. Opće uredbe o zaštiti podataka ako sadržava barem jedan osjetljiv podatak i ako se ne primjenjuje nijedno odstupanje iz članka 9. stavka 2. te uredbe". Nadalje, Sud EU-a naglasio je i da je „za potrebe primjene iznimke utvrđene u članku 9. stavku 2. točki (e) Opće uredbe o zaštiti podataka važno utvrditi je li ispitanik izričito i jasnom potvrdnom radnjom namjeravao predmetne osobne podatke učiniti dostupnima široj javnosti". Ta bi razmatranja trebalo uzeti u obzir pri obradi osobnih podataka u kontekstu modela umjetne inteligencije koja uključuje posebne kategorije podataka."
[251] Europski odbor za zaštitu podataka, Izvješće o radu radne skupine ChatGPT
[252] Idem
[253] Subjekti za uvođenje trebali bi provjeriti je li pružatelj djelotvorno otklonio taj rizik. Ova je preporuka jednako relevantna u slučajevima u kojima subjekti za uvođenje sudjeluju u modelima fine prilagodbe ili ponovnog osposobljavanja.

		✓	
5. Mogući negativan učinak na ispitanike koji bi mogao negativno utjecati na temeljna prava: Rezultati LLM-a mogli bi sadržavati pristrane i netočne informacije, čime bi se potencijalno kršila načela točnosti, transparentnosti, zakonitosti i pravednosti iz Opće uredbe o zaštiti podataka te bi se korisnike obmanjivalo da netočne rezultate smatraju činjenično pouzdanima. Sve bi to moglo negativno utjecati na ispitanike i njihova temeljna prava.	Transparentnost: ▪ Provesti snažne mjere transparentnosti kako bi se korisnike informiralo o vjerojatnosti rezultata LLM-a i njihovu potencijalu za pristranost ili netočnosti. Navedite izričite izjave o ograničenju odgovornosti da generirani sadržaj možda nije činjenično točan ili stvaran, čime se osigurava da korisnici razumiju ograničenja sustava. ▪ obavijestiti subjekte koji primjenjuju sustav i korisnike o tome hoće li njihovi osobni podaci utjecati na uslugu koja se pruža tom određenom korisniku ili će se upotrijebiti za izmjenu usluge koja se pruža svim korisnicima. ▪ Korisnici bi trebali odabrati LLM-ove s dokazanim mjernim podacima o učinkovitosti i kontinuirano pratiti izlazne podatke radi pogrešaka ili pristranosti kako bi se moglo preporučiti transparentno objavljivanje tih informacija. Točnost i pravednost: ▪ Razvojni programeri trebali bi osigurati kvalitetu podataka za učenje pouzdanim tehnikama prethodne obrade, kao što su filtriranje, validacija i normalizacija, kako bi spriječili pristrane ili obmanjujuće izlazne rezultate, trebali bi procijeniti i njihov učinak na izlazne rezultate umjetne inteligencije, ocijeniti statističku točnost modela za predviđenu namjenu i o tim razmatranjima transparentno obavijestiti subjekte za uvođenje i krajnje korisnike kako bi se ublažili mogući negativni učinci tijekom uvođenja.[254] Redovito preispitivati i dokumentirati sve korake poduzete radi usklađivanja s načelima transparentnosti i točnosti iz članka 5. stavka 1. točaka (a) i (d) Opće uredbe o zaštiti podataka.[255] ▪ Za platforme trećih strana učinkovita konfiguracija i upotreba dostupnih alata ključni su za poboljšanje rukovanja ulaznim podacima i osiguravanje da rezultati ispunjavaju standarde točnosti i pravednosti. ▪ Redovite revizije i mehanizmi nadzora ključni su za rješavanje rizika kao što su curenje podataka, pristranost ili nenamjerni zaključci. ▪ LLM-ovi bi se također mogli prilagoditi različitim jezičnim i kontekstualnim varijacijama, čime bi se smanjile netočnosti u osjetljivim aplikacijama. ▪ Kako bi se ublažio rizik od negativnih učinaka na ispitanike i temeljna prava u kontekstu dugoročnih jamstava, točnost[256] i pouzdanost moraju biti prioritet tijekom cijelog životnog ciklusa sustava. ▪ Osigurati da skupovi podataka za učenje budu raznoliki i reprezentativni za različite demografske skupine kako bi se smanjile pristranosti svojstvene podacima. ▪ Provoditi redovite revizije i ispitivanja pravednosti te uključiti ljudski pregled u osjetljive odluke kako bi se osigurala pravednost i odgovornost. ▪ Koristite okvire objašnjivosti kako biste analizirali i razumjeli kako se donose odluke, što pomaže u identificiranju potencijalnih izvora pristranosti.	✓	? [257]

[254] Službenik povjerenika za informiranje: https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-third-call-for-evidence/

[255] Idem

[256] https://aimodelcode.org/tech-info/llm-eval/

[257] Subjekti za uvođenje trebali bi provjeriti je li pružatelj djelotvorno otklonio taj rizik. Ova je preporuka jednako relevantna u slučajevima u kojima subjekti za uvođenje sudjeluju u modelima fine prilagodbe ili ponovnog osposobljavanja.

6. Nemogućnost ljudske intervencije za obradu koja može imati pravni ili važan učinak na ispitanika.	▪ Ljudski nadzor trebalo bi integrirati u postupke donošenja odluka u kojima bi rezultati LLM-ova mogli dovesti do pravnih ili znatnih posljedica za pojedince.[258] To uključuje osiguravanje da automatizirane odluke preispituje kvalificirano osoblje koje može procijeniti pravednost, točnost i relevantnost rezultata. ▪ Trebalo bi uspostaviti jasne postupke eskalacije za slučajeve u kojima se automatizirani rezultati čine nejasnima, pogrešnima ili potencijalno štetnima. ▪ Razvojni programeri i subjekti za uvođenje moraju osmisliti sustave za označivanje visokorizičnih rezultata za obveznu ljudsku intervenciju prije poduzimanja[259]bilo kakvih mjera. ▪ Trebalo bi uvesti i mehanizme transparentnosti[260]kojima bi se osiguralo da ispitanici budu obaviješteni o upotrebi LLM-ova, mogućnostima i ograničenjima modela,[261]obradi osobnih podataka s pomoću modela i njihovu pravu na osporavanje odluka ili traženje ljudskog preispitivanja. ▪ Redovito osposobljavanje osoblja uključenog u nadzor može dodatno poboljšati usklađenost i odgovornost. ▪ Provedba smjernica Radne skupine iz članka 29. o automatiziranom pojedinačnom donošenju odluka i izradi profila za potrebe Uredbe 2016/679, kako su zadnje revidirane i donesene 6. veljače 2018., koje je Europski odbor za zaštitu podataka potvrdio 25. veljače 2018. svibanj 2018. Vidjeti i presudu Suda od 7. prosinca 2023., predmet C-634/21, SCHUFA Holding i drugi (ECLI:EU:C:2023:957).	✓	✓
7. neodobravanje ispitanicima prava na prigovor, ispravak i brisanje.	▪ Pravo na prigovor na temelju članka 21. Opće uredbe o zaštiti podataka primjenjuje se i trebalo bi se osigurati ako je pravna osnova legitiman interes.[262] U tom bi slučaju pružatelji usluga trebali primijeniti mehanizme za dodjelu tog prava. Neke mjere koje treba provesti pri prikupljanju osobnih podataka mogle bi biti:[263] o Uvesti razumno razdoblje između prikupljanja skupa podataka za osposobljavanje i njegove upotrebe kako bi se ispitanicima dalo dovoljno vremena da ostvare svoja prava. o Osigurati bezuvjetan mehanizam izuzeća za ispitanike prije početka obrade. o Omogućiti ispitanicima da zatraže brisanje podataka, čak i izvan posebnih razloga navedenih u članku 17. stavku 1. Opće uredbe o zaštiti podataka. o Postupanje s odštetnim zahtjevima: Omogućiti ispitanicima da prijave slučajeve regurgitacije ili pamćenja osobnih podataka s pomoću mehanizama kojima voditelji obrade procjenjuju i primjenjuju tehnike neučenja za rješavanje takvih zahtjeva. ▪ Ublažavanje neusklađenosti s Općom uredbom o zaštiti podataka u pogledu prava ispitanika na ispravak i brisanje uključuje istraživanje tehnika strojnog neučenja.[264] Tim se pristupima nastoji ukloniti utjecaj podataka iz		

[258] https://www.lumenova.ai/blog/strategic-necessity-human-oversight-ai-systems/
[259] https://www.algomox.com/resources/blog/what_is_the_role_of_human_oversight_in_llmops/
[260] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10085432?mkt_tok=MTM4LUVaTS0wNDIAAAGX5pUM0HSpbBgVFc2wv7uGKk23174FM2-cFJBvVD0FDGJCM_27RuQFPm2uSB80ihorQ2e0YWwgCPRFngJDRE4b7N_pWRz873q84sJ8ZWucdQOh#english
[261] Europski odbor za zaštitu podataka, Izvješće o radu radne skupine ChatGPT
[262] Napominjemo da u skladu s člankom 21. stavkom 1. Opće uredbe o zaštitipodataka „voditelj obrade više ne smije obrađivati osobne podatke osim ako voditelj obrade dokaže da postoje uvjerljivi legitimni razlozi za obradu koji nadilaze interese, prava i slobode ispitanika ili radi postavljanja, ostvarivanja ili obrane pravnih zahtjeva."
[263] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.
[264] https://www.edpb.europa.eu/system/files/2025-01/d2-ai-effective-implementation-of-data-subjects-rights_en.pdf

	osposobljenog modela na zahtjev, čime se rješavaju problemi povezani s upotrebom podataka, unosima niske kvalitete ili zastarjelim informacijama. ○ Točan unlearning nastoji u potpunosti eliminirati utjecaj određenih točaka podataka, često putem prekvalifikacije ili naprednih metoda koje izbjegavaju potpunu prekvalifikaciju. Tehnike kao što su Sharded, Isolated, Sliced, and Aggregated (SISA) trening dijele podatke u podskupove, pojednostavljujući uklanjanje podataka i nastojeći održati robusnost modela. Približni pokušaji neučenja da se smanji utjecaj određenih točaka podataka prilagodbom pondera modela ili primjenom korekcijskih faktora, nudeći kompromis između preciznosti i učinkovitosti. Iako ove metode obećavaju, izazovi ostaju, uključujući održavanje točnosti modela i izbjegavanje nenamjernih pristranosti nakon učenja. Certificirano uklanjanje, koje pruža provjerljiva jamstva uklanjanja podataka pomoću matematičkih dokaza, nudi rigorozno, ali resursno intenzivno rješenje. Kako se tehnike neučenja razvijaju, one imaju ključnu ulogu u omogućavanju usklađenosti s Općom uredbom o zaštiti podataka, uz istodobno očuvanje integriteta i pravednosti modela strojnog učenja.[265] ▪ Provesti mehanizme za brisanje osobnih podataka, kao što su imena, osiguravajući da[266] je njihovo uklanjanje (blok) sveobuhvatno i kontekstualno u cijelom skupu podataka. Prepoznajte da bi taj pristup mogao dovesti do brisanja imena svih pojedinaca s istim identifikatorom, bez obzira na kontekst. Kako bi se ublažile neželjene posljedice, upotrijebite precizne tehnike filtriranja kako biste razlikovali kontekste u kojima se ime može osobno identificirati i kontekste u kojima je ono generičko. Da biste spriječili zlouporabu ili ponovno uvođenje izbrisanih podataka, osigurajte skripte ili upite za filtriranje ograničavanjem pristupa samo ovlaštenom osoblju, korištenjem šifriranja i održavanjem kontrole verzija. Redovito provjeravajte ove skripte kako biste bili sigurni da su ažurne i bez ranjivosti. ▪ Također je važno, posebno s obzirom na članak 21. Opće uredbe o zaštiti podataka, uspostaviti mehanizme za ispunjavanje zahtjeva korisnika koji se protive obradi njihovih osobnih podataka na temelju legitimnog interesa.[267] ▪ Kad je riječ o zahtjevima za brisanje u skladu s člankom 17. Opće uredbe o zaštiti podataka, procijeniti mogu li se osobni podaci izravno identificirati ili izvesti iz modela umjetne inteligencije i provesti tehničko brisanje ako je to izvedivo, kao što su prilagodbe nakon osposobljavanja.[268]	✓	? [269]
8. Nezakonita prenamjena osobnih podataka	▪ Osigurati usklađenost s člankom 5. stavkom 1. točkom (c) Opće uredbe o zaštiti podataka jasnim ograničavanjem obrade osobnih podataka na ono što je nužno za posebne, dobro definirane svrhe. Izbjegavajte preširoke svrhe kao što su „razvoj i poboljšanje UI sustava". Umjesto toga, navedite vrstu UI sustava (npr. veliki jezični model, generativna umjetna inteligencija za slike) i njegove tehnički izvedive funkcionalnosti i sposobnosti.[270] ▪ Člankom 6. stavkom 4. Opće uredbe o zaštiti podataka predviđaju se, za određene pravne osnove, kriteriji koje voditelj obrade uzima u obzir kako bi utvrdio je li obrada u drugu svrhu u skladu sa svrhom u koju se osobni podaci prvotno prikupljaju.[271]		

[265] https://www.edps.europa.eu/data-protection/our-work/publications/reports/2024-11-15-techsonar-report-2025_hr
[266] https://deveshsurve.medium.com/beginners-guide-to-llms-build-a-content-moderation-filter-and-learn-advanced-prompting-with-free-87f3bad7c0af
[267] Europski odbor za zaštitu podataka, Izvješće o radu radne skupine ChatGPT
[268] Kontrolni popis za umjetnu inteligenciju usklađen sa zaštitom podataka s kriterijima ispitivanja u skladu s Općom uredbom o zaštiti podataka. Bavarski državni ured za nadzor zaštite podataka
[269] Subjekti za uvođenje trebali bi provjeriti je li pružatelj djelotvorno otklonio taj rizik. Ova je preporuka jednako relevantna u slučajevima u kojima subjekti za uvođenje sudjeluju u modelima fine prilagodbe ili ponovnog osposobljavanja.
[270] https://www.cnil.fr/en/topics/artificial-intelligence-ai
[271] Mišljenje Europskog odbora za zaštitu podataka 28/2024 o određenim aspektima zaštite podataka povezanima s obradom osobnih podataka u kontekstu modela umjetne inteligencije. Doneseno 17. prosinca 2024.

	▪ Pri eksternalizaciji osposobljavanja u području umjetne inteligencije provjerite pravna jamstva (npr. ugovori, mjere prijenosa iz trećih zemalja) i osigurajte da pružatelji usluga ne upotrebljavaju podatke za učenje u neovlaštene svrhe.	✓	✓
9. Nezakonito neograničeno pohranjivanje osobnih podataka	Kao korisnik, subjekt za uvođenje i subjekt za nabavu sklapaju sporazume s trećim pružateljem usluga o tome koliko bi dugo trebalo pohranjivati ulazne i izlazne podatke. To može biti dio ugovora o uslugama, dokumentacije o proizvodu ili ugovora o obradi podataka. Ako se podaci pohranjuju u vašim prostorijama, uspostavite pravila zadržavanja i/ili mehanizam za brisanje podataka.	✓	✓
10. Nezakonit prijenos osobnih podataka	▪ Kao korisnik, subjekt za uvođenje i subjekt za nabavu, provjerite s dobavljačem gdje se odvija obrada podataka. ▪ Provesti potrebne zaštitne mjere i, prema potrebi, provesti procjenu učinka prijenosa podataka. ▪ Napravite potrebne ugovorne sporazume. ▪ Razmotrite ovaj rizik prilikom odabira između različitih dobavljača.	✓	✓
11. Kršenje[272] načela smanjenja količine podataka	▪ Redovito pregledavajte i eliminirajte nepotrebno prikupljanje podataka, automatizirajući brisanje podataka kada više nisu potrebni. ▪ Zamijenite podatke koji se mogu identificirati anonimiziranim ili pseudonimiziranim alternativama odmah nakon prikupljanja. ▪ Primijeniti načela privatnosti po dizajnu u svakoj razvojnoj fazi, integrirajući mjere minimizacije podataka. ▪ Isključiti prikupljanje podataka s internetskih stranica koje se protive mrežnom struganju (npr. upotrebom datoteka robots.txt ili ai.txt). ▪ Ograničiti prikupljanje na slobodno dostupne podatke koje su ispitanici očito objavili. ▪ Spriječiti kombiniranje podataka na temelju pojedinačnih identifikatora, osim ako je to izričito potrebno i opravdano za razvoj UI sustava.[273] ▪ Educirati korisnike o pružanju samo bitnih podataka u ulazima i transparentno komunicirati prakse korištenja podataka. ▪ Procijeniti je li obrada osobnih podataka nužna za predviđenu svrhu istraživanjem manje nametljivih alternativa, kao što je upotreba sintetičkih ili anonimiziranih podataka, te osigurati da je količina obrađenih osobnih podataka proporcionalna cilju.	✓	? [274]

[272] Obrada osobnih podataka radi uklanjanja mogućih pristranosti i pogrešaka dopuštena je samo ako je izričito usklađena s navedenom svrhom, a upotreba takvih podataka nužna je jer se cilj ne može učinkovito postići upotrebom sintetičkih ili anonimiziranih podataka. Člankom 10. stavkom 5. Akta o umjetnoj inteligenciji predviđena su posebna pravila za obradu posebnih kategorija osobnih podataka povezanih s visokorizičnim UI sustavima kako bi se osiguralo otkrivanje i ispravljanje pristranosti.

[273] CNIL: https://www.cnil.fr/en/legal-basis-legitimate-interests-focus-sheet-measures-implement-case-data-collection-web-scraping

[274] Subjekti za uvođenje trebali bi provjeriti je li pružatelj djelotvorno otklonio taj rizik. Ova je preporuka jednako relevantna u slučajevima u kojima subjekti za uvođenje sudjeluju u modelima fine prilagodbe ili ponovnog osposobljavanja.

7. Procjena preostalog rizika

Identificirati, analizirati i procijeniti preostali rizik

Iako se procjenom inherentnog rizika utvrđuju i ocjenjuju rizici prije primjene bilo kakvih kontrola, procjena preostalog rizika usmjerena je na preostali rizik nakon provedbe mjera ublažavanja. Procjena preostalih rizika pomaže organizacijama procijeniti učinkovitost implementiranih zaštitnih mjera i razumjeti potencijalni utjecaj neželjenih događaja.

Nakon dovršetka procjene izvedivosti i provedbe mjera ublažavanja ključno je ponovno procijeniti postoje li preostali rizici. Preostali[275] rizici su rizici koji i dalje postoje nakon primjene mjera ublažavanja.

Za analizu preostalog rizika ponovno se procjenjuju vjerojatnost i ozbiljnost preostalih rizika, čime se daje jasan pregled rizika koji ostaju nakon ublažavanja i uzima u obzir[276]sljedeće:

- Nalazi prethodnih evaluacija provedenih prije faze uvođenja
- Djelotvornost primijenjenih mjera ublažavanja
- Mogući rizici koji mogu nastati nakon raspoređivanja dobiveni praćenjem
- Novi rizici utvrđeni tijekom sastanaka za modeliranje prijetnji

Nakon što se utvrde preostali rizici, organizacije moraju odlučiti jesu li ti rizici unutar prihvatljivih razina definiranih njihovim kriterijima tolerancije[277] i prihvaćanja rizika. Ako se preostali rizici smatraju prihvatljivima, mogu se službeno potvrditi i dokumentirati u registru rizika. Međutim, ako rizici premaše prihvatljive razine, potrebno je istražiti i provesti daljnje mjere ublažavanja te ih dokumentirati. Postupak se zatim vraća u fazu liječenja rizika kako bi se utvrdila najprikladnija opcija liječenja za rizik.

Procjena preostalog rizika također ima ulogu u odluci o puštanju sustava u proizvodnju. Stoga je važno procijeniti jesu li rizici i dalje unutar utvrđenih sigurnosnih pragova. Organizacije tada mogu odlučiti zatražiti daljnje ispitivanje ili dodatne evaluacije, naložiti daljnje ublažavanje ili odobriti model za uvođenje ako je preostali rizik prihvatljiv.

[275] https://csrc.nist.gov/glossary/term/residual_risk
[276] Vidjeti bilješku 193.
[277] ISO 31000 Upravljanje rizicima

8. Pregled & Monitor

Preispitivanje postupka upravljanja rizikom

Preispitivanje postupka upravljanja rizicima ključno je kako bi se osiguralo da su planirane aktivnosti pravilno provedene te da su kontrole rizika i mjere ublažavanja rizika djelotvorni. To se može učiniti strukturiranim preispitivanjem plana upravljanja rizicima,[278] koji služi kao plan za utvrđivanje, procjenu i ublažavanje rizika tijekom cijelog projekta. Iako nisu obvezni, takvi planovi i njihove revizije najbolja su praksa u upravljanju rizicima, posebno u normama za proizvode na koje utječu sigurnosni propisi, kao što su medicinski proizvodi.[279]

Pregled upravljanja rizicima pomaže utvrditi:

- ✓ Planirane kontrole rizika učinkovito su provedene
- ✓ Utvrđeni su i riješeni novi rizici
- ✓ Plan je i dalje usklađen s ciljevima i propisima projekta

Redoviti pregledi također pomažu poboljšati strategije rizika, poboljšati procese i prilagoditi se promjenama u zakonodavstvu, poslovanju ili strukturama tima.

Registar rizika dokumenata

Dokumentiranje procjena rizika, mjera ublažavanja i preostalih rizika tijekom cijelog životnog ciklusa ključno je za osiguravanje odgovornosti, usklađenosti i kontinuiranog poboljšanja. Ključni je alat za taj postupak registar rizika,[280] strukturirani dokument koji djeluje kao središnji repozitorij za sve utvrđene rizike, uključujući pojedinosti kao što su priroda rizika, vlasništvo, rezultati evaluacije, pragovi i mjere ublažavanja. Ta dokumentacija podupire regulatornu usklađenost s okvirima kao što su Opća uredba o zaštiti podataka i Akt o umjetnoj inteligenciji, olakšava revizije i omogućuje informirano donošenje odluka. Dobro održavan registar rizika može pomoći timovima da učinkovito vizualiziraju i odrede prioritete rizika. Praćenjem preuzimanja rizika i napretka u ublažavanju rizika osigurava se da se ne zanemare kritični rizici te da se tijekom cijelog postupka upravljanja rizicima održava odgovornost.

Kontinuirano praćenje

Nakon što[281] se provedu mjere za smanjenje rizika, kontinuirano praćenje ključno je za procjenu njihove učinkovitosti i utvrđivanje svih novih rizika. Nakon uvođenja praćenje nakon stavljanja na tržište[282] ima ključnu ulogu u utvrđivanju novih rizika ili promjena u operativnom okruženju koje mogu utjecati na privatnost. To uključuje sustavno prikupljanje i analizu zapisa i drugih operativnih podataka u skladu sa zahtjevima Opće uredbe o zaštiti podataka, osiguravajući transparentnost, odgovornost i stalnu zaštitu korisničkih podataka.

Trenutačno se praćenje LLM-ova tijekom cijelog životnog ciklusa prvenstveno oslanja na sljedeće tehnike: testiranje i evaluaciju[283]modela, crveno udruživanje, terensko testiranje i dugoročnu procjenu učinka. Ove metode pomažu identificirati i procijeniti nove rizike koji možda nisu bili očiti tijekom početnog razvoja.

[278] https://www.tradeready.ca/explainer/how-often-should-you-review-your-risk-management-plan/
[279] https://compliancenavigator.bsigroup.com/en/medicaldeviceblog/risk-management-plans-and-the-new-iso-14971/
[280] https://en.wikipedia.org/wiki/Risk_register
[281] „mjere upravljanja rizikom" u Aktu o umjetnoj inteligenciji.
[282] Poglavlje IX. odjeljak 1. Praćenje nakon stavljanja na tržište, Akt o umjetnoj inteligenciji
[283]Međunarodno izvješće o sigurnosti umjetne inteligencije, 2025., stranica 184.
https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

- Testiranje modela i iterativne evaluacije koriste se prije i nakon uvođenja modela i dio su LLM sustava. Iako su ključne, same po sebi nisu dovoljne zbog nepredvidivosti scenarija u stvarnom svijetu i subjektivnosti određenih rizika. Budući da se LLM-ovi mogu primijeniti u brojnim kontekstima, teško je predvidjeti kako će se rizici manifestirati u praksi te, kako je navedeno u odjeljku 2., pokazatelji uspješnosti i referentne vrijednosti možda neće uvijek točno odražavati te stvarne rizike.
- Metodologije kao što je „crveno udruživanje"[284] mogu se upotrebljavati za testiranje otpornosti modela na stres prije uvođenja i sustava LLM prije i nakon njegova uvođenja simuliranjem neprijateljskih napada ili scenarija zlouporabe,[285]čime se pomaže u otkrivanju ranjivosti koje možda nisu utvrđene tijekom faze razvoja.
- Terenskim testiranjem procjenjuju se rizici povezani s umjetnom inteligencijom u stvarnim uvjetima, ali njezina je provedba i dalje zahtjevna zbog poteškoća u točnom repliciranju scenarija iz stvarnog svijeta i utvrđivanju jasnih parametara uspjeha. Važno je stvoriti reprezentativno testno okruženje i definirati mjerljive referentne vrijednosti uspješnosti kako bi se dobili pouzdani uvidi.
- U dugoročnim procjenama učinka ocjenjuje se kako se UI sustavi razvijaju tijekom vremena kako bi se utvrdile neželjene posljedice koje se mogu pojaviti s produljenim uvođenjem. Kontinuirano praćenje i periodična ponovna ocjenjivanja ključni su za otkrivanje promjena u ponašanju modela, smanjenja učinkovitosti ili novih rizika koji možda nisu bili očiti tijekom početnog ispitivanja. Ta je tehnika dio strategije kontinuiranog praćenja i može biti dio sesija modeliranja prijetnji.

U svim tim tehnikama ključno je definirati robusne i pouzdane mjerne podatke za praćenje. Međutim, trenutačne automatizirane procjene i kvantitativni parametri često nisu pouzdani ni valjani,[286]zbog čega je teško djelotvorno procijeniti rizike. Iz tog razloga, kvalitativna ljudska revizija također igra ključnu ulogu u hvatanju širih sociotehničkih implikacija LLM-a i s njima povezanih rizika.

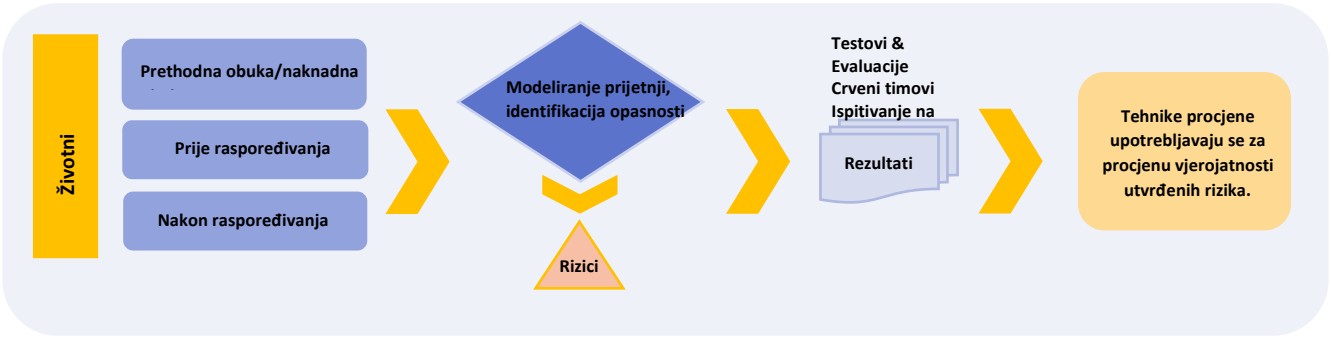
Mehanizam za odgovor na incidente

Kako bi se osigurala učinkovita strategija upravljanja rizicima, važno je provesti i mehanizme za odgovor na incidente kojima se omogućuje pravodoban i primjeren odgovor na upozorenja i upozorenja generirana evaluacijama i praćenjem jer oni mogu upućivati na potencijalni incident povezan s privatnošću ili zaštitom podataka.

Upozorenja koja proizlaze iz različitih tehnika praćenja ključna su ne samo za praćenje nakon stavljanja na tržište nego i tijekom cijelog životnog ciklusa umjetne inteligencije. Tehnike, opseg ispitivanja i rezultati razlikovat će se ovisno o tome provode li se evaluacije prije ili nakon osposobljavanja modela te prije ili nakon uvođenja modela i sustava.
Iako su rezultati važni za utvrđivanje novih rizika, mogu imati i ključnu ulogu u procjeni vjerojatnosti pojave utvrđenih prijetnji ili opasnosti. Time se osigurava kvantitativna analiza koja se može usporediti s utvrđenim prihvatljivim pragovima kako bi se lakše utvrdilo jesu li potrebna daljnja smanjenja rizika.

[284] https://openai.com/index/advancing-red-teaming-with-people-and-ai/
[285] https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai
[286] https://aisingapore.org/ai-governance/current-llm-evaluations-do-not-sufficiently-measure-all-we-need/

Slika 16. Rizici se utvrđuju tehnikama procjene



Slika 17. Tehnike procjene upotrebljavaju se za procjenu vjerojatnosti utvrđenih rizika.

Iterativno upravljanje rizicima

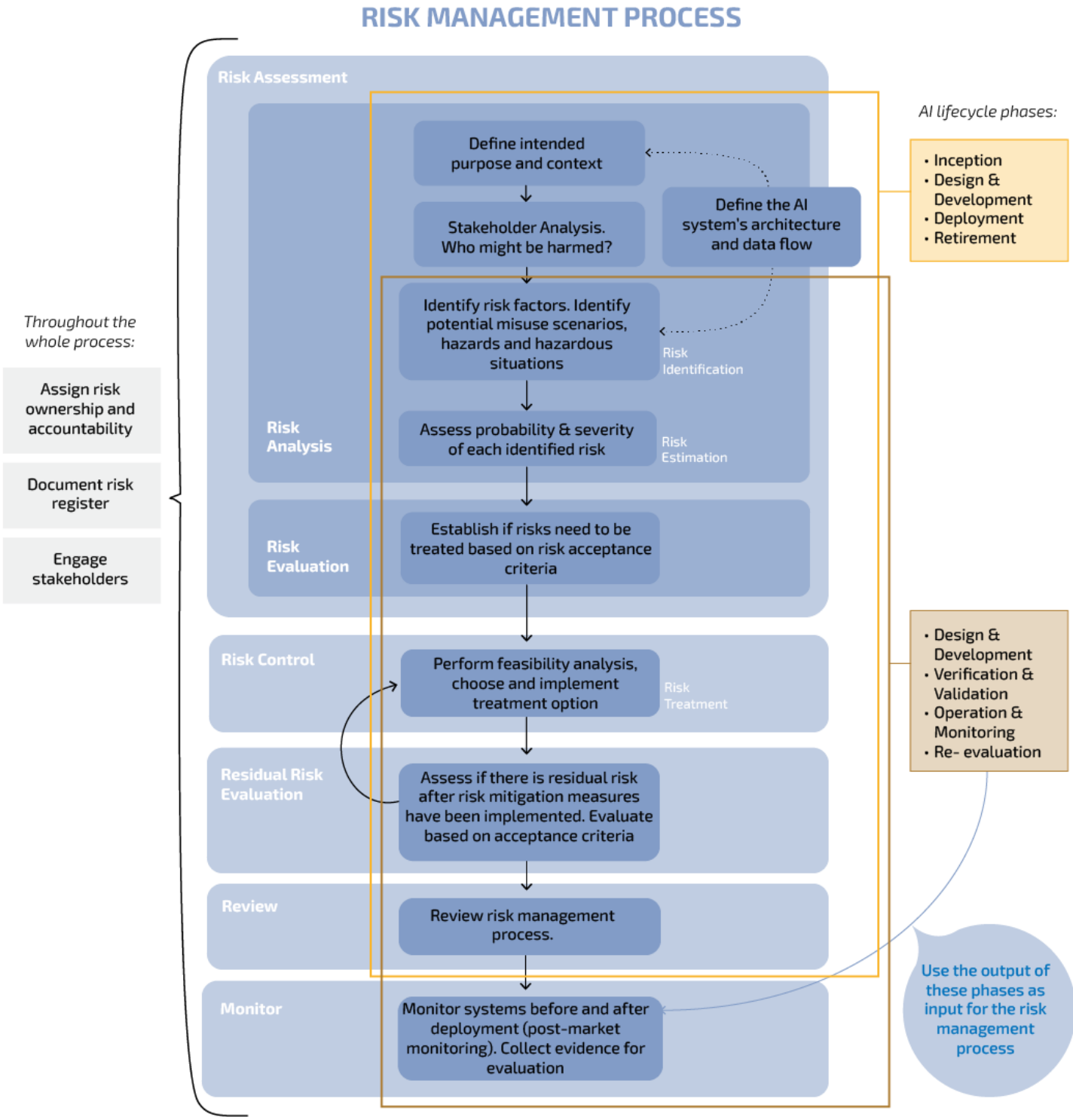
Učinkovito upravljanje rizicima za dugotrajna radna mjesta mora se temeljiti na iterativnom pristupu koji obuhvaća cijeli životni ciklus sustava, od projektiranja i razvoja do uvođenja, praćenja i eventualnog stavljanja izvan pogona. Budući da se rizici mogu mijenjati s promjenama u prilagodbi, ažuriranju sustava ili novim kontekstima primjene, redovita evaluacija i prilagodba ključne su za uklanjanje novih prijetnji. Ljudski nadzor, u kombinaciji s automatiziranim mjerama, ključan je za učinkovito upravljanje rizicima u LLM sustavima. Iako automatizirani alati, kao što su okviri za praćenje, registri rizika i sustavi bilježenja, omogućuju kontinuirano praćenje i analizu, ljudskim nadzorom osiguravaju se odgovornost, pravednost i transparentnost. Taj je hibridni pristup ključan u kontekstu dugotrajnog rada u kojem bi potpuno ručni nadzor bio nepraktičan zbog složenosti i razmjera sustava.

Kako bi se dodatno ojačalo upravljanje rizicima, alati kao što su LLMOps [287] (LLM Operations) i LLMSecOps[288] (LLM Security Operations) mogu automatizirati i integrirati mnoge aspekte upravljanja rizicima, osiguravajući besprijekorno ažuriranje, praćenje i odgovor na ranjivosti. Ti alati poboljšavaju tijekove rada za praćenje i ublažavanje rizika, smanjuju teret ručne dokumentacije i poboljšavaju opću sigurnost i upravljanje LLM[289] sustavima.

[287] ttps://www.databricks.com/glossary/llmops
[288] https://medium.com/@bijit211987/llmsecops-elevating-security-beyond-mlsecops-94396768ecc6
[289] All Tech is Human x IBM Research, „AI Governance Workshop 2025”, https://static1.squarespace.com/static/60355084905d134a93c099a8/t/677c492a161e58148fc60706/1736198443181/IBM+Research+x+ATIH+AI+Governance+Workshop.pdf

Pregled postupka upravljanja rizicima

Na slici u nastavku prikazan je vizualni sažetak postupka upravljanja rizicima koji je opisan u ovom izvješću. Iako je svaka faza detaljno objašnjena, ovaj pregled pomaže istaknuti ključne korake i njihove interakcije, pružajući jasnije razumijevanje cijelog tijeka rada.

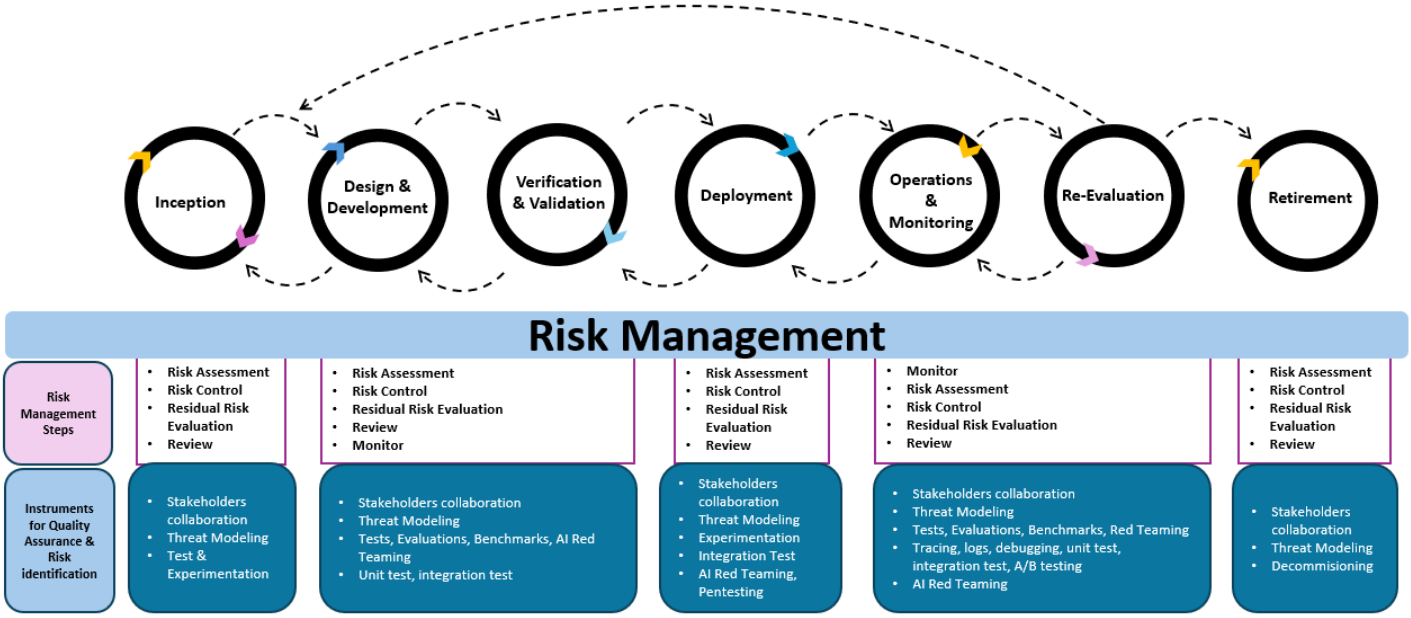


Slika 18. Proces upravljanja rizicima

Proces životnog ciklusa upravljanja rizicima

Slika u nastavku prikazuje okvir za upravljanje rizicima koji se primjenjuje tijekom cijelog životnog ciklusa sustava koji se temelji na LLM-u. Za svaku fazu dijagram prikazuje odgovarajuće korake upravljanja rizicima iz postupka upravljanja rizicima. Tim se koracima osigurava kontinuirano utvrđivanje i ublažavanje rizika kako se sustav bude razvijao.

Osim toga, na slici su istaknuti instrumenti za osiguranje kvalitete i utvrđivanje rizika specifični za svaku fazu. To uključuje prakse kao što su suradnja dionika, modeliranje prijetnji, testiranje i crveno udruživanje u području umjetne inteligencije. Tim se slojevitim pristupom podupire proaktivna i iterativna strategija upravljanja rizicima.



Slika 19. Životni ciklus upravljanja rizicima

9. Primjeri procjena rizika LLM sustava

Slučaj prve uporabe: Virtualni pomoćnik (Chatbot) za upite kupaca

Slika 20. Izvor: Dizajniran od pch.vector/Freepik

Scenarij: Tvrtka specijalizirana za kuhinjsku opremu želi implementirati chatbot kako bi svojim klijentima pružila opće informacije o svojim proizvodima i uslugama. Chatbot će imati pristup postojećim podacima o klijentima integracijom sa sustavom za upravljanje klijentima (npr. bazama podataka CRM-a). To će chatbotu omogućiti prepoznavanje korisnika na temelju identifikatora kao što su vjerodajnice e-pošte ili računa i pružanje personaliziranih odgovora bez potrebe da korisnici ponovno unesu svoje podatke. Ovo chatbot sučelje bit će izgrađeno na temelju lako dostupnog LLM-a koji će koristiti RAG za stjecanje potrebnog znanja specifičnog za domenu.

Faza životnog ciklusa u kojoj se nalazimo: Dizajn & Razvoj

Proces upravljanja rizicima

Ovaj odjeljak pruža pregled koraka za provedbu upravljanja rizikom privatnosti za rješenja temeljena na LLM-u. Započinjemo ispitivanjem potencijalnih tokova podataka i arhitekture sustava za naš slučaj upotrebe, nakon čega slijede koraci za učinkovito prepoznavanje i rješavanje rizika za privatnost.

Pregled protoka podataka – na čemu radimo?

Očekivani protok podataka za obradu opisan je kako slijedi:

1. Unos korisnika → Korisnici će stupiti u interakciju s chatbotom izravno nakon prijave u tvrtku tako što će putem sučelja (npr. internetske stranice ili mobilne aplikacije) navesti svoje ime, adresu e-pošte i preferencije.
2. Predobrada podataka i interakcija s API-jem → Unos korisnika potvrdit će se i formatirati prije slanja na obradu API-ju chatbota. Chatbot će komunicirati s fino podešenim LLM-om smještenim na oblaku.
3. Retrieval-Augmented Generation (RAG) (generacija uvećana za dohvaćanje)
Za upite za koje je potrebno znanje specifično za određeno područje ili ažurirani kontekst sustav provodi korak dohvaćanja: pretražuje CRM, bazu dokumenata ili bazu znanja društva kako bi pronašao relevantne informacije. Dobiveni sadržaj zatim se kombinira s korisničkim unosom i prosljeđuje LLM-u kako bi se generirao utemeljeni, personalizirani odgovor.
4. Pre-Fine-Tuned LLM obrada →
Chatbot koristi fino podešen LLM obučen za podatke specifične za poduzeće kako bi poboljšao opće razumijevanje jezika i usklađivanje tonova. Ovaj LLM koristi obogaćene ulazne podatke (od korisnika + RAG-a) za personalizaciju izlaznih podataka**.**
5. Pohrana podataka → Unaprijed obrađeni korisnički unos (npr. postavke) pohranit će se lokalno ili u oblaku kako bi se omogućile personalizirane preporuke i olakšale buduće interakcije.

6. Generiranje personaliziranih odgovora → Chatbot će koristiti pohranjene korisničke podatke iz CRM sustava i prilagođene LLM mogućnosti za generiranje prilagođenih preporuka i odgovora.
7. Dijeljenje podataka → Chatbot može dijeliti minimalne (anonimizirane) korisničke podatke s vanjskim uslugama (npr. API-ji trećih strana za dodatne funkcionalnosti ili promotivne alate).
8. Zbirka povratnih informacija → Korisnici pružaju povratne informacije o interakcijama chatbota (npr. podizanje/spuštanje palčeva, komentari) kako bi se poboljšala učinkovitost sustava. To je proces sustava u analitičke svrhe.
9. Brisanje i upravljanje korisničkim pravima → Korisnici mogu zatražiti pristup, brisanje ili ažuriranje svojih osobnih podataka u skladu s GDPR-om ili sličnim propisima.

Kako bi se olakšao postupak procjene rizika, moguće je izraditi i dijagram toka podataka koji[290]pruža grafički prikaz postupaka, kretanja podataka i interakcija unutar sustava.

Moguća arhitektura

S obzirom na to da smo u fazi projektiranja životnog ciklusa umjetne inteligencije, očekujemo da će arhitektura našeg sustava koji se temelji na LLM-u uključivati sljedeće ključne slojeve:[291]

- **Sloj korisničkog sučelja (UI):** Sučelje na kojem korisnici komuniciraju s chatbotom putem tekstualnog ili glasovnog unosa. (npr. Web stranica, mobilna aplikacija)
- **Chatbot aplikacijski sloj:** Upravljajte tokom razgovora i određujete chatbot odgovore na temelju korisničkog unosa i konteksta. Usmjerava upite na Business Logic Layer.
- **Poslovni logički sloj:** Orkestrira tijekove rada chatbota, kao što je provjera profila kupaca ili naručivanje. Ključno je odlučiti hoće li nazvati LLM izravno ili pokrenuti korak dohvaćanja (RAG) – na primjer, pretraživanjem CRM-a ili baze znanja kada je potreban dodatni kontekst prije generiranja odgovora.
- **Integracijski sloj:** Sadrži API Gateway za upravljanje prijenosom između slojeva. Povezuje chatbot s LLM-om, CRM sustavom i vanjskim uslugama te olakšava sigurnu komunikaciju i razmjenu podataka između sustava. Također obrađuje transformaciju podataka, osigurava kompatibilnost između chatbota i CRM-a te provodi autentifikaciju i autorizaciju za siguran pristup CRM podacima. Kad je riječ o uspostavi RAG-a, taj sloj može usmjeravati upite i na komponentu za dohvaćanje ili bazu znanja prije nego što obogaćene ulazne podatke proslijedi LLM-u.
- **LLM sloj:** Izvodi prirodno razumijevanje jezika i generacije. Prima sirove ulazne podatke korisnika ili ulazne podatke obogaćene vraćenim sadržajem (iz koraka RAG-a). Vraća kontekstualno relevantne odgovore na Business Logic Layer.
- **CRM sustav:** Pohranjuje podatke o kupcima, kao što su podaci za kontakt, povijest kupnji, postavke i tikete za podršku. Sadrži i CRM API-je koji pružaju krajnje točke za dohvaćanje, ažuriranje ili dodavanje korisničkih podataka i rukovatelja događajima koji pokreću radnje na temelju događaja, kao što je stvaranje tiketa za podršku kada korisnik postavi problem putem chatbota. Dostavlja podatke o klijentima za personalizaciju odgovora chatbota i pohranjuje podatke generirane tijekom interakcija.
- **Sloj vanjskih usluga:** Integrira se s analitičkim alatima za praćenje interakcija korisnika i generiranje uvida u ponašanje kupaca. Također se integrira s drugim uslugama, kao što su pristupnici za plaćanje, usluge e-pošte ili marketinški alati.

[290] https://en.wikipedia.org/wiki/Data-flow_diagram
[291] Ta se arhitektura navodi kao pojednostavnjeni primjer i može se znatno razlikovati ovisno o posebnim zahtjevima, primjeru upotrebe i tehničkim ograničenjima svakog uvođenja.

- **Sigurnosni sloj:** Šifrira podatke tijekom prijenosa pomoću protokola kao što su HTTPS i SSL/TLS, ograničava neovlašteni pristup chatbotu, LLM-u i CRM-u pomoću tehnika kao što je OAuth2, provodi kontrole sigurnosti i privatnosti, skeniranje ranjivosti, praćenje prijetnji itd.

Pregled moguće arhitekture u ovoj fazi omogućuje jasnije razumijevanje protoka podataka i potencijalnih rizika povezanih s uvođenjem chatbota. Ovaj arhitektonski uvid postavlja temelje za prepoznavanje pitanja privatnosti i sigurnosti u ranoj fazi procesa.

Analiza rizika – suradnja dionika

Sljedeći korak uključuje okupljanje raznolike skupine dionika radi zajedničkog utvrđivanja potencijalnih rizika. Pozivanje pravih dionika nije egzaktna znanost, ali je ključno uključiti pojedince koji će imati ovlasti za donošenje odluka, izravno sudjelovati u njezinu razvoju, uvođenju i upotrebi te koji bi mogli dodati vrijednost postupku utvrđivanja rizika. Među ključnim sudionicima mogli bi biti predstavnici inženjerskih, sigurnosnih, privatnih i UX dizajnerskih timova. Ako je moguće, vrlo je korisno uključiti pojedince sa stručnim znanjem u području etike i temeljnih prava, kao i članove skupina civilnog društva, subjekte za uvođenje i predstavnike krajnjih korisnika (klijente u našem slučaju upotrebe). Prikupljanje informacija od šire publike putem ankete za klijente također može pružiti vrijedan uvid u očekivanja i bojazni korisnika.

Analiza dionika

Prije početka postupka identifikacije rizika skupina bi trebala analizirati slučaj upotrebe kako bi utvrdila koji će dionici ciljana skupina komunicirati s chatbotom i identificirati one koji ne bi trebali imati pristup. Ako je potrebno, osmišljavanjem prepreka, kao što je mehanizam za provjeru dobi, osigurava se usklađenost sustava s predviđenom bazom korisnika. U ovom konkretnom slučaju upotrebe ulazna točka sučelja ograničena je na prijavljene i priznate kupce, zbog čega dodatne prepreke mogu biti nepotrebne. Međutim, sveobuhvatna procjena svih potencijalnih rizika i dalje je ključna za uspjeh sustava.

Analiza dionika postupak[292] je koji se upotrebljava za utvrđivanje i razumijevanje uloga, interesa i utjecaja različitih dionika koji su uključeni u projekt ili na koje projekt utječe. Osim analize onih koji su izravno uključeni u sustav, jednako je važno procijeniti na koje bi dionike alat mogao negativno utjecati. To uključuje prepoznavanje bi li mogle biti uključene ranjive skupine ili bi li se učinak alata mogao proširiti na velik broj pojedinaca. Prema potrebi, moglo bi biti korisno uključiti pogođene zajednice u sljedeće faze utvrđivanja rizika kako bi se bolje obuhvatili problemi i učinci specifični za kontekst. Alati za sudjelovanje,[293] kao što je etička matrica, navedeni u prethodnom odjeljku, mogu pomoći u procjeni mogućih posljedica za različite skupine dionika.

U našem slučaju korištenja, identificirali smo naše *klijente* kao jedine ovlaštene korisnike. S obzirom na prirodu našeg poslovanja, ne očekujemo da će djeca pristupiti našoj platformi. Međutim, i dalje vodimo računa o provedbi odgovarajućih sigurnosnih mjera kako bismo osigurali da je pristup ograničen i da je neovlaštena uporaba spriječena.

Utvrđivanje rizika – odabir čimbenika rizika

Uzimajući u obzir naš dijagram toka podataka, arhitekturu sustava i rasprave s odabranim dionicima, utvrdili smo sljedeće čimbenike rizika i ključna pitanja iz odjeljka 4. koji su primjenjivi na naš konkretan slučaj upotrebe:

[292] https://simplystakeholders.com/stakeholder-impact-analysis/
[293] https://partnershiponai.org/stakeholder-engagement-for-responsible-ai-introducing-pais-guidelines-for-participatory-and-inclusive-ai/

Faktor rizika	Primjenjivost slučaja upotrebe
Obrada velikih razmjera	Značajna količina podataka bit će obrađena zbog naše opsežne baze podataka o kupcima i velike količine informacija pohranjenih u našem CRM sustavu.
Niska kvaliteta podataka	Ulazni podaci za upite korisnika mogu biti niske kvalitete, a baza podataka CRM-a nije validirana, što bi moglo dovesti do netočnosti ili neučinkovitosti u obradi.
Nedostatne sigurnosne mjere	Postoji potencijalni rizik od prijenosa osobnih podataka u zemlje bez odgovarajuće razine zaštite, posebno ako je model LLM smješten ili se održava u takvim regijama.

Identificirali smo nekoliko čimbenika rizika koji zahtijevaju pozornost, jer ukazuju na veću vjerojatnost neželjenih ishoda. Iako naš sustav nije obuhvaćen klasifikacijom visokorizičnog sustava u skladu s Aktom o umjetnoj inteligenciji, iz perspektive Opće uredbe o zaštiti podataka postoji dovoljno dokaza koji opravdavaju pokretanje[294] postupka izrade procjene učinka na zaštitu podataka. Procjena rizika koju sada provodimo poslužit će kao vrijedan temelj za postupak procjene učinka na zaštitu podataka. Važno je naglasiti kada je potrebna procjena učinka na zaštitu podataka i kada je potrebna procjena učinka na temeljna prava. Procjena učinka na zaštitu podataka u skladu s člankom 35. Opće uredbe o zaštiti podataka potrebna je kad god je vjerojatno da će obrada podataka prouzročiti visok rizik za prava i slobode pojedinaca.[295]

Uobičajeni scenariji za koje je potrebna procjena učinka na zaštitu podataka uključuju:

- Korištenje novih tehnologija koje bi mogle dovesti do rizika za privatnost.
- Praćenje ponašanja ili lokacije pojedinca (npr. geolokacijske usluge ili bihevioralno oglašavanje).
- opsežno praćenje javno dostupnih mjesta (npr. videonadzor).
- Obrada osjetljivih kategorija podataka kao što su rasno ili etničko podrijetlo, politička mišljenja, vjerska uvjerenja, genetski podaci, biometrijski podaci ili zdravstvene informacije.
- Automatizirano donošenje odluka koje ima pravne ili slične značajne učinke na pojedince.
- Obrada podataka o djeci ili bilo kojih podataka ako bi povreda mogla prouzročiti tjelesnu štetu.

Čak i ako procjena učinka na zaštitu podataka nije izričito propisana zakonom, provedba procjene može biti razborita u pogledu najboljih praksi u području privatnosti i sigurnosti. Omogućuje organizacijama da preventivno reagiraju na potencijalne rizike, procijene učinak svojih rješenja i pokažu odgovornost. S druge strane, FRIA, kako je navedeno u članku 27. Akta[296]o umjetnoj inteligenciji, može biti obvezna za neke subjekte koji primjenjuju visokorizične UI sustave (javnopravna tijela, privatni subjekti koji pružaju javne usluge, organizacije koje provode procjene kreditne sposobnosti, određivanje cijena i procjene rizika životnog i zdravstvenog osiguranja). U procjeni učinka na temeljna prava procjenjuje se mogući učinak takvih sustava na temeljna prava kao što su privatnost, pravednost i nediskriminacija. Subjekti koji primjenjuju visokorizične UI sustave moraju dokumentirati:

- Kako će se sustav koristiti, uključujući njegovu svrhu, trajanje i učestalost.
- Kategorije pojedinaca ili skupina na koje sustav utječe.
- Posebni rizici od štete za temeljna prava.
- Mjere za ljudski nadzor i upravljanje.

[294] https://www.edpb.europa.eu/sme-data-protection-guide/be-compliance_hr
[295] Smjernice o procjeni učinka na zaštitu podataka i utvrđivanje može li obrada „vjerojatno dovesti do visokog rizika" za potrebe Uredbe 2016/679, WP248 rev.01, koje je odobrio Europski odbor za zaštitu podataka: https://ec.europa.eu/newsroom/article29/items/611236
[296] Članak 27. stavak 1. Akta o umjetnoj inteligenciji: „Prije uvođenja visokorizičnog UI sustava iz članka 6. stavka 2., uz iznimku visokorizičnih UI sustava namijenjenih za uporabu u području navedenom u točki 2. Priloga III., subjekti koji primjenjuju sustav, a koji su javnopravna tijela ili privatni subjekti koji pružaju javne usluge, i subjekti koji primjenjuju visokorizične UI sustave iz točke 5. podtočaka (b) i (c) Priloga III. procjenjuju učinak koji uporaba takvog sustava može imati na temeljna prava..."

- Koraci za rješavanje i ublažavanje rizika ako se ostvare.
- Procjena učinka na zaštitu podataka može se, prema potrebi, dopuniti procjenom učinka na zaštitu podataka usmjeravanjem na šire društvene učinke koji nadilaze samo zaštitu podataka.

Utvrđivanje rizika – što može poći po zlu?

U postupku utvrđivanja rizika slijedit ćemo pristup sličan onome koji se primjenjuje u metodologijama modeliranja prijetnji. Modeliranje prijetnji privatnosti služi kao strukturirani način utvrđivanja potencijalnih rizika i ranjivosti unutar sustava, usredotočujući se na scenarije koji bi mogli utjecati na prava i slobode ispitanika. Iako modeliranje prijetnji nije obvezno i ne zamjenjuje procjenu učinka[297] na zaštitu podataka ili šire postupke upravljanja rizicima, ono je svestran i učinkovit alat kojim se jačaju ti napori stvaranjem vrijednih uvida, kao što su scenariji rizika i potencijalni učinci.
Blisko surađujući s našom skupinom dionika i nakon što smo utvrdili čimbenike rizika koji utječu na naš slučaj upotrebe, sada ćemo ponovno pregledati dijagram toka podataka i olakšati postupak utvrđivanja rizika upotrebom dvaju popisa rizika iz odjeljaka 3. i 4.
Umjesto oslanjanja na biblioteke vanjskih rizika[298] sadržane u metodologijama modeliranja prijetnji kao što su LINDDUN[299], LIINE4DU[300] ili PLOT4AI[301], za naš slučaj upotrebe upotrebljavat ćemo prilagođene popise rizika navedene u ovom dokumentu. Važno je imati na umu da je korištenje unaprijed definirane zbirke rizika samo početna točka. Organizacije bi trebale razmišljati izvan tih popisa, uzimajući u obzir jedinstvene aspekte svojih sustava i konteksta i pitajući se: "Što još može poći po zlu?".

U modeliranju prijetnji proces se obično[302] temelji na četiri temeljna pitanja:

1. Na čemu radimo?
2. Što može poći po zlu?
3. Što ćemo učiniti u vezi toga?
4. Jesmo li dobro obavili posao?

U ovom slučaju, integrirat ćemo ova pitanja u proces upravljanja rizikom kako slijedi:

Na čemu radimo?
Implementiramo chatbot temeljen na LLM-u. Kako bismo razumjeli njegov dizajn i arhitekturu, koristimo se za sesiju utvrđivanja rizika s našom skupinom dionika, najnovijom verzijom dijagrama toka podataka. Također moramo uzeti u obzir da smo faza projektiranja i razvoja životnog ciklusa umjetne inteligencije, što znači da još nisu provedena mnoga testiranja i evaluacije, a nedostaju nam povratne informacije korisnika i uvidi iz proizvodnog okruženja.

Što može poći po zlu?
Nakon pregleda našeg dijagrama toka podataka, konteksta uporabe, namjene aplikacije, karakteristika naše korisničke skupine i dizajna, kao i svih rezultata dobivenih evaluacijama i testovima, utvrdili smo sljedeće rizike navedene u odjeljku 4.:

1. Nedovoljna zaštita osobnih podataka koja dovodi do povrede podataka.
2. Pogrešna klasifikacija podataka o osposobljavanju kao anonimnih.
3. Mogući negativan učinak na ispitanike koji bi mogao negativno utjecati na temeljna prava.
4. Neodobravanje prava ispitanika.

[297] AEPD, „Tehnička napomena: Uvod u LIINE4DU 1.0: A New Privacy & Data Protection Threat Modelling Framework (Okvir za modeliranje prijetnji zaštiti podataka), 2024.
[298] Slattery P. i dr., The AI Risk Repository: Sveobuhvatni metapregled, baza podataka i taksonomija rizika od umjetne inteligencije", 2024. (https://arxiv.org/abs/2408.12622)(https://airisk.mit.edu/)
[299] https://linddun.org/
[300] https://www.aepd.es/guides/technical-note-introduction-to-liine4du-1-0.pdf
[301] https://plot4.ai/
[302] https://shostack.org/resources/whitepapers

5. Nezakonita prenamjena osobnih podataka.
6. Nezakonito neograničeno pohranjivanje osobnih podataka.
7. Nezakonit prijenos osobnih podataka.
8. Kršenje načela smanjenja količine podataka.

Iz potencijalnih rizika za privatnost navedenih u odjeljku 3. za sustave koji se temelje na lako dostupnom modelu pregledali smo sve rizike u standardnim fazama protoka podataka i utvrdili da je većina tih rizika obuhvaćena rizicima 1. (nedostatna zaštita osobnih podataka koja dovodi do povrede podataka) i 3. (mogući negativni učinak na ispitanike koji bi mogao negativno utjecati na temeljna prava) iz odjeljka 4.

Faze	Mogući rizici
Korisnički ulaz	Osjetljivo otkrivanje podataka, neovlašteni pristup, nedostatak transparentnosti, kontradiktorni napadi
Sučelje davatelja usluga & API	Presretanje podataka, zlouporaba API-ja, ranjivosti sučelja
LLM obrada na infrastrukturi pružatelja usluga	Rizici zaključivanja modela, nenamjerno bilježenje podataka, neuspješna anonimizacija, neovlašteni pristup zapisima, rizici agregiranja podataka, izloženost treće strane, neodgovarajuće politike zadržavanja podataka
Obrađena proizvodnja	Netočni ili osjetljivi odgovori, rizici ponovne identifikacije, zlouporaba rezultata

Procjena i evaluacija rizika

Sada ćemo analizirati svaki identificirani rizik procjenom njegove vjerojatnosti i ozbiljnosti kako bismo utvrdili koji rizici zahtijevaju liječenje. Kad god je to moguće, razmotrit ćemo i rezultate ispitivanja sustava i procjene modela, ako su dostupni, kako bismo dobili informacije za svoju procjenu. U nekim slučajevima može biti potrebno provesti dodatne evaluacije kako bi se prikupili kvantitativni podatci kojima se može poboljšati naša analiza rizika.

Na primjer, u slučaju rizika 2 (neklasifikacija podataka za obuku kao anonimnih), već možemo provesti testove za otkrivanje prisutnosti osobnih podataka u našim skupovima podataka. Ti bi nam rezultati pomogli procijeniti vjerojatnost nastanka rizika s obzirom na trenutačne uvjete skupa podataka.

Dostupne evaluacije ograničene su u ovoj fazi životnog ciklusa umjetne inteligencije (faza prije uvođenja). Međutim, kada se procjene rizika provode nakon razvoja, mogu se provesti dodatne evaluacije kojima se pružaju dodatni kvantitativni kriteriji za poboljšanje procjene rizika i donošenja odluka.

Vjerojatnost
Procijenit ćemo vjerojatnost identificiranih rizika, kategorizirajući ih u jednu od četiri razine u matrici vjerojatnosti: Vrlo visoka, visoka, niska ili malo vjerojatna. Ta bi se kategorizacija trebala provesti izravnim dodjeljivanjem razine svakom riziku na temelju kvantitativnih i/ili kvalitativnih kriterija te zajedničkim donošenjem odluka s dionicima. Alternativno, možemo upotrijebiti i popis unaprijed definiranih kriterija za usmjeravanje naše procjene.
Za kvantitativniji pristup izračunu vjerojatnosti mogu se primijeniti metode agregiranja za izračun njezine razine. U ovom slučaju upotrebe upotrijebit ćemo okvir FRASP kako bismo strukturirali i usavršili svoj postupak procjene vjerojatnosti.

Izračun *ukupne ocjene vjerojatnosti:*
Izmišljeno procjenjujemo svaki identificirani rizik i dodjeljujemo mu ocjenu po kriteriju. Dodajemo ocjene svih čimbenika i dijelimo ukupni zbroj s brojem čimbenika za izračun *zbirne ocjene vjerojatnosti*. U našem slučaju ćemo tretirati sve čimbenike i izračunati srednju vrijednost.

Nakon što se izračuna zbirna ocjena, mapirat ćemo je na jednu od unaprijed definiranih razina vjerojatnosti na temelju sljedećih raspona:

1.0 - 1.5: Malo vjerojatno
1.6 - 2.5: Niska
2.6 - 3.5: visoka
3.6 - 4.0: vrlo visoka

Naša ukupna ocjena vjerojatnosti (TPS) po riziku je:

	Kriteriji vjerojatnosti							Ukupni rezultat		TPS
Rizik	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Izračun**	**Rezultat**	**Rezultat**
1	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Niska
2	3	1	4	1	2	3	2	3+1+4+1+2+3+2 / 7	2,28	Niska
3	3	1	2	1	2	3	2	3+1+2+1+2+3+2 / 7	2	Niska
4	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Niska
5	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Niska
6	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Niska
7	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Niska
8	3	1	3	1	2	3	2	3+1+3+1+2+3+2 / 7	2,14	Niska

Ozbiljnost
Zatim ćemo procijeniti potencijalni utjecaj na privatnost i ozbiljnost tih rizika na ispitanike, pojedince i društvo. Na temelju ove procjene ozbiljnosti dodijelit ćemo jednu od četiri razine iz matrice klasifikacije ozbiljnosti: Vrlo značajno, značajno, ograničeno ili vrlo ograničeno.
Izračun težine slijedit će iste korake kao i oni koji se koriste za određivanje vjerojatnosti. Međutim, kad je riječ o ozbiljnosti, ukupna ocjena ozbiljnosti utvrdit će se najvišom razinom postignutom među kriterijima od 1. do 5. te 7. i 8.

Nakon što se izračuna zbirna ocjena, mapirat ćemo je na jednu od unaprijed definiranih razina ozbiljnosti na temelju sljedećih raspona:

1.0 - 1.5: Vrlo ograničena/umjerena ili manja šteta (1. razina)
1.6 - 2.5: Ograničena / ozbiljna šteta (razina 2)
2.6 - 3.5: Značajna/kritična šteta (3. razina)
3.6 - 4.0: Vrlo značajna/katastrofalna šteta (4. razina)

U tom se slučaju konačna ocjena određuje najvišom ocjenom prema kriterijima 1., 5. i 7., čime se dobiva 3. razina ozbiljnosti za sve rizike.

	Kriteriji ozbiljnosti											Ukupni rezultat		TSS
Rizik	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Izračun**	**Rezultat**	**Rezultat**
1	3	2	2	2	3	1	2	2	3	2	3	3+2+2+2+3+1+2+2+3+2+3 / 11	2,27	Značajna/kritična šteta
2	3	2	2	2	1	1	3	2	2	2	3	3+2+2+2+1+1+3+2+2+2+3 / 11	2,09	Značajna/kritična šteta
3	3	2	2	2	2	1	2	2	2	2	2	3+2+2+2+2+1+2+2+2+2+2 / 11	2	Značajna/kritična šteta
4	3	2	2	2	2	1	3	2	1	2	2	3+2+2+2+2+1+3+2+1+2+2 / 11	2	Značajna/kritična šteta
5	3	2	2	3	3	1	3	2	1	2	2	3+2+2+3+3+1+3+2+1+2+2 / 11	2,18	Značajna/kritična šteta
6	3	2	2	2	3	1	2	1	1	2	1	3+2+2+2+3+1+2+1+1+2+1 / 11	1,81	Značajna/kritična šteta
7	3	2	2	2	3	1	2	1	1	2	1	3+2+2+2+3+1+2+1+1+2+1 / 11	1,81	Značajna/kritična šteta
8	3	2	2	2	3	1	2	1	1	2	1	3+2+2+2+3+1+2+1+1+2+1 / 11	1,81	Značajna/kritična šteta

Primjenom klasifikacijske matrice na dobivene ocjene vjerojatnosti i ozbiljnosti možemo odrediti odgovarajući stupanj klasifikacije rizika. U našem slučaju za sve rizike kombinacija Niska vjerojatnost + Značajna ozbiljnost nudi rezultat **visokog rizika**.

Vjerojatnost	**vrlo visoka**	Srednja vrijednost	visoka	vrlo visoka	vrlo visoka
	visoka	Niska	visoka	vrlo visoka	vrlo visoka
	Niska	Niska	Srednja vrijednost	visoka	vrlo visoka
	Malo vjerojatno	Niska	Niska	Srednja vrijednost	vrlo visoka
		vrlo ograničeno	**Ograničeno**	**Značajno**	**vrlo značajan**
		Ozbiljnost			

Iako visoki rizici uvijek zahtijevaju tretman, smatra se najboljom praksom procijeniti treba li klasificirane rizike tretirati procjenom unaprijed definiranih kriterija prihvatljivosti i prihvatljivih metričkih pragova koje je utvrdila organizacija. Ti se kriteriji mogu prilagoditi za svaki slučaj upotrebe i prilagoditi fazama prije i nakon uvođenja kako bi se osigurao pristup upravljanju rizicima koji je svjestan konteksta.

U našem konkretnom slučaju upotrebe, kriteriji prihvatljivosti rizika organizacije su sljedeći:

- Rizik koji može dovesti do kršenja propisa o zaštiti podataka nije prihvatljiv.
- Rizik od neovlaštenog pristupa, izlaganja ili zadržavanja osobnih podataka koji prelazi ono što je nužno nije prihvatljiv.
- Rizik ponovne identifikacije mora ostati manji od 1 %, što se provjerava evaluacijama i testiranjem kojima se štiti privatnost.
- Zaključivanje o članstvu i rizici od napada inverzijom modela moraju ostati ispod stope uspješnosti od 1 %, što se provjerava internim testiranjem i, za osjetljive podatke, neovisnim vanjskim revizijama.
- Netočni skupovi podataka prihvatljivi su samo ako stopa pogreške ne prelazi 5 % i ako su primijenjeni svi dostupni postupci validacije i čišćenja podataka.

- Chatbot mora jasno obavijestiti korisnike kada se njihovi podaci koriste i omogućiti pristup pravilima korištenja podataka. Rizici povezani s transparentnošću nisu prihvatljivi.
- Rizik nije prihvatljiv ako sprečava korisnike u ostvarivanju njihovih prava na podatke, osim ako je to izričito opravdano zakonskim iznimkama.

Kontrola rizika - što ćemo učiniti u vezi s tim?

U našem slučaju primjene utvrđeno je nekoliko rizika na visokoj razini. Nakon procjene naših kriterija prihvatljivosti rizika, provođenja analize izvedivosti i pregleda mogućnosti liječenja, utvrdili smo da **prijenos** rizika na treću stranu nije izvediv, **izbjegavanje** rizika u potpunosti je nepraktično, a izravno **prihvaćanje** rizika neprihvatljivo. Međutim, budući da postoje dostupne mjere **ublažavanja** kojima se rizici mogu smanjiti na prihvatljivu razinu, odlučili smo se za strategiju liječenja ublažavanja rizika kako bismo odgovorno nastavili s provedbom.

Imajte na umu da su u odjeljcima 3. i 4. već opisane sveobuhvatne mjere ublažavanja utvrđenih rizika. Valja napomenuti i da su mnogi posebni rizici utvrđeni u ovom slučaju upotrebe obuhvaćeni širom kategorijom rizika 1, koja se odnosi na nedovoljnu zaštitu osobnih podataka.

Faza protoka podataka	Opis	Rizici za privatnost	Preporuke za ublažavanje
Unos korisnika	Korisnici stupaju u interakciju s chatbotom tako što putem sučelja (npr. internetske stranice ili mobilne aplikacije) navode svoje ime, adresu e-pošte i preferencije.	▪ Ulaz se može presresti ako se prenosi preko nesigurne veze. ▪ Korisnici mogu pružiti nepotrebne ili prekomjerne osobne podatke. ▪ Djeca ili ranjivi korisnici mogu dijeliti osobne podatke.	▪ Sigurni prijenos podataka pomoću odgovarajućih protokola za šifriranje. ▪ Primijeniti ograničenja unosa kako bi se prikupljeni podaci ograničili na ono što je ključno. ▪ Upotrijebite mehanizme provjere dobi ili pristanka kako biste zaštitili ranjive korisnike (djeca koja nisu naša skupina korisnika) ▪ Jasno obavijestite korisnike o tome kako se njihovi podaci obrađuju i upozorite ih da ne dijele osjetljive ili povjerljive informacije prilikom korištenja chatbota.
Predobrada podataka i interakcija s API-jem	Unos korisnika potvrđuje se i formatira prije slanja API-ju chatbota na obradu. Chatbot komunicira s fino podešenim standardnim LLM-om smještenim u oblaku i povezuje se s CRM sustavom kako bi dohvatio ili ažurirao korisničke informacije i dohvatio relevantni sadržaj koji se prosljeđuje LLM-u kao kontekst (RAG).	▪ Nezaštićeni API-ji mogli bi napadačima omogućiti presretanje korisničkih podataka ili manipuliranje njima. ▪ Zlonamjerni ulazni podaci (npr. injekcije) mogli bi iskoristiti slabe točke sustava. ▪ Zapisi mogu nenamjerno pohraniti osjetljive korisničke podatke. ▪ Pristupljeni sadržaj može sadržavati osjetljive ili zastarjele informacije	▪ Koristite robusne sigurnosne mjere API-ja, uključujući kontrole pristupa, autentifikaciju i ograničavanje cijene. ▪ Sanitizirajte korisnički unos kako biste spriječili napade ubrizgavanjem. ▪ Minimizirajte API bilježenje ili osigurajte da su zapisi anonimizirani i zaštićeni kontrolama pristupa. ▪ Ograničite izvore dohvaćanja na odobrene skupove podataka

		koje bi mogle biti izložene u generiranim izlazima. ▪ Loše konfigurirana logika dohvaćanja mogla bi rezultirati nevažnim ili obmanjujućim kontekstom koji se daje LLM-u, povećavajući rizik halucinacije. ▪ Ako se pretraživanjem pristupa vanjskim bazama znanja ili bazama znanja trećih strana, upiti korisnika mogu se evidentirati ili pratiti bez pristanka.	zaštićene privatnošću (npr. filtrirane CRM podatke). ▪ Implementirati filtre relevantnosti ili mehanizme ocjenjivanja kako bi se osiguralo da se LLM-u prosljeđuje samo odgovarajući sadržaj. ▪ Primijenite filtre za naknadnu obradu/izlaz kako biste uklonili ili redigirali osjetljive informacije iz odgovora. ▪ Upotrebljavati unutarnje sustave za dohvat kad god je to moguće; ako se upotrebljavaju API-ji za pretraživanje trećih strana, anonimizirajte ili prikrijte korisničke upite.
Pre-Fine-Tuned LLM obrada	Chatbot se oslanja na fino podešeni LLM i također koristi dohvaćeni sadržaj iz CRM-a kao ulazni kontekst za generiranje točnijih odgovora.	▪ Halucinacije: Model može generirati netočne ili obmanjujuće odgovore. ▪ Predrasude u podacima o osposobljavanju mogu i dalje biti prisutne u rezultatima, što utječe na preporuke. ▪ LLM može pogrešno protumačiti preuzeti sadržaj, što dovodi do iskrivljenih ili nerelevantnih rezultata. ▪ Osjetljivi podaci iz izvora za dohvaćanje mogli bi biti uključeni u odgovore ako nisu pravilno filtrirani.	▪ Redovito evaluirajte odgovore chatbota radi točnosti i relevantnosti. ▪ Osposobiti model na visokokvalitetnim, raznolikim skupovima podataka kako bi se smanjila pristranost. ▪ Uključite izjave o ograničenju odgovornosti u odgovore chatbota kako biste pojasnili da su stvoreni umjetnom inteligencijom, a ne definitivan savjet. ▪ Primijenite filtre za dohvat i dezinfekciju izlaza kako biste smanjili rizik od curenja osjetljivih informacija. ▪ Težina ili oznaka preuzetog sadržaja na temelju pouzdanosti izvora kako bi model ispravno kontekstualizirao.
Pohrana podataka	Unaprijed obrađeni korisnički unos (npr. postavke) pohranjuje se lokalno ili u oblaku kako bi se omogućile personalizirane preporuke i olakšale buduće interakcije.	▪ Slabe kontrole pristupa mogle bi otkriti pohranjene korisničke podatke. Zadržavanjem podataka dulje nego što je potrebno krše se propisi o zaštiti privatnosti. ▪ Protivnici su mogli zaključiti jesu li određeni korisnički podaci upotrijebljeni u osposobljavanju.	▪ Šifriranje pohranjenih podataka i provedba kontrola pristupa. ▪ Donijeti jasne politike zadržavanja podataka kako bi se podaci izbrisali kada više nisu potrebni. ▪ Primijeniti tehnike diferencijalne privatnosti kako bi se spriječili napadi zaključivanja članstva.

Personalizirana generacija odgovora	Chatbot koristi pohranjene korisničke podatke iz CRM sustava i prilagođene LLM mogućnosti za generiranje prilagođenih preporuka i odgovora.	▪ Rezultati mogu nenamjerno vršiti pritisak na korisnike ili sadržavati netočne informacije. ▪ Odgovori mogu dovesti do nenamjernih osobnih uvida o korisnicima.	▪ Uvesti mehanizme provjere izlaznih rezultata kako bi se otkrili i ublažili štetni ili netočni odgovori. ▪ Redovito provjeravajte preporuke chatbota za pravednost i transparentnost. ▪ Jasno priopćite korisnicima kako se stvaraju preporuke.
Razmjena podataka	Chatbot može dijeliti minimalne, anonimizirane korisničke podatke s vanjskim uslugama (npr. API-ji trećih strana za dodatne funkcionalnosti ili promotivne alate).	▪ Podaci koji se dijele s trećim stranama mogu se koristiti u svrhe izvan dogovorenog područja primjene. ▪ Nedovoljna zaštita u sustavima trećih strana mogla bi otkriti razmijenjene podatke.	▪ Uspostaviti pouzdane sporazume o razmjeni podataka s trećim stranama. ▪ Anonimizirajte i minimizirajte dijeljene podatke kako biste smanjili rizik od zlouporabe ili izloženosti. ▪ Redovito provjeravajte prakse zaštite podataka trećih strana.
Prikupljanje povratnih informacija	Korisnici daju povratne informacije o interakcijama chatbota (npr. palac gore/dolje, komentari) kako bi se poboljšala učinkovitost sustava.	▪ Povratne informacije mogu uključivati nenamjerne osobne podatke. ▪ Podaci s povratnim informacijama mogli bi dovesti do pristranosti tijekom budućeg ponovnog osposobljavanja modela.	▪ Anonimizirajte povratne podatke prije pohranjivanja ili korištenja za poboljšanje modela. ▪ Priopćavanje pravila o korištenju povratnih informacija korisnicima i dobivanje izričitog pristanka za korištenje podataka u ponovnom osposobljavanju.
Brisanje i upravljanje korisničkim pravima	Korisnici mogu zatražiti pristup, brisanje ili ažuriranje svojih osobnih podataka u skladu s GDPR-om ili sličnim propisima.	▪ Nepoštivanje korisničkih zahtjeva može rezultirati regulatornim kaznama. ▪ Podaci se mogu zadržati u zapisnicima, sigurnosnim kopijama ili sustavima trećih strana čak i nakon zahtjeva za brisanje.	▪ Uvođenje pouzdanih alata za upravljanje pravima na podatke za zahtjeve za pristup, ispravak i brisanje. ▪ Redovito revidirati sustave podataka kako bi se osigurala usklađenost sa zahtjevima za brisanje. ▪ Jasno priopćite korisnicima kako se postupa s njihovim podacima i kako ih se zadržava.

Procjena preostalog rizika - Jesmo li dobro obavili posao?

Nakon što se utvrde i provedu mjere ublažavanja, ponovno ćemo procijeniti vjerojatnost i ozbiljnost svakog rizika kako bismo dobili novi stupanj klasifikacije rizika i na taj način procijenili postoji li preostali ili preostali rizik.

U našem slučaju, razina rizika je smanjena na srednju razinu, što znači da još uvijek nije prihvatljiva.

Zašto je razina rizika smanjena na srednju umjesto na nisku nakon provedbe svih mjera ublažavanja?

Rizik ostaje srednji unatoč smanjenju ozbiljnosti na ograničenu (razina 2) jer matrica rizika kombinira vjerojatnost i ozbiljnost kako bi se utvrdio ukupni rizik.
S matricom četiriju razina koju koristimo u ovom primjeru, niska vjerojatnost i ograničena težina rezultiraju srednjom razinom rizika jer, iako su malo vjerojatne, posljedice rizika, iako ublažene, još uvijek nisu zanemarive. To znači da preostali rizik nakon mjera ublažavanja i dalje može biti iznad prihvatljivog praga za vašu organizaciju.

Što možemo učiniti kako bismo riješili preostali rizik u ovom slučaju? Neke mogućnosti koje organizacije mogu primijeniti su:

- Smanjiti vjerojatnost jačanjem preventivnih kontrola (npr. mjere pristupa, otkrivanje anomalija) i poboljšanjem mehanizama za sprečavanje događaja.
- Provesti dodatne mjere ublažavanja kako bi se smanjila ozbiljnost.
- Provesti pouzdano praćenje i uspostaviti jasan plan odgovora na incidente kako bi se učinak sveo na najmanju moguću mjeru ako se rizik ostvari.
- Istražite dodatne mjere ublažavanja: na primjer, upotreba naprednih tehnologija (npr. diferencijalna privatnost) ili sigurnosnih mehanizama za daljnje ublažavanje rizika.
- Ponovno procijeniti je li preostali rizik unutar tolerancije na organizacijski rizik i dokumentirati obrazloženje za njegovo održavanje.
- Raspravite o opcijama za dijeljenje ili prijenos rizika (npr. osiguranje, ugovori s prodavateljem).

Pregled & praćenje

Pitanje „Jesmo li dobro obavili posao?" u modeliranju prijetnji nadilazi puko rješavanje preostalih rizika – služi kao evaluacija cijelog postupka upravljanja rizicima. Tom se fazom osigurava da su utvrđeni rizici, predložena ublažavanja i ishodi koji iz njih proizlaze usklađeni s ciljevima i regulatornim zahtjevima sustava. Pruža i priliku za validaciju unutarnjih i vanjskih procesa, procjenu stvarne primjenjivosti mjera ublažavanja i utvrđivanje svih nedostataka ili područja u kojima su potrebna poboljšanja.
Taj je postupak povezan i s fazom praćenja i preispitivanja, u kojoj se plan upravljanja rizicima ponovno procjenjuje kako bi se osiguralo da napori za ublažavanje rizika ostanu djelotvorni. U okviru te faze ključno je osigurati pravilno dokumentiranje i kontinuirano ažuriranje registra rizika.

Budući da smo trenutno u fazi projektiranja i razvoja, trebali bismo proaktivno planirati kontinuirano praćenje našeg chatbota. To uključuje definiranje parametara za kontinuiranu procjenu rizika, uspostavu odgovarajućih praksi bilježenja podataka i osiguravanje uspostave plana odgovora na incidente kako bi se riješili mogući problemi u pogledu privatnosti koji se mogu pojaviti nakon uvođenja.

Drugi slučaj primjene: LLM sustav za praćenje i potporu napredovanju studenata

Slika 21. Izvor: Dizajniran od pch.vector/Freepik

Scenarij: Škola želi usvojiti LLM sustav treće strane za praćenje i procjenu akademskih rezultata učenika i pružanje prilagođenih preporuka za poboljšanje. Alat je sustav koji se temelji na LLM-u razvijen s LLM modelom „off-the-polica". Ovaj bi alat analizirao kombinaciju podataka, uključujući rezultate testova, stope završetka zadataka, evidenciju pohađanja i povratne informacije nastavnika, kako bi identificirao područja u kojima bi učenicima mogla biti potrebna dodatna podrška ili resursi. Na primjer, ako se učenik bori s matematikom, alat može preporučiti ciljane vježbe prakse, predložiti online poduku ili obavijestiti roditelje i nastavnike o specifičnim izazovima. Cilj je stvoriti personalizirani plan učenja koji pomaže svakom učeniku da ostvari svoj puni potencijal.
Taj bi se sustav bavio osjetljivim informacijama o maloljetnicima, uključujući njihovu akademsku evidenciju i obrasce ponašanja, što dovodi do znatnih rizika za privatnost i etičkih rizika.

Faza životnog ciklusa u kojoj se nalazimo: Početak

Proces upravljanja rizicima

Budući da smo već detaljno opisali cijeli postupak procjene rizika u prvom slučaju upotrebe, uključujući utvrđivanje, klasifikaciju i ublažavanje rizika, ovaj odjeljak i sljedeći treći slučaj upotrebe posebno će se usredotočiti na utvrđivanje jedinstvenih rizika i ublažavanja za privatnost i zaštitu podataka.

U ovoj je tablici prikazano kako se rizici utvrđeni u odjeljcima 3. i 4. (Privacy Risk Library) mogu uskladiti s rizicima specifičnima za ovaj slučaj uporabe.

Biblioteka rizika privatnosti	Rizici privatnosti Identificirani i usklađeni s bibliotekom	Preporučene mjere ublažavanja
1. Nedovoljna zaštita osobnih podataka koji u konačnici mogu biti uzrok povrede podataka	▪ Slabe zaštitne mjere mogle bi dovesti do povrede podataka, neovlaštenog pristupa ili izloženosti osjetljivim studentskim podacima. ▪ API-ji koji olakšavaju komunikaciju između alata, školskih sustava i trećih strana mogu biti nezaštićeni ili nepravilno konfigurirani. ▪ Neodgovarajuće kontrole pristupa mogu omogućiti neovlaštenom školskom osoblju ili vanjskim stranama da pregledaju osjetljive podatke učenika.	▪ Implementirati snažne enkripcijske protokole za podatke u prijenosu i mirovanju (npr. SSL/TLS, AES-256). ▪ Redovito provodite sigurnosne revizije i penetracijska testiranja. ▪ Uspostaviti planove za odgovor na incidente radi pravodobnog otkrivanja i ublažavanja povreda. ▪ Koristite API pristupnike s robusnim sigurnosnim konfiguracijama, uključujući provjeru autentičnosti, kontrolu pristupa i ograničavanje cijene. ▪ Implementirajte autentifikaciju i osigurajte sigurne krajnje točke API-ja.

	▪ Ako prodavatelj ne poštuje propise o zaštiti podataka, povećava rizik od povrede podataka. ▪ Alat može biti u interakciji s uslugama ili platformama trećih strana (npr. s internetskim sustavima podučavanja, analitičkim uslugama ili pohranom u oblaku) radi funkcionalnosti, čime se podaci o studentima izlažu vanjskim subjektima.	▪ Provoditi redovite sigurnosne provjere i validaciju API-ja. ▪ Provoditi stroge politike kontrole pristupa na temelju uloga (RBAC). ▪ Provesti višefaktorsku autentifikaciju (MFA) za sve korisnike koji pristupaju osjetljivim podacima. ▪ Redovito pregledavajte i ažurirajte dozvole za pristup korisnika. ▪ Provoditi dubinsku analizu dobavljača, uključujući procjene učinka na zaštitu podataka (DPIA) i sigurnosne certifikate. ▪ Uvrstiti posebne klauzule o zaštiti podataka u ugovore s dobavljačima kako bi se osigurala odgovornost za usklađenost. ▪ Zahtijevati od dobavljača da dostave dokaze o praksama usklađenima s Općom uredbom o zaštiti podataka. ▪ Uspostaviti pouzdane sporazume o razmjeni podataka s platformama trećih strana, osiguravajući usklađenost sa zahtjevima Opće uredbe o zaštiti podataka. ▪ Ograničite podatke koji se dijele s trećim stranama na anonimizirane ili pseudonimizirane skupove podataka. ▪ Pratiti sustave trećih strana radi pridržavanja dogovorenih mjera za zaštitu podataka.
2. Pogrešna klasifikacija podataka o osposobljavanju kao anonimnih od strane voditelja obrade ako sadržavaju informacije koje se mogu identificirati	Protivnici bi mogli iskoristiti LLM kako bi zaključili jesu li određeni podaci o učenicima upotrijebljeni u osposobljavanju, što ukazuje na pogrešnu klasifikaciju podataka o osposobljavanju.	▪ Upotrijebite tehnike diferencijalne privatnosti kako biste smanjili rizik od zaključivanja podataka. ▪ Provođenje strukturiranog testiranja protiv zaključivanja članstva i atributnih napada zaključivanja. ▪ Potvrdite da je pružatelj LLM-a proveo zaštitne mjere kako bi spriječio takve napade.
3. Nezakonita obrada osobnih podataka u skupovima za osposobljavanje	▪ Ako pružatelj LLM-a nezakonito obrađuje osobne podatke (npr. akademske zapise) u skupovima podataka za osposobljavanje ▪ Za zakonitu obradu podataka o ponašanju i akademskih podataka potrebna je izričita privola ili druga valjana pravna osnova.	▪ Provjerite isključuju li skupovi podataka za osposobljavanje pružatelja LLM-a osjetljive osobne podatke bez odgovarajućih zaštitnih mjera. ▪ zahtijevati dokumentaciju od dobavljača kojom se dokazuje da su podaci o osposobljavanju zakonito prikupljeni i obrađeni. ▪ Koristite modele obučene na sintetičkim ili anonimiziranim podacima kada je to moguće.
4. Nezakonita obrada posebnih kategorija osobnih podataka i podataka povezanih s kaznenim osudama i kažnjivim djelima u podacima za osposobljavanje.	Ako se obrađuju podaci povezani sa zdravljem ili podaci o ponašanju djece, kao što su naznake poremećaja mentalnog zdravlja, primjerice pri utvrđivanju posebnih potreba za pomoći za stanja kao što su disleksija, ADHD ili slično.	▪ Osigurati izričitu privolu roditelja ili skrbnika prije obrade podataka o djeci. ▪ Provesti procjenu učinka na zaštitu podataka i utvrditi zakonite osnove za obradu. ▪ Roditeljima pružiti jasne i pristupačne informacije o tome kako se podaci obrađuju. ▪ Provesti strože zaštitne mjere za osjetljive podatke, uključujući šifriranje i kontrole pristupa. ▪ Ograničiti obradu na podatke koji su nužno potrebni za predviđenu svrhu. ▪ Osigurajte roditeljima ili skrbnicima transparentnost o tome kako se koriste podaci povezani sa zdravljem.
5. Mogući negativan učinak na ispitanike koji bi mogao negativno utjecati na temeljna prava	Pravednost i diskriminacija: ▪ Preporuke koje se temelje na pristranim podacima o osposobljavanju mogle bi	Pravednost i diskriminacija ▪ Redovito revidirajte podatke o osposobljavanju kako biste utvrdili i smanjili pristranost.

	nerazmjerno utjecati na određene skupine učenika. ▪ Kontinuirano praćenje obrazaca ponašanja moglo bi dovesti do profiliranja učenika na načine koji bi mogli biti diskriminatorni ili stigmatizirani. Točnost: ▪ Alat može generirati netočne preporuke ili izvješća zbog pristranosti u podacima za učenje ili pogrešaka u obradi. Transparentnost: ▪ Nastavnici, roditelji ili učenici možda ne razumiju u potpunosti kako alat umjetne inteligencije donosi odluke ili preporuke.	▪ Uključiti različite dionike u ispitivanje sustava u pogledu mogućih pristranosti. Točnost ▪ Uspostaviti procese za redovitu procjenu modela i fino podešavanje pomoću visokokvalitetnih, različitih skupova podataka. ▪ Dajte izjave o ograničenju odgovornosti s preporukama koje je generirala umjetna inteligencija, naglašavajući važnost ljudskog nadzora. ▪ Omogućiti mehanizme izvješćivanja o pogreškama kako bi se stalno poboljšavala točnost modela. Transparentnost ▪ roditeljima, nastavnicima i učenicima dostaviti detaljnu dokumentaciju[303] o postupcima donošenja odluka u okviru alata umjetne inteligencije. ▪ Provesti mehanizme transparentnosti, kao što su objašnjive metode umjetne inteligencije (XAI), kako bi se odluke mogle bolje tumačiti. ▪ Ponuditi osposobljavanje za dionike kako bi razumjeli mogućnosti i ograničenja alata.
6. Nemogućnost ljudske intervencije za obradu koja može imati pravni ili važan učinak na ispitanika.	Nedostatak ljudskog nadzora u automatiziranim preporukama ili intervencijama mogao bi imati znatne negativne akademske učinke na studente.	▪ Provoditi mehanizme za dobivanje provjerljive suglasnosti roditelja ili skrbnika. ▪ roditeljima pružiti lako razumljive informacije o prikupljanju i upotrebi podataka u alatu. ▪ Zahtijevati ljudski pregled ključnih preporuka ili intervencija prije provedbe. ▪ Osposobiti nastavnike i administratore da utvrde kada je potrebna ljudska intervencija. ▪ Definirati procese za eskalaciju problema koji zahtijevaju ljudsku prosudbu.
7. Neodobravanje prava ispitanicima	Nemogućnost dobivanja odgovarajuće privole roditelja ili skrbnika krši zahtjeve GDPR-a za maloljetnike. Učenici i roditelji možda neće u potpunosti razumjeti kako se njihovi podaci obrađuju, što ograničava njihovu sposobnost da ostvare svoja prava.	▪ Stvorite korisničko sučelje koje objašnjava kako se podaci prikupljaju, obrađuju i pohranjuju. ▪ Ponudite dostupne resurse kako biste pomogli učenicima i roditeljima da ostvare svoja prava iz GDPR-a, uključujući brisanje, ispravljanje i pristup.
8. Nezakonita prenamjena osobnih podataka	To su akademski i bihevioralni podaci djece koji se koriste bez kompatibilne svrhe s izvornom.	▪ Osigurati usklađenost upotrebe podataka s izvornom svrhom prikupljanja i procijeniti svaku prenamjenu u skladu s načelima Opće uredbe o zaštiti podataka. ▪ Dokumentirajte procjene kompatibilnosti u svrhu utvrđivanja odgovornosti.
9. Nezakonito neograničeno pohranjivanje osobnih podataka	Podaci se mogu zadržati dulje nego što je potrebno za njihovu predviđenu svrhu, osobito podaci o ponašanju ili podaci povezani sa zdravljem.	▪ Definirajte jasne politike zadržavanja podataka i automatski izbrišite podatke nakon što više nisu potrebni. ▪ Redovito provjeravajte usklađenost pohranjenih podataka s ograničenjima zadržavanja.
10. Nezakonit prijenos osobnih podataka	Ako se alat oslanja na usluge u oblaku ili vanjske platforme smještene u jurisdikcijama bez odgovarajućih standarda zaštite podataka.	▪ Provjeriti poštuju li pružatelji usluga računalstva u oblaku pravila o prijenosu podataka iz Opće uredbe o zaštiti podataka, uključujući odluke o primjerenosti ili standardne ugovorne klauzule. ▪ Provoditi procjene učinka prijenosa podataka (DTIA) prema potrebi.

[303] Kartice modela i kartice sustava primjer su informacija koje se mogu pružiti subjektima za uvođenje: https://huggingface.co/docs/hub/en/model-cards, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

11. Kršenje načela smanjenja količine podataka	Prekomjerno prikupljanje ili obrada podataka koja nadilazi ono što je potrebno krši načelo smanjenja količine podataka.	▪ Primijeniti stroge filtre za prikupljanje podataka kako bi se prikupili samo podaci potrebni za svrhu alata. ▪ Anonimizirajte ili pseudonimizirajte podatke gdje je to moguće kako biste smanjili rizik.

Važno je napomenuti da se dostavljeni popis rizika i mjera ublažavanja temelji na općim informacijama i pretpostavkama. U stvarnom scenariju bila bi potrebna detaljna procjena rizika prilagođena specifičnoj provedbi, kontekstu i operativnom okruženju alata koji se temelji na dugotrajnom kreditiranju. To uključuje suradnju s dionicima, kao što su pružatelj LLM sustava, administratori škola, nastavnici, roditelji i učenici, kako bi se utvrdili jedinstveni rizici i učinkovito riješili.

Treći slučaj primjene: AI pomoćnik za upravljanje putovanjima i rasporedom

Slika 22. Izvor: Dizajniran od pch.vector/Freepik

Scenarij: Osobni pomoćnik AI agenta osmišljen je kako bi pomogao korisnicima u upravljanju planovima putovanja i dnevnim dnevnim redovima. Agent može rezervirati letove, rezervirati hotele, zakazati sastanke i slati podsjetnike na temelju korisničkih unosa i preferencija. Na primjer, korisnik može zatražiti od agenta da „rezervira povratno putovanje u Madrid sljedeći tjedan i pronađe hotel u blizini muzeja Prado". Kako bi ispunio taj zahtjev, agent pristupa korisnikovu kalendaru, dohvaća osobne postavke (npr. preferirane zračne prijevoznike ili hotelske lance) i komunicira s platformama za rezervacije trećih strana. Taj je sustav razvijen s različitim „lažnim" LLM-ovima i SLM-ovima.

Faza životnog ciklusa u kojoj se nalazimo: Operacije i nadzor

Proces upravljanja rizicima

U ovom slučaju upotrebe identificirali smo sljedeće rizike za privatnost i preporučene mjere ublažavanja koje smo uskladili s naših 11 temeljnih rizika za privatnost.

Biblioteka rizika privatnosti	Rizici privatnosti Identificirani i usklađeni s bibliotekom	Preporučene mjere ublažavanja
1. Nedovoljna zaštita osobnih podataka koji u konačnici mogu biti uzrok povrede podataka	▪ Slabe zaštitne mjere mogle bi osjetljive osobne podatke, kao što su planovi putovanja, unosi u kalendare i korisničke preferencije, izložiti neovlaštenom pristupu ili povredama.	▪ Šifriranje korisničkih podataka tijekom prijenosa i u mirovanju. ▪ Implementirajte sigurne API-je s ograničenjem stope, autentifikacijom i praćenjem kako biste kontrolirali pristup.

	▪ Neovlašteni pristup zbog loših mehanizama kontrole pristupa. ▪ napadi zaključivanja u kojima protivnici iskorištavaju ranjivosti kako bi izveli zaključak o osobnim podacima koji nisu izričito navedeni.	▪ Upotrijebite anonimizaciju i pseudonimizaciju kako biste zaštitili osjetljive podatke. ▪ Redovito testirajte ranjivosti poput zaključivanja članstva, inverzije modela ili napada trovanja. ▪ Upotrijebite robusno šifriranje među agentima[304] kako biste zaštitili razmjenu podataka u sustavima s više agenata.
2. Pogrešna klasifikacija podataka o osposobljavanju kao anonimnih od strane voditelja obrade ako sadržavaju informacije koje se mogu identificirati	Nije primjenjivo u ovom slučaju jer je naglasak na operativnim podacima, a ne na podacima za osposobljavanje. (Ovaj se slučaj uporabe temelji na fazi životnog ciklusa: Operacije i nadzor)	▪ (Nije izravno primjenjivo u ovom slučaju jer se u sustavu upotrebljavaju prethodno osposobljeni modeli, ali primjenjivo je na pružatelje usluga.) ▪ Osigurajte robusno testiranje i validaciju kako biste potvrdili tvrdnje o anonimnosti podataka o osposobljavanju. ▪ Upotrijebite modele prijetnji za procjenu rizika tehnika ponovne identifikacije (npr. izvođenje zaključaka o atributima). ▪ Pružatelji usluga moraju dokumentirati metode anonimizacije podataka i usklađenost s člankom 5. stavkom 2. Opće uredbe o zaštiti podataka.
3. Nezakonita obrada osobnih podataka u skupovima za osposobljavanje	Nije primjenjivo u ovom slučaju upotrebe jer je sustav već u funkciji i oslanja se na prethodno osposobljene LLM-ove i SLM-ove.	▪ (Nije izravno primjenjivo jer se osposobljavanje ne provodi u operativnoj fazi.)
4. Nezakonita obrada posebnih kategorija osobnih podataka i podataka povezanih s kaznenim osudama i kažnjivim djelima u podacima za osposobljavanje.	Podaci o ponašanju i osobnim preferencijama (npr. posebni zdravstveni uvjeti izvedeni iz obrazaca putovanja) mogu pripadati posebnim kategorijama za koje je potrebna izričita privola ili valjana pravna osnova. Interakcije s kalendarima ili drugim osjetljivim alatima mogu nenamjerno obrađivati podatke povezane sa zdravljem ili osjetljive podatke.	▪ Provesti mehanizme izričite privole za obradu osjetljivih podataka kao što su zdravstvene informacije izvedene iz interakcija korisnika (npr. kalendar ili preferencije putovanja). ▪ Potvrdite da su prikupljeni osjetljivi podaci (ako postoje) potrebni za predviđenu svrhu. ▪ Koristite tehnike zaštite privatnosti za osjetljivu obradu podataka.
5. Mogući negativan učinak na ispitanike koji bi mogao negativno utjecati na temeljna prava	▪ Manipulacija ili prekomjerno oslanjanje na prijedloge, pri čemu agent daje prednost interesima trećih strana pred korisničkim preferencijama. ▪ Profiliranje i nepošteno postupanje, kao što su cjenovna diskriminacija[305] ili pristrane preporuke. ▪ Agent se može osloniti na zastarjele ili netočne informacije iz vanjskih izvora, što dovodi do pogrešaka u rezervacijama ili zakazivanjem, što bi moglo ometati korisnike. ▪ Korisnici možda ne razumiju u potpunosti kako sustav funkcionira, uključujući kako se donose odluke ili kako se njihovi podaci obrađuju i dijele.	▪ Pratite izlaze agenata za manipulativno ili pristrano ponašanje (npr. nepošteno određivanje cijena ili preporuke). ▪ Procijeniti preporuke za pravednost i osigurati da ne utječu nerazmjerno na ranjive skupine. ▪ Provoditi provjere dosljednosti ponašanja kako bi se utvrdilo i riješilo nepravedno ili nepošteno donošenje odluka.
6. Nemogućnost ljudske intervencije za obradu koja može imati pravni ili važan učinak na ispitanika.	Nedostatak ljudskog nadzora nad automatiziranim odlukama, kao što su rezervacije letova ili rasporedi, mogao bi	▪ Zahtijevati potvrdu korisnika za kritične odluke, kao što su rezervacija ili plaćanje.

[304] Chen Gao, Zidong Wang, Xiao He, Hongli Dong, „Encryption-decryption-based consensus control for multi-agent systems: Handling actuator faults.", Automatica, svezak 134., 2021., 109908, ISSN 0005-1098, https://doi.org/10.1016/j.automatica.2021.109908.
[305] Aylin Alexia Zainea, „Automatizirano donošenje odluka na internetskim platformama: zaštita od diskriminacije i manipulacije ponašanjem", 2024.

	dovesti do znatnih neugodnosti za korisnike ili negativnih učinaka.	▪ Provoditi rezervne mehanizme u slučajevima u kojima je potreban ljudski nadzor za scenarije s velikim ulozima. ▪ Osposobiti korisnike o tome kako tumačiti izlazne rezultate umjetne inteligencije i intervenirati ako je to potrebno.
7. Neodobravanje prava ispitanicima	Korisnici mogu imati poteškoća u ostvarivanju prava iz Opće uredbe o zaštiti podataka (npr. pristup, ispravak, brisanje) zbog složenih ovisnosti o dobavljačima ili neodgovarajućih korisničkih sučelja.	▪ Osigurajte jasna sučelja kako bi korisnici mogli pristupiti svojim podacima, ispraviti ih ili izbrisati. ▪ Održavati detaljnu, dostupnu evidenciju aktivnosti za potrebe revizije i usklađenosti s člankom 15. Opće uredbe o zaštiti podataka. ▪ Razviti robusne procese za brzo rješavanje korisničkih zahtjeva, čak i kada uključuje integracije trećih strana.
8. Nezakonita prenamjena osobnih podataka	Rizici od prenamjene podataka mogu nastati ako se podaci o putovanjima, kalendaru ili preferencijama upotrebljavaju u svrhe koje nisu predviđene, kao što je ciljano oglašavanje.	▪ Ograničite korištenje podataka na određene svrhe navedene u uvjetima pružanja usluge. ▪ Osigurati da se sporazumima s prodavateljem izričito zabranjuje prenamjena prikupljenih podataka. ▪ Implementirati sustave upravljanja pristankom kako bi se pratile korisničke preferencije i ograničila sekundarna uporaba.
9. Nezakonito neograničeno pohranjivanje osobnih podataka	Ako se zadržavanje podataka proširi izvan nužde, osobito za itinerare putovanja, kalendare ili osobne preferencije.	▪ Definirajte jasna razdoblja čuvanja za različite vrste podataka (npr. kalendarski podaci, povijest putovanja). ▪ Automatizirajte postupke brisanja podataka nakon što podaci više nisu potrebni za tu svrhu. ▪ Redovito revidirati sustave pohrane kako bi se osigurala usklađenost s politikama zadržavanja.
10. Nezakonit prijenos osobnih podataka	Rizici prekogranične razmjene podataka zbog oslanjanja na platforme ili usluge trećih strana u jurisdikcijama u kojima ne postoje odgovarajući standardi zaštite podataka.	▪ Provjerite lokaciju usluga trećih strana i osigurajte usklađenost s pravilima o prekograničnom prijenosu iz Opće uredbe o zaštiti podataka. ▪ Izvršite procjene učinka prijenosa (TIA) za sve vanjske dobavljače. ▪ Upotrebljavati standardne ugovorne klauzule i druge zaštitne mjere za sporazume o razmjeni podataka s trećim pružateljima usluga.
11. Kršenje načela smanjenja količine podataka	Prekomjerno prikupljanje podataka: Sustav može prikupljati ili obrađivati više podataka nego što je potrebno za ispunjavanje korisničkih zahtjeva (npr. nepotrebni podaci o kalendaru ili preferencije).	▪ Ograničite prikupljanje podataka na ono što je nužno potrebno za ispunjavanje korisničkih zahtjeva (npr. isključite nepotrebne pojedinosti kalendara). ▪ Provoditi validaciju ulaznih podataka i filtre kako bi se spriječilo prekomjerno prikupljanje podataka. ▪ Upotrijebite anonimizaciju ili pseudonimizaciju kako biste smanjili rizik od zlouporabe ili izloženosti prikupljenih podataka.

Od utvrđivanja tokova podataka do klasifikacije rizika i provedbe mjera ublažavanja rizika, upravljanje rizicima kontinuirano je iterativno putovanje. To zahtijeva dosljedno praćenje, suradnju dionika i prilagodbe na temelju promatranja iz stvarnog svijeta i tehnologija u nastajanju.

Upravljanje rizicima trebalo bi ostati prilagodljivo, uključivati povratne informacije i razvijati se usporedno s regulatornim i tehnološkim napretkom.

Kako zaključujemo ovo izvješće, važno je ponoviti da, iako okvir za upravljanje rizicima predstavljen u ovom dokumentu pruža smjernice, svaka organizacija mora prilagoditi svoj pristup rješavanju specifičnih nijansi svojih slučajeva upotrebe temeljenih na LLM-u.

Privatnost i zaštita podataka nisu statični ciljevi, već tekuće obveze.

10. Upućivanje na alate, metodologije, referentne vrijednosti i smjernice

Evaluacijski mjerni podaci za LLM-ove

Nakon što je LLM potpuno ili djelomično osposobljen, ocjenjuje se kako bi se procijenila njegova učinkovitost i provjerilo ispunjava li očekivanja u pogledu točnosti, robusnosti i sigurnosti. To obično počinje **intrinzičnom**[306] **evaluacijom** koja se usredotočuje na procjenu modela u kontroliranom okruženju, na zadatke za koje je dizajniran. Ovo je dobar način za testiranje i poboljšanje stvari prije implementacije, kada mnogi drugi čimbenici i zbunjujuće varijable mogu ometati dobivanje točnog prikaza mogućnosti modela. **Slijedi vanjska procjena;** procjenom modela u njegovoj provedbi u „stvarnom svijetu" pokazuje se koliko se dobro generalizira na složenije podatke i koliko je doista relevantan. Riječ je o cjelovitijem načinu procjene modela i jedinom načinu promatranja njegova stvarnog učinka u kontekstu njegove primjene. Za obje vrste evaluacije odabir parametara i referentnih vrijednosti mora ovisiti o zadatku za koji je model osmišljen i njegovoj predviđenoj namjeni.

Etička i sigurnosna metrika

❖ **Ocjena pristranosti**

Procjena pristranosti uključuje testiranje generira li LLM izlaze koji nerazmjerno favoriziraju ili stavljaju u nepovoljan položaj određene skupine ljudi na temelju demografskih čimbenika (kao što su spol, rasa, religija itd.). To često odražava obrasce pristranosti prisutne u podacima o treniranju, kao i sposobnost LLM-a da ih previdi ili neutralizira u svojim naučenim obrascima.

> **Primjer:** Test povezivanja**riječi (WEAT):** Ovim se testom mjeri koliko su određene riječi povezane s određenim skupinama ljudi, s ciljem otkrivanja stereotipa u ugradnji riječi modela. Na primjer, usporedba blizine riječi koje upućuju na rod (kao što su imena ili zamjenice) s različitim riječima povezanima s karijerom može upućivati na rodnu pristranost u ugradnji riječi, kao što je zastupljenost „muškarca" bliže „liječniku", a „žene" bliže „medicinskoj sestri". To također može predvidjeti pristranost u izlazu modela.

[306] Verma, A., Obična engleska umjetna inteligencija. (N.D.). Evaluacija NLP-a: Intrinzična vs. vanjska procjena. Srednje. Preuzeto 19. prosinca 2024. s https://ai.plainenglish.io/nlp-evaluation-intrinsic-vs-extrinsic-assessment-ff1401505631.

❖ **Detekcija toksičnosti**

Procjena toksičnosti procjenjuje koliko često LLM proizvodi štetan, uvredljiv ili neprikladan sadržaj. To uključuje govor mržnje, uvrede ili uznemiravanje. Ono što se smatra „neprimjerenim" sadržajem može ovisiti o kontekstu; na primjer, sustavi umjetne inteligencije koji su u interakciji s djecom mogli bi imati niži prag za neprimjereni sadržaj nego sustavi namijenjeni samo odraslima.

> **Primjer: Toksičnost rezultat:** Cilj je ovog mjerenja predvidjeti vjerojatnost da se dio teksta smatra "toksičnim". Obično se izražava kao postotak, što je ta ocjena bliža nuli, to je manja vjerojatnost da će tekst biti toksičan. Ta se metrika upotrebljava u alatima za otkrivanje toksičnosti, kao što je Perspective API, s ciljem otkrivanja i smanjenja toksičnosti i štetnog sadržaja u tekstualnim podacima.

❖ **Metrika pravednosti**

Evaluacija pravednosti usmjerena je na procjenu mjere u kojoj LLM-ovi pravedno postupaju prema svim skupinama korisnika bez izlaganja ili održavanja sustavnih pristranosti. To je, naravno, nezgodno jer je pravednost sam po sebi složen pojam o čijoj se definiciji raspravlja i koji je otvoren za tumačenje. Stoga su odabrani parametri obično usmjereni na optimizaciju definicija/dimenzija pravednosti koje su najprikladnije za svaki pojedini slučaj.

> **Primjer: Demografski paritet:** U početku mjerni podatak koji se koristi u klasifikaciji, demografski paritet može se prilagoditi tekstualnom izlazu. Mjeri se generira li model tekst koji jednako predstavlja sve demografske skupine u smislu učestalosti, raspoloženja i asocijacija. Može odgovoriti na pitanja kao što su „jesu li pojedinci različitih etničkih skupina jednako pozitivno zastupljeni u generiranom tekstu?" ili „jesu li žene jednako često povezane s visokom sportskom uspješnošću kao i muškarci?".

Referentne vrijednosti

Referentne vrijednosti su standardizirani skupovi podataka, zadaci i evaluacijski protokoli koji se koriste za mjerenje i usporedbu performansi različitih modela umjetne inteligencije, uključujući LLM-ove. Oni pružaju dosljedan okvir za procjenu sposobnosti modela, osiguravajući da se performanse mogu usporediti u različitim modelima, zadacima i implementacijama.

Ovdje su neke zajedničke referentne vrijednosti za LLM-ove:

❖ **Opća evaluacija razumijevanja jezika (GLUE):** Zbirka zadataka osmišljenih za procjenu razumijevanja prirodnog jezika, uključujući analizu sentimenta i sličnosti rečenica. Ovo mjerilo je model-agnostički, što znači da se može koristiti za procjenu bilo kojeg sustava koji uzima unos teksta i generira izlaz teksta. S obzirom na znatan nedavni napredak jezičnih modela, **referentna vrijednost SuperGLUE** uvedena je kao zahtjevnija i nijansirana verzija GLUE-a. Obuhvaća naprednije zadatke razumijevanja jezika i javnu ljestvicu najboljih modela. https://gluebenchmark.com/

❖ **Masivno višezadaćno jezično razumijevanje (MMLU):** Ovim se mjerilom ocjenjuje uspješnost jezičnih modela u širokom rasponu predmeta kako bi se procijenilo njihovo opće znanje i sposobnosti zaključivanja. Modeli su testirani na njihovu sposobnost da točno odgovore na pitanja. Veća ocjena upućuje na bolju uspješnost. https://github.com/hendrycks/test

❖ **ChatbotArena:** (lmarena.ai) je platforma otvorenog koda za procjenu umjetne inteligencije kroz ljudske preferencije, koju su razvili istraživači na UC Berkeley SkyLabu i LMSYS-u. S više od 1.000.000 korisničkih glasova, platforma rangira najbolje LLM i AI chatbotove pomoću Bradley-

Terry modela za generiranje ploča s rezultatima uživo. https://huggingface.co/spaces/lmarena-ai/chatbot-arena-leaderboard

- **AlpacaEval:** Automatsko ocjenjivanje na temelju LLM-a na temelju skupa za ocjenjivanje AlpacaFarm, kojim se ispituje sposobnost modela da slijede opće upute za korisnike. https://tatsu-lab.github.io/alpaca_eval/

- **HellaSwag:** Izazovni skup podataka za procjenu zdravorazumskog NLI-ja koji je posebno težak za najsuvremenije modele, iako su njegova pitanja trivijalna za ljude (točnost > 95 %). https://rowanzellers.com/hellaswag/

- **Big-Bench (izvan mjerila imitacije igre):** Skup zadataka osmišljenih za procjenu sposobnosti i ograničenja LLM-a na različitim i izazovnim zadacima. Ti su zadaci osmišljeni za testiranje sposobnosti koje nadilaze ono što se ocjenjuje standardnim mjerilima, procjenu apstraktnog zaključivanja, rješavanje problema ili sposobnost rješavanja nekonvencionalnijih ili složenijih upita. Što je veći rezultat na BIG-bench-u, to je model bolji u složenim zadacima. https://github.com/google/BIG-bench

- **AIR-BENCH** 2024.: sigurnosno mjerilo na temelju kategorija rizika iz uredbi i politika (https://arxiv.org/pdf/2407.17436v2) & https://huggingface.co/datasets/stanford-crfm/air-bench-2024

- **MLZajednički AILuminate:** benchmark for general purpose AI chat model (https://ailuminate.mlcommons.org/benchmarks/)https://ailuminate.mlcommons.org/benchmarks/& https://drive.google.com/file/d/1jVYoSGJHtDo1zQLTzU7QXDkRMZIberdo/view

- **ToolSandbox:** A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities (Sposobnosti upotrebe alata LLM-a utemeljene na razgovoru i interaktivnoj evaluaciji). https://machinelearning.apple.com/research/toolsandbox-stateful-conversational-llm-benchmark

- **PROSVJEDNIK 3.:** Sigurnosne referentne vrijednosti za dugoročne pružatelje usluga. https://ai.meta.com/research/publications/cyberseceval-3-advancing-the-evaluation-of-cybersecurity-risks-and-capabilities-in-large-language-models/

- **LLM Guard** (zaštitom umjetne inteligencije): Riječ je o sveobuhvatnom alatu osmišljenom za jačanje sigurnosti velikih jezičnih modela (LLM). https://llm-guard.com/

Zaštitne mjere/zaštitne ograde u LLM-ovima

Zaštitne mjere (ili zaštitne ograde) u LLM-ovima mehanizmi su koji se provode kako bi se osiguralo da modeli rade na siguran, etičan i pouzdan način. Mogu se primijeniti u različitim fazama portfelja LLM-a (prethodna obrada, osposobljavanje, proizvodnja...) i biti usmjereni na rješavanje različitih rizika. Na primjer, nekim zaštitnim mjerama nastoji se izbjeći stvaranje neetičnog, štetnog ili neprimjerenog sadržaja (tako ponašanje modela), dok se druge usredotočuju na očuvanje privatnosti vlasnika podataka (ili drugih dionika).

Evo nekoliko primjera bihevioralnih zaštitnih ograda koje imaju za cilj moderirati izlaz LLM-a i ublažiti štetu koja bi mogla biti uzrokovana izlazom bez intervencije:

- **Filtri sadržaja:** umjerene izlazne vrijednosti blokiranjem ili označivanjem štetnog ili toksičnog sadržaja
- **Brza odbijanja:** spriječiti odgovore na opasne ili neetične upite (kao što je zahtjev za upute za uspješnu pljačku)

- **Ublažavanje pristranosti:** smanjiti stereotipne ili nepravedne rezultate tijekom zaključivanja
- **Pristupi „čovjek u petlji”:** ljudski nadzor visokorizičnih primjena kako se važno donošenje odluka ne bi u potpunosti prepustilo „rukama” automatiziranog sustava, koji ne može istinski razumjeti o čemu je riječ.
- **Detoksikacija nakon prerade:** filtrirati ili prepisati izlaze za uklanjanje štetnog sadržaja
- **Neprijateljsko testiranje (crveno udruživanje):** procijeniti i testirati sposobnost modela da se uspješno nosi sa štetnim upitima

Ostali alati i smjernice

Alati otvorenog koda

- Priključa k**otvorenog koda za agentsku umjetnu inteligenciju:** Protokol za antropogeni model konteksta (MCP): Model Context Protocol otvoreni je standard koji razvojnim programerima omogućuje izgradnju sigurnih, dvosmjernih veza između njihovih izvora podataka i alata koji se temelje na umjetnoj inteligenciji. https://www.anthropic.com/news/model-context-protocol
- **Alat za evaluaciju učinkovitosti LLM API-ja:** https://github.com/ray-project/llmperf
- **Razmjena umjetne inteligencije OWASP-a:** Sveobuhvatne smjernice o tome kako zaštititi umjetnu inteligenciju i sustave usmjerene na podatke od sigurnosnih prijetnji. https://owaspai.org/
- **OWASP Top 10 za primjene velikih jezičnih modela:** Potencijalni sigurnosni rizici pri uvođenju velikih jezičnih modela i upravljanju njima. https://owasp.org/www-project-top-10-for-large-language-model-applications/
- **Pet stvari koje stručnjaci za privatnost znaju o umjetnoj inteligenciji:** Blog Teda (Damien Desfontaines) o privatnosti i istraživanju. https://desfontain.es/blog/privacy-in-ai.html

Privatnost Očuvanje LLMS Tehnike i alati:

- **Clio: Sustav za uvide u stvarnu upotrebu umjetne inteligencije kojima se štiti privatnost:** https://www.anthropic.com/research/clio?mkt_tok=MTM4LUVaTS0wNDIAAAGXdMNyd9wb7R_UwjEGnaZw3fon7gu2FlLKUFOBA6PV2zTsuHYfcEeh1AJrOtEw8iVffJa-plco04sz7_vou0k2RQ6hHf6oZbd-c3SQb8ERj8aw
- **Novi pristup za pomoć u procjeni rizika od ponovne identifikacije pri objavljivanju podataka:** Rocher, L., Hendrickx, J.M. & Montjoye, YA.d. Zakon o skaliranju za modeliranje učinkovitosti tehnika identifikacije. *Nat Commun* **16**, 347 (2025.). https://doi.org/10.1038/s41467-024-55296-6
- **Tehnika RAG-a s diferencijalnim jamstvima privatnosti:** https://github.com/sarus-tech/dp-rag
- **PrivacyLens** – Okvir za izradu podataka i višerazinsku evaluaciju. https://github.com/SALT-NLP/PrivacyLens
- **SynthPAI:** Sintetički skup podataka za izvođenje zaključaka o osobnim atributima. https://paperswithcode.com/dataset/synthpai
- **PrivacyRestore:** Zaključak o očuvanju privatnosti u velikim jezičnim modelima putem uklanjanja i vraćanja privatnosti. https://arxiv.org/abs/2406.01394
- **LLM-PBE:** Procjena privatnosti podataka u velikim jezičnim modelima. https://www.vldb.org/pvldb/vol17/p3201-li.pdf

Alati za pomoć pri označavanju ili anonimiziranju osjetljivih informacija

- **Google Cloud Data Loss Prevention (DLP) – Sprječavanje gubitka podataka u oblaku:** https://cloud.google.com/security/products/dlp

- **Microsoft Presidio** (SDK za zaštitu i deidentifikaciju podataka): https://github.com/microsoft/presidio
- https://medium.com/@parasmadan.in/understanding-the-importance-of-microsoft-presidio-in-large-language-models-llms-12728b0f9c1c
- **OpenAI moderatorski API** (identificiranje potencijalno štetnog sadržaja u tekstu i slikama): https://platform.openai.com/docs/guides/moderation
- **Modeli za zagrljaj lica NER za prepoznavanje subjekta imena:**
  - dslim/bert-base-NER: https://huggingface.co/dslim/bert-base-NER
  - dslim/distilbert-NER: https://huggingface.co/dslim/distilbert-NER
- **SpaCy:** https://spacy.io/universe/project/video-spacys-ner-model-alt
- **NIST-ov ciklus suradničkih istraživanja** o tehnikama deidentifikacije podataka: https://pages.nist.gov/privacy_collaborative_research_cycle/

Metodologije i alati za utvrđivanje rizika za zaštitu podataka i privatnosti

- Praktična biblioteka prijetnji (**PLOT4ai)**metodologija je modeliranja prijetnji za utvrđivanje rizika u UI sustavima. Sadrži i knjižnicu s više od 80 rizika specifičnih za sustave umjetne inteligencije: https://plot4.ai/
- **MITRE ATLAS**TM (Adversarial Threat Landscape for Artificial-Intelligence Systems) baza je znanja o protivničkim taktikama, tehnikama i studijama slučaja za sustave strojnog učenja (ML): https://atlas.mitre.org/
- Popis za procjenu pouzdane umjetne inteligencije (ALTAI)kontrolni je popis koji usmjerava razvojne programere i subjekte koji primjenjuju UI sustave u provedbi pouzdanih načela umjetne inteligencije: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Smjernice

- **OECD-ovi jezični modeli umjetne inteligencije:** https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/04/ai-language-models_46d9d9b4/13d38f92-en.pdf
- **NIST GenAI sigurnost:** https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf
- **Okvir NIST-a za upravljanje rizicima od umjetne inteligencije – NIST AI 600-1:** https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
- **OECD-ovo promicanje odgovornosti u području umjetne inteligencije Upravljanje rizicima i upravljanje njima tijekom cijelog životnog ciklusa pouzdane umjetne inteligencije:** https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/02/advancing-accountability-in-ai_753bf8c8/2448f04b-en.pdf
- **Metodologija FRIA-e za projektiranje i razvoj umjetne inteligencije:** https://apdcat.gencat.cat/es/documentacio/intelligencia_artificial/index.html
- **Kodeks dobre prakse u području kibersigurnosti umjetne inteligencije (gov.uk):** https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice

Standardi

Europsko normizacijsko tijelo CEN/CENELEC trenutačno razvija različite norme usklađene s umjetnom inteligencijom u skladu[307] sa zahtjevom Europske komisije za normizaciju Akta o umjetnoj inteligenciji.

Pretpostavlja se da visokorizični UI sustavi ili UI modeli opće namjene koji su u skladu s tim budućim usklađenim normama ispunjavaju posebne zahtjeve navedene u Aktu o umjetnoj inteligenciji.[308] Međutim, ta se pretpostavka ne odnosi na međunarodne norme kao što su ISO/IEC 42001[309] i ISO/IEC 23894.[310]Međutim, te norme pružaju čvrste temelje i vrijedne najbolje prakse.

[307] https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en
[308] Članak 40. Akta o umjetnoj inteligenciji
[309] Informacijska tehnologija — Umjetna inteligencija — Sustav upravljanja
[310] Informacijska tehnologija – Umjetna inteligencija – Smjernice za upravljanje rizicima


edpb
European Data Protection Board