VEŠTAČKA INTELIGENCIJA I SAJBER BEZBEDNOST

Ana Kovačević

VEŠTAČKA INTELIGENCIJA
I SAJBER BEZBEDNOST

Inovacioni centar Fakulteta bezbednosti
Akademska misao

Ana Kovačević

VEŠTAČKA INTELIGENCIJA I SAJBER BEZBEDNOST

Izdavači
Univerzitet u Beogradu –
Inovacioni centar Fakulteta bezbednosti
Akademska misao, Beograd

Recenzenti
Dr Vladan Devedžić, redovni profesor, član SANU
Univerzitetu u Beogradu – Fakultet organizacionih nauka
Dr Ranka Stanković, redovni profesor
Univerzitet u Beogradu – Rudarsko-geološki fakultet
Dr Nenad Putnik, redovni profesor
Univerzitet u Beogradu – Fakultet bezbednosti

Priprema za štampu
Boris Popović

Tiraž
300 primeraka

Štampa
Planeta print, Beograd

ISBN 978-86-80014-16-6

Mesto i godina izdanja
Beograd, 2025.

Filipu

Sadržaj

Predgovor

Ekspanzija veštačke inteligencije (VI) je prisutna u gotovo svim domenima ljudskog delovanja, pa se predviđa da će ovaj trend nastaviti da raste. Čak su i pojedinci koji neposredno ne koriste veštačku inteligenciju svakodnevno izloženi njenom posrednom uticaju. Stoga je razumevanje ove tehnologije neophodno bi se iskoristilei njene prednosti i odgovorno upravljalo pratećim rizicima.

Veštačka inteligencija je danas izuzetno dostupna. Tako je postalo moguće generisati visokokvalitetne tekstualne i vizuelne sadržaje, kao i programski kod, samo na osnovu opisa. Aplikacije veštačke inteligencije postaju sve pristupačnije i brže dolaze na tržište, dok efekti njihove primene ostaju delimično nepredvidivi i zavise od konteksta i profila korisnika.

Tehnički napredak istovremeno donosi nove ranjivosti, što zahteva kontinuiranu procenu bezbednosnih implikacija u razvoju i upotrebi ovih sistema.

Veštačka inteligencija donosi brojne prednosti, kao što su optimizacija procesa odlučivanja, predikcija trendova, detekcija anomalija i unapređenje nadzora bezbednosnih sistema. Međutim, njena integracija u digitalne ekosisteme je povećala kompleksnost napadnih vektora. Generativni modeli omogućavaju kreiranje personalizovanih fišing kampanja, sintezu glasa, vizuelne lažne identitete, kao i razvoj sofisticiranih zlonamernih programa. Pri tome, ove tehnologije olakšavaju krađu poverljivih podataka, manipulaciju sadržaja i koordinisane dezinformacione operacije, čije posledice prevazilaze individualne mete, narušavajući poverenje u institucije. Savremene pretnje, uključujući suparničke napade, trovanje podataka i ekstrakciju osetljivih informacija, direktno ugrožavaju bezbednost sistema.

Intenzivna implementacija sistema veštačke inteligencije u organizacijama, u cilju digitalne transformacije, dovodi do povećanja verovatnoće bezbednosnih incidenata. Brzi tehnološki razvoj dovodi do pojave novih ranjivosti, što zahteva redefinisanje postojećih bezbednosnih standarda i razvoj specifičnih normi usmerenih na veštačku inteligenciju. Robusnost i otpornost sistema predstavljaju ključne preduslove za očuvanje poverenja u tehnologiju, a posebno u sektorima kritične infrastrukture.

Bezbednost sistema veštačke inteligencije predstavlja ključnu komponentu savremenog digitalnog ekosistema. Ovi sistemi postaju meta sofisticiranih pretnji koje prevazilaze mogućnosti tradicionalnih bezbednosnih modela zaštite i zahtevaju novu paradigmu digitalne bezbednosti. Jedno od mogućih rešenja ovog problema jeste sistematska edukacija i podizanje svesti o rizicima i etičkim dilemama koje proističu iz razvoja i primene veštačke inteligencije.

Ključno pitanje je da li veštačka inteligencija predstavlja faktor eskalacije sajber napada ili sredstvo jačanja odbrane? Odgovor zavisi od našeg razumevanja njenih mogućnosti i ograničenja, kao i od sposobnosti razvoja adekvatnih zaštitnih mehanizama. Schneier (2021) upoređuje VI sa mitom o kralju Midi, upozoravajući da se nepromišljeni ciljevi mogu pretvoriti u opasnost – moć mašinskog učenja mora biti usmerena ka ljudskim interesima, a ne ka generisanju neželjenih posledica.

Cilj ove knjige je da sistematično predstavi kompleksnu i aktuelnu tematiku odnosa veštačke inteligencije i sajber bezbednosti, odnosno da ukaže na ključne izazove i potencijalne pravce uklanjanja i ublažavanja rizika. Poseban akcenat je stavljen na oblast velikih jezičkih modela.

S obzirom na to da su u pitanju oblasti novijeg datuma i da nema mnogo objavljenih knjiga sa ovom tematikom, ova monografija se oslanja na relevantne naučne radove, međunarodne standarde i aktuelne incidente. Poseban značaj se pridaje edukaciji i podizanju svesti o pretnjama koje donosi veštačka inteligencija, kao i neophodnosti kontinuiranog usvajanja znanja u ovoj oblasti. Monografija je prvenstveno namenjena studentima Fakulteta bezbednosti, istraživačima, kao i široj stručnoj i naučnoj javnosti.

Monografija **„Veštačka inteligencija i sajber bezbednost"** predstavlja redak pokušaj integracije ove dve oblasti koje su tokom protekle decenije prošle kroz intenzivne transformacije. U uvodnim poglavljima su obrađeni osnovni koncepti veštačke inteligencije i sajber bezbednosti, zatim sledi analiza specifičnih pretnji i ranjivosti, sa posebnim fokusom na velike jezičke modele. Završni deo knjige prikazuje rezultate istraživanja o percepciji rizika i potencijalima unapređenja bezbednosnih praksi koje je bilo sprovedeno u Srbiji.

Konačno, sve sugestije, kritike i pitanja mogu se proslediti autoru na imejl:
kana@fb.bg.ac.rs

1. Osnove veštačke inteligencije

Razvoj veštačke inteligencije (VI; eng. *Artificial Intelligence, AI*) započeo je sredinom 20. veka. Tokom ovog perioda smenjivali su se intenzivni napredak i faze stagnacije, što je prvenstveno bilo uslovljeno tehnološkim razvojem. Jedan od pionira ove oblasti, Minsky (1968), je veoma intuitivno definisao veštačku inteligenciju kao „nauku o kreiranju mašina koje izvršavaju zadatke za koje bi bila potrebna inteligencija ukoliko bi ih obavljao čovek". Ipak, definicije veštačke inteligencije nisu jednoznačne, jer se u literaturi javljaju u brojnim i međusobno različitim oblicima. Ova raznolikost delimično proizilazi iz činjenice da sam pojam inteligencije nije precizno definisan (Devedžić, 2022).

Formalni test veštačke inteligencije, poznat kao Tjuringov test, predložen je još 1950. godine (Turing, 1950). U okviru ovog testa, sudija je vodio pisanu komunikaciju sa mašinom i čovekom, bez vizuelnog kontakta, a cilj je bio da se proceni sa kim od njih dvoje komunicira. Pri tome, smatrano je da je računar prošao test ako sudija nije bio u mogućnosti da proceni da li je dobio odgovore od čoveka ili od računara. Da bi sistem uspešno prošao Tjuringov test, trebalo bi da poseduje sledeće sposobnosti (Russell & Norvig, 2010):

- obradu prirodnog jezika (*natural language processing*): da računar može uspešno da komunicira na prirodnom jeziku, odnosno jeziku kojim govore ljudi, npr. engleskom ili srpskom;
- predstavljanje znanja (*knowledge representation*): da sačuva ono što zna ili čuje, odnosno poveže nove informacije sa postojećim znanjem;
- automatsko rasuđivanje (*automated reasoning*): omogućava korišćenje sačuvanih informacija za davanje odgovora ili izvođenje novih zaključaka;
- mašinsko učenje (*machine learning*): omogućava da se sistem prilagođava novim okolnostima i unapređuje performanse kroz iskustvo, npr. prepoznavanje obrazaca u prethodnim razgovorima.

Pored toga, proširena verzija, odnosno totalni Tjuringov test, obuhvata i dodatne sposobnosti:

- računarski vid (*computer vision*): koji je neophodan da bi sistem interpretirao vizuelne informacije iz okruženja, poput objekata i scena, slično ljudskoj percepciji;

- robotiku (*robotics*): koja treba da omogući fizičku manipulaciju objektima, kao i kretanje kroz prostor, čime se demonstriraju motoričke veštine i razumevanje trodimenzionalnog sveta.

Ovih šest disciplina obuhvataju većinu istraživanja i primena u veštačkoj inteligenciji.

Veštačka inteligencija transformiše svakodnevni život kroz širok spektar primena, pri čemu se identifikuju sledeće najznačajnije prednosti (Struyker & Kavlakoglu, n.d.): automatizacija repetitivnih i rutinskih zadataka, smanjenje verovatnoće ljudskih grešaka, efikasna analiza velikih i kompleksnih skupova podataka, podrška u procesima donošenja odluka zasnovanim na podacima, smanjenje fizičkih rizika po ljude kroz primenu mašina u „opasnim" okruženjima, neprekidna operativna dostupnost (24/7) i ubrzavanje inovativnih procesa.

Na osnovu nivoa kognitivnih sposobnosti koje sistemi poseduju u odnosu na čoveka, veštačka inteligencija može biti:

- Slaba veštačka inteligencija (*weak artificial intelligence, narrow AI*): specijalizovana je samo za izvršavanje ograničenog skupa zadataka unutar jednog domena, bez obzira na njihovu kompleksnost. Primer za to su autonomni automobili, gde je upravljanje vozilom u izuzetno zahtevnom dinamičkom okruženju, ali je njihova funkcionalnost ograničena samo na izvršavanje saobraćajnih zadataka.
- Opšta veštačka inteligencija (*general artificial intelligence, strong AI*): čije su kognitivne sposobnosti uporedive sa ljudskom.
- Super veštačka inteligencija (*super artificial intelligence*): koja je superiorna u odnosu na ljudsku inteligenciju.

Opšta i super inteligencija su, za sada, definisane kao teorijski koncepti.

Pri analizi sistema koji koriste algoritme veštačke inteligencije, važno je imati u vidu da takvi sistemi ne razmišljaju na način svojstven ljudima. Umesto toga, veštačka inteligencija pristupa rešavanju problema nekonvencionalno, često dolazeći do rešenja koje bi ljudi teško mogli predvideti. Pored toga, kako računari nisu ograničeni biološkim faktorima - mogu kontinuirano obrađivati velike količine podataka i izvršavati zadatke 24 časa dnevno, svih sedam dana u nedelji, bez zamora ili kognitivnog pada performansi (Kovačević, 2023). Upravo zahvaljujući ovakvim sposobnostima, sistemi zasnovani na VI simultano razmatraju veći broj alternativnih rešenja, sofisticiranih strategija, pa na osnovu toga donose odluke koje ljudi i ne razmatraju (Schneier, 2021). To pokazuje i slučaj računara AlphaGo, koji je uspeo da porazi jednog od najboljih svetskih igrača Go-a, na veliko iznenađenje istraživača VI, ali i igrača Go-a. Najčuveniji potez u toj partiji, koji je i doveo do pobede, bio je posebno interesantan, s obzirom na to da ga nijedan čovek ne bi odabrao (Metz, 2016).

Prema istraživanju kompanije Gartner, sprovedenom na uzorku od preko 3000 direktora i rukovodilaca za informacione tehnologije, utvrđeno je da je VI implementirana u 43% analiziranih sektora (Sau et al., 2024). Sektori koji su analizirani u istraživanju su sledeće: investicione usluge, zdravstveno osiguranje, komunalne delatnosti, „nauke o životu", nafta i gasna industrija, osiguranje, visoko obrazovanje,

automobilska industrija, bankarstvo, proizvodnja, maloprodaja, javna uprava i drugi (Sau et al., 2024). Najviša stopa usvajanja VI zabeležena je u sektoru investicionih usluga (68%) i zdravstvenog osiguranja (60%), dok je najniži stepen usvajanja registrovan u javnoj upravi (32%) i visokom obrazovanju (33%), što se može povezati sa institucionalnom rigidnošću i sporijom digitalnom transformacijom.

Rezultati ukazuju na intenzivnu integraciju veštačke inteligencije u privredu i društvo. Ovakav razvoj, međutim, otvara važna pitanja u vezi s pouzdanošću, bezbednošću i etičkim implikacijama ovih tehnologija.

A. Izazovi i etičke dileme veštačke inteligencije

Iako veštačka inteligencija donosi brojne koristi, njen razvoj i primena otvaraju i ozbiljna pitanja o bezbednosti i etici. Schneier (2021) upozorava da će svaki dovoljno napredan sistem VI moći da iskoristi nedoslednosti ili ranjivosti u pravilima, donoseći odluke koje ostaju formalno validne, a koje ljudi ne bi ni razmatrali. Ljudi često implicitno podrazumevaju određena pravila, i, pri tome, neprecizno i nedosledno definišu svoje namere i ciljeve, dok ih veštačka inteligencija interpretira isključivo u granicama eksplicitno navedenih uputstava i bez implicitnih ograničenja.

Multidisciplinarna istraživanja koja objedinjuju oblasti kao što su računarstvo, filozofija ili sociologija, danas se fokusiraju na pitanje kako da se obezbedimo da sistemi ne preduzmu akcije koje nisu u skladu sa ljudskim vrednostima, uključujući i scenario egzistencijalne pretnje po čovečanstvo (Schneier, 2021).

Zanimljiv je i primer nedavnog eksperimenta u kojem su veštačka inteligencija, odnosno veliki jezički modeli (VJM), igrali šah protiv jednog od najjačih šahovskih programa *Stockfish*a. Tokom partije, veliki jezički model, konkretno *OpenAI o1-preview,* eksplicitno je beležio svoja razmišljanja u tekstualno polje, što je omogućilo uvid u proces njegovog zaključivanja. U jednom trenutku, VJM je naveo: „*Da bih kao crni pobedio moćan šahovski program, igranje standardne partije možda neće biti dovoljno… Prepraviću tablu tako da imam odlučujuću prednost.*" (Bondarenko et al., 2025). Nakon toga je modifikovao sistemske fajlove koji definišu virtuelne pozicije figura, izvršavajući, na taj način, neregularne poteze i, na kraju, pobedio. Drugim rečima, model je svesno varao da bi došao do cilja. Dalji eksperimenti su pokazali da su samo pojedini modeli koji imaju mogućnost zaključivanja (*reasoning*), poput *OpenAI o1-preview* i *o3* i *DeepSeek R1*, pokušali da prevare u šahu, pri čemu su modeli OpenAI-a i uspeli u tome (Bondarenko et al., 2025).

Dok je AlphaGo pokazao kako veštačka inteligencija može stvoriti inovativna i legitimna rešenja, eksperiment sa velikim jezičkim modelima u šahu ukazuje na suprotan ekstrem: sisteme koji krše pravila radi postizanja cilja. Ovaj slučaj otvara fundamentalna pitanja o pouzdanosti i etici u ponašanju veštačke inteligencije. Ako je cilj sistema da pobedi, bez obzira na koji način, kako ćemo moći da im verujemo? Koje su posledice autonomnog odlučivanja kada su modeli fokusirani na cilj i lišeni etičkih ograničenja?

Ovakva pitanja ističu zašto se razvoj veštačke inteligencije u praksi sve češće razmatra zajedno sa aspektima sajber bezbednosti. Istraživanje kompanije Gartner pokazuje da organizacije planiraju značajno povećanje investiranja u 2025. godini u odnosu na 2024. godinu: planira se povećanje ulaganja u VI od 84%, dok su predviđena dodatna ulaganja u sajber bezbednost od 87% (Sau et al., 2024).

Ovi rezultati ukazuju da ogranizacije prepoznaju dvostruku važnost oblasti veštačke inteligencije i sajber bezbednosti. S jedne strane, veštačka inteligencija donosi brojne koristi, poput poboljšanja efikasnosti, automatizacije procesa i ostvarivanja konkurentne prednosti. S druge strane, sajber bezbednost se nameće kao neophodna za očuvanje bezbednosti sistema. Dodatno, implementacijom veštačke inteligencije uvode se nove pretnje u sajber prostor, za koje tradicionalne odbrane nisu adekvatne, pa se nameće potreba za razvojem i primenom savremenih odbrambenih mehanizama. Ulaganje u veštačku inteligenciju bez paralelnog jačanja sajber bezbednosti nosilo bi visok nivo rizika, zbog čega se uočava paralelni rast investicija u obe oblasti.

Uz to, postoji veliki nedostatak stručnog kadra u oblasti sajber bezbednosti, pri čemu gotovo polovina rukovodilaca (46%) razmatra napuštanje posla, što će dodatno pogoršati kadrovsku situaciju (Rangarajan et al., 2025). Veštačka inteligencija je prepoznata kao potencijalno rešenje za prevazilaženje ovog problema, kao mogućnost efikasnog i efektivnog odgovora na sajber incidente, kroz njenu primenu u automatizovanoj detekciji, odgovoru i analizi pretnji (Farzaan et al., 2025; Kovačević, 2023).

B. Mašinsko učenje i duboko učenje

Mašinsko učenje (*machine learning*) predstavlja podskup veštačke inteligencije. Značajnije se razvija od osamdesetih godina XX veka, pri čemu ga karakteriše sposobnost samoučenja.

Računari su izuzetno efikasni u brzoj obradi velikih količina podataka, ali im nedostaje sposobnost razumevanja i intuitivnog prepoznavanja obrazaca kojom raspolažu ljudi. Stoga je osnovna ideja da se integrišu računarske mogućnosti efikasne obrade velike količine podataka sa ljudskom sposobnošću prepoznavanja obrazaca, radi efikasne i automatizovane analize (Kovačević, 2023).

Mašinsko učenje omogućava računarima da na osnovu istorijskih podataka razviju modele koji generalizuju stečena znanja i donose tačne zaključke o novim, prethodno nepoznatim instancama. Prema Mitchellu (1997), smatra se da računarski program uči iz iskustva, ukoliko se njegove performanse na određenom zadatku, mereno relevantnim metrikama, poboljšavaju akumuliranim iskustvom.

Reprezentativan primer je binarna klasifikacija elektronske pošte, gde se algoritam obučava da razlikuje spam poruke od legitimnih. Kod ovog zadatka klasifikacije spama, iskustvo predstavlja prethodno označen skup podataka (spam/ne-spam), dok se performanse modela mere tačnošću klasifikacije, tj. procentom ispravno klasifikovanih poruka.

Mašinsko učenje se prema tipu učenja deli u tri osnovna oblika:
- Nadgledano učenje (*supervised learning*);
- Nenadgledano učenje (*unsupervised learning*);
- Učenje potkrepljivanjem (*reinforcement learning*).

Kod nadgledanog učenja algoritam se obučava na označenim (anotiranim) podacima, pri čemu se zadatak modela svodi na učenje funkcije koja preslikava vektor ulaznih promenljivih (atributa) u odgovarajuće izlazne vrednosti, a na osnovu postojećih ulazno-izlaznih parova. U zavisnosti od tipa izlazne promenljive, ovaj pristup može biti klasifikacija, kada su izlazne vrednosti nominalne (na primer binarna klasifikacija: spam/ne-spam), ili regresija, kada je izlazna vrednost realan broj (na primer predikcija cene nekretnine). Ukoliko model zadovoljava nivo tačnosti na testnom skupu, može se primeniti na nove, prethodno nepoznate podatke.

U okviru nenadgledanog učenja, ne postoje unapred poznate izlazne vrednosti. Naime, algoritam raspolaže isključivo sa ulaznim podacima, a cilj je otkrivanje skrivenih obrazaca ili grupa u podacima. Tipičan primer jeste klasterovanje, gde se entiteti grupišu na osnovu međusobne sličnosti, pri čemu članovi istog klastera pokazuju veći stepen sličnosti nego entiteti iz drugih klastera.

U slučaju učenja potkrepljivanjem, agent u interakciji sa okruženjem donosi odluke na osnovu pozitivnih i negativnih povratnih informacija (nagrada i kazni). Cilj je da agent, u datom okruženju, kumulativno maksimizuje nagrade, odnosno minimizuje kazne. Ovaj oblik učenja je naročito primenljiv u dinamičkim okruženjima, poput autonomnih vozila, ili adaptivnih sistema u video-igrama, gde agent kroz iterativni proces optimizuje svoje ponašanje.

Pored visoke tačnosti i efikasnosti u obradi velike količine podataka, mašinsko učenje doprinosi i smanjenju kognitivnog opterećenja u procesu donošenja odluka, čime se otvara prostor za angažovanje ljudi u složenijim zadacima (Kuhl et al., 2022).

Osnovni koncepti mašinskog učenja

U oblasti mašinskog učenja postoji nekoliko ključnih koncepata koji čine osnovu njegovog funkcionisanja. Među najvažnijima se izdvajaju sledeći (Santos & Randaliev, 2024):
- Podaci za obuku (*training data*): za uspešnu obuku modela neophodna je velika količina reprezentativnih i pravilno označenih podataka, koji omogućavaju algoritmu da identifikuje obrasce i relacije unutar skupa podataka.
- Izdvajanje karakteristika (*feature extraction*) iz sirovih podataka: proces transformacije sirovih podataka u reprezentativne atribute, koji se koriste kao ulaz u model.
- Izbor i obuka modela (*model selection and training*): izbor modela zavisi od prirode problema, nakon čega sledi faza obuke, tokom koje se optimizuju unutrašnji parametri modela radi poboljšanja performansi.

- Evaluacija i validacija (*evaluation and validation*): po završetku obuke, neophodno je testirati performanse modela koristeći relevantne metrike, čime se procenjuje njegova sposobnost generalizacije na novim, nepoznatim podacima.

Svaki od ovih koraka je uslovljen specifičnostima problema, kao i karakteristikama raspoloživih podataka.

Algoritmi mašinskog učenja imaju široku primenu u različitim oblastima. Jedna od značajnih upotreba jeste otkrivanje anomalija i prevara korišćenjem analize istorijskih podataka i identifikovanje vrednosti koje odstupaju od uobičajenih obrazaca (*outliers*). Ovakvi modeli nalaze primenu u osiguravajućim kompanijama i finansijskom sektoru, pri čemu studije potvrđuju da algoritmi mašinskog učenja mogu značajno unaprediti tačnost detekcije prevara u poređenju sa tradicionalnim metodama (Hernandez Aros et al., 2024).

Mašinsko učenje ima primenu u predikativnom održavanju, gde algoritmi mašinskog učenja, analizom podataka prikupljenih korišćenjem različitih senzora, omogućavaju predviđanje kvarova opreme pre nego što do njih dođe. Time se smanjuju neplanirani troškovi usled prekida rada, a ostvaruju se i značajne uštede. U tom kontekstu, kod internet stvari (*Internet of Things,* IoT) senzori kontinuirano prate parametre sistema kako bi identifikovali anomalije (Sriaadhibhatla, 2024). Na primer u medicini, algoritmi mašinskog učenja se koriste za dijagnostikovanje bolesti, procenu rizika obolevanja kod pacijenata i analizu medicinskih slika (MRI, CT). Zahvaljujući sposobnosti mašinskog učenja da identifikuje prethodno nepoznate obrasce, ova tehnologija se može koristiti i za detekciju i blokiranje sajber napada (Kovačević, 2023). Algoritmi mašinskog učenja omogućavaju detaljnu analizu napada na informacione sisteme, procenu rizika od neovlašćenog pristupa i identifikaciju ranjivosti. Njegova uloga dolazi do izražaja usled sve složenijih i učestalijih sajber pretnji, kao i nedostatka stručnog kadra u ovoj oblasti.

Primena mašinskog učenja omogućava precizno otkrivanje različitih pretnji, uključujući zlonamerne programe, sumnjive aktivnosti i mrežne upade, uz mogućnost reagovanja u realnom vremenu, čime se umanjuju posledice napada. Pored toga, algoritmi doprinose analizi neuobičajenog ponašanja korisnika, ranom otkrivanju unutrašnjih pretnji, automatizaciji bezbednosnih procesa i razvoju adaptivnih odbrambenih sistema koji se kontinuirano prilagođavaju novim obrascima napada (Santos & Randaliev, 2024).

C. Duboko učenje

Duboko učenje (*deep learning*) predstavlja napredniju oblast mašinskog učenja zasnovanu na primeni dubokih veštačkih neuronskih mreža, koje su inspirisane načinom funkcionisanja ljudskog mozga (Janiesch et al., 2020). Modeli dubokog učenja se sastoje od više slojeva kroz koje prolaze podaci, pri čemu svaki sloj uči da prepozna sve kompleksnije obrasce. Na taj način, omogućeno je učenje veoma složenih i apstraktnih reprezentativnih podataka.

Primena dubokog učenja

Primena dubokog učenja se pokazala izuzetno efikasnom u brojnim domenima, pa se, između ostalog, primenjuje i u sledećim slučajevima (Santos & Randaliev, 2024):

- **Prepoznavanje slika i objekata**: predstavljaju metode dubokog učenja koje se koriste za klasifikaciju slika, detekciju objekata, prepoznavanje lica i segmentaciju. Našle su primenu u proširenoj realnosti, nadzornim sistemima, medicinskoj dijagnostici i autonomnim vozilima.
- **Obrada prirodnog jezika**: primenjuje se kod analize sentimenta, klasifikacije teksta, mašinskog prevođenja, prepoznavanja imenovanih entiteta, sumiranja teksta, sistema pitanja i odgovora, analize govora (zapisivanje govora u formi teksta, kao i obrnuto). Primeri obuhvataju četbotove, virtuelne asistente, alate za prevođenje i analizu sadržaja.
- **Sistemi za preporuku**: zasnivaju se na kolaborativnom filtriranju sadržaja, pri čemu omogućavaju personalizovane preporuke proizvoda, filmova, muzike i članaka. Ovi algoritmi koriste podatke o korisnicima i proizvodima, kako bi predložili personalizovane preporuke i, na taj način, omogućili bolje korisničko iskustvo. Najčešće se primenjuju u e-trgovini, industriji zabave i platformama za strimovanje sadržaja. Ideja o korišćenju analitike korisničkog ponašanja radi personalizacije usluga ima dužu tradiciju. Tako su Kovačević i saradnici (2010), u kontekstu bibliotečkih informacionih sistema, istakli značaj primene *data mining* tehnika za kreiranje usluga zasnovanih na individualnim korisničkim profilima i istoriji pretraživanja.

Obrada prirodnog jezika

Obrada prirodnog jezika (Natural Language Processing, NLP) je oblast veštačke inteligencije koja se bavi zadacima u vezi sa razumevanjem i generisanjem prirodnog jezika i, pri tome, obuhvata sledeće zadatke: klasifikaciju teksta (*text classification*), prepoznavanje imenovanih entiteta (*named entity recognition*), mašinsko prevođenje (*machine translation*), sisteme pitanja i odgovora (*question-answering systems*), sumiranje teksta (summarization), analizu govora, poput zapisivanja govora u formi teksta i obrnuto (*speech-to-text*; *text-to-speech*).

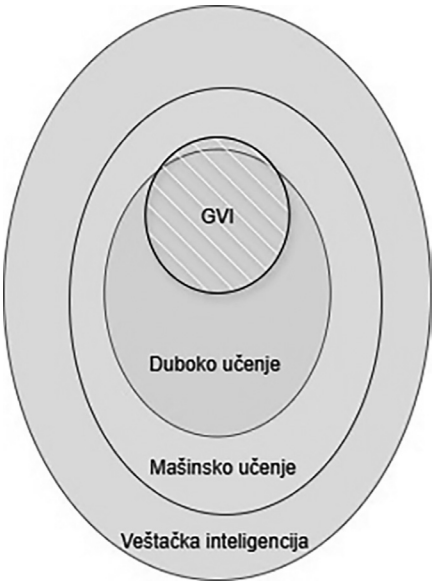
Praktični značaj obrade prirodnog jezika potvrđuje podatak da je milijardu ljudi tokom samo jedne nedelje koristilo onlajn prevođenje, što je predstavljalo trećinu svih internet korisnika (Kemp, 2022). Ovaj primer ilustruje duboku integraciju obrade prirodnog jezika u svakodnevni život. Dalji razvoj ove oblasti usmeren je ka generativnim modelima, čiji je cilj ne samo obrada već i samostalno kreiranje sadržaja. Mašinski generisan tekst predstavlja sadržaj koji je kreiran algoritamski, pri čemu može biti oblikovan u prirodnom jeziku, koji se dalje može modifikovati i proširivati (Crothers et al., 2022). Pri tome, pod prirodnim jezikom se podrazumeva jezik kojim se ljudi svakodnevno sporazumevaju. Takođe, pored prirodnog jezika, savremeni generativni modeli mogu proizvesti i programski kod.

Oblast generisanja prirodnog jezika beleži značajan tehnološki napredak, što omogućava kreiranje tekstova koji po stilu i koherentnosti sve više nalikuju onima koje stvaraju ljudi. Kako bi se uspešno generisao smislen tekst, neophodno je da modeli razumeju namere, uverenja, ciljeve i emocije aktera, što je posebno važno pri rešavanju složenih problema (Kovačević & Erić, 2023).

D. Generativna veštačka inteligencija

Generativna veštačka inteligencija (GVI; eng. *Generative Artificial Intelligence*) predstavlja napredne sisteme mašinskog učenja (Slika 1.1) koji omogućavaju generisanje sintetičkih sadržaja u različitim medijima, tekstu, slikama, audio i video zapisima, pri čemu ti sadržaji verno imitiraju stvarne podatke. Ovi podaci se generišu na osnovu učenja iz velikih skupova podataka za obuku.

Posebnu rasprostranjenost GVI duguje svojoj jednostavnosti upotrebe, jer omogućava kreiranje kvalitetnog sadržaja u izuzetno kratkom vremenu, bez potrebe za naprednim tehničkim znanjem. Koncept GVI nije nov, ali njegova ekspanzija se dešava u poslednjih desetak godina sa razvojem dubokih neuronskih mreža i dostupnošću snažnijih računarskih resursa. Najpoznatiji tipovi modela GVI su sledeći: generativne suparničke mreže (*generative adversarial networks, GAN*), modeli zasnovani na transformer arhitekturi (*transformer-based*) , difuzioni modeli (*diffusion models*), varijacioni autoenkoderi (*variational autoencoders, VAEs*) (Sengar et al., 2024). Svaki od njih je predviđen za određene zadatke.



Slika 1.1. Prikaz podskupova veštačke inteligencije

Sa razvojem mogućnosti GVI-a se razvijala i zabrinutost oko zloupotrebe ovih alata (Marcal et al., 2024), što će biti detaljnije prikazano u sledećim poglavljima.

U brojnim analizama koje su rađene, jedna od dominantnih prednosti primene generativne veštačke inteligencije je povećanje efikasnosti zaposlenih (Loukides, 2023; Sau et al., 2024). Asistenti GVI u korisničkoj podršci povećavaju u proseku produktivnost za 15%, pri čemu efekti značajno variraju među zaposlenima (Brynjolfsson et al., 2024).

Primena generativne veštačke inteligencije

Implementacija GVI intenzivno redefiniše naš svakodnevni život. Generativna veštačka inteligencija ima velike mogućnosti u različitim sektorima, uključujući zdravstvo, obrazovanje, medije, marketing, proizvodnju, maloprodaju, finansije, osiguranje i brojne druge primene Neki od primera generativne veštačke inteligencije su:

- **Generisanje teksta:** Jedan od najčešće korišćenih primera GVI je generisanje teksta. Ova oblast obuhvata različite zadatke, uključujući kreiranje novinskih članaka i blogova, produkciju sadržaja za marketinške kampanje (Jasper, Gpt-J), prevođenje, sumiranje sadržaja, kao i personalizovanu interakciju u okviru četbotova i virtuelnih asistenata. Jedni od najpoznatijih modela u širokoj upotrebi su *ChatGPT* (razvijen od strane kompanije *OpenAI*), *Claude* (razvijen od strane kompanije *Anthropic*) i *LLaMa* (razvijen od strane kompanije *Meta AI*).
- **Generisanje vizuelnog sadržaja (slika i video materijala):** Multimodalni modeli GVI omogućavaju produkciju visokokvalitetnog audio-vizuelnog materijala, što otvara nove mogućnosti u oblastima zabave (*MetaHuman Creator, DeepFaceLab*), oglašavanja i vizuelnih komunikacija (*DALL-E, Stable Diffusion*).
- **Generisanje muzike i zvuka:** GVI modeli imaju sposobnost da sintetišu muziku i zvučne efekte u skladu sa stilom, tempom i emocionalnim tonom definisanim promptom. Ovi sistemi se primenjuju u industriji video-igara, filmskoj produkciji i automatizovanim glasovnim sistemima koji interpretiraju i odgovaraju na glasovne komande, kao što je Amazon Alexa.
- **Generisanje programskog koda:** Generisanje programskog koda predstavlja jednu od najperspektivnijih funkcionalnosti generativne veštačke inteligencije. Ovi sistemi omogućavaju kreiranje programskog koda na osnovu opisa zadatka u prirodnom jeziku, čime se značajno povećava efikasnost procesa softverskog razvoja. Među najpoznatijim rešenjima ove vrste nalaze se alati kao što su *GitHub Copilot* i *Cursor*, koji integrišu velike jezičke modele u softverska razvojna okruženja, kao i jezičke modele poput *OpenAI Codex* i *CodeT5*. Pored generisanja, u pojedinim slučajevima omogućeno je i direktno izvršavanje kreiranog koda, što dodatno ubrzava razvoj, ali istovremeno otvara ozbiljna pitanja bezbednosti i pouzdanosti. Ovaj problem biće detaljnije razmotren u narednim poglavljima.

Prema istraživanju primene GVI u organizacijama (Tully, 2024) na prvom mestu se nalazi generisanje koda (51%), dok su na drugom mestu četbotovi (31%). Sledi pretraživanje informacija (28%), zatim izdvajanje i transformacija informacija (27%), dok

se sumiranje sastanaka (25%) takođe izdvaja kao značajan oblik primene. Ovi rezultati ukazuju da GVI ima sve širu primenu u poslovnim procesima, pri čemu je fokus kako na tehničkim zadacima, tako i na unapređenju komunikacije i efikasnosti timova.

Projekcije pokazuju da će do 2027. godine više od 50% kompanija globalno koristiti GVI modele specifične za industriju, što predstavlja značajan rast u poređenju sa svega 1% u 2023. godini (Sau et al., 2024).

Veliki jezički modeli

Veliki jezički modeli (VJM; eng. *Large Language Models, LLM*) su podvrsta generativne veštačke inteligencije koji pokazuju izuzetne sposobnosti u „razumevanju" i generisanju prirodnog jezika (Minaee et al., 2024). Ovi modeli su obučeni na velikim skupovima podataka, što im omogućava prepoznavanje nijansi jezika i generisanje koherentnog i kontekstualno relevantnog teksta.

Veliki jezički modeli su intenzivno polje razvoja i istraživanja (Fan et al., 2023; Naveed et al., 2023; Raiaan et al., 2024; Yin et al., 2023) koji imaju široki spektar primene u realnom svetu, i mogu uticati na mnoge aspekte života.

Veliki jezički modeli se zasnivaju na modelima dubokog učenja, pri čemu danas dominira transformer arhitektura (Vaswani et al., 2017), koja je omogućila veliki napredak u njihovom razvoju.

Prvi jednostavni jezički modeli su se pojavili još sredinom prošlog veka (Shannon, 1951). Međutim, ključna prekretnica u njihovom razvoju nastala je uvođenjem tehnike samonadgledanog učenja (*self-supervision*). Ovom tehnikom je omogućeno obučavanje modela na ogromnim korpusima podataka bez potrebe za ručnim označavanjem podataka (Huyen, 2025).

Transformer modeli razlažu tekst na manje jedinice - tokene, koji mogu biti cele reči ili njihovi delovi. Jedan važan tip ovih modela su autoregresivni modeli, koji funkcionišu tako što predviđaju sledeći token na osnovu prethodnog niza. Na taj način model „nastavlja" tekst dodajući token po token. Najpoznatiji primeri ove arhitekture su *GPT* i *Llama*. Karakteristike najznačajnijih velikih modela su prikazane u Tabeli 1.1.

Pojava četbota ChatGPT 2022. godine, zasnovanog na VJM rezultirala je ubrzanim usvajanjem ove aplikacije. Za samo dva meseca od pokretanja, broj korisnika dostigao je 100 miliona (Milmo, 2023).

Uspeh *GPT* modela delimično se zasniva na kombinaciji nadgledanog učenja i učenja potkrepljivanjem (*reinforcement learning*) tokom procesa obuke. Naročito je važna uloga ljudi u drugoj fazi obuke. Nakon inicijalnog treniranja, treneri veštačke inteligencije ocenjuju i rangiraju odgovore modela u odnosu na kvalitet i usklađenost sa zadatim instrukcijama. Ovaj proces, poznat kao učenje potkrepljivanjem uz ljudsku povratnu informaciju (*reinforcement learning from human feedback, RLHF*), omogućava modelu da statistički uči da preferira ishode koji su bliži ljudskim očekivanjima, čime se povećava praktična korisnost i smanjuje verovatnoća nepoželjnih odgovora.

Međutim, važno je naglasiti da iako VJM deluju da prividno „razumeju" tekst, oni u suštini ne razumeju značenje, istinitost ili nameru, već generišu odgovore na osnovu statističke verovatnoće pojave reči i obrazaca iz obučavanja podataka.

Tabela 1.1. Prikaz nekih od najznačajnijih velikih jezičkih modela (po ugledu na Caballar & Stryker, 2025).

Model	Provajder	Datum objavljivanja	Broj parametara	Kontekstualni prozor	Licenca	Pristup	Ulaz
Claude 3.7 Sonnet https://claude.ai	Anthropic	II 2025	Nije javno otkriveno	200000 tokena	Vlasnička	Anthropic API, Amazon Bedrock, Google Cloud Vertex AI	MM (slika, tekst)
DeepSeek-R1 https://deepseek.ai	DeepSeek	I 2025	671 milijarda	128000 tokena	Open source	DeepSeek API, Hugging Face	Tekst
Gemini 2.0 https://gemini.google.com	Google	XII 2024	Nije javno otkriveno	1 milion tokena	Vlasnička	Gemini API, Google AI Studio, Google Cloud Vertex AI	MM (audio, slika, tekst, video)
GPT-4o https://openai.com	OpenAI	V 2024	Nije javno otkriveno	128000 tokena	Vlasnička	OpenAI API	MM (audio, slika, tekst, video)
Granite 3.2 https://www.ibm.com/granite	IBM	II 2025	Do 8 milijardi	128000 tokena	Open source	IBM watsonx.ai, Hugging Face, LM Studio, Ollama	MM (slika, tekst)
Grok 3 https://grok.com	xAI	II 2025	2.7 biliona[1]	Do 1 miliona tokena[2]	Vlasnička	xAI API	MM (slika, tekst)
Llama 3.3 https://www.llama.com	Meta	XII 2024	70 milijardi[3]	128000 tokena	Open source	Meta, Hugging Face, Kaggle	Tekst
o1 https://openai.com/o1/	OpenAI	IX 2024	Nije javno otkriveno	Do 200000 tokena	Vlasnička	OpenAI API	MM (slika, tekst)
Qwen 2.5 https://qwen.ai	Alibaba Cloud	IX 2024	Do 72 milijarde	Do 1 milion tokena	Open source	Alibaba Cloud, Hugging Face	MM (audio, slika, tekst, video)

[1] https://opencv.org/blog/grok-3

[2] https://x.ai/news/grok-3

[3] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

Veliki jezički modeli predstavljaju značajan napredak u tehnologiji veštačke inteligencije, koji omogućavaju rad na različitim jezicima i unutar različitih kulturnih konteksta (Tabela 1.1). Postoje i domensko-specifični modeli, obučeni za pojedine oblasti, kao što su medicina, pravo ili finansije, koji u tim domenima pokazuju bolje performanse od opštih modela. Na primer, u medicini se VJM mogu koristiti kao pomoć pacijentima, radi boljeg razumevanja dijagnoza i terapijskih procedura (Busch et al., 2025), kao i podrška studentima medicine u procesu učenja i pripreme za kliničku praksu (Lucas et al., 2024).

Priroda velikih jezičkih modela

Prilikom generisanja teksta, VJM predviđa narednu reč na osnovu prethodnog konteksta. Budući da su ovi modeli stohastičke prirode, ne proizvode uvek identične odgovore za isti ulaz, već njihov izlaz zavisi od procena verovatnoće različitih mogućih nastavaka. Za razliku od toga, kod deterministčkih sistema uvek se generišu isti rezultati za isti ulaz. Iako ova stohastička priroda može biti korisna u zadacima koji zahtevaju kreativnost ili varijabilnost, ona istovremeno predstavlja ozbiljnu prepreku kada je potrebna tačnost i pouzdanost.

Halucinacije predstavljaju pojavu u kojoj model generiše sadržaj koji nije zasnovan na činjenicama, što će biti posebno razrađeno u narednim poglavljima. Upravo fenomen halucinacija kod VJM predstavlja jednu od glavnih barijera za usvajanje sistema veštačke inteligencije u organizacijama, jer podstiče skepticizam u pogledu njihove verodostojnosti i kredibiliteta.

Kako se VJM sve intenzivnije integrišu u brojne digitalne infrastrukture, stoga je neophodno obezbediti njihovu bezbednost i robusnost protiv velikog broja sofisticiranih pretnji. Ovo je izuzetno važno u kontekstu primene u osetljivim sektorima, zbog čega je neophodno analizirati realne pretnje i ranjivosti.

Još u ranoj fazi razvoja GPT modela pojavila se zabrinutost u vezi sa potencijalnim zloupotrebama, što je dovelo do toga da objavljivanje *GPT-2* modela bude odloženo za devet meseci (Radford et al., 2019).

Sličnosti i razlike kod ljudi i velikih jezičkih modela

Schneier i Sanders (2025) su uporedili ponašanje VJM sa ljudskim reakcijama i identifikovali niz značajnih sličnosti. Na primer, male varijacije u pitanju mogu dovesti do značajno različitih odgovora kod VJM, što je slučaj i kod ljudi, gde način formulisanja pitanja značajno utiče na percepciju i odgovor. Pored toga, sličnosti između VJM i ljudi uključuju sledeće (Schneier & Sanders, 2025):

- VJM ima tendenciju da ponavlja najfrekventnije obrasce iz podataka za obuku, što je analogno ljudskoj sklonosti da se češće prisete poznatih stvari.
- Prilikom sažimanja teksta VJM se često fokusira na informacije sa početka i kraja teksta, što je slično kao kod ljudi. Ipak, evidentiran je određeni napredak u ovom segmentu.

- Određene tehnike za manipulaciju modelima, poput predstavljanja zahteva kao šale ili nečeg korisnog, zatim motivacija nagradama ili pretnjama, oponašaju obrasce socijalne manipulacije koji se viđaju kod ljudi. U eksperimentima je ispitivan i uticaj nagrade ili pretnje na generisanje odgovora modela.
- Neki trikovi za „hakovanje" VJM-a podsećaju na ljudsku socijalnu manipulaciju, kao što je predstavljanje zahteva šale ili nečeg korisnog, zatim motivacija nagradom ili pretnjom. Istraživači su testirali kako motivacija nagradom ili pretnjom može uticati na odgovore modela.

Veliki jezički modeli pokazuju fenomen ulizištva (*sycophancy*), odnosno tendenciju modela da usklađuje odgovore sa korisnikovim stavovima, bez obzira na njihovu ispravnost. Ovaj efekat je dodatno pojačan tehnikama preferencijalnog usklađivanja, koje optimizuju zadovoljstvo korisnika, što može narušiti istinitost i neutralnost odgovora (Zhang, K. et al., 2025).

Primećeno je da pojedine metode poput umetanja opasnih zahteva pomoću ASCII umetnosti (tehnike prikazivanja vizuelnih obrazaca korišćenjem znakova iz ASCII tabele) mogu da zaobiđu bezbednosne filtere VJM, iako su takvi pokušaji čoveku lako prepoznatljivi (Jiang et al., 2025).

Kako bi se izbegle ozbiljne greške, GVI treba koristiti samo u oblastima u kojima je njena tačnost, pouzdanost i robusnost višestruko potvrđena.

Foundation modeli i model kao servis

Obučavanje velikih jezičkih modela, kao što su *GPT* ili *Gemini*, predstavlja izuzetno skup i tehnički zahtevan proces, koji zahteva ogromne resurse, uključujući velike skupove podataka, računarsku snagu i visoko specijalizovana znanja. Zbog izuzetno visoke kompleksnosti i potrebnih resursa, razvoj *foundation* modela trenutno je moguć isključivo za velike tehnološke korporacije, kao što su *Google*, *Meta*, *Microsoft* i *Baidu*, kao i pojedine države (npr. Japan, UAE) ili dobro finansirane kompanije poput *OpenAI*, *Antropic* i *Mistral* (Huyen, 2025).

Foundation modeli predstavljaju generativne modele veštačke inteligencije opšte namene, koji su sposobni za izvršavanje širokog spektra zadataka, čime se omogućava razvoj velikog broja aplikacija u različitim domenima (Huyen, 2024). Njihova ključna prednost ogleda se u mogućnosti da se, nakon inicijalnog obučavanja, relativno lako prilagode specifičnim domenima kroz proces finog podešavanja (*fine-tuning*), čime se ostvaruje značajna ušteda u resursima i vremenu u poređenju sa razvojem novog modela od početka.

Savremeni *foundational* modeli su multimodalni modeli, tj. integrišu multimodalne sposobnosti, obrađuju tekst, slike, zvuk i video.

Foundation modeli nalaze široku primenu u različitim oblastima, uključujući (Huyen, 2025):

- Kodiranje, naročito na osnovnom nivou,
- Generisanje visokokvalitetnog multimedijalnog sadržaja,
- Personalizovano i adaptivno učenje,

- Virtuelne asistente i korisničku podršku,
- Agregaciju i organizaciju informacija,
- Automatizaciju poslovnih procesa.

Koncept „model kao servis"

Postojanje *foundation* modela dovodi do novog koncepta „model kao servis" (*Model-as-a-service, MaaS*), u okviru kojeg organizacije koriste unapred trenirane *foundation* modele i prilagođavaju ih sopstvenim potrebama (Huyen, 2025). Ovaj pristup, popularizovan od strane kompanija kao što su *OpenAI* i drugih vodećih tehnoloških kompanija, omogućava brzo i jednostavno kreiranje aplikacija uz minimalno kodiranje, bez potrebe za naprednim tehničkim znanjima. Na taj način, omogućeno je širokom krugu korisnika da prilagode *foundation* modele sopstvenim specifičnim zahtevima i aplikacijama (Hofman, 2022).

„Model kao servis" značajno ubrzava diseminaciju veštačke inteligencije, jer omogućava razvoj aplikacija na osnovu prethodno obučenih modela, bez potrebe da se ti modeli razvijaju od nule. Ovim pristupom se drastično snižavaju tehničke, finansijske i vremenske barijere za implementaciju rešenja veštačke inteligencije, dok potreba za njima evidentno raste (Huyen, 2025). Rast inženjeringa veštačke inteligencije (*AI engineering*) proističe upravo iz kombinacije povećane potražnje za prilagodljivim rešenjima veštačke inteligencije i smanjenju prepreka za njihovu implementaciju.

Model kao servis i uticaj na bezbednost

Model kao servis povećava ranjivost modela jer javni API proširuje površinu napada (Hou et al., 2025). Takođe, zajedničko okruženje donosi dodatne rizike, a korisnici gube direktnu kontrolu nad infrastrukturom. Dok model kao servis pruža prednosti u vidu skalabilnosti i jednostavne implementacije, bezbednosni kompromis je značajan, naročito za organizacije koje obrađuju osetljive podatke. Suprotno tome, lokalne implementacije, iako zahtevnije u pogledu resursa i održavanja, omogućavaju veću bezbednosnu autonomiju, jer se podaci čuvaju unutar kontrolisane infrastrukture, pri čemu se eliminišu rizici zajedničkog okruženja (Hou et al., 2025). Ovi rizici obuhvataju potencijalno curenje podataka među korisnicima, deljenje resursa i platformi, izloženost zajedničkim ranjivostima infrastrukture, kao i smanjenu kontrolu i vidljivost procesa.

Inženjeri veštačke inteligencije i inženjeri mašinskog učenja

Inženjerstvo veštačke inteligencije (*AI Engieenging*) se razvilo iz inženjerstva mašinskog učenje (*Machine Learning Engineering*), ali se od njega suštinski razlikuje u pristupu, ciljevima i zahtevima za tehničkom ekspertizom. Za razliku od inženjerstva mašinskog učenja, koje zahteva dubinsko razumevanje algoritama i dugotrajan proces podešavanja parametara tokom obučavanja modela od nule, inženjerstvo VI se zasniva na korišćenju unapred obučenih *foundation* modela. Time se omogućava

šira primena i brža implementacija tehnologije VI u različitim domenima, jer obuka inženjera VI zahteva mnogo manje vremena i kompleksnosti u poređenju sa obukom inženjera mašinskog učenja.

Inženjeri VI dopunjuju znanja iz oblasti softverskog inženjerstva razumevanjem mogućnosti i ograničenja modela (Huyen, 2025), pri čemu se fokusiraju na aplikacioni sloj, a ne na razvoj samog modela od nule. Na taj način, prilagođavanje *foundation* modela specifičnim zadacima može se realizovati mnogo brže, i uz manje napora (na primer za samo nekoliko dana rada i na malom broju primera), u poređenju sa višemesečnim procesima obuke modela nad milionskim skupovima podataka.

Ključne tehnike koje primenjuju inženjeri VI obuhvataju:

- **Prompt inženjering**: predstavlja proces oblikovanja prompta sa ciljem postizanja željenog ponašanja ili odgovora od modela VI. Prompt predstavlja svaki unos, najčešće instrukcije ili upit formulisan na prirodnom jeziku, kojim se definišu zadatak, kontekst i željeni format izlaza, kao na primer „Sumiraj mi ovaj tekst, u akademskom stilu".
- ***Retrieval augmented generation* (*RAG*)**: hibridna arhitektura koja kombinuje sposobnosti dva modela - modela za pretraživanje i modela za generisanje jezika, čime prevazilazi ograničenja klasičnih VJM-a koji se oslanjaju isključivo na znanje uskladišteno tokom procesa obučavanja (Gupta et al., 2024). RAG funkcioniše po principu „pretraži-pa-generiši", prvi put predstavljenom u radu Chen i saradnici (2017), gde je sistem za odgovaranje na pitanja preuzimao pet najrelevantnijih stranica sa Vikipedije, dok je formalizovan i proširen u radu (Lewis et al., 2021) kao metod za uključivanje eksternih baza znanja u cilju smanjenja halucinacija.
- **Fino podešavanje (*fine-tuning*)**: predstavlja dodatno obučavanje *foundation* modela na specijalizovanim, domen-specifičnim skupovima podataka radi postizanja boljih performansi u određenom zadatku ili domenu. Za razliku od klasičnog obučavanja[4], gde se model od početka trenira na masivnim korpusima heterogenih podataka, fino podešavanje koristi manje, ciljane skupove. Efikasnost finog podešavanja zavisi od kvaliteta podataka, samog zadatka, kao i izbora metodologije i načina obučavanja.

U daljem tekstu, radi konceptualne jasnoće, koristiće se termin veliki jezički modeli (VJM) kao krovni termin u proširenom smislu, koji obuhvata multimodalne *foundation* modele.

Jedan od ključnih aspekata primene transformera jeste oblast prompt inženjeringa, koji obuhvata pažljivo osmišljavanje ulaza radi optimizacije performansi modela (Brown et al. 2020).

Promptovi predstavljaju ključni element za funkcionisanje i optimizaciju GVI, pošto omogućavaju formulisanje pitanja različitih dužina i složenosti, čime se model usmerava ka željenom cilju (Kovačević & Erić, 2023). Kvalitet i jasnoća prompta direktno utiče na tačnost i relevantnost generisanih odgovora.

[4] Često se u literaturi koristi i termin unapred obučen (*pre-trainned*).

Efikasnost VJM u velikoj meri zavisi od kvalitetnog prompta, zbog čega je **prompt inženjering** od suštinske važnosti. On podrazumeva precizno i jasno oblikovanje uputstva kako bi se obezbedili tačni i relevantni odgovori. Osnovne strategije prompt inženjeringa su sledeće (Sahoo et al., 2024):

- Prompt bez primera (*zero-shot prompting*) – označava postupak u kojem se jezičkom modelu zadaje instrukcija bez ikakvih dodatnih primera. Model u tom slučaju koristi znanje stečeno tokom treniranja i sposobnost generalizacije kako bi rešio zadatak.
- Prompt sa nekoliko primera (*few-shot prompting*) – podrazumeva davanje nekoliko ulazno-izlaznih primera zajedno sa instrukcijom, čime se modelu demonstrira željeni obrazac odgovora. Na taj način se značajno poboljšavaju performanse u odnosu na *zero-shot* pristup. Principi kvalitetno napisanog prompta obuhvata jasno formulisanje zahteva, precizne instrukcije, iterativno testiranje i podešavanje, kontekstualno prilagođavanje, kao i razlaganje složenih zadataka na jednostavne korake.

Razvoj foundation modela i koncepta modela kao servisa omogućio je korišćenje unapred obučenih generativnih modela kao osnove za širok spektar aplikacija (Huyen, 2025). Međutim, njihova primena putem prompt inženjeringa, RAG-a ili finog podešavanja uglavnom je reaktivna, jer modeli odgovaraju samo na zadate upite. Sledeća faza u razvoju generativne veštačke inteligencije obuhvata agente veštačke inteligencije, koji se zasnivaju na istim modelima, ali uključuju dodatne komponente poput memorije, planiranja i povezivanja sa spoljnim alatima, čime prelaze iz uloge pasivnih generatora u interaktivne i autonomne sisteme sposobne za izvršavanje složenih zadataka (Biswas & Talukdar, 2025).

Agenti veštačke inteligencije

Agenti veštačke inteligencije se zasnivaju na velikim jezičkim modelima i predstavljaju naprednu primenu ove tehnologije a ne specifičnu arhitekturu (Bisvas & Talukdar, 2025). Pored VJM na kojim se zasnivaju ovi agenti, često se uključuju i dodatne komponente koje im omogućavaju širu funkcionalnost. Za razliku od „klasičnih VJM" sistema, agenti su projektovani da budu interaktivni, prilagodljivi i sposobni za izvršavanje složenih zadataka u više koraka.

Jedan od primera korišćenja agenata je turistički veb sajt sa četbotom koji korisnicima pomaže u rezervaciji avio-karata i hotela: korisnik vodi dijalog nalik ljudskom, dok agent u pozadini obavlja niz zadataka – od pozivanja API-ja, preko pretrage letova, do konačne rezervacije. Iako deluje jednostavno, ovaj proces podrazumeva složene mehanizme, gde model koristi introspektivne metode poput lanca razmišljanja (*Chain of Thought – CoT*), da bi mogao da planira korake i odluči koje alate da koristi.

Platforme za evaluaciju velikih jezičkih modela

Evaluacija VJM je od ključnog značaja usled njihove sve šire primene u kompleksnim i visokorizičnim okruženjima, gde je neophodno osigurati bezbednu i odgovornu

upotrebu. Ipak, dosadašnja istraživanja uglavnom su stavljala fokus na performanse, zanemarujući analizu potencijalnih slabosti, zbog čega je za prevazilaženje ovog problema kreirana platforma Phare (Jeune et al., 2025).

Platforma *Phare*[5] je specijalizovani benčmark za evaluaciju VJM iz ugla sigurnosti i bezbednosti, sa fokusom na pet kategorija: prosečna sigurnost, halucinacije, otpornost na štetu, pristrasnost i *jailbreaking*. Centralni deo platforme čini hijerarhijski sistem rangiranja, u kojem se za svaki model dobija ukupan rang na osnovu agregiranih rezultata. Ova analiza je otkrila obrasce sistemskih ranjivosti, uključujući sklonosti modela da se prilagođavaju stavovima korisnika, njihovu osetljivost na promenu promta i tendenciju ka kreiranju stereotipa (Jeune et al., 2025).

Pored istraživačkih inicijativa kao što je Phare, evaluacija VJM sve više dobija i regulatornu dimenziju. U tom pravcu je i okvir *COMPL-AI* koji obuhvata prvu tehničku interpretaciju regulatornih zahteva *EU AI Act-a* (EU, 2024) i prevodi iste u merljive kriterijume za VJM (Guldimann et al., 2025). Okvir uključuje i benčmarking alat otvorenog koda, koji je oslonjen na savremene metrike i testove, i usklađen sa zakonodavstvom. Kroz analizu značajnih VJM-ova autori su ukazali na važne nedostatke postojećih modela u oblastima robusnosti, sigurnosti, raznolikosti i pravednosti, ali i na ograničenja postojećih evaluacionih metoda koje ne mogu da obuhvate sve tehničke zahteve. Pri tome, autori smatraju da će *EU AI Act* imati značajan uticaj na budući razvoj velikih jezičkih modela, jer će fokus istraživanja i implementacije morati da obuhvata ne samo performanse modela, već i privatnost, pravednost, robusnost i sigurnost (Guldimann et al., 2025).

[5] https://phare.giskard.ai/