PREDGOVOR

Veštačka inteligencija oduvek je bila vođena ambicijom da se razviju sistemi koji ne samo da izvršavaju unapred definisane zadatke, već i umeju da razmišljaju, prilagođavaju se i deluju samostalno. Sa usponom generativne veštačke inteligencije i agentnih sistema, nalazimo se na prekretnici – u trenutku kada veštački agenti mogu da razmišljaju o sopstvenim iskustvima, razvijaju strategije i sarađuju prirodno i efikasno. Ova knjiga dolazi upravo u tom ključnom momentu, nudeći celovit uvid u ovu novu i sve značajniju oblast.

Kroz višedecenijsko iskustvo u radu sa veštačkom inteligencijom – od osnovnih istraživanja do razvoja proizvoda – u ulozi tehničkog lidera u istraživačkoj diviziji kompanije Microsoft, Microsoft Research, i u mladim tehnološkim kompanijama koje su podržali investicioni fondovi, imao sam priliku da doprinesem projektima koji su ostavili snažan trag, kao i da učestvujem u razvoju velikih sistema veštačke inteligencije koji se koriste širom sveta, kako u komercijalnom, tako i u poslovnom okruženju. Moj angažman u okviru IEEE organizacije, kroz različite liderske uloge u oblasti veštačke inteligencije, pružio mi je dodatnu perspektivu o važnosti etičkog i odgovornog razvoja ove tehnologije.

Autore ove knjige, Vrika Talukdara i Anđanavu Bisvasa, upoznao sam tokom zajedničkog rada na globalnim IEEE inicijativama posvećenim razvoju veštačke inteligencije. Njihovo znanje i iskustvo potvrđeno je značajnim ulogama u kompaniji AWS, radom na projektima iz oblasti mašinskog učenja i veštačke inteligencije, kao i kroz akademski doprinos. Obojica su prepoznati kao vodeći stručnjaci u oblasti generativne veštačke inteligencije, sa bogatim iskustvom i dubokim razumevanjem složenih tehnoloških i etičkih izazova. Njihovo znanje čini ih kvalifikovanim autorima za knjigu ove vrste.

Knjiga je podeljena na tri dela. Prvi deo sadrži jasna objašnjenja osnovnih pojmova generativne veštačke inteligencije i ključnih karakteristika agentnih sistema. Čitaoci se upoznaju sa osnovnim konceptima – od varijacionih autokodera i generativnih suparničkih mreža do složenih arhitektura agenata – što omogućava i početnicima i onima sa više iskustva da se lakše snađu u ovoj složenoj oblasti.

Drugi deo posvećen je praktičnim metodima za razvoj svesnih inteligentnih agenata koji su u stanju da se prilagođavaju, odlučuju i uspešno sarađuju sa drugim agentima. Poglavlja posvećena refleksiji, introspekciji, planiranju i saradnji među agentima pružaju korisne smernice i uvide za razvoj sistema koji neprekidno uče i usavršavaju se.

Treći deo bavi se širim pitanjima koja prate primenu generativnih agentnih sistema u realnim okruženjima. Ističu se teme kao što su transparentnost, jasnoća tumačenja, bezbednost i etički nadzor, čime knjiga jasno ukazuje na odgovornosti koje donosi upotreba naprednih tehnologija veštačke inteligencije. Ponuđen je okvir koji pomaže u izgradnji poverenja i obezbeđuje da razvoj ovih moćnih alata ostane u skladu sa društvenim vrednostima i etičkim načelima.

Ova knjiga je mnogo više od tehničkog priručnika – ona je vodič za oblikovanje budućnosti generativne veštačke inteligencije i inteligentnih agenata. Namenjena je istraživačima, industrijskim liderima i pronalazačima koji žele da učestvuju u oblikovanju sledeće etape razvoja ove tehnologije – na način koji je odgovoran i društveno koristan.

Iskreno preporučujem ovu sadržajnu i aktuelnu knjigu. Pozivam vas da se udubite u naredna poglavlja i priključite se zajednici koja aktivno učestvuje u razvoju uzbudljivog i dinamičnog sveta veštačke inteligencije.

Metju R. Skot
tehnički direktor, Minset.ai
Proleće 2025

PREDGOVOR 2

Svet je bio zatečen kada je u novembru 2022. lansiran ChatGPT. Bio je to trenutak iznenadnog uvida — kao da razgovarate sa stvarnom osobom. Odgovori nisu bili samo tečni, već i informativni i korisni. Od tada se ova tehnologija razvija neverovatnom brzinom, a stotine miliona ljudi koriste ChatGPT i slične sisteme kao pomoć pri pisanju ili za pronalaženje odgovora koji se teško mogu dobiti uz pomoć klasičnih veb pretraživača.

Međutim, kada nešto zatražimo od ljudskog asistenta — na primer, da rezerviše putovanje — on to i uradi: pronađe letove i rezerviše hotel. U tome se ogleda suština agentnih sistema. Veliki jezički modeli (LLM), poput onih koji pokreću ChatGPT, ne služe samo za generisanje teksta, već mogu da preduzimaju konkretne radnje — da pišu kod, pozivaju API interfejse i komuniciraju sa veb servisima. Kao što je Veb nekada služio samo za prikaz statičnog sadržaja, a kasnije omogućio interaktivne aplikacije, agentni sistemi danas donose sličnu revoluciju u našem odnosu prema veštačkoj inteligenciji.

Za razliku od drugih knjiga koje se pretežno bave teorijskim osnovama agentnih sistema, ova knjiga stavlja akcenat na praktičnu primenu. Autori nude konkretne primere koda, stvarne slučajeve upotrebe i detaljna uputstva za integraciju agentnih radnih tokova u poslovno okruženje radi rešavanja konkretnih izazova. Ova knjiga je pristupačan, ali sveobuhvatan izvor znanja za softverske inženjere, inženjere mašinskog učenja, tehničke rukovodioce, direktore i druge donosioce odluka koji žele da upoznaju agentne sisteme.

Svi koji su koristili asistente zasnovane na veštačkoj inteligenciji, kao što je ChatGPT, znaju koliko su korisni kao pomoć u pisanju ili programiranju — uz uslov da korisnik, pre upotrebe, proveri dobijeni rezultat. Ipak, veštačka inteligencija ponekad generiše netačne informacije s toliko uverenja da joj korisnik može olako poverovati. Za razliku od ljudi, koji znaju da izraze sumnju rečima poput „Nisam siguran, ali mislim...“, sistemi veštačke inteligencije nemaju mehanizam kojim bi ukazali na sopstvenu nesigurnost. Želimo da naši agenti na bazi veštačke inteligencije postupaju ispravno — i da nas pitaju kad nisu sigurni. Zbog toga se ova knjiga bavi ključnim pitanjima kao što su poverenje, transparentnost, jasnoća tumačenja i pouzdanost, ali i etičkim razmatranjima poput pristrasnosti i bezbednosti. Opisuje kako agentni sistemi mogu biti prilagođeni korisnicima i svesni konteksta, te kako mogu da planiraju i reaguju u skladu s posledicama svojih postupaka.

Kao neko ko se decenijama bavi veštačkom inteligencijom, verujem da će ova knjiga biti dragocen resurs za studente, istraživače i stručnjake, jer pruža jasan pregled puta napred u doba u kojoj inteligentni sistemi oblikuju naše okruženje. Vrika Talukdara sam upoznao na jednom panelu u okviru IEEE konferencije i odmah me je impresionirala njegova sposobnost da poveže teoriju i praksu. U knjizi Izgradnja agentnih sistema veštačke inteligencije, on i Anđanava Bisvas upravo to i čine — nude praktičan vodič za budućnost automatizacije zasnovane na veštačkoj inteligenciji.

Dr Aleks Asero,
član Nacionalne inženjerske akademije,
IEEE počasni član

SARADNICI

O AUTORIMA

Anđanava Bisvas je višestruko nagrađivani viši arhitekta rešenja iz oblasti veštačke inteligencije, sa više od 17 godina iskustva u industriji. Stručnjak je za mašinsko učenje, generativnu veštačku inteligenciju, obradu prirodnog jezika, duboko učenje, obradu podataka i arhitekturu oblaka. Sarađuje sa velikim kompanijama na razvoju i skaliranju naprednih sistema veštačke inteligencije u oblaku. Prepoznat je po doprinosu primenjenoj veštačkoj inteligenciji. Objavio je radove u više naučnih časopisa i aktivno doprinosi projektima otvorenog koda u oblasti veštačke inteligencije i mašinskog učenja. Nosilac je zvanja počasnog člana britanskog društva BCS, britanskog IET i indijskog IETE, i viši član IEEE organizacije. Čest je govornik na stručnim skupovima i bio je na rukovodećim pozicijama u kompanijama kao što su IBM i Oracle. Poreklom iz Indije, danas živi u San Dijegu, u Kaliforniji, sa suprugom i sinom, i nastavlja da daje doprinos i bude inspiracija tehničkoj zajednici.

Želim da izrazim duboku zahvalnost svima koji su me pratili na ovom putu razvoja generativne veštačke inteligencije. Mojim kolegama, sadašnjim i bivšim, koji su me podsticali da razmišljam drugačije i s kojima sam delio uzbuđenje otkrivanja ove oblasti koja se ubrzano razvija — vaša saradnja bila je neprocenjiva. Mojim mentorima, koji su u mojim idejama prepoznali potencijal i vodili me sa mudrošću i strpljenjem — vaša vera u mene učinila je ovu knjigu mogućom.

Mome izuzetnom sinu, čija jedanaestogodišnja maštovitost svakodnevno podseća da budućnost pripada onima koji umeju da je zamisle. Tvoja pitanja me podstiču da složene pojave objašnjavam jednostavno — što je možda i najvrednija veština kada je reč o generativnoj veštačkoj inteligenciji. I mojoj supruzi, mojoj tihoj saputnici kroz besane noći i trenutke otkrovenja — tvoje strpljenje je temelj na kom su ove ideje izgrađene.

Mojoj majci, čija mudrost nadilazi tehniku i čija podrška ne poznaje granice. I mome ocu, čije odsustvo duboko osećam, ali čiji uticaj prepoznajem u svakom izazovu koji prihvatam. Znam da bi bio ponosan na ovu knjigu.

Kao što je Ada Lavljes jednom zapisala: „Analitička mašina plete algebarske obrasce kao što Žakardov razboj plete cveće i listove.“ Današnja generativna veštačka inteligencija nastavlja tu čudesnu igru između ljudske kreativnosti i računarskih mogućnosti — ples koji sam imao privilegiju da zabeležim.

Ova knjiga je namenjena svima koji veruju da tehnologija treba da služi čoveku — da unapređuje ljudskost, a ne da je potiskuje.

– Anđanava Bisvas

Vrik Talukdar je vizionarski lider u oblasti generativne veštačke inteligencije u kompaniji Amazon, sa više od dve decenije međunarodnog iskustva u veštačkoj inteligenciji, računarstvu u oblaku i razvoju proizvoda. Kao pionir u transformacijama zasnovanim na veštačkoj inteligenciji, predvodio je brojne projekte modernizacije velikih razmera koji podstiču razvoj kompanija i utiču na milione korisnika širom sveta. Bio je na čelu razvoja i primene nagrađivanih tehnologija iz oblasti veštačke inteligencije i mašinskog učenja, koje se danas u velikim razmerama koriste u kompanijama sa Fortune 500 liste, oblikujući primenu veštačke inteligencije u stvarnom svetu.

Prepoznat kao značajan istraživač u ovoj oblasti, Vrikov rad u generativnoj veštačkoj inteligenciji, multimodalnim sistemima, obradi prirodnog jezika i računarskom vidu često se citira. Kao viši član IEEE organizacije, predsedavajući i panelista na brojnim industrijskim forumima, savetuje međunarodne odbore kao što su CTSoc Industry Forums i NIC, gde učestvuje u definisanju industrijskih standarda i oblikovanju uloge veštačke inteligencije u budućnosti. Svoja dostignuća često predstavlja na prestižnim konferencijama kao što su Svetski tehnološki samit, IEEE HKN, ICCE, CERAWeek i ADIPEC, gde povezuje najnovija istraživanja sa praktičnim rešenjima koja ubrzavaju inovacije u industriji.

Sa čvrstim uporištem u oblasti računarstva, kao ko-predsedavajući IEEE NIC organizacije, posvećen je osnaživanju mladih stručnjaka. Kao autor i inteligentni lider, nastavlja da pomera granice veštačke inteligencije i inspiriše buduće pronalazače. Vrik živi u Kaliforniji sa svojom porodicom.

Kao što je rekla Merijan Rajt Edelman: „Obrazovanje služi tome da poboljša tuđe živote i ostavi zajednicu i svet boljim nego što ih je zateklo."

Duboko sam zahvalan svima koji su me podržavali, usmeravali i inspirisali tokom ovog putovanja.

Mojoj porodici – roditeljima, sinu i supruzi – vaša nepokolebljiva vera u mene je temelj moje snage.

Mome sinu, čija beskrajna radoznalost da sazna „na čemu tata radi" svakodnevno me je podsećala na to koliko je istraživanje važno. Tvoja pitanja su budila nove ideje onda kada su mi najviše trebale, a tvoja mašta me podsećala na neograničene mogućnosti koje ovaj rad otvara. Naučio si me da inspiracija često dolazi u obliku jednostavnog pitanja od nekoga ko veruje da možeš da objasniš univerzum.

Mojim roditeljima, čija me podrška, vera i želja da razumeju svet koji se brzo menja stalno podsećaju da je znanje tu da se deli, a ne da se gomila. Vaša posvećenost učenju tokom celog života bila mi je moralna smernica, a vaš ponos moj veliki podstrek.

Mojoj supruzi, čija tiha snaga, strpljenje i neprekidna podrška nisu samo omogućili da ovo putovanje bude moguće, već su mu dale i duboko značenje. Tvoje prisustvo tokom besanih noći, nepokolebljiva vera u mene kroz sve izazove i tvoja smirujuća prisutnost davali su mi jasnoću da privedem kraju ono što sam započeo. Bila si moj oslonac i inspiracija.

Zajednici istraživača u oblasti veštačke inteligencije – vaša neprestana potraga za inovacijama nadahnjuje me i podseća na pravu vrednost ovog rada. Nakon godina pisanja naučnih radova, članaka i blogova, pisanje prve knjige bilo je transformativno iskustvo – ono koje me je nateralo da složenost pretočim u jasnoću, a ideje u konkretnu vrednost.

I svim prijateljima i onima koji su mi želeli dobro – vaše prisustvo u trenucima izazova pretvaralo je prepreke u prekretnice. Bilo da ste uputili reč podrške, podelili koristan uvid ili jednostavno bili tu kada je bilo najvažnije – podsetili ste me da nijedan put nije usamljen.

Ova knjiga je posvećena svima vama. Hodali ste uz mene, verovali u moju viziju i pomogli da svaka stranica dobije svoj oblik.

– Vrik Talukdar

O RECENZENTIMA

Mandžit Čakraborti je iskusni tehnološki lider sa bogatim iskustvom u vođenju digitalne transformacije u sektoru finansijskih usluga. Kao arhitekta poslovnih rešenja u kompaniji Amazon Web Services, predvodi inicijative modernizacije nasleđenih sistema i osmišljava inovativna rešenja zasnovana na oblaku za najveće finansijske institucije širom sveta. Zahvaljujući dokazanom uspehu u poslovnoj i tehničkoj arhitekturi, Mandžit se ističe u primeni analize zasnovane na podacima, nudeći primenljive uvide i rešenja. Njegova stručnost obuhvata širok spektar oblasti, uključujući migraciju u oblak, strategije modernizacije velikih računarskih sistema, integraciju sistema, hibridna rešenja, obradu podataka i poslovnu inteligenciju. Mandžit je tražen govornik i učestvovao je na brojnim uglednim tehničkim forumima. Objavio je više stručnih tekstova i doprineo brojnim tehnološkim publikacijama, deleći svoje znanje sa širom tehnološkom zajednicom. Istaknuti je recenzent koji redovno ocenjuje radove iz industrije i naučne publikacije, čime značajno doprinosi razvoju tehnologije i akademske misli. Živi u američkoj saveznoj državi Floridi.

Martin Janev je ostvareni softverski inženjer sa skoro deset godina iskustva u različitim industrijama, uključujući vazduhoplovstvo i medicinsku tehnologiju. Tokom svoje uspešne karijere specijalizovao se za razvoj i integraciju najsavremenijih softverskih rešenja za ključne oblasti kao što su upravljanje vazdušnim saobraćajem i sistemi za hromatografiju. Cenjeni je predavač i profesor računarskih nauka na Državnom univerzitetu u Ficburgu, sa dubokim razumevanjem OpenAI programskih interfejsa i visokim nivoom znanja u izgradnji, obučavanju i podešavanju sistema veštačke inteligencije. Kao priznati autor, Martin je kroz svoja dela pomogao drugima da savladaju složenost razvoja veštačke inteligencije. Sa izuzetnim rezultatima i višestrukim veštinama, i dalje pokreće inovacije i doprinosi transformaciji softverskog inženjeringa.

O RECENZENTU ZA SRPSKO IZDANJE

Miroslav Ristić je redovni profesor na Prirodno-matematičkom fakultetu Univerziteta u Nišu, sa preko 25 godina iskustva u razvoju statističkog softvera. Posebno se ističe njegov rad na razvoju grafičkog korisničkog interfejsa R Commander za programski jezik R. Dugi niz godina recenzirao je značajan broj knjiga za izdavačku kuću Springer i časopis Journal of Applied Statistics. Od 2023. godine aktivno recenzira najaktuelnija izdanja izdavačke kuće „Kompjuter biblioteka". Nakon prevođenja, svako izdanje prolazi kroz njegovo stručno vrednovanje i recenziju prevoda, sa ciljem da se osigura da prevodi budu ne samo jasni, precizni i prilagođeni čitaocima, već i da održe visok kvalitet i stručnu relevantnost knjiga.

UVOD

Knjiga *Izgradnja agentnih sistema veštačke inteligencije* pruža teorijske osnove i praktične smernice za razumevanje i primenu generativne veštačke inteligencije i agentnog pristupa. Ove dve oblasti danas predvode novi talas u razvoju veštačke inteligencije, menjajući način na koji se automatizuju procesi, donose odluke i podstiče kreativnost — u oblastima koje su do juče bile van domašaja tehnologije. Sistemi sposobni da samostalno stvaraju tekst, slike ili čak planove delovanja, dok istovremeno rasuđuju i prilagođavaju se, već danas menjaju prakse u zdravstvu, finansijama i robotici.

U uvodnim poglavljima razmatraju se osnovni koncepti generativne veštačke inteligencije, kroz prikaz modela kao što su **varijacioni autokoderi** (**VAE**), **generativne suparničke mreže** (**GAN**) i autoregresivni modeli. Prikazane su njihove praktične primene u kreiranju sadržaja, dizajnu i naučnom istraživanju, ali i ograničenja s kojima se susreću.

Nakon toga, pažnja se usmerava na agentne sisteme. Objašnjavaju se pojmovi kao što su delovanje, autonomija i saradnja više agenata. Upoređuju se različite arhitekture agenata — od savetodavnih, preko reaktivnih do hibridnih — i analizira kako agenti mogu da sarađuju, koordiniraju se i zajednički ostvaruju ciljeve.

Kada su teorijske osnove postavljene, prelazi se na praktične aspekte: kako agenti mogu da preispituju sopstveno razmišljanje, planiraju i koriste spoljne alate. Opisane su konkretne tehnike za meta-zaključivanje, samoobjašnjavanje, strateško planiranje i rad u timovima agenata. Posebna pažnja posvećena je načinu projektovanja pouzdanih i odgovornih agenata, uz balansiranje između autonomije i kontrole, kao i etičkim pitanjima razvoja veštačke inteligencije.

U završnim poglavljima predstavljene su primene u praksi — od obrade prirodnog jezika i robotike, do sistema za podršku odlučivanju i optimizaciji. Razmatraju se i ključna pitanja poverenja, transparentnosti, pristrasnosti i bezbednosti, neophodna za pouzdan rad sistema veštačke inteligencije.

Čitava knjiga prožeta je primerima iz prakse, zadacima za vežbu i strategijama implementacije koje povezuju teoriju sa konkretnom primenom. Namenjena je svima koji se bave veštačkom inteligencijom — istraživačima, inženjerima, programerima i tehničkim liderima — i pruža alate i znanja potrebna za izgradnju autonomnih i prilagodljivih agenata sposobnih da uče, sarađuju i razvijaju se.

Krenimo zajedno na ovo putovanje, oblikujući budućnost inteligentnih sistema – agenta po agenta.

Kome je ova knjiga namenjena

Ova knjiga namenjena je stručnjacima koji se bave veštačkom inteligencijom — istraživačima, inženjerima, programerima i tehnološkim liderima — koji žele da razumeju, osmisle i izgrade agente zasnovane na veštačkoj inteligenciji, sposobne za autonomno delovanje, prilagođavanje i učenje. Bilo da razvijate sisteme koji integrišu generativne modele, ili se bavite arhitekturom naprednih agentnih rešenja, knjiga pruža čvrstu teorijsku osnovu i praktične smernice za primenu u realnim uslovima.

Šta je obrađeno u knjizi

Poglavlje 1, Osnove generativne veštačke inteligencije - Objašnjavaju se osnovni pojmovi **generativne veštačke inteligencije**, vrste modela (varijacioni autokoderi, generativne suparničke mreže, autoregresivni modeli), njihove praktične primene, ali i ograničenja, pristrasnosti i etičke dileme koje ih prate.

Poglavlje 2, Osnovni principi agentnih sistema - Definišu se ključni pojmovi kao što su delovanje, autonomija i osnovne osobine agenata — proaktivnost, reaktivnost i sposobnost saradnje. Razmatraju se različite arhitekture agenata, kao i organizacija rada u multi-agentnim sistemima.

Poglavlje 3, Ključni elementi inteligentnih agenata - U fokusu su reprezentacija znanja, razmišljanja, učenje, odlučivanje i uloga generativne veštačke inteligencije u unapređivanju sposobnosti agenata.

Poglavlje 4, Refleksija i introspekcija u radu agenata - Prikazuje se kako agenti mogu da analiziraju sopstveni način razmišljanja, uče iz prethodnog iskustva i razvijaju složenije obrasce odlučivanja, uz pomoć tehnika kao što su meta-zaključivanje, samoobjašnjavanje i samomodelovanje.

Poglavlje 5, Upotreba alata i planiranje u agentnim sistemima - Obrađuju se načini na koje agenti koriste spoljne alate i planiraju zadatke, uz kombinovanje alata i strategija odlučivanja radi bolje efikasnosti i ostvarivanja ciljeva.

Poglavlje 6, Pristup sa koordinatorom, izvršiocem i delegatom - Predstavljen je CWD model koji opisuje kako agenti mogu da preuzimaju uloge koordinatora, izvršioca ili delegata, radi bolje raspodele zadataka i resursa u timskom radu.

Poglavlje 7, Proverene tehnike za dizajn agentnih sistema - Opisane su preporučene prakse u dizajniranju agenata: od formulacije ciljeva i zadataka, preko definisanja granica autonomije, do obezbeđivanja transparentnosti i odgovornosti u radu sistema.

Poglavlje 8, Izgradnja poverenja u sisteme zasnovane na generativnoj veštačkoj inteligenciji - Analiziraju se pristupi koji doprinose poverenju korisnika u sisteme veštačke inteligencije — mogućnost objašnjenja odluka, kontrola neizvesnosti i pristrasnosti, kao i projektovanje sistema koji su pouzdani i razumljivi.

Poglavlje 9, Bezbednosni i etički aspekti u radu sa agentima veštačke inteligencije - Poglavlje se bavi rizicima povezanim sa generativnom veštačkom inteligencijom, kao i načelima odgovornog razvoja, etičkim smernicama, i pitanjima zaštite privatnosti i podataka u praksi.

Poglavlje 10, Praktične primene i najčešći slučajevi upotrebe - Prikazuju se konkretne primene generativne veštačke inteligencije u stvarnom svetu — od kreiranja kreativnog sadržaja i agenata za komunikaciju, do robotike i sistema za podršku odlučivanju.

Poglavlje 11, Zaključak i pogled u budućnost - Sažimaju se ključni koncepti obrađeni u knjizi, razmatraju savremeni trendovi u razvoju generativne veštačke inteligencije i agentne inteligencije, uvodi se pojam **opšte veštačke inteligencije** (**AGI**), kao i izazovi i mogućnosti koji predstoje u ovoj oblasti.

Kako da izvučete najviše iz ove knjige

Biće vam znatno lakše da pratite sadržaj ako posedujete sledeće:

- **Osnovno poznavanje veštačke inteligencije i mašinskog učenja**: Iako knjiga pokriva osnovne principe, poznavanje pojmova iz oblasti veštačke inteligencije/mašinskog učenja, dubokog učenja i rada sa programskim jezikom Python pomoći će vam da jasnije razumete naprednije teme.
- **Praktičan rad**: Isprobajte priložene primere koda i okvire za izgradnju generativne veštačke inteligencije i agentnih sistema. Bilo da radite u lokalnom razvojnom okruženju ili u oblaku, praksa će značajno doprineti boljem razumevanju materijala.
- **Kritički odnos prema etici i bezbednosti u sistemima veštačke inteligencije**: Dok budete istraživali generativnu veštačku inteligenciju i autonomne agente, razmišljajte o poverenju, pristrasnosti i odgovornom dizajnu, kako biste gradili sisteme koji se oslanjaju na etičke principe.

KORIŠĆENI SOFTVER	OPERATIVNI SISTEMI
Python, Jupyter Notebooks, CrewAI	Windows, macOS, Linux

Preuzimanje primera koda

Paketi sa kodom dostupni su na GitHub stranici knjige:
`https://github.com/PacktPublishing/Building-Agentic-AI-Systems`

Takođe možete pronaći i druge resurse iz bogate kolekcije knjiga i video materijala na:
`https://github.com/PacktPublishing`

Konvencije u knjizi

U celom tekstu korišćene su određene oznake radi lakšeg snalaženja.

Podebljan tekst označava nov pojam, ključnu reč ili izraz koji se pojavljuje na ekranu. Na primer: „**Prilagođeni plan uvođenja**: Na osnovu definisanih ciljeva i potreba, kreirajte plan koji precizno definiše korake, prekretnice i vremenski okvir za postizanje zadatih rezultata.“

> **Saveti ili važne napomene**
> pojavljuju se ovako.

Kontakt

Bilten
Ako želite da ostanete u toku sa najnovijim dostignućima u oblasti generativne veštačke inteligencije i velikih jezičkih modela (LLM), možete se prijaviti na naš nedeljni bilten *AI_Distilled* putem linka: `https://packt.link/817Ur`

Povratne informacije naših čitalaca su uvek dobrodošle.

Povratne informacije

Ako imate bilo kakva pitanja, sugestije ili komentare u vezi sa ovom knjigom, pišite nam na: `customercare@packtpub.com` Ne zaboravite da u naslovu poruke navedete naziv knjige.

Greške i ispravke

Uprkos pažljivoj pripremi sadržaja, moguće je da se potkrade neka greška. Ukoliko uočite propust, bićemo vam zahvalni ako ga prijavite putem stranice:
`www.packtpub.com/support/errata`

Pitanje autorskih prava

Ako naiđete na nelegalne kopije ove knjige na internetu, molimo vas da nas obavestite o tome. Pošaljite lokaciju ili link na: `copyright@packt.com`

Zainteresovani ste da postanete autor?

Ako ste stručni u nekoj oblasti i želite da pišete ili doprinesete razvoju knjige, posetite:
`authors.packtpub.com`

Podelite svoje utiske

Kada završite sa čitanjem knjige *Izgradnja agentnih sistema veštačke inteligencije*, biće nam drago da čujemo vaše utiske! Kliknite ovde da biste direktno otišli na Amazon stranicu za recenziju ove knjige i podelili svoje mišljenje.

Vaša recenzija nam je dragocena — ne samo kao povratna informacija, već i kao doprinos široj tehnološkoj zajednici. Pomažete nam da nastavimo sa objavljivanjem kvalitetnog i korisnog sadržaja.

Deo 1

Osnove generativne veštačke inteligencije i agentnih sistema

Ovaj deo knjige postavlja osnovne pojmove generativne veštačke inteligencije i agentnih sistema, pružajući čitaocu pouzdanu teorijsku osnovu za razumevanje i razvoj inteligentnih, autonomnih agenata.

Sadržaj ovog dela čine sledeća poglavlja:

- *Poglavlje 1, Osnove generativne veštačke inteligencije*
- *Poglavlje 2, Osnovni principi agentnih sistema*
- *Poglavlje 3, Ključni elementi inteligentnih agenata*

1

Osnove generativne veštačke inteligencije

Generativna veštačka inteligencija u veoma kratkom vremenu postala je jedna od najuticajnijih tehnologija u oblasti **veštačke inteligencije** i mašinskog učenja. Temeljno menja način na koji se pristupa kreativnim procesima i rešavanju problema u različitim industrijama i oblastima primene. Istovremeno, ona pomera granice autonomije u agentnim sistemima zasnovanim na veštačkoj inteligenciji. U ovom poglavlju upoznaćemo se sa osnovama generativne veštačke inteligencije. Istražićemo šta podrazumevamo pod tim pojmom, kao i kratku istoriju razvoja modela generativne veštačke inteligencije. Potom ćemo se osvrnuti na različite vrste modela — među kojima su **varijacioni autokoderi** (**VAE**), **generativne suparničke mreže** (**GAN**), autoregresivni modeli i modeli zasnovani na transformatorima. Nakon toga razmotrićemo primene generativne veštačke inteligencije, a poglavlje ćemo zaključiti kratkom analizom izazova i ograničenja koje ova tehnologija nosi.

U ovom uvodnom poglavlju obrađujemo sledeće teme, koje čine osnovu za dalje razumevanje autonomnih inteligentnih agenata zasnovanih na generativnoj veštačkoj inteligenciji:

- Uvod u generativnu veštačku inteligenciju
- Vrste modela generativne veštačke inteligencije
- Primene generativne veštačke inteligencije
- Izazovi i ograničenja generativne veštačke inteligencije

Po završetku ovog poglavlja, imaćete celovit uvid u generativnu veštačku inteligenciju — njene osnovne pojmove, široku lepezu primena i savremene izazove. Takođe ćete razumeti njen potencijal i ograničenja, sa posebnim osvrtom na ključnu ulogu koju ima u razvoju inteligentnih agenata i agentnih sistema.

Uvod u generativnu veštačku inteligenciju

Generativna veštačka inteligencija odnosi se na skup tehnologija veštačke inteligencije koje su sposobne da generišu različite oblike sadržaja — tekst, slike, zvuk ili video. Ovi sistemi veštačke inteligencije mogu da stvaraju novi sadržaj na osnovu podataka na kojima su obučavani i

ulaznih parametara, koji najčešće uključuju tekstualne upite, ali mogu podrazumevati i druge oblike ulaza, poput slika. Posebnu pažnju u poslednje vreme izazvala je jednostavnost s kojom savremeni interfejsi zasnovani na generativnoj veštačkoj inteligenciji mogu proizvesti kvalitetan tekst, grafiku ili video — i to u roku od svega nekoliko sekundi.

U najjednostavnijim crtama, generativna veštačka inteligencija zasniva se na stvaranju novih podataka koji podsećaju na one na kojima je model obučavan. Drugim rečima, učenje skrivenih obrazaca, struktura i raspodela ulaznih podataka omogućava modelu da generiše nove primere na sličan način. Na primer, ukoliko je model obučavan na skupu podataka koji sadrži ljudska lica, biće sposoban da stvara potpuno nova lica ljudi koji ne postoje u stvarnosti, ali izgledaju veoma realistično. U svojoj suštini, modeli generativne veštačke inteligencije funkcionišu tako što uče raspodelu verovatnoća skupa podataka, a zatim iz te raspodele generišu nove uzorke. Takav pristup razlikuje se od diskriminacionih modela, čiji je cilj da nauče kako da razlikuju klase podataka. Na primer, diskriminacioni model može naučiti da razlikuje slike mačaka i pasa, dok će generativni model biti obučen da stvara sasvim nove slike mačaka ili pasa.

Koncept generativne veštačke inteligencije postoji već duže vreme — zapravo, potiče još iz perioda kada je mašinsko učenje bilo u začecima. Ipak, tek u poslednjih desetak godina ovo područje dolazi u centar pažnje, zahvaljujući razvoju neuronskih mreža i naglom napretku računarske snage. Rani pokušaji generativnog modelovanja, poput modela Gausovih mešavina (GMM) i skrivenih modela Markova (HMM), oslanjali su se na jednostavnije statističke tehnike. Metodi za modelovanje složenih raspodela podataka danas imaju daleko veći potencijal da se kombinuju sa metodima koji su ranije bili dominantni, ali su ih u velikoj meri potisnule tehnike dubokog učenja. Upravo je duboko učenje označilo prekretnicu u razvoju generativne veštačke inteligencije. Modeli kao što su varijacioni autokoderi (VAE), koji su se pojavili početkom 2010-ih, koristili su duboke neuronske mreže kako bi predstavili složenije raspodele podataka. U istom periodu razvijene su i generativne suparničke mreže, koje su prvi put omogućile da dve neuronske mreže međusobno sarađuju u generisanju podataka, koristeći principe teorije igara. U širem smislu, ovi proboji predstavljaju veliki korak napred ka mogućnosti stvaranja realističnih, visokokvalitetnih i uverljivih sadržaja.

Takvi modeli otvaraju prostor za inovacije u gotovo svim industrijama — od zdravstva, finansija i obrazovanja, do turizma, proizvodnje i mnogih drugih. U kreativnim industrijama, generativna veštačka inteligencija pomaže umetnicima i dizajnerima da razmišljaju izvan ustaljenih okvira, podstičući nastanak novih, originalnih ideja. U domenu zdravstva, omogućava brže otkrivanje lekova i personalizovanu medicinu. U poslovnom svetu, koristi se za unapređenje odnosa sa korisnicima putem personalizovanog sadržaja i prilagođenih iskustava. Pored toga, nameće se i pitanje same prirode kreativnosti i originalnosti: kada mašine postaju sposobne da u velikim količinama stvaraju sadržaj koji deluje kao originalan, postavljamo sebi nova pitanja — šta je zapravo kreativnost, ko je autor, i gde su etičke granice u stvaranju sadržaja pomoću veštačke inteligencije.

Nakon što smo sagledali osnovni pojam i istorijat generativne veštačke inteligencije, naredni korak je upoznavanje sa različitim vrstama modela generativne veštačke inteligencije.

Vrste modela generativne veštačke inteligencije

Generativna veštačka inteligencija predstavlja uzbudljivu oblast u okviru veštačke inteligencije, usmerenu na stvaranje novih, veštačkih podataka učenjem obrazaca iz postojećih skupova podataka. Cilj je generisanje izlaza koji ima slične statističke osobine i karakteristike podataka korišćenih za obuku. U nastavku sledi pregled najpoznatijih i najuticajnijih vrsta modela generativne veštačke inteligencije: varijacionih autokodera, generativnih suparničkih mreža i autoregresivnih modela.

Varijacioni autokoderi

Jedan od najpoznatijih modela generativne veštačke inteligencije je varijacioni autokoder. Suštinska ideja iza varijacionog autokodera zasniva se na učenju preslikavanja zasnovanog na verovatnoći između podataka i latentnog prostora, i obrnuto. To znači da model uči kako da pretvori stvarne podatke u pojednostavljenu, sažetu reprezentaciju, a zatim iz te reprezentacije rekonstruiše podatke koji izgledaju realistično. Varijacioni autokoderi su konstruisani tako da zadrže veliku verodostojnost generisanih podataka, dok u isto vreme čuvaju uređenu strukturu latentnog prostora, što omogućava generisanje novih primera sličnih podacima na kojima je model obučavan. Najčešće varijante varijacionih autokodera uključuju sledeće:

- **Varijacioni autokoder**: osnovna verzija modela koja vrši sažimanje i rekonstrukciju podataka, pružajući okvir za generisanje novih primera iz naučenog latentnog prostora. Drugim rečima, varijacioni autokoder je model generativne veštačke inteligencije koji ne samo da uči kako da sažme i rekonstruiše podatke, već istovremeno i kako da modeluje raspodelu verovatnoća latentnog prostora. Zamislite varijacioni autokoder kao vešt algoritam koji razume suštinu podataka i ume da ih ponovo stvori. Kao da imate umetnika koji ume da pretvori složenu sliku u jednostavan skicirani oblik (kodiranje), a zatim iz te skice rekonstruiše originalnu sliku (dekodiranje). Ono što čini varijacioni autokoder posebnim jeste to što ne kopira doslovno, već uči stil, obrazac i strukturu, što mu omogućava da stvori nove, originalne primere u istom duhu. U praksi, varijacioni autokoderi se koriste u mnogim kreativnim oblastima. Na primer, u procesu otkrivanja novih lekova, istraživači iz kompanije AstraZeneca koristili su varijacioni autokoder kako bi istražili hemijski prostor i predložili nove molekulske strukture sa željenim svojstvima — što može značajno ubrzati razvoj lekova [1].
- **Beta-varijacioni autokoder**: proširena verzija osnovnog varijacionog autokodera, koja uvodi hiperparametar pomoću kojeg se eksplicitno kontroliše odnos između kvaliteta rekonstrukcije i stepena razdvajanja u latentnom prostoru. Cilj ovog modela je da stvori interpretabilniju reprezentaciju podataka, u kojoj su ključne karakteristike jasno razdvojene. Pojednostavljeno rečeno, beta-varijacioni autokoder je fleksibilnija osnovna verzija varijacionog autokodera. Omogućava istraživačima da podešavaju koliko model treba da se fokusira na preciznu rekonstrukciju podataka, a koliko na razumevanje suštinskih karakteristika. Možemo ga zamisliti kao nastavnika likovnog koji ne uči učenike da samo kopiraju slike, već da uoče i odvoje osnovne elemente slike — boju, oblik, stil. Upravo zbog sposobnosti da razdvaja karakteristike podataka, beta-varijacioni autokoderi se posebno koristi u oblastima kao što su računarski vid i robotika. Na primer, istraživači su koristili beta-varijacione autokodere da nauče robote da lakše razumeju objekte, razdvajajući karakteristike poput veličine, boje i položaja. Time se postiže veća prilagodljivost i efikasnost robota u različitim situacijama [2].
- **Uslovni varijacioni autokoder**: varijanta koja uslovljava proces generisanja dodatnim informacijama, najčešće oznakama kategorija. Umesto da generiše proizvoljan novi sadržaj, uslovni varijacioni autokoder omogućava generisanje kontrolisanih primera — primera koji pripadaju tačno određenim kategorijama. Uslovni varijacioni autokoder možete zamisliti kao kontrolisanu verziju varijacionog autokodera, poput umetnika koji može da slika u različitim stilovima po zahtevu. Umesto da stvara slike na slučajan način, ovom umetniku se može reći: „Nacrtaj pejzaž“, ili „Napravi portret“. Zahvaljujući tom nivou kontrole, uslovni varijacioni autokoder ima široku primenu u praksi. Na primer, u razvoju video igara, uslovni varijacioni autokoder se koristi za proceduralno generisanje sadržaja — kao što su dizajn likova, raspored nivoa, muzika, zvučni efekti itd. Kada se modelu zada uslov: „Napravi šumski nivo“ ili „Napravi pustinjski nivo“,

uslovni varijacioni autokoder može generisati raznovrsna i vizuelno zanimljiva okruženja, čime štedi vreme dizajnerima i istovremeno unapređuje iskustvo igrača [3].

Generativne suparničke mreže

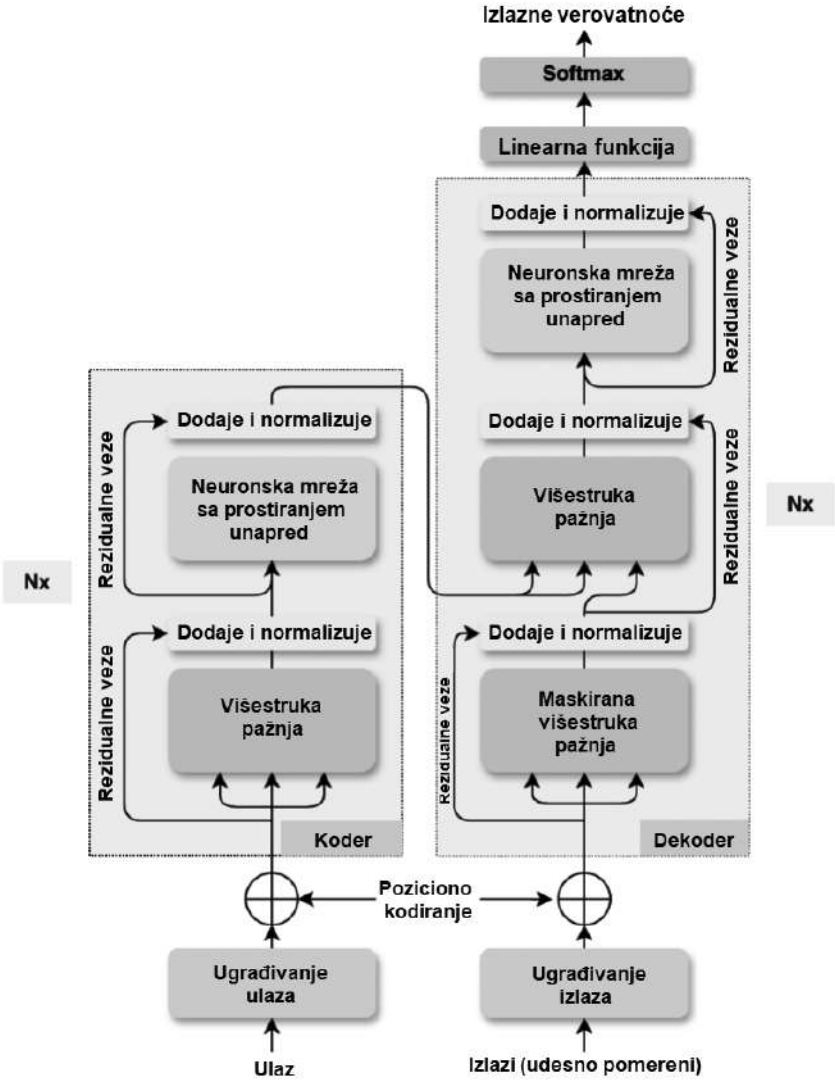
Generativna suparnička mreža sastoji se od dve neuronske mreže: **generatora** i **diskriminatora**. Generator stvara veštačke primere podataka, dok je zadatak diskriminatora da nauči da razlikuje stvarne podatke od generisanih. Ove dve mreže obučavaju se istovremeno, međusobnoj suprostavljene: generator pokušava da prevari diskriminator, dok diskriminator pokušava da tačno klasifikuje šta je „pravo", a šta „lažno". U toj vrsti takmičenja, generator postaje sve uspešniji u stvaranju uverljivih, lažnih podataka. Neki od najpoznatijih tipova generativnih suparničkih mreža su:

- **Generativna suparnička mreža**: osnovna verzija u kojoj se generator i diskriminator obučavaju u suparničkom odnosu. Ovaj model predstavlja osnovu za većinu kasnijih inovacija u oblasti generativnog modelovanja. Možete ga zamisliti kao igru između falsifikatora i detektiva — generator pokušava da napravi lažni novac, dok ga diskriminator pokušava prepoznati kao falsifikat. Kako se „igra" nastavlja, obojica postaju sve veštiji: falsifikator pravi sve uverljivije novčanice, a detektiv postaje sve bolji u otkrivanju prevara. Ovakva struktura omogućava generativnim suparničkim mrežama da stvore izuzetno realistične lažne podatke — bilo da je reč o slikama, zvuku ili nekom drugom tipu sadržaja [4].
- **Duboke konvolucione generativne suparničke mreže (DCGAN)**: predstavlja unapređenje osnovne generativne suparničke mreže kroz primenu dubokih konvolucionih neuronskih mreža. Duboka konvoluciona generativna suparnička mreža je trenutno jedna od najboljih arhitektura za generisanje slika visoke rezolucije. Možemo je uporediti sa umetnikom koji umesto olovke koristi čitav digitalni atelje. Ovaj model može da stvara znatno detaljnije i vizuelno uverljivije slike. Duboke konvolucione generativne suparničke mreže se posebno ističu u prepoznavanju i reprodukovanju složenih obrazaca na slikama [5].
- **Vaserštajnova generativna suparnička mreža (WGAN)**: Vaserštajnova generativna suparnička mreža uvodi novu funkciju gubitka zasnovanu na tzv. Vaserštajnovom rastojanju. Ova promena rešava određene probleme prilikom obučavanja generativnih suparničkih mreža, čineći proces obuke stabilnijim i poboljšavajući kvalitet generisanih primera. Ona omogućava bolje merenje rastojanja između dve raspodele verovatnoća. Možemo je zamisliti kao unapređeni dijalog između umetnika (generatora) i kritičara (diskriminatora): umesto jednostavnog „dobro" ili „loše", komunikacija postaje nijansiranija — „bliži si", „daleko si", „skoro si pogodio". Takva preciznija povratna informacija dovodi do konzistentnijeg napretka i kvalitetnijih rezultata. U praksi, Vaserštajnove generativne suparničke mreže su našle primenu u medicinskom snimanju, gde se koriste za generisanje veštačkih medicinskih slika u svrhu obučavanja dijagnostičkih sistema veštačke inteligencije. Time se stvaraju veći i raznovrsniji skupovi podataka koji mogu doprineti većoj tačnosti u otkrivanju bolesti na osnovu skeniranja i rendgenskih snimaka [6].
- **StyleGAN**: modeli poznati po tome što generišu slike izuzetno visokog kvaliteta i realističnosti. Njihova posebnost leži u sposobnosti da stil i sadržaj tretiraju odvojeno, čime omogućavaju veću kontrolu nad generisanim rezultatima. Dalje unapređene verzije — StyleGAN2 i StyleGAN3 — dodatno poboljšavaju vernost i prirodnost prikazanih slika. Model možete zamisliti kao vrhunskog digitalnog umetnika koji ne samo da može da naslika realistične prizore, već može da kombinuje različite stilove i sadržaje. Na primer, uzme stil Van Goga i primeni ga na pejzaž moderne urbane

sredine. Ova fleksibilnost čini StyleGAN izuzetno moćnim i svestranim alatom za različite kreativne i profesionalne primene [7].

Autoregresivni modeli i arhitektura transformatora

Autoregresivni modeli generišu svoje tačke podataka jednu po jednu, uslovljavajući svaku tačku prethodnom. Iznenađujuće, ovo se pokazuje kao veoma korisno kod zadataka u kojima je redosled ili struktura podataka važna, kao što su generisanje teksta i slika. Arhitektura transformatora, predstavljena u radu *Attention Is All You Need* [8], predstavlja arhitekturu modela koja je revolucionisala mnoge zadatke nad sekvencijalnim podacima, posebno u oblasti **obrade prirodnog jezika** (**NLP**). Njene ključne komponente prikazane su na *slici 1.1*:



Slika 1.1 – Arhitektura transformatora

Ove ključne komponente uključuju sledeće:

- **Mehanizam samopažnje**: Računska tehnika koja omogućava modelu da se dinamički fokusira na različite delove ulaza prilikom obrade svakog elementa
- **Višestruka pažnja**: Više mehanizama pažnje koji se izvršavaju paralelno, omogućavajući modelu da se istovremeno fokusira na različite aspekte ulaznih podataka
- **Poziciono kodiranje**: Dodaje informacije o poziciji svakog elementa u sekvenci
- **Neuronske mreže sa prostiranjem unapred**: Obrađuju izlaz iz slojeva pažnje
- **Normalizacija slojeva i rezidualne veze**: Poboljšavaju stabilnost obuke i protok informacija

Transformatori se mogu koristiti u autoregresivnim i ne-autoregresivnim konfiguracijama, što ih čini svestranim za različite zadatke. Slede neki primeri:

- **PixelCNN**: Autoregresivno generiše slike piksel po piksel, uslovljavajući svaki piksel prethodno generisanim. Ovaj model veoma uspešno generiše slike visokog kvaliteta sa finim detaljima [9].
- **PixelSNAIL**: Unapređena verzija modela PixelCNN koja u model uvodi mehanizme pažnje radi hvatanja jakih međuzavisnosti između piksela, čime se postiže bolji kvalitet slike [10].
- **GPT**: Skraćenica za Generativni unapred obučeni transformator, označava liniju modela specijalizovanih za generisanje teksta — tačnije, za predviđanje koje reči slede. Među njima su GPT-2, GPT-3 i GPT-4, koji su ostvarili ogroman napredak u generisanju koherentnog i kontekstualno relevantnog teksta [11] [12].
- **BERT**: Za razliku od GPT modela, BERT, što je skraćenica za Dvosmerna koderska reprezentacija iz transformatora, dizajniran je da razume kontekst iz oba pravca u tekstu. Koristi koderski deo arhitekture transformatora i obično se koristi za zadatke koji zahtevaju razumevanje celokupnog ulaza, umesto za autoregresivno generisanje teksta [13].
- **Transformator za prenos teksta u tekst (T5)**: Ovaj model sve zadatke u obradi prirodnog jezika postavlja u tekst-u-tekst formatu. Koristi kompletnu arhitekturu transformatora (i koder i dekoder) i može da obradi različite zadatke generisanja teksta [14].

Iako sve zavisi od konkretnog zadatka, ovi modeli pokazuju kako se arhitektura transformatora može prilagoditi i za autoregresivne (kao što je GPT) i za ne-autoregresivne (kao što je BERT) zadatke, demonstrirajući svoju svestranost u radu sa sekvencijalnim podacima. Snage se razlikuju među generativnim modelima. Ova klasa modela obuhvata mogućnosti koje se protežu od generisanja fotorealističnih slika, preko koherentnog teksta, do potpuno novih primera podataka.

Na osnovu uspeha modela kao što je GPT, istraživači su skalirali ove arhitekture kako bi stvorili **velike jezičke modele** (**LLM**). Ovi modeli, često obučavani na ogromnim količinama otvorenih tekstova sa interneta i drugim slobodno dostupnim tekstualnim podacima, pokazali su izuzetnu veštinu u razumevanju i generisanju teksta nalik ljudskom u najrazličitijim primenama, demonstrirajući njihovu svestranost i napredne jezičke sposobnosti. Primeri velikih jezičkih modela uključuju GPT-3, GPT-4, PaLM i BERT-large. Ovi modeli su pomerili granice mogućeg u obradi prirodnog jezika, pokazujući visoku uspešnost u zadacima koji uključuju odgovaranje

na pitanja, sažimanje, kreativno pisanje i generisanje programskog koda. Sledi pregled najčešćih tipova velikih jezičkih modela:

- Autoregresivni veliki jezički modeli: Ovi jezički modeli obično su sposobni da generišu tekst sekvencijalno, to jest, token po token. Neki od najčešćih zadataka za koje se ovi modeli koriste uključuju generisanje teksta, dopunjavanje teksta i kreativno pisanje. Neki poznati primeri autoregresivnih velikih jezičkih modela su serija GPT modela (GPT-3, GPT-4) i PaLM.
- **Veliki jezički modeli zasnovani samo na koderu:** Ovi modeli su specijalizovani za zadatke **razumevanja prirodnog jezika (NLU)**, koji podrazumevaju analizu i razumevanje ulaznog teksta bez generisanja novog teksta. Veliki jezički modeli zasnovani samo na koderu, kao što su BERT i njegove varijante poput RoBERTa, fokusirani su na razumevanje konteksta i značenja ulaznog teksta. Ovi modeli obrađuju celu ulaznu sekvencu istovremeno, što im omogućava da „uhvate" kontekst, što pomaže kod zadataka kao što su klasifikacija teksta, **prepoznavanje imenovanih entiteta** (**NER**) i analiza osećanja.
- **Veliki jezički modeli sa koder-dekoder arhitekturom:** Ovo je kombinacija kodera i dekodera, gde modeli mogu ne samo da razumeju kontekst ulaznog teksta (NLU), već i da generišu tekstualni izlaz. Model T5, koji je ranije pomenut, veoma je poznat primer velikih jezičkih modela ove vrste, zajedno sa modelima kao što je BERT. Ovi modeli su sposobni za zadatke kao što su jezičko prevođenje, sažimanje sadržaja i odgovaranje na pitanja.
- **Multimodalni veliki jezički modeli:** Multimodalnost je koncept u kojem model veštačke inteligencije može da obrađuje ne samo tekst, već i druge modalitete, kao što su slike, video i zvuk. Multimodalni veliki jezički modeli mogu da obrađuju i generišu sadržaj u različitim modalitetima — tekst, slike, zvuk, video — ili njihovim kombinacijama. Ovi modeli su relativno novijeg datuma, a najnoviji napredak u računarskoj snazi omogućio je njihovu obuku. Neki od poznatijih multimodalnih velikih jezičkih modela su DALL-E, Stable Diffusion, Flamingo, GPT-4 i LlaVa.
- **Veliki jezički modeli prilagođeni instrukcijama:** Veliki jezički modeli se često obučavaju na velikim skupovima javno dostupnih veb stranica i tekstova. Kao takvi, u određenim situacijama ovi modeli možda neće uspešno obavljati zadatke koji koriste podatke sa kojima se nisu ranije susreli, te mogu da ne isprate nameru korisnika. Na primer, uputstvo za korišćenje specijalizovanog uređaja u proizvodnji može sadržati terminologiju specifičnu za određenu oblast, s kojom se model možda nije susreo tokom obuke. U takvim slučajevima, veliki jezički modeli mogu se dodatno obučavati (fino podešavati) na specifičnim podacima z oblasti, kako bi naučili da prate tačno određene instrukcije i upite (odnosno, tekstualne ulaze modela). Ovo specijalno fino podešavanje omogućava modelu da, pored svojih uobičajenih sposobnosti, obavlja i zadatke specifične za oblast za koju je podešen. Zamislite to kao lokalizaciju lične kopije modela koja je osposobljena da izvršava zadatke koji su vama važni. Jedan od poznatih primera ovakvog modela je InstructGPT, gde su istraživači fino podesili GPT-3 model tako da prati konkretne instrukcije. Ova vrsta modela koristi se u mnogim različitim zadacima, kao što su četbotovi, virtuelni asistenti i aplikacije specijalizovane za određene zadatke.

- **Veliki jezički modeli specifični za oblast:** Kao što je ranije pomenuto, iako se veliki jezički modeli često obučavaju na velikim skupovima otvorenih i opšte dostupnih podataka sa interneta, oni možda neće davati dobre rezultate u specifičnim oblastima. Dok su modeli prilagođeni instrukcijama jedan od načina da se ovi opšti modeli primene na specifične zadatke, istraživači su razvili i pristup obučavanja velikih jezičkih modela specifičnih za određenu oblast od samog početka. To znači da su ovi modeli prethodno obučeni na visoko specijalizovanom znanju iz određene oblasti. Na primer, model BioBERT je obučavan na strogo medicinskim podacima, dok je LegalBERT obučen na pravnim dokumentima. Ovi modeli su korisni za zadatke u specijalizovanim oblastima kao što su medicina, pravo i finansije.

Agenti zasnovani na velikim jezičkim modelima

Svaka od prethodno opisanih vrsta velikih jezičkih modela ima svoje prednosti i nedostatke i odgovara različitim vrstama primene. Međutim, fokus ove knjige biće na detaljnijem istraživanju mogućnosti agenata zasnovanih na velikim jezičkim modelima i na tome kako se mogu graditi napredni sistemi koji ih koriste. Važno je naglasiti da agenti zasnovani na velikim jezičkim modelima ne pripadaju jasno definisanoj kategoriji, jer predstavljaju naprednu primenu tehnologije velikih jezičkih modela, a ne određeni tip arhitekture modela. Agenti se obično grade na osnovu velikih jezičkih modela prilagođenih instrukcijama, koji se zatim kombinuju sa dodatnim komponentama i mogućnostima, o čemu će biti više reči u nastavku knjige.

Pogledajmo jednostavan primer koji ilustruje mogućnosti agentnog sistema zasnovanog na velikom jezičkom modelom. U pitanju je turistički veb sajt koji koristi **četbota za rezervaciju letova** —interfejs komunikacije zasnovan na velikom jezičkom modelu. On može da pomaže korisnicima pri rezervaciji avio-karata kroz razgovore nalik ljudskim. U pozadini, četbot izvršava različite zadatke dok razgovor napreduje. Evo kako može da izgleda primer razgovora sa korisnikom koji želi da rezerviše avio-kartu:

ASISTENT ZA REZERVACIJU LETOVA	
	Zdravo, želim da rezervišem let.
	Zdravo! Biće mi zadovoljstvo da vam pomognem oko rezervacije leta. Da bismo počeli, možete li mi reći grad iz kojeg polazite i odredište?
	Želim da putujem iz San Dijega za SFO.
	Odlično, hvala vam na tim informacijama. Planirate let iz San Dijega (SAN) za međunarodni aerodrom u San Francisku (SFO). Da bih pronašao najbolje opcije letova, potrebno mi je još nekoliko detalja: **1.** Koji vam je željeni datum putovanja? **2.** Da li je u pitanju putovanje u jednom smeru ili povratno putovanje? **3.** Koliko će putnika putovati?

	Želim da idem sledećeg petka i da se vratim u nedelju. Putujem sam.
	Hvala vam na dodatnim informacijama. Da rezimiram ono što za sada imam: • Polazak: San Dijego (SAN) Odredište: San Francisko (SFO) • Datum polaska: Sledeći petak – 23. 8. 2024. • Datum povratka: Nedelja – 25. 8. 2024. • Putnici: 1 odrasla osoba Pre nego što potražim letove, imam još nekoliko pitanja: **1.** Da li vam odgovara određeno doba dana za let (ujutru, popodne ili uveče)? **2.** Imate li omiljenu avio-kompaniju ili vam odgovaraju sve opcije? **3.** Koju klasu usluge preferirate (ekonomska, premium ekonomska, biznis ili prva klasa)?
	Preferiram jutarnje letove. Nemam omiljenu avio-kompaniju. Ekonomska klasa je u redu.
	Hvala na tim detaljima. Pretražiću jutarnje letove u ekonomskoj klasi kod svih dostupnih avio-kompanija za vaše putovanje. Samo trenutak dok pronađem najbolje opcije. [pretraga letova... molimo sačekajte…]
	Pronašao sam nekoliko opcija za vaše putovanje. Evo najboljih jutarnjih letova: Polazak (SAN do SFO) – Sledeći petak: **1.** United Airlines UA1234: Polazak u 7.15, Dolazak u 8.45 **2.** Alaska Airlines AS5678: Polazak u 8.30, Dolazak u 10.00 Povratak (SFO do SAN) – Nedelja: **1.** Delta Air Lines DL9876: Polazak u 9.00, Dolazak u 10.30 **2.** Southwest Airlines WN5432: Polazak u 10.15, Dolazak u 11.45 Ukupna cena povratne karte, sa svim taksama i naknadama, iznosi: • Opcija 1 (United/Alaska): 325 USD • Opcija 2 (Delta/Southwest): 310 USD Koju opciju biste želeli? Ili želite da potražim još letova?

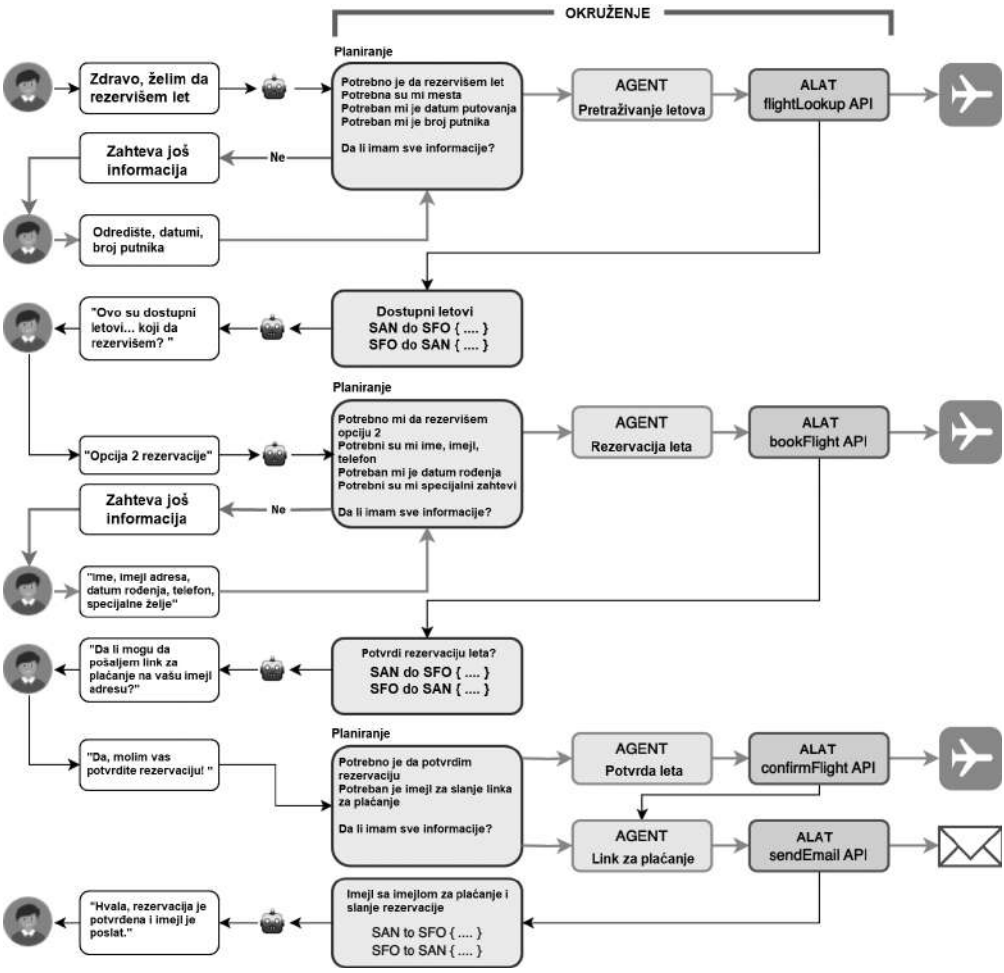
	Druga opcija mi se dopada. Možete li to da rezervišete za mene?
	Naravno! Biće mi zadovoljstvo da rezervišem drugu opciju za vas. Da potvrdimo, rezervišem sledeću rutu: Polazak: Alaska Airlines AS5678 • Datum: 23. 8. 2024. • Polazak iz San Dijega u 8.30, dolazak u San Francisko u 10.00 Povratak: Southwest Airlines WN5432 • Datum: 25. 8. 2024. • Polazak iz San Franciska u 10.15, dolazak u San Dijego u 11.45 Ukupna cena: 310 USD Da bismo nastavili sa rezervacijom, potrebne su mi sledeće informacije: **1.** Vaše puno ime, kao što stoji u zvaničnom dokumentu sa fotografijom **2.** Vaš datum rođenja **3.** Vaša imejl adresa (radi slanja potvrde) **4.** Vaš broj telefona **5.** Posebni zahtevi (npr. preferencije za obrok, sedište itd.) Takođe, imajte u vidu da će za završetak rezervacije biti potreban način plaćanja. Da li želite da nastavim?

	Da, spreman sam. Zovem se Džon Smit, rođen 15. 5. 1985. Moja imejl adresa je john.smith@email.com, a broj telefona 555-123-4567. Nemam posebnih zahteva.
	Hvala što ste dali te informacije, gospodine Smit. Unosim vaše podatke u sistem za rezervaciju. Evo sažetka onoga što imam: • Ime: Džon Smit • Datum rođenja: 15. 5. 1985. • Imejl adresa: john.smith@email.com • Telefon: 555-123-4567 • Posebni zahtevi: Nema Poslednji korak je obrada plaćanja u iznosu od 310 USD za vaše letove. Iz bezbednosnih razloga, ne mogu da prihvatim podatke o kreditnoj kartici direktno u ovom razgovoru. Umesto toga, poslaću vam bezbedni link za plaćanje na vašu imejl adresu (john.smith@email.com). Kada primite imejl, pratite link kako biste završili proces plaćanja. Nakon uspešnog plaćanja, dobićete još jedan imejl sa potvrdom leta i elektronskim kartama. Da li mogu da pošaljem link za plaćanje na vašu imejl adresu?
	Da, u redu je. Pošaljite link za plaćanje.
	Odlično, gospodine Smit. Pokrenuo sam proces slanja bezbednog linka za plaćanje na vašu imejl adresu (john.smith@email.com). Trebalo bi da ga dobijete u narednih nekoliko minuta. Proverite i „Inbox“ i „Spam“ direktorijume ako ga ne vidite odmah. Kada završite plaćanje, sistem će automatski obraditi vašu rezervaciju i poslati imejl sa potvrdom i elektronskim kartama. Taj imejl će sadržati sledeće informacije: **1.** Broj potvrde rezervacije **2.** Detaljnu rutu leta **3.** Uputstva za prijavu na let **4.** Informacije o dozvoljenom prtljagu **5.** Kontakt podatke avio-kompanije Nakon što primite potvrdu, proverite sve podatke da biste se uverili da je sve ispravno. Ako primetite bilo kakav problem ili imate dodatna pitanja, slobodno nam se obratite. Da li vam je potrebna još neka pomoć u vezi sa vašim putovanjem ili procesom rezervacije?

	Ne, to je sve. Hvala vam na pomoći!
	Nema na čemu, gospodine Smit! Drago mi je što sam mogao da vam pomognem oko rezervacije leta iz San Dijega za San Francisko. Ako vam bude potrebna dodatna pomoć ili budete imali bilo kakva pitanja, slobodno nam se obratite — dostupni smo 24/7. Želim vam prijatan put u San Francisko sledeće nedelje!

Na prvi pogled, ovaj primer deluje prilično jednostavno. Međutim, u pozadini se odvija čitav niz procesa. Prvo, korisnik započinje sa neodređenim zahtevom da rezerviše let. Asistent zasnovan na veštačkoj inteligenciji razume da, kako bi mogao da nastavi s obradom zahteva, mora da prikupi još nekoliko informacija, poput grada polaska i dolaska, datuma putovanja i slično.

Kako razgovor odmiče, asistent zasnovan na veštačkoj inteligenciji postavlja dodatna pitanja kako bi mogao uspešno da završi rezervaciju. Informacije koje prikuplja u svakoj fazi koriste se kao parametri za pozivanje funkcija API interfejsa u pozadini koje se odnose na rezervaciju letova; na primer, grad, datumi putovanja i broj putnika potrebni su da bi se pronašle informacije o dostupnim letovima putem (izmišljenog) API interfejsa pod nazivom **flightLookup**, dok su ime korisnika, telefon i imejl adresa potrebni za samu rezervaciju putem API interfejsa **bookFlight** i tako dalje. Na *slici 1.2* dat je pregled odluka koje model donosi, toka interakcije i alata (funkcija API interfejsa) koje poziva da bi završio svaki korak:



Slika 1.2 – Četbot za rezervaciju letova sa višestrukim agentima zasnovanim na velikim jezičkim modelima

Pored samog pozivanja alata, primetićete i da model vrši izvesnu introspekciju, poznatu i kao **lanac razmišljanja** (**CoT**), već na početku razgovora. To znači da model u pozadini osmišljava pristup korak po korak kako bi izvršio zadatak, ali ujedno uočava koje informacije nedostaju. Potom korisniku odgovara sa zahtevom da dostavi potrebne informacije. Na osnovu unetih podataka, model potom samostalno donosi određene odluke i poziva specifične alate pomoću agenta da bi završio zadatak.

Pored eksternih alata i API interfejsa, postoji još niz aspekata koje treba uzeti u obzir prilikom projektovanja ovakvog sistema. Osnovne komponente svakog od tih aspekata biće obrađene detaljno u narednim poglavljima. Za sada je dovoljno znati da agenti predstavljaju naprednu primenu velikih jezičkih modela (kao što je ovaj asistent zasnovan na velikim jezičkim modelima za rezervaciju putovanja) i da čine novoformirajuće polje koje kombinuje elemente različitih tipova velikih jezičkih modela sa dodatnim tehnikama veštačke inteligencije, poput učenja pojačavanjem, planiranja i korišćenja alata. Kako budete učili više u narednim poglavljima, primetićete da su agenti projektovani da budu interaktivniji, prilagodljiviji i sposobniji da izvršavaju složene zadatke u više koraka, u poređenju sa standardnim velikim jezičkim modelima, što ih čini pogodnim za najrazličitije složene zadatke i radne tokove.

Ali pre nego što se detaljnije upustimo u detalje agentnih sistema, osvrnućemo se na neke od primena generativne veštačke inteligencije.

Primene generativne veštačke inteligencije

Generativna veštačka inteligencija ima inovativne sposobnosti koje obuhvataju širok spektar oblasti. Kao što je ranije pomenuto, industrije poput zdravstva, finansija, obrazovanja, medija i zabave, marketinga, proizvodnje, maloprodaje i drugih, sve imaju potencijal da imaju koristi od mogućnosti generativne veštačke inteligencije. U nastavku je prikazan pregled samo nekih od glavnih primena tehnologije generativne veštačke inteligencije:

- **Generisanje slika i video zapisa**: Multimodalni modeli generativne veštačke inteligencije (odnosno veliki jezički modeli sposobni da obrađuju slike, tekst, zvuk i video) omogućavaju generisanje realističnog audio-vizuelnog sadržaja putem različitih platformi i alata. Na primer, u industriji medija i zabave, generativna veštačka inteligencija može pomoći u stvaranju vizuelnih efekata, dizajniranju virtuelnih slika korisnika i razvoju sadržaja za virtuelnu stvarnost. U modnoj i dizajnerskoj industriji koristi se za osmišljavanje novih modela odeće, pripremu virtuelnih modnih revija i predviđanje modnih trendova. U oblasti marketinškog oglašavanja i promocije, generativna veštačke inteligencije se koristi za kreiranje prilagođenih reklama, kampanja, marketinške komunikacije i drugih kreativnih materijala kao što su slike i logotipi.
- **Generisanje teksta i sadržaja**: Generativna veštačka inteligencija ostvarila je značajan napredak u zadacima koji se zasnivaju na tekstu. Na primer, pisanje sadržaja – kao što je brzo i jednostavno generisanje članaka, blog postova, marketinških i promotivnih tekstova za proizvode – već predstavlja neke od najčešćih primena. Četbotovi i virtuelni asistenti pružaju korisničku podršku na skoro ljudski način kada je reč o informacijama i upitima. Generativna veštačka inteligencija takođe pomaže u prevođenju teksta, sažimanju dokumenata i lokalizaciji sadržaja, čime se povećava pristupačnost informacija na platformama kao što su onlajn učenje i specijalizovani pretraživači.
- **Generisanje muzike i zvuka**: Kreiranje originalne muzike, zvučnih efekata i sinteze glasa predstavlja još jedan oblik multimedijalnog generisanja sadržaja koji omogućava generativna veštačka inteligencija. U praksi, ovakvi sadržaji se koriste u industriji video-igara i zabave za izradu audio materijala za obuku, automatizovanih asistenata u kol centrima, kao i u internet pametnim uređajima poput Amazon Alexe ili Google Nextra, koji mogu primati glasovne komande i odgovarati glasom kako bi izvršili automatizovane zadatke.

- **Zdravstvo i otkrivanje lekova**: Generativna veštačka inteligencija nalazi primenu u oblasti zdravstva. Na primer, koristi se za dizajniranje novih lekova i predviđanje njihove efikasnosti u lečenju bolesti ili drugih zdravstvenih stanja; za personalizovanu medicinu, gde se planovi lečenja prilagođavaju individualno, za svakog pacijenta ponaosob; kao i u medicinskom snimanju, gde se poboljšava kvalitet slika i generišu veštačke slike u istraživačke svrhe.
- **Generisanje koda**: Generisanje koda je nova funkcionalnost velikih jezičkih modela koja značajno pomaže programerima u razvoju softvera. Ovi modeli često mogu da generišu precizne delove koda, a ponekad i čitave funkcije, na osnovu tekstualnih uputstava. Napredne primene generativne veštačke inteligencije u generisanju koda uključuju dodatke za različita integrisana razvojna okruženja (IDE), poput Visual Studio Code (VS Code), koji mogu da razumeju kompletan kontekst kodne baze koda, identifikuju greške, kreiraju dokumentaciju za kod i generišu skriptove za jedinačna testiranja. Još jedna primena je pretvaranje teksta u upit, gde se prirodnije tekstualno pitanje prevodi u SQL upit, koji se zatim izvršava nad bazom podataka kako bi se dobili željeni rezultati. Međutim, ovo je oblast u kojoj je potrebna posebna pažnja, naročito pre izvršavanja koda koji generiše veliki jezički model, kako bi se izbegli rizici od pokretanja zlonamernog koda. Često se, u takvim slučajevima, generisani kod izvršava u izolovanom okruženju i prethodno se proverava da bi se utvrdilo da li je bezbedan i pogodan za izvršavanje u autonomnom okruženju.
- **Autonomni radni tokovi i robotika**: Kao što smo ranije ukratko pomenuli, agenti veštačke inteligencije zasnovani na velikim jezičkim modelima predstavljaju naprednu primenu generativne veštačke inteligencije, pri čemu veliki jezički modeli čine njihovu osnovu. Ovi agenti mogu obavljati različite zadatke u raznovrsnim slučajevima upotrebe. Na primer, veb sajt za rezervaciju putovanja sa virtuelnim četbot asistentom može implementirati agente koji automatski izvršavaju proces rezervacije putovanja i hotela za korisnika. U takvim slučajevima, sistem zasnovan na agentima može preduzimati konkretne radnje korišćenjem alata (poput pozivanja API interfejsa za rezervaciju letova), jednostavno razumevajući kontekst razgovora koji korisnik vodi sa četbotom. Napredniji agentni sistemi posebno se primenjuju u robotici, gde su radnje robota prvenstveno kontrolisane putem agenata. Agentni robot donosi odluke na osnovu svoje okoline ili specifičnog zadatka koji mu je dodeljen. Agent zatim izvršava radnje koje sam veliki jezički model ne može — na primer, agent može uključiti ili isključiti mehanizam za pokretanje ruke robota. Veštačka inteligencija u robotici je novo i otvoreno istraživačko polje, sa brojnim inovacijama koje se sve više pojavljuju u industriji, naročito u oblasti proizvodnje.

Ovo su samo neki od uobičajenih primera upotrebe generativne veštačke inteligencije zasnovane na velikim jezičkim modelima. Drugi specifični primeri uključuju primenu u modi i dizajnu, generisanje veštačkih podataka, personalizovani obrazovni sadržaj, finansijsko modeliranje i predviđanje, kao i predikciono održavanje. Sledeća lista prikazuje kratak pregled komercijalnih alata i alata otvorenog koda koji koriste generativnu veštačku inteligenciju za različite slučajeve upotrebe koje smo prethodno pomenuli:

SLUČAJEVI UPOTREBE	KOMERCIJALNI ALATI	ALATI OTVORENOG KODA
Vizuelni efekti i dizajn virtuelnih slika korisnika	• **MetaHuman Creator (Unreal Engine)**: Omogućava izradu visoko realističnih digitalnih likova za potrebe video-igara i filmske produkcije. • **NVIDIA Omniverse**: Platforma za trodimenzionalni dizajn, saradnju i simulaciju, korisna za kreiranje vizuelnih efekata	• **DeepFaceLab**: Koristi se za zamenu lica u video sadržajima i izradu digitalnih slika korisnika • **StyleGAN**: Može da generiše izuzetno realistična lica i prilagodi se za izradu virtuelnih slika korisnika.
Razvoj sadržaja za virtuelnu realnost	• **Unity**: Iako nije namenjen isključivo za virtuelnu stvarnost, poseduje snažne mogućnosti za razvoj sadržaja virtuelne stvarnosti • **Adobe Aero**: Omogućava kreiranje iskustava proširene stvarnosti	• **A-Frame**: Veb okvir za izgradnju iskustava virtuelne stvarnosti • **Godot**: Alat za razvoj igara otvorenog koda sa podrškom za virtuelnu stvarnost
Dizajn odeće i virtuelne modne revije	• **CLO3D**: Program za trodimenzionalni dizajn garderobe sa mogućnošću organizovanja virtuelnih modnih revija • **Browzwear**: Nudi trodimenionalna dizajn rešenja za modnu industriju	• **Blender**: Iako je prvenstveno alat za trodimenzionalno modelovanje, može se koristiti za modni dizajn i virtuelne revije
Predviđanje modnih trendova	• **WGSN**: Koristi veštačku inteligenciju za predviđanje modnih trendova • Heuritech: Pruža predviđanje trendova na osnovu veštačke inteligencije	• **Nema**
Marketing – generisanje oglasa i kreiranje kampanja	• **Jasper**: Asistent zasnovan na veštačkoj inteligenciji za pisanje koji može pomoći u kreiranju marketinških tekstova • **Midjourney**: Alat zasnovan na veštačkoj inteligenciji za pravljenje slika, pogodan za marketinške potrebe	• **GPT-J**: Model jezika otvorenog koda, koji se može fino podesiti za kreiranje marketinških tekstova • **Stable Diffusion**: Model za generisanje slika otvorenog koda, sposoban da pravi marketinške vizuale

SLUČAJEVI UPOTREBE	KOMERCIJALNI ALATI	ALATI OTVORENOG KODA
Kreiranje logotipa i slika	• **DALL-E 2**: Može generisati jedinstvene slike i logotipe na osnovu tekstualnih opisa • **Canva**: Iako nije u potpunosti zasnovan na veštačkoj inteligenciji, uključuje funkcije veštačke inteligencije za pomoć pri dizajnu	• **Stable Diffusion**: Može se koristiti za generisanje logotipa i slika • **Craiyon (nekadašnji DALL-E mini)**: Alternativa alata DALL-E otvorenog koda
Generisanje teksta i sadržaja	• ChatGPT kompanije OpenAI • Claude AI kompanije Anthropic • Jasper • Copy.ai • Anyword • Writer • WriteSonic I još mnogo drugih…	• **Mistral 7B**: Veliki jezički model otvorenog koda, prepoznatljiv po efikasnosti i tačnosti u zadacima generisanja teksta • **LlaMA**: Kolekcija prethodno obučenih i fino podešenih generativnih tekstualnih modela otvorenog koda, koji nude verzije sa različitim veličinama parametara • **BLOOM**: Veliki model veštačke inteligencije orvorenog pristupa razvijen od strane više od 1.000 istraživača veštačke inteligencije, poznat po svojoj robusnosti i svestranosti u zadacima generisanja teksta I još mnogo drugih…
Generisanje koda	• GitHub Copilot • Amazon Q for Developers • Tabnine • OpenAI Codex I još mnogo drugih…	• **Code T5**: Model veštačke inteligencije otvorenog koda obučen za zadatke vezane za kod. Može generisati delove koda, kompletne funkcije, pa čak i prevoditi između programskih jezika • **Polycoder**: Alternativa otvorenog koda komercijalnim generatorima koda. Obučavan je na velikoj kodnoj bazi koja pokriva više programskih jezika i može generisati kod za različite zadatke

SLUČAJEVI UPOTREBE	KOMERCIJALNI ALATI	ALATI OTVORENOG KODA
Autonomni radni tokovi i robotika	• **UiPath** integriše generativnu veštačku inteligenciju u svoju platformu za robotizovanu automatizaciju procesa (RPA). Koristi veštačku inteligenciju za otkrivanje mogućnosti automatizacije i unapređenje zadataka kao što su razumevanje dokumenata i analiziranje komunikacija. • **Automation Anywhere** integriše generativnu veštačku inteligenciju u svoju platformu za automatizaciju. Pruža funkcije kao što je Automation Co-Pilot za automatizatore, koji koristi generativnu veštačku inteligenciju za ubrzanje produktivnosti programera. • **NVIDIA** pruža radne tokove veštačke inteligencije koji koriste njihove okvire veštačke inteligencije, uključujući sposobnosti generativne veštačke inteligencije, za razvoj inovativnih rešenja u robotici i autonomnim sistemima.	• **OpenAI Gym** pruža alat za razvoj i upoređivanje algoritama za učenje pojačavanjem, koji se mogu kombinovati sa velikim jezičkim modelima za naprednije primene u robotici. • **Hugging Face** nudi biblioteke otvorenog koda koje se mogu koristiti za implementaciju velikih jezičkih modela u robotici i aplikacijama za autonomne radne tokove. • **LangChain** je Python biblioteka otvorenog koda koja se koristi za razvoj aplikacija sa velikim jezičkim modelima, a može se primeniti za kreiranje inteligentnijih autonomnih radnih tokova i robotskih sistema.

Tabela 1.1 – Primeri komercijalnih alata i alata otvorenog koda koji koriste generativnu veštačku inteligenciju

Ova lista ni na koji način ne predstavlja iscrpan pregled svih komercijalnih alata i alata otvorenog koda koji su trenutno dostupni na tržištu, a situacija se svakodnevno menja. Svedoci smo sve većeg broja startap kompanija koje nude nove i inovativne načine za rešavanje stvarnih problema pomoću generativne veštačke inteligencije, kao i pojave novih proizvođača modela koji razvijaju najsavremenije velike jezičke modele, sposobnije od prethodnih i znatno jeftinije za upotrebu. Sve to nam pokazuje koliko je oblast generativne veštačke inteligencije dinamična i uzbudljiva.

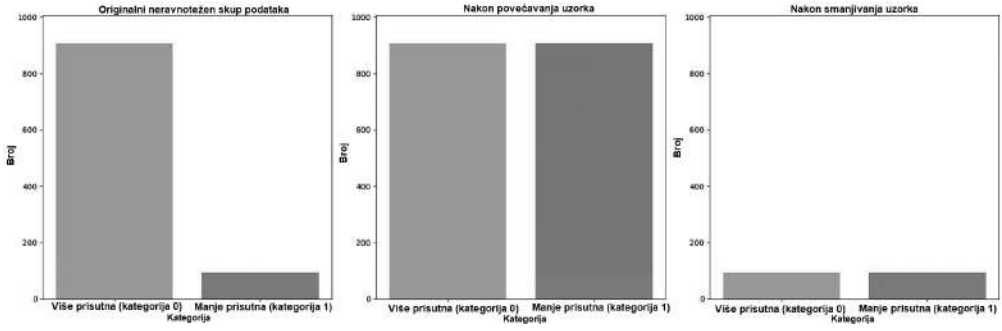
Izazovi i ograničenja generativne veštačke inteligencije

Iako generativna veštačka inteligencija donosi brojne prednosti, nije lišena izazova i ograničenja. Neka od tih ograničenja treba pažljivo razmotriti prilikom razmatranja njene primene u konkretnim slučajevima. U nastavku je kratak prikaz najčešćih problema u vezi sa velikim jezičkim modelima i nekih načina za njihovo prevazilaženje.

Kvalitet podataka i pristrasnost

Generativni modeli u velikoj meri zavise od kvaliteta i raznolikosti podataka u skupu za obuku. Svaki model koji se obučava na pristrasnim ili nereprezentativnim podacima proizvodiće rezultate sa istim vrstama pristrasnosti, čime se postojeće pristrasnosti učvršćuju ili dolazi do marginalizacije jedne ili više grupa, u slučaju da pristrasnost postoji u skupu podataka za obuku.

Jedan od načina da se ovaj problem ublaži jeste da se obezbede raznoliki i kvalitetniji podaci, kako bi se u skupu podataka za obuku obuhvatio širok spektar perspektiva. Kao i kod svakog problema mašinskog učenja, analiza podataka i poznavanje raspodele podataka po karakteristikama često pomažu. Analiza podataka može da otkrije neravnoteže koje mogu uneti pristrasnost u model. Postoji više algoritamskih načina da se smanji pristrasnost u podacima za obuku (na primer, tehnikama povećavanja ili smanjivanja uzoraka), ali svaki od tih metoda ima svoje prednosti i nedostatke. Na primer, zamislimo dve kategorije podataka u skupu za obuku, pri čemu se nalazi više primera podataka iz kategorije 0 nego iz kategorije 1, što izaziva prirodnu neravnotežu u skupu podataka. Kada se model obučava na takvom skupu, on će se „prekomerno obučiti" na podatke iz kategorije 0 i postaće efikasniji u obradi, pa čak i memorisanju podataka te kategorije, dok neće dobro obrađivati podatke koji pripadaju kategoriji 1. Ovo u osnovi dovodi do lošijih performansi modela i izražene pristrasnosti. Na sledećoj slici prikazani su efekti povećavanja i smanjivanja uzorka na takav skup podataka kako bi se ublažila pristrasnost:



Slika 1.3 – Neravnomerna raspodela podataka i efekti povećavanja i smanjivanja uzorka

Privatnost podataka

Brojni eksperimenti i istraživanja pokazali su da veliki jezički modeli imaju sklonost da odavaju podatke na kojima su obučavani. Ovo je naročito problematično ako su modeli obučavani na velikim količinama privatnih ili vlasničkih informacija. Posebne tehnike i načini zadavanja upita velikim jezičkim modelima pokazali su da je dovoljno naterati model da generiše podatke koji sadrže doslovan tekst iz njegovog skupa za obuku. Ove tehnike zadavanja upita nisu sofisticirane i često predstavljaju veoma jednostavan i jeftin napad, kojim se model može prisiliti da otkrije informacije. Na primer, kao što je prikazano u radu pod nazivom *Scalable Extraction of Training Data from (Production) Language Models* [15], istraživači su uspeli da potroše svega 200 dolara vrednosti poziva prema API interfejsu modela GPT-3.5 kompanije OpenAI i da pomoću tehnike ubrizgavanja upita nateraju model da otkrije privatne informacije. Otkrivene informacije uključuju imena ljudi, imejl adrese, fizičke adrese i brojeve telefona, koji su se nenamerno našli u skupu za obuku modela. Iako neke od ovih problema aktivno rešavaju kompanije koje razvijaju

modele, kao što je OpenAI, oni i dalje mogu predstavljati izazov za organizacije koje odluče da obučavaju sopstvene modele interno.

Neki od uobičajenih mehanizama za ublažavanje ovog problema uključuju primenu anonimizacije ili pseudonimizacije nad podacima za obuku, kako bi se uklonile ili prikrile lične, privatne ili poverljive informacije pre samog obučavanja modela. Obe ove tehnike uključuju korišćenje manjih, bržih i jeftinijih modela veštačke inteligencije koji obavljaju klasifikaciju i prepoznavanje entiteta radi identifikacije prisustva ličnih ili privatnih podataka u skupu za obuku. Kada se identifikuju, privatne informacije mogu biti prikrivene, maskirane ili u potpunosti uklonjene iz skupa podataka za obuku. Važno je napomenuti da primena ovih tehnika može imati određene posledice u zavisnosti od konkretne namene modela, pa je neophodna detaljna analiza kako bi se osiguralo da se time ne narušavaju performanse modela.

Računarski resursi

Obuka složenih generativnih modela zahteva izuzetno velike računarske resurse i predstavlja ozbiljno opterećenje po pitanju potrošnje energije i troškova — što često čini obuku velikih jezičkih modela ekonomski neisplativom. Iako su postignuti brojni hardverski napreci koji omogućavaju obučavanje ovakvih modela, pristup tim hardverskim resursima i dalje je ograničen, skup i rezervisan za privilegovane korisnike. Na primer, **grafički procesori** (**GPU**), prvobitno osmišljeni za prikaz visokokvalitetne grafike, postali su ključni za obuku, podešavanje i izvođenje modela veštačke inteligencije. CUDA arhitektura NVIDIA kompanije, predstavljena 2006. godine, omogućila je da se grafički procesori koriste za opštu obradu podataka, a savremeni grafički procesori poput NVIDIA A100 i H100 modela optimizovani su upravo za radna opterećenja veštačke inteligencije. Ipak, ovi vrhunski grafički procesori mogu koštati i desetine hiljada dolara po jedinici.

Smatra se da sama obuka modela kao što je GPT-3 košta nekoliko miliona dolara samo u računskoj potrošnji. Tačna cena zavisi od više faktora, uključujući veličinu modela, trajanje obuke i efikasnost hardvera. Na primer, procena je da je za obuku GPT-3 modela potrošeno oko 4–5 miliona dolara u računarskim resursima, dok su noviji modeli poput GPT-4 i PaLM verovatno još skuplji za obučavanje. Osim samog hardvera, postoje i značajni infrastrukturni troškovi povezani sa potrošnjom električne energije, hlađenjem i prostorom u centrima podataka. Zbog ovako visokih troškova, obukom velikih jezičkih modela uglavnom se bave velike tehnološke kompanije ili dobro finansirane istraživačke institucije. Danas su preko servisa u oblaku dostupni unapred obučeni modeli i mogućnosti za njihovo dodatno podešavanje, što omogućava da i manji timovi i organizacije pristupe određenim aspektima rada s velikim jezičkim modelima. Jedan od načina da se prevaziđe ovaj izazov leži u nedavnim inovacijama sa **malim jezičkim modelima** (**SLM**) koji predstavljaju znatno manje modele generativne veštačke inteligencije, obučavane za izvršavanje konkretnih i specijalizovanih zadataka. Iako su ovakvi modeli ograničeni na usko definisane oblasti, njihova obuka je znatno jeftinija jer zahteva daleko manje računarskih resursa.

Etičke i društvene posledice

Napredak u razvoju generativne veštačke inteligencije otvorio je niz složenih etičkih i društvenih pitanja, a sledeći izazovi predstavljaju samo deo šire slike:

- **Lažni multimedijalni sadržaji i dezinformacije**: Generativna veštačka inteligencija može da stvara vrlo realističan, veštački generisan sadržaj. Ako se takav sadržaj koristi u loše namere, može da dovede do stvaranja lažnog multimedijalnog materijala ili širenja dezinformacija, što predstavlja ozbiljnu pretnju po privatnost, bezbednost i poverenje javnosti.
- **Intelektualna svojina**: Kreiranje sadržaja koji je sličan već postojećem otvara pitanja o autorskim pravima i pravima intelektualne svojine. Pravno regulisanje originalnosti i vlasništva nad sadržajem generisanim veštačkom inteligencijom ostaje složen i nejasan izazov.
- **Gubitak radnih mesta**: Iako ekonomisti često umanjuju zabrinutost u vezi sa gubicima radnih mesta u sektorima generisanja sadržaja i automatizovanih zadataka, realnost je da su promene neminovne. Zato je važno paralelno razvijati strategije za prekvalifikaciju i osposobljavanje pogođenih radnika, kao i podsticanje preduzetničkih inicijativa.

Kada se govori o ublažavanju društvenih i etičkih posledica, odgovor je više filozofske nego tehničke prirode. Postoje određene tehničke mere, poput detekcije lažnih multimedijalnih slika i alata za prepoznavanje ličnih podataka, ali šira društvena pitanja ostaju u domenu javne rasprave i zakonske regulative o ulozi veštačke inteligencije u društvu.

Generalizacija i kreativnost

Jedan od ključnih problema modela generativne veštačke inteligencije jeste slabija sposobnost generalizacije. Tačnije, modeli retko kada generišu sadržaj koji je zaista različit od podataka na kojima su obučavani. Iako briljiraju u prepoznavanju i reprodukovanju obrazaca iz postojećih podataka, njihova sposobnost za istinsku originalnost – kreativnost s velikim *K* – ostaje ograničena.

Kako se generativna veštačka inteligencija dalje razvija, izvesno je da će sa sobom donositi nove izazove. Trenutno već postoji čitav niz istraživačkih napora koji su usmereni na to da se obezbedi odgovorna i smisleno korisna primena ove tehnologije.

Rezime

Generativna veštačka inteligencija predstavlja oblast naglog razvoja, sa ogromnim potencijalom da transformiše industrije i svakodnevni život. Neki od najvažnijih primera primene uključuju realističnu generaciju slika i teksta, ubrzano otkrivanje lekova i unapređene oblike kreativnog izražavanja.

Ključni elementi za razumevanje ove oblasti uključuju poznavanje različitih generativnih modela, poput varijacionih autokodera), generativnih suparničkih mreža i autoregresivnih modela, koji otkrivaju kako tehnologije funkcionišu i gde mogu da se primene. Ipak, jednako je važno razumeti i ograničenja generativne veštačke inteligencije – od kvaliteta podataka i potrebnih računarskih resursa, do etičkih problema i kreativnih ograničenja.

U sledećem poglavlju upoznaćemo se sa principima agentnih sistema, uključujući pojmove delovanja i autonomije, osobine inteligentnih agenata, različite arhitekture inteligentnih sistema, kao i multi-agentne sisteme.

Pitanja

1. Modeli generativne veštačke inteligencije mogu da primaju samo tekstualne podatke kao ulaz. Tačno ili netačno?
2. Koje su različite vrste modela generativne veštačke inteligencije?
3. Koje su neke od etičkih i društvenih posledica generativne veštačke inteligencije?
4. Koji su neki od metoda za ublažavanje pristrasnosti u podacima za obuku?
5. Koji su neki od velikih jezičkih modela otvorenog koda za generisanje teksta i sadržaja?

Odgovori

1. Netačno. Modeli generativne veštačke inteligencije mogu da primaju tekst, slike, video i audio podatke kao ulaz.
2. Varijacioni autokoderi, generativne suparničke mreže i autoregresivni modeli.
3. Lažni multimedijalni snimci, dezinformacije, pitanja autorskih prava ili prava intelektualne svojine i gubitak radnih mesta su neke od etičkih i društvenih posledica generativne veštačke inteligencije.
4. Neki od uobičajenih metoda za ublažavanje pristrasnosti u podacima za obuku su povećavanje uzorka i smanjivanje uzorka.
5. Mistral, LLaMA i Bloom su neki od velikih jezičkih modela otvorenog koda za generisanje teksta i sadržaja.

Dodatna literatura

- *Mastering Machine Learning Algorithms – drugo izdanje* – Đuzepe Bonakorso
- *Machine Learning for Imbalanced Data* – Kumar Abishek i dr Munir Abdelaziz
- *Generative AI with Python and TensorFlow 2* – Džozef Bebkok i Ragav Bali

Literatura

1. *Auto-Encoding Variational Bayes*: `https://arxiv.org/abs/1312.6114`
2. *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*: `https://openreview.net/forum?id=Sy2fzU9gl`
3. *Learning Structured Output Representation using Deep Conditional Generative Models*: `https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html`

4. *Generative Adversarial Networks*: https://arxiv.org/abs/1406.2661
5. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*: https://arxiv.org/abs/1511.06434
6. *Wasserstein GAN*: https://arxiv.org/abs/1701.07875
7. *A Style-Based Generator Architecture for Generative Adversarial Networks*: https://ieeexplore.ieee.org/document/8953766
8. *Attention Is All You Need*: https://arxiv.org/abs/1706.03762
9. *Conditional Image Generation with PixelCNN Decoders*: https://arxiv.org/abs/1606.05328
10. *PixelSNAIL: An Improved Autoregressive Generative Model*: https://www.researchgate.net/publication/322114155_PixelSNAIL_An_Improved_Autoregressive_Generative_Model
11. *Language Models are Few-Shot Learners*: https://arxiv.org/abs/2005.14165
12. *GPT-4 Technical Report*: https://arxiv.org/abs/2303.08774
13. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*: https://arxiv.org/abs/1810.04805
14. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*: https://arxiv.org/abs/1910.10683
15. *Scalable Extraction of Training Data from (Production) Language Models*: https://arxiv.org/abs/2311.17035

Pridružite se našim zajednicama na platformama Discord i Reddit

Imate pitanja o knjizi ili želite da učestvujete u diskusijama o generativnoj veštačkoj inteligenciji i velikim jezičkim modelima? Pridružite se našem Discord serveru na https://packt.link/I1tSU i našem Reddit kanalu na https://packt.link/ugMW0 kako biste se povezali, delili iskustva i sarađivali sa entuzijastima sličnih interesovanja.