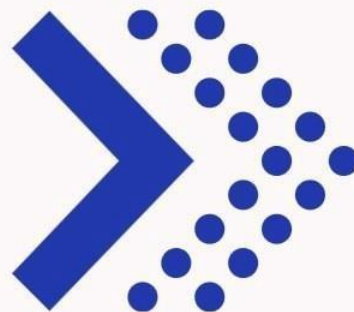




Halucinacije veštačke inteligencije

Halucinacije veštačke inteligencije (AI hallucinations) predstavljaju fascinantn, ali i problematičan fenomen koji se javlja kod generativnih modela, kao što su veliki jezički modeli (LLM) poput ChatGPT-a ili GPT-4. Halucinacije se odnose na situacije u kojima AI generiše odgovore koji zvuče logično i uverljivo, ali su potpuno netačni ili izmišljeni. Ove greške izazivaju posebnu pažnju, jer mogu ugroziti poverenje u AI tehnologije, širiti dezinformacije i zbunjivati korisnike.



Ilustracija imagebot.ai

Šta su AI halucinacije?

Pojam „halucinacije“ u ovom kontekstu preuzet je iz neurologije, gde se odnosi na perceptivne obmane – viđenje ili doživljaj nečega što nije realno prisutno. Kod AI modela, halucinacije su odgovori koji nisu zasnovani na podacima na kojima je model treniran niti na stvarnim informacijama. Umesto toga, AI kreira sadržaj koji može zvučati verodostojno, ali ne odražava stvarnost.

Na primer, generativni tekstualni modeli mogu proizvesti izmišljene informacije o istorijskim ličnostima, citatima, naučnim činjenicama ili čak generisati detalje koji ne postoje, poput naslova knjiga ili članaka. Kod vizuelnih AI modela, halucinacije mogu uključivati prikaz predmeta, pejzaža ili ljudi koji ne odgovaraju stvarnim podacima, već su rezultat algoritamskih interpretacija ili grešaka.

Primeri halucinacija

Jedan poznat primer dogodio se 2023. godine kada je Googleov AI model Bard proizveo tvrdnju da je James Webb teleskop snimio prvu fotografiju egzoplanete. Ova informacija bila je netačna, ali je predstavljena na ubedljiv način. Ovo je izazvalo zabrinutost u naučnim i tehnološkim krugovima, jer takve greške mogu dovesti do širenja dezinformacija, posebno kada dolaze iz pouzdanih izvora.

U domenu programiranja, AI modeli koji generišu kod često prave slične greške. Na primer, model može generisati sintaksu koja deluje pravilno, ali je nefunkcionalna ili neefikasna. Slični problemi se javljaju i kod AI sistema koji obrađuju slike ili zvuk, gde rezultati mogu biti nepravilni ili neprikladni.

Kako i zašto dolazi do AI halucinacija?

Halucinacije su rezultat nekoliko faktora povezanih s načinom na koji AI modeli funkcionišu:

1. Statistička priroda modela

Generativni modeli poput GPT-a rade na osnovu prepoznavanja obrazaca u velikim skupovima podataka. Kada model nema konkretne informacije, on generiše odgovore na osnovu statističke verovatnoće, odnosno na osnovu reči ili pojmova koji su najčešće povezani u prethodno obrađenim podacima. Ovo može rezultirati logički strukturiranim, ali činjenično netačnim odgovorima.

2. Nedovoljni ili pristrasni podaci

Ako su podaci na kojima je AI model treniran nepotpuni, zastareli ili pristrasni, to će se odraziti na generisane odgovore. Na primer, modeli trenirani na internetu mogu usvojiti dezinformacije ili pristrasne perspektive koje kruže na određenim platformama.

3. Pretpostavka koherentnosti

Jedan od glavnih ciljeva AI modela je da generišu odgovore koji su gramatički ispravni i koherentni. Međutim, koherentnost ne znači nužno i tačnost. Model može „zvučati“ uverljivo, čak i kada daje netačne informacije.

4. Kompleksnost zadataka

Kada se model suoči s pitanjima koja zahtevaju dubinsko razumevanje ili precizne informacije, može doći do grešaka. Na primer, ako modelu postavite pitanje o složenoj pravnoj ili medicinskoj temi, velika je verovatnoća da će generisani odgovor sadržati netačne ili izmišljene informacije.

Problematicnost AI halucinacija

Halucinacije nisu samo tehnički problem; one imaju šire društvene implikacije. U informativnim i obrazovnim kontekstima, širenje netačnih podataka može imati ozbiljne posledice. Na primer, učenik koji koristi AI za istraživanje može usvojiti netačne informacije, što utiče na njegovu edukaciju i kasnije odluke.

U nauci i medicini, netačni odgovori AI modela mogu izazvati pogrešne interpretacije podataka, što može ugroziti istraživanja ili čak dovesti do opasnih situacija. Širenje dezinformacija putem AI modela takođe može narušiti poverenje javnosti u tehnologiju.

Kako rešiti problem AI halucinacija?

Različiti pristupi mogu smanjiti učestalost i posledice AI halucinacija:

1. Bolja obuka modela

Treniranje AI na pouzdanijim, tačnijim i raznovrsnijim podacima može značajno smanjiti rizik od halucinacija. Takođe, kontinuirano ažuriranje podataka omogućava modelima da budu u toku s novim saznanjima i činjenicama.

2. Provera činjenica

Implementacija sistema za automatsku proveru činjenica u okviru AI modela može smanjiti učestalost halucinacija. Na primer, AI može preusmeriti korisnika na pouzdane izvore ili naznačiti da nema dovoljno informacija za tačan odgovor.

3. Specijalizovani modeli

Ograničavanje domena u kojem AI model operiše može smanjiti greške. Na primer, AI specijalizovan za medicinu može biti precizniji od opšteg modela kada se radi o medicinskim pitanjima.

4. Transparentnost

AI modeli bi trebalo da jasno naznače kada nisu sigurni u tačnost svog odgovora. Ovo omogućava korisnicima da bolje procene pouzdanost informacija koje dobijaju.

Pozitivne strane halucinacija

Iako su AI halucinacije uglavnom percipirane kao problem, one mogu imati i pozitivne aspekte, posebno u kreativnim industrijama. Na primer, halucinacije vizuelnih modela često rezultiraju nadrealnim slikama koje inspirišu umetnike. Slično tome, tekstualni modeli mogu proizvesti inovativne ideje koje ljudi možda ne bi spontano osmislili.

Stephen Levy, autor članka „In Defence of AI Hallucinations“ u časopisu *Wired*, ističe da halucinacije ne treba isključivo posmatrati kao neuspehe. On sugerše da ovaj fenomen može podstaći kreativnost i otvoriti nove mogućnosti u umetnosti, dizajnu i istraživanju.

Zaključak

Halucinacije veštačke inteligencije su složen problem koji se ne može u potpunosti eliminisati, ali se može ublažiti kroz poboljšanje algoritama, obuke i sistema za proveru činjenica. Dokle god korisnici razumeju prirodu AI halucinacija i koriste generativne modele s kritičkim pristupom, ove tehnologije mogu ostati koristan alat u svakodnevnom životu i radu.

Halucinacije nas, ujedno, podsećaju na ključnu ulogu ljudske inteligencije – da proverava, analizira i preispituje informacije. U kombinaciji sa AI, ljudska intuicija i kritičko razmišljanje mogu doprineti efikasnijem i odgovornijem korišćenju veštačke inteligencije.