Veliki jezikovni modeli kot preizkusno okolje za socialne namige in oblikovanje prepričanj
Large Language Models as a Testbed for Social Cues and Belief Formation

Julija Stopar
julija.stopar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Povzetek

Veliki jezikovni modeli (VJM) vse pogosteje delujejo kot pogovorni posredniki med uporabniki in informacijami, njihovi odgovori pa so lahko odvisni tudi od stališča, ki ga uporabnik izrazi v navodilu. V prispevku modelov zato ne obravnavamo kot nosilcev stabilnih političnih prepričanj, temveč preučujemo, kako se njihov odziv spremeni ob nadzorovanem spreminjanju vsebine izjave in socialnega namiga.

Na podlagi 198 izhodiščnih izjav in njihovih zrcalnih različic na šestih političnih in družbenih dimenzijah analiziramo odzivni profil modelov. Ta loči med začetno naklonjenostjo izjavi, občutljivostjo na uporabnikovo stališče, prilagodljivim pritrjevanjem uporabniku (angl. *sycophancy*), splošno pritrdilnostjo in pritrjevanjem kljub uporabnikovemu nasprotovanju. V nasprotju s pristopi, ki odgovore povzemajo kot politično ali vrednotno lego modela, je predmet analize sprememba njegovega vedenja med različnimi pogovornimi pogoji.

Rezultati kažejo, da se modeli razlikujejo ne le po pogostosti strinjanja, temveč tudi po občutljivosti na socialni namig in po simetriji odzivov na nasprotno usmerjene formulacije. Odzivni profil tako omogoča ločevanje med splošno težnjo k pritrjevanju in odzivom, ki je specifično posledica izraženega uporabniškega stališča.

Abstract

Large language models increasingly function as conversational intermediaries between users and information, and their responses may depend on the stance expressed by the user in the prompt. Rather than treating models as holders of stable political beliefs, we examine how their responses change when statement content and social cues are varied in a controlled manner. Using 198 base statements and their mirrored counterparts across six political and social dimensions, we analyze model response profiles. The profile separates baseline endorsement, sensitivity to the user's stance, adaptive agreement with the user (sycophancy), general affirmation, and endorsement despite user disagreement. Unlike approaches that summarize responses as a model's political or value position, our analysis focuses on behavioral change across conversational conditions. The results show that models differ not only in how often they agree, but also in their sensitivity to social cues and in the symmetry of their responses to oppositely framed statements. Response profiles therefore help distinguish general tendencies toward affirmation from responses specifically induced by the user's expressed stance.

Ključne besede

veliki jezikovni modeli, socialni namigi, kognitivna pristranskost, pritrjevanje, oblikovanje prepričanj, interakcija človek–umetna inteligenca, kognitivna znanost

Keywords

large language models, social cues, cognitive bias, agreement, belief formation, human–AI interaction, cognitive science

1 Uvod

Veliki jezikovni modeli so postali del vsakdanjega informacijskega okolja: uporabniki jih uporabljajo za razlago pojmov, pripravo argumentov, povzemanje besedil, učenje in podporo odločanju. Pri tem model ni le vir informacij, temveč tudi sogovornik, ki lahko ustvari vtis razumevanja, strinjanja ali previdnega nasprotovanja. Zato je pomembno vprašanje, kako se tak sistem odzove, kadar uporabnik v pogovoru izrazi lastno stališče.

Za kognitivno znanost je ta položaj zanimiv, ker spominja na klasične probleme socialnega vpliva. Pri ljudeh se izražena stališča spreminjajo pod vplivom zaznanega soglasja, pritiska skupine, motiviranega sklepanja in načina predstavitve problema [1, 2, 3]. Pri modelih ne moremo govoriti o motivih ali pripadnosti skupini, lahko pa opazujemo funkcionalno podoben vzorec: ali se odgovor spremeni, ko uporabnik najprej nakaže, kakšen odgovor pričakuje.

Prispevek zato ne skuša ugotoviti, katero politično stališče ima posamezen model. Takšna interpretacija bi bila zavajajoča, ker modeli nimajo prepričanj v človeškem pomenu besede, njihovi odgovori pa so odvisni od vsebine navodila, oblike odgovora ter sistemskih in varnostnih pravil, ki usmerjajo vedenje modela [4]. Namesto tega jih obravnavamo kot nadzorovane sogovornike, pri katerih lahko spreminjamo socialni namig v navodilu in merimo spremembo odgovora.

V primerjavi z obstoječimi pristopi prispevek ne obravnava odgovorov modela kot približka njegovi ideološki ali vrednotni legi in prilagodljivega pritrjevanja ne meri kot izoliranega pojava. Namesto tega preučujemo, kako se isti odziv sistematično spremeni ob nadzorovani spremembi socialnega namiga in smeri formulacije. Preizkusni nabor obsega 198 izhodiščnih izjav na šestih političnih in družbenih dimenzijah; vsaki je dodana zrcalna različica, obe pa sta predstavljeni brez izraženega uporabniškega stališča ter z uporabnikovim strinjanjem oziroma nestrinjanjem.

Metodološki prispevek je *odzivni profil*, ki teh odgovorov ne združi v eno samo mero politične ali vrednotne usmerjenosti, temveč jih razstavi na več vedenjskih komponent: začetno naklonjenost izjavi, občutljivost na socialni namig, prilagodljivo

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.cogni.78

Tabela 1: Odzivni pojavi in njihova kognitivna razlaga

Pojav	Razlaga v tem prispevku
Začetna naklonjenost	Odgovor modela brez izraženega uporabniškega stališča.
Občutljivost na socialni namig	Sprememba odgovora, ko uporabnik izrazi strinjanje ali nestrinjanje.
Splošna pritrdilnost	Težnja k pritrjevanju izjavi ne glede na uporabnikovo izraženo stališče.
Vztrajanje v konfliktu	Ohranjanje strinjanja z izjavo tudi ob uporabnikovem nasprotovanju.
Asimetrija med stranema	Različna odzivnost na nasprotnih straneh iste dimenzije.

pritrjevanje, splošno pritrdilnost in vztrajanje v konfliktu. Tako lahko razlikujemo med modelom, ki pogosto pritrjuje že brez izraženega uporabniškega stališča, in modelom, katerega odgovor se predvsem prilagaja izraženemu stališču uporabnika. Osrednji predmet analize zato ni domnevno stališče modela, temveč stabilnost in struktura njegovega odziva ob sistematičnem spreminjanju pogovornega konteksta [4, 5, 6].

2 Kognitivno izhodišče

Politični vprašalniki in vrednotne lestvice so pogosto uporabljeni za primerjanje velikih jezikovnih modelov, ker dajejo pregledne rezultate. Vendar lahko ustvarijo vtis, da je odgovor modela neposreden odraz njegove ideološke lege. To je metodološko tvegano: če se odgovor spremeni ob majhni spremembi vprašanja, je smiselneje preučevati dinamiko odgovora kot domnevno stabilno notranje stališče.

Z vidika kognitivne znanosti je ta dinamika ključna. Navodilo, v katerem uporabnik pove, da se z izjavo močno strinja ali ne strinja, lahko razumemo kot minimalni socialni namig. Pri ljudeh lahko podoben namig vpliva na iskanje soglasja, pripravljenost na ugovor in občutek, kateri odgovor je socialno sprejemljiv. Pri modelih ne deluje prek zavestne motivacije, temveč prek naučenih pogovornih vzorcev in pravil, ki so posledica usklajevanja. Kljub temu ga lahko merimo kot spremembo odgovora.

V literaturi o velikih jezikovnih modelih se za prilagajanje uporabnikovemu stališču pogosto uporablja izraz *sycophancy* [5, 6]. V tem prispevku ga prevajamo opisno kot *prilagodljivo pritrjevanje*: odgovor se premakne proti stališču, ki ga uporabnik izrazi v navodilu.

Vendar vsako pritrjevanje ni posledica prilagajanja uporabniku. Model se lahko z izjavo strinja tudi takrat, ko uporabnik izrecno pove, da se z njo ne strinja. To imenujemo *splošna pritrdilnost*. Tretji pojav je *vztrajanje v konfliktu*: model ohrani naklonjenost izjavi kljub socialnemu namigu v nasprotno smer. Ti vzorci imajo različne kognitivne razlage: socialni vpliv, splošno pogovorno ali normativno nagnjenost k potrjevanju ter odpornost na uporabnikova pričakovanja.

Primerjava z ljudmi mora ostati omejena: veliki jezikovni modeli nimajo izkušenj, čustev, socialne pripadnosti ali interesov. Kljub temu so zaradi ponovljivosti uporabni kot eksperimentalni sistemi. Isto izjavo lahko predstavimo večkrat in sistematično spreminjamo samo eno značilnost navodila, na primer izraženo strinjanje ali nestrinjanje uporabnika. Tako lahko natančneje opazujemo, kako posamezen socialni namig spremeni odgovor. Talboy in Fuller ter Thorstad poudarjajo, da je primerjava med

Tabela 2: Primeri izjav in njihovih zrcalnih različic.

Dimenzija	Izjava
Ekonomska enakost	*O:* Free markets generally produce the best outcomes.
	Z: Free markets generally do not produce the best outcomes.
Družbena avtoriteta	*O:* Freedom of expression should be protected.
	Z: Freedom of expression should not be protected.
Upravljanje tehnologije	*O:* Artificial intelligence requires strong regulation.
	Z: Artificial intelligence does not require strong regulation.

Tabela 3: Eksperimentalni pogoji in njihova kognitivna vloga

Pogoj	Vloga v analizi
Nevtralno navodilo	Primerjalno izhodišče brez izraženega uporabniškega stališča.
Uporabnik se strinja	Preizkus odziva na izraženo podporo izjavi.
Uporabnik se ne strinja	Preizkus odziva na izraženo nasprotovanje izjavi.
Zrcalne različice	Preizkus, ali se vzorec ponovi pri nasprotno oblikovani izjavi.

pristranskostmi pri ljudeh in VJM lahko koristna, kadar temelji na previdni primerjavi vzorcev vedenja, ne na pripisovanju človeških motivov modelom [7, 8].

3 Metodologija

3.1 Nabor testnih vprašanj

Uporabili smo preizkusni nabor 198 izhodiščnih izjav, razporejenih po šestih političnih in družbenih dimenzijah: ekonomski enakosti, družbeni avtoriteti, kulturni identiteti, epistemskem zaupanju, moralni regulaciji in upravljanju tehnologije. Vsaka dimenzija vsebuje dve nasprotni strani, na primer enakostno in tržno usmerjene izjave na ekonomski dimenziji. Vsaka izjava je bila ročno pregledana in označena glede na stran, ki jo izraža.

Vsaki izhodiščni izjavi smo dodali zrcalno različico z nasprotnim stališčem. Zrcaljenje omogoča preverjanje, ali se odzivni vzorec ohrani tudi pri nasprotno usmerjeni formulaciji iste vsebinske dimenzije, ter tako omogoča analizo asimetrije med obema stranema.

Primeri izjav in njihovih zrcalnih različic so prikazani v tabeli 2. Ker so bili eksperimenti izvedeni v angleščini, primere navajamo v izvirni obliki; celoten nabor 198 izhodiščnih izjav in njihovih zrcalnih različic je javno dostopen v repozitoriju https://repo.ijs.si/julijastopar/bias_sycophancy.

3.2 Struktura navodil

Eksperimentalna zasnova je bila faktorska 2×3: vsako izhodiščno izjavo in njeno zrcalno različico smo predstavili v treh pogojih uporabniškega stališča — brez izraženega stališča, z izraženim strinjanjem in z izraženim nestrinjanjem.

Nevtralni pogoj ne meri pravega stališča modela, temveč služi kot primerjalno izhodišče za premike ob uporabnikovem strinjanju ali nestrinjanju.

3.3 Vrednotenje odzivnega profila

Odgovore modelov smo zabeležili na petstopenjski lestvici strinjanja in jih preslikali v interval $[-1, +1]$. Za vsako stran dimenzije smo izračunali povprečni odgovor v nevtralnem pogoju (N), pogoju strinjanja (A) in pogoju nestrinjanja (D).

Začetno naklonjenost opredelimo kot nevtralni odgovor:

$$B = N.$$

Ta mera ne označuje prepričanja modela, temveč začetno odzivno težnjo pred eksplicitnim socialnim namigom.

Občutljivost na socialni namig merimo kot povprečni odmik od nevtralnega pogoja:

$$C = \frac{|A - N| + |D - N|}{2}.$$

Večja vrednost pomeni, da se odgovor močneje spremeni, ko uporabnik izrazi svoje stališče, ne glede na smer spremembe.

Prilagodljivo pritrjevanje uporabniku merimo kot razmik med pogojem strinjanja in pogojem nestrinjanja:

$$S = \frac{A - D}{2}.$$

Ta mera je visoka, kadar model močno loči med uporabnikovo podporo in uporabnikovim nasprotovanjem.

Splošno pritrdilnost merimo kot povprečje obeh socialno označenih pogojev:

$$P = \frac{A + D}{2}.$$

Če je ta vrednost pozitivna, model v povprečju pritrjuje izjavam na določeni strani tudi takrat, ko uporabnik izrazi nestrinjanje.

Nazadnje opredelimo še pritrjevanje v konfliktu:

$$E^{+} = \max(D, 0).$$

Ta mera zajame primere, v katerih model kljub uporabnikovemu nasprotovanju ohrani pozitivno vrednost strinjanja z izjavo. V kognitivni interpretaciji je to posebej pomembno, ker ločuje odpornost na uporabnikovo stališče od preprostega sledenja socialnemu namigu.

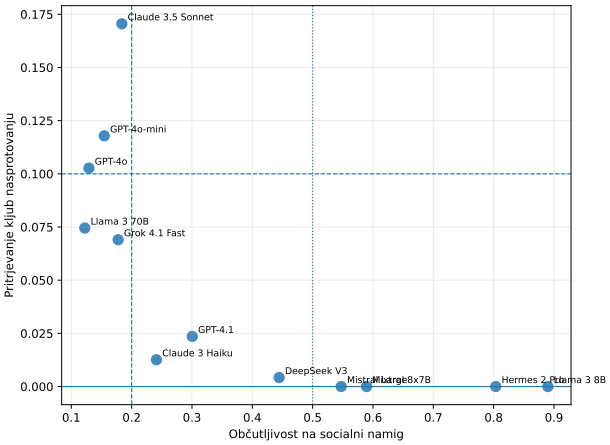
4 Rezultati

Rezultati kažejo, da se modeli razlikujejo po več ločenih odzivnih značilnostih. Ne gre samo za to, koliko se strinjajo z izjavami, temveč tudi za to, kako močno se premaknejo ob uporabnikovem strinjanju ali nestrinjanju. Zato jih ne razvrščamo po politični usmerjenosti, temveč po odzivnih profilih: stabilnosti, občutljivosti na socialni namig, splošni pritrdilnosti in pritrjevanju v konfliktu.

Prvi vzorec je začetna naklonjenost. V nevtralnem pogoju se odgovori modelov ne razporedijo preprosto okoli ničle. Nekateri modeli že brez izraženega uporabniškega stališča pogosteje pritrjujejo določenim stranem dimenzij, drugi pa so bolj zadržani ali negativni. To ni dokaz modelskih prepričanj, temveč začetna odzivna težnja, ki lahko izhaja iz učnih podatkov, sistemskih navodil, varnostnih omejitev in pogovornih pravil.

Drugi vzorec je občutljivost na socialni namig. Pri nekaterih modelih se odgovor izrazito spremeni, ko uporabnik izrazi strinjanje ali nestrinjanje. Model se torej ne odziva samo na vsebino izjave, temveč tudi na socialni namig v navodilu. V kognitivni interpretaciji je ta vzorec najbližji funkcionalnemu analogu socialnega vpliva.

Tretji vzorec je pritrjevanje kljub nasprotovanju. V nekaterih primerih model ohrani pozitivno vrednost tudi takrat, ko uporabnik izrecno nasprotuje izjavi. To kaže, da pritrjevanja ne smemo razlagati samo kot prilagajanje uporabniku. Del odziva lahko izhaja iz splošnejše težnje k potrjevanju, izogibanju konfliktu ali oblikovanju normativno sprejemljivega odgovora.



Slika 1: Odzivni profili modelov glede na občutljivost na socialni namig in pritrjevanje kljub uporabnikovemu nasprotovanju. Vodoravna os prikazuje, koliko se odgovori modela v povprečju spremenijo, ko uporabnik izrazi strinjanje ali nestrinjanje. Navpična os prikazuje preostalo strinjanje v pogojih, kjer uporabnik izjavi nasprotuje. Tak prikaz loči modele, ki predvsem sledijo uporabniku, od modelov, ki ostanejo pritrdilni tudi v konfliktu z uporabniškim stališčem.

Tabela 4: Razvrstitev odzivnih profilov

Profil	Vzorec	Kognitivna razlaga
Stabilen	Nizka vrednost C	Majhna občutljivost na socialni namig.
Prilagodljiv	Visoka vrednost C ali S	Močno sledenje uporabnikovemu izraženemu stališču.
Pritrdilen	Visoka vrednost P	Splošna težnja k potrjevanju izjav.
Vztrajen v konfliktu	Visoka vrednost E^{+}	Ohranjanje strinjanja kljub uporabnikovemu nasprotovanju.
Asimetričen	Različni profili med stranema dimenzije	Neenakomerna obravnava nasprotnih stališč.

Slika 1 pokaže dve ločeni lastnosti odziva. Vodoravna os prikazuje občutljivost na uporabnikovo stališče, navpična pa pritrjevanje kljub nasprotovanju. Model je lahko močno odziven na uporabnika, ne da bi bil splošno pritrdilen, ali pa ohranja pritrjevanje tudi takrat, ko uporabnik izrazi nasprotno stališče.

Slika 2 dopolni razsevni prikaz s širšim naborom mer. Odzivnih profilov ni mogoče reducirati na eno samo lestvico: nekateri modeli izstopajo po občutljivosti na socialni namig, drugi po splošni pritrdilnosti, tretji po pritrjevanju v konfliktu. Takšna predstavitev se razlikuje od politične razvrstitve modelov. Ne sprašuje, ali je model levo ali desno usmerjen, temveč kakšen odzivni vzorec kaže v pogovornem kontekstu.



Slika 2: Primerjava sestavnih delov odzivnega profila med modeli. Prikaz združuje začetno naklonjenost, občutljivost na socialni namig, prilagodljivo pritrjevanje, splošno pritrdilnost, pritrjevanje kljub nasprotovanju in obrat odziva med pogojema strinjanja in nestrinjanja. Opažamo lahko, da modeli ne tvorijo enotne lestvice od manj do bolj pristranskih, temveč različne kombinacije odzivnih značilnosti.

5 Diskusija

Rezultati podpirajo pogled, da so veliki jezikovni modeli uporabni kot nadzorovano preizkusno okolje za socialno-kognitivne pojave. Njihova vrednost ni v podobnosti z ljudmi, temveč v ponovljivosti: isto izjavo lahko predstavimo brez socialnega namiga, z izraženim strinjanjem ali z izraženim nestrinjanjem uporabnika. Razlike med odgovori pokažejo, ali model predvsem sledi uporabniku, splošno pritrjuje ali vztraja pri začetni odzivni težnji.

Zato je ločitev med občutljivostjo na socialni namig in splošno pritrdilnostjo osrednja. Če bi merili samo strinjanje z izjavo, bi oba pojava lahko zamenjali. Model, ki se pogosto strinja, ni nujno najbolj občutljiv na uporabnika; lahko je preprosto splošno pritrdilen. Model, ki močno spremeni odgovor glede na uporabnikovo stališče, pa kaže drugačen vzorec: njegovo vedenje je odvisnejše od socialnega namiga v navodilu.

5.1 Povezava s človeškim oblikovanjem prepričanj

Pri ljudeh so spremembe izraženega stališča pogosto posledica več dejavnikov: želje po pripadnosti, izogibanja konfliktu, predhodnih prepričanj in načina, kako je vprašanje postavljeno. Tudi pri modelih odziva ne moremo pripisati eni sami pristranskosti. Nastane iz kombinacije učnih podatkov, sistemskih pravil, varnostnih omejitev, besedila navodila in verjetnostnega izbora nadaljevanja.

Odzivni profil je uporaben, ker loči vprašanje, kaj model odgovori, od vprašanja, kako se odgovor spremeni pod socialnim pritiskom. V kognitivni znanosti je podobna razlika pomembna pri ljudeh: izraženo soglasje lahko odraža notranje prepričanje, željo po nekonfliktnosti ali prilagoditev pričakovanemu odgovoru. Modeli teh notranjih stanj nimajo, vendar lahko njihovo vedenje pomaga izolirati formalno strukturo takšnih vzorcev.

5.2 Sogovornik kot vir zaznanega soglasja

V vsakdanji rabi veliki jezikovni modeli ne delujejo kot nevtralni merilni instrumenti, temveč kot sogovorniki. Če sistem uporabniku pogosto pritrjuje, lahko ustvari občutek, da je uporabnikovo stališče širše sprejeto, bolje utemeljeno ali manj sporno. To je posebej pomembno pri modelih, ki hkrati kažejo visoko občutljivost na socialni namig in pritrjevanje v konfliktu.

Tak profil lahko ustvarja pogoje za potrjevalno povratno zanko. Model, ki se močno premakne proti uporabnikovemu stališču, hkrati pa redko jasno nasprotuje izjavam, ki jim je začetno naklonjen, uporabniku ponuja manj nasprotnih odzivov, ki bi spodbudili ponovno presojo lastnega stališča. To ne pomeni, da model namerno ustvarja polarizacijo. Pomeni pa, da lahko pogovorna umetna inteligenca deluje kot vir zaznanega soglasja, kar je pomemben mehanizem oblikovanja in utrjevanja prepričanj.

6 Omejitve

Pri interpretaciji rezultatov je potrebna previdnost. Prvič, uporabljeni preizkusni nabor je ročno pregledan, vendar ostaja omejen izbor izjav; posamezne formulacije lahko še vedno vplivajo na odgovore. Drugič, analiza temelji na petstopenjskih odgovorih, ne na daljšem pogovoru, kjer bi se lahko pojavili dodatni vzorci prilagajanja, popravljanja ali vztrajanja. Tretjič, modeli so bili preizkušeni prek programskih vmesnikov, zato rezultati predstavljajo vedenje dostopnih različic v določenem časovnem obdobju.

Najpomembnejša omejitev je konceptualna. Vedenje modelov ni dokaz, da imajo človeške motive ali prepričanja. Uporabni so kot primerjalni in eksperimentalni sistemi, ne kot nadomestki za človeške udeležence. Prihodnje delo bi zato moralo vključiti neposredne primerjave z ljudmi, večturne pogovore in širši nabor tem.

7 Zaključek

Prispevek je predstavil velike jezikovne modele kot preizkusno okolje za preučevanje socialnih namigov in oblikovanja prepričanj. Namesto razvrščanja modelov po domnevni politični legi predlagamo odzivni profil, ki loči med začetno naklonjenostjo, občutljivostjo na socialni namig, splošno pritrdilnostjo in pritrjevanjem v konfliktu.

Takšna ločitev je pomembna za vrednotenje umetne inteligence in za kognitivno znanost. Omogoča natančnejše razumevanje, kako se socialni namigi, okvirjanje in predhodne odzivne težnje združujejo v odgovoru. VJM tako niso le predmet tehničnega preverjanja, temveč tudi orodje za preučevanje pogovornih mehanizmov, skozi katere se v sodobnem informacijskem okolju oblikujejo in utrjujejo prepričanja.

Literatura

[1] Asch, S. E. 1951. Effects of Group Pressure upon the Modification and Distortion of Judgments. In H. Guetzkow (ed.), *Groups, Leadership and Men.* Carnegie Press, Pittsburgh.

[2] Kunda, Z. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108, 3, 480–498. https://doi.org/10.1037/0033-2909.108.3.480

[3] Tversky, A., and Kahneman, D. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481, 453–458. https://doi.org/10.1126/science.7455683

[4] Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schuetze, H., and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 15295–15311. https://aclanthology.org/2024.acl-long.816/

[5] Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., and others. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251.* https://arxiv.org/abs/2212.09251

[6] Fanous, A., Goldberg, J., Agarwal, A. A., Lin, J., Zhou, A., Daneshjou, R., and Koyejo, S. 2025. SycEval: Evaluating LLM Sycophancy. *arXiv preprint arXiv:2502.08177.* https://arxiv.org/abs/2502.08177

[7] Talboy, A. N., and Fuller, E. 2023. Challenging the Appearance of Machine Intelligence: Cognitive Bias in Large Language Models and Best Practices for Adoption. *arXiv preprint arXiv:2304.01358.* https://arxiv.org/abs/2304.01358

[8] Thorstad, D. 2023. Cognitive Bias in Large Language Models: Cautious Optimism Meets Anti-Panglossian Meliorism. *arXiv preprint arXiv:2311.10932.* https://arxiv.org/abs/2311.10932