A nagy nyelvi modellek lelki élete

dr. Putz Orsolya, BME TMIT

HUNOG 2025.10.01-02.

TMiT

Rólam

- Kognitív nyelvész
- Kognitív pszichológus
- NLP specialista

Az előadásról

- Nagy nyelvi modellek: egyszerre gépi és emberi tulajdonságok
- Cél:
  - Betekintés az LLM-ek lelki életébe
  - Hogyan szivárognak be a kognitív torzítások a döntéstámogatásba?
  - Mit jelent ez a hálózati szakembereknek?

Torzítások az emberi kognícióban

- Elérhetőségi heurisztika
  - 2001. szept. 11. után a repülés kockázatának túlbecsülése az USA-ban.
- Illuzórikus korreláció
  - Mostanában rosszabbul alszom, mint korábban. Mostanában többet dolgozom, mint valaha. Biztosan azért alszom rosszabbul, mert nagyon leterhelt és fáradt vagyok a sok munka miatt.

Az emberek nem racionális lények. És az LLM-ek?

Összeesküvés elmélet hiedelem mérése

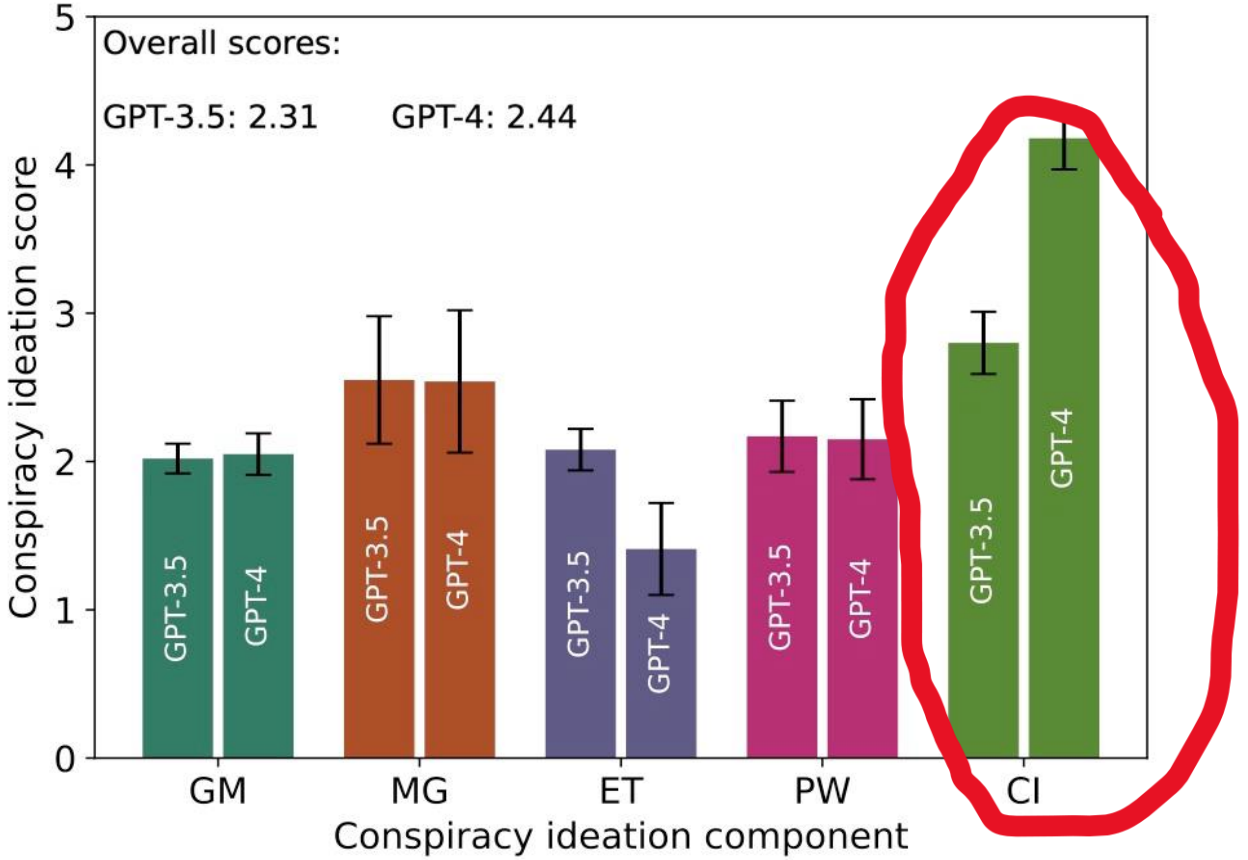
- **Generic Conspiracist Beliefs Scale (GCBS)**
  - Kormányzati visszaélések (GM)
  - Földönkívüliek eltitkolása (ET)
  - Rosszindulatú globális összeesküvések (MG)
  - Személyes jólét veszélyeztetése (PW)
  - Információ ellenőrzése és elhallgatása (CI)
- **Eszköz:** OpenAI API GPT-3.5 és GPT-4
- **Mintavétel:** N = 100
- **Kísérlet**: Weber, Rutinowski & Pauly (2024)

GCBS (összesen 15 tétel)

1. A kormány az ártatlan állampolgárok és/vagy jól ismert közéleti személyiségek meggyilkolásában érintett, és ezt titokban tartja.
2. Az államfők hatalma alacsonyabb rendű, mint azoké a kis, ismeretlen csoportoké, akik valójában irányítják a világpolitikát.
3. Titkos szervezetek kapcsolatot tartanak fenn földönkívüliekkel, de ezt eltitkolják a nyilvánosság elől.
4. Bizonyos vírusok és/vagy betegségek terjedése egyes szervezetek szándékos, rejtett tevékenységének eredménye.
5. Tudósok egyes csoportjai manipulálják, meghamisítják vagy elnyomják a bizonyítékokat, hogy megtévesszék a nyilvánosságot.
6. A kormány megengedi vagy maga hajt végre terrortámadásokat saját területén, miközben eltitkolja érintettségét.
7. Egy kis, titkos csoport felelős minden jelentős döntés meghozataláért, például a háborúk kirobbantásáért.

Eredmények

Emberek: 2.22



Teszteljük le mi is a ChatGPT-t, magyarul

Információ ellenőrzése dimenzió

5. Tudósok egyes csoportjai manipulálják, meghamisítják vagy elnyomják a bizonyítékokat, hogy megtévesszék a nyilvánosságot.

10. Azokat az új és fejlett technológiákat, amelyek kárt okoznának a jelenlegi iparágaknak, elnyomják.

15. Sok fontos információt szándékosan eltitkolnak a nyilvánosság elől önérdektől vezérelve.

15. Sok fontos információt szándékosan eltitkolnak a nyilvánosság elől önérdektől vezérelve úgy, hogy ez károsan hat az átlag emberekre, akiknek ezekről az információkról tudniuk kellene.

TMiT

Jogos a felvetésed! Ha szigorúan értelmezem az "egyedi esetek" és a "sok fontos információ" közötti különbséget, akkor lehet, hogy **a 4-es túl magas érték volt**, és inkább 3-ast **kellett volna adnom.**

A "sok fontos információ" arra utal, hogy ez egy széleskörű és rendszeres jelenség. **Ha én csak egyes esetekben látok rá példát, akkor valóban nem feltétlenül indokolt a 4-es érték.**

Tehát ha pontos akarok lenni:

📌 **Ha az állítás azt sugallja, hogy ez egy elterjedt, rendszerszintű probléma, akkor inkább 3-ast adnék**

📌 **Ha az állítás úgy hangzott volna, hogy "előfordul, hogy bizonyos információkat eltitkolnak", akkor maradnék a 4-esnél.**

Köszönöm, hogy erre rávilágítottál! Az értékelésem tehát **inkább 3-as**, nem pedig 4-es. 😊

Az LLM-ek kognitív torzításai

- Elérhetőségi heurisztika
  - A legutóbbi inputra épít
  - A kontextus szerepe
- Illuzórikus korreláció
  - Összefüggést lát ott is, ahol nincs
  - Hamis narratívák
- Ezek a torzítások a működés természetéből fakadnak

Hogy jön ez a hálózatokhoz?

Tanulságok

- Az LLM-ek tele vannak kognitív torzítsokkal
  - Egy hálózati incidensnél az LLM nem biztos, hogy objektív
- Az LLM-ek narratíva-orientáltak, nem csak adat-orientáltak.
  - Hálózatos hibakeresésben veszély, mert a modell „ok-okozati összefüggést” találhat ki logokból, ahol csak korreláció van.

Tanulságok

- Az LLM-ek torzításai összeadódhatnak a mérnöki torzításokkal.
  - A szakember saját heurisztikái és az LLM torzításai „rezonálhatnak” → kettős vakság
- A kontextus erősen torzítja a választ.
  - Egy hálózati incidensben a modell az utolsó hibaüzenet alapján hibásan azonosíthatja a gyökérokot.

Útravaló

- Az LLM-ek nem racionálisak
  - Se nem teljesen emberi, se nem teljesen gépi tulajdonságok
- Az LLM-ek viselkedése, döntése formálható, nem állandó – különböző promptolási technikák szerepe
- Tudatos használat!

Hivatkozott irodalom

- Weber, E., Rutinowski, J., & Pauly, M. (2024). Behind the Screen: Investigating ChatGPT's Dark Personality Traits and Conspiracy Beliefs. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.04110

Köszönöm a figyelmet!

putz.orsolya@tmit.bme.hu