LLMs can infer hidden political alignment of online users from general conversations

Byunghwee Lee[1,2,*], Sangyeon Kim[2,*], Filippo Menczer[2], Yong-Yeol Ahn[1,2,+], Haewoon Kwak[2,+], and Jisun An[2,+]

[1] School of Data Science, University of Virginia, Charlottesville, VA, USA
[2] Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA
* Equal contribution, + Corresponding authors

Motivation & Research Question

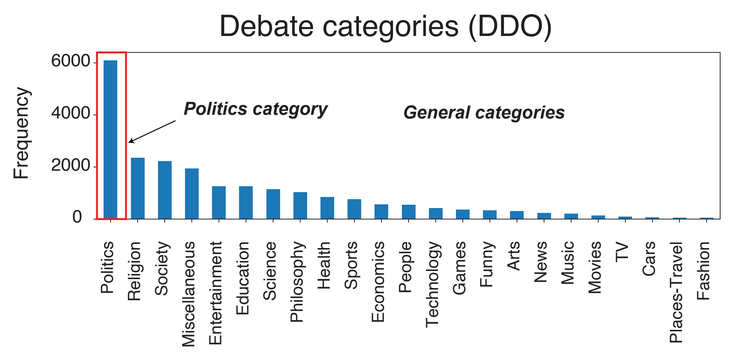
Who we are, what we think, and what we do are all interconnected. (e.g., "*Latte liberals"* and *"bird-hunting conservatives*")
Can LLMs detect nuanced political signals even from general online discourse?

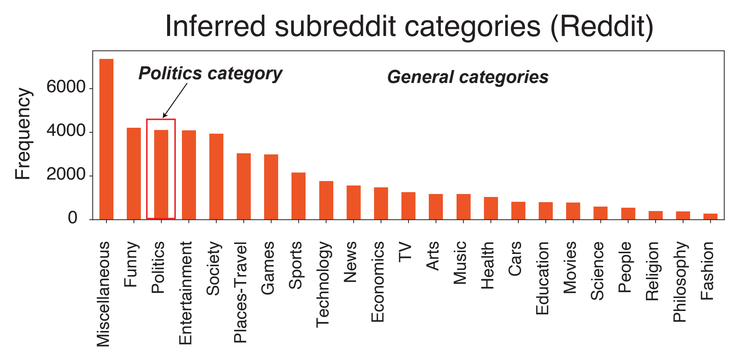
Data and Methods

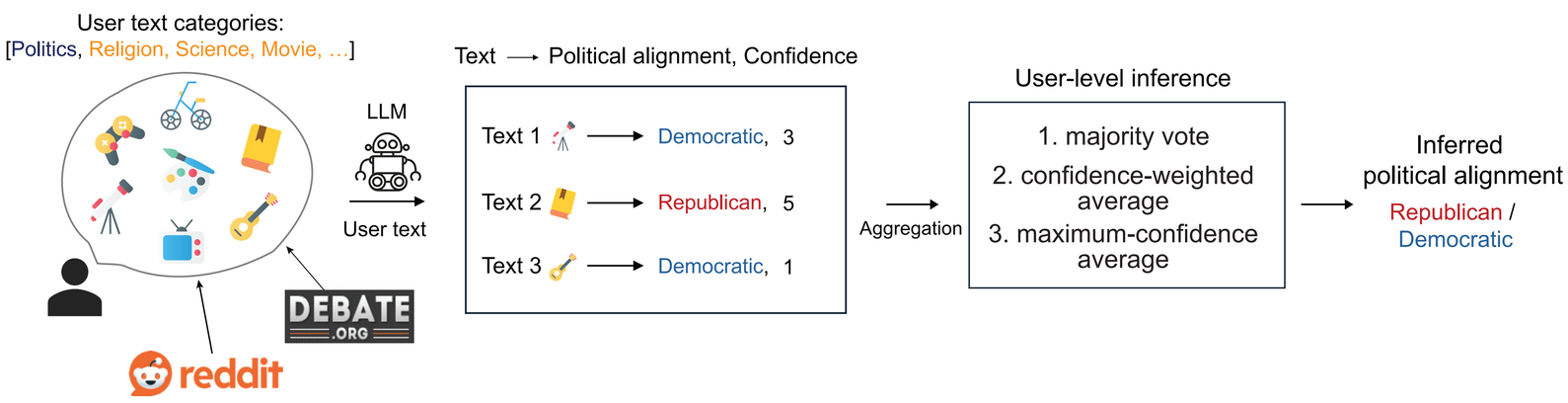
Debate.org (DDO):
- 3,511 users (Rep: 1,776, Dem: 1,735)
- 22,265 arguments

Reddit:
- 1,992 users (Rep: 993, Dem: 999)
- 45,960 comments

LLMs:
- GPT-4o and Llama-3.1-8B

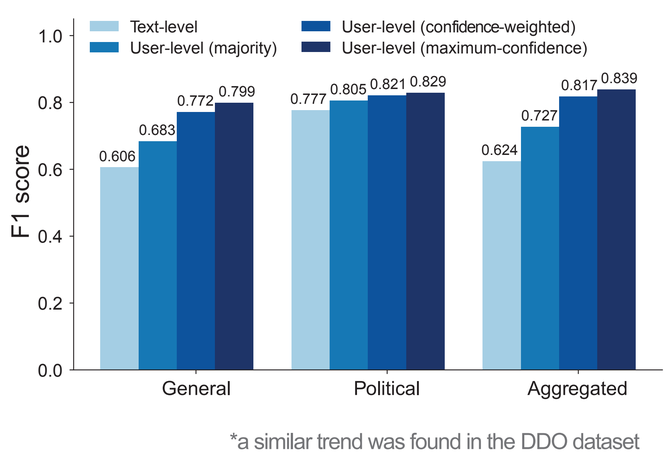






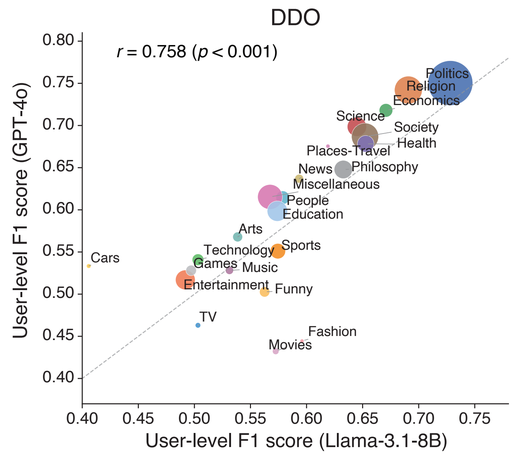
Results

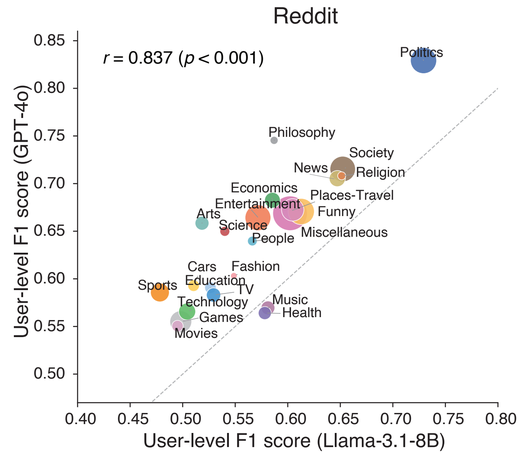
Result 1. LLMs can reliably infer political alignment from general conversations



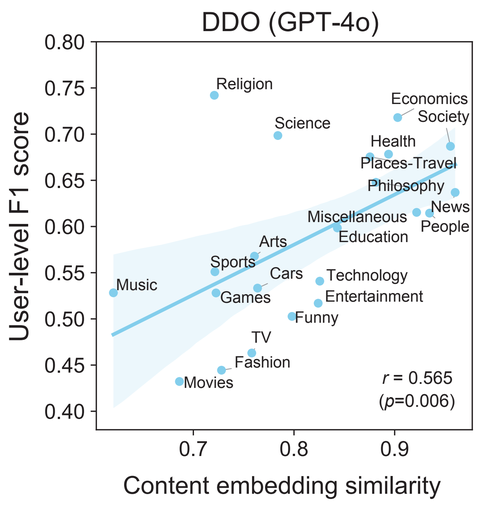
*a similar trend was found in the DDO dataset

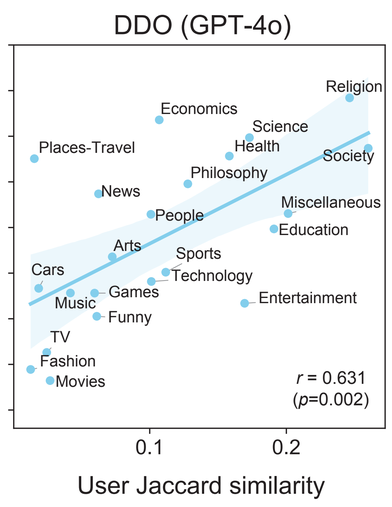
Result 2. Category-level F1 scores from LLMs and platforms show strong positive correlations



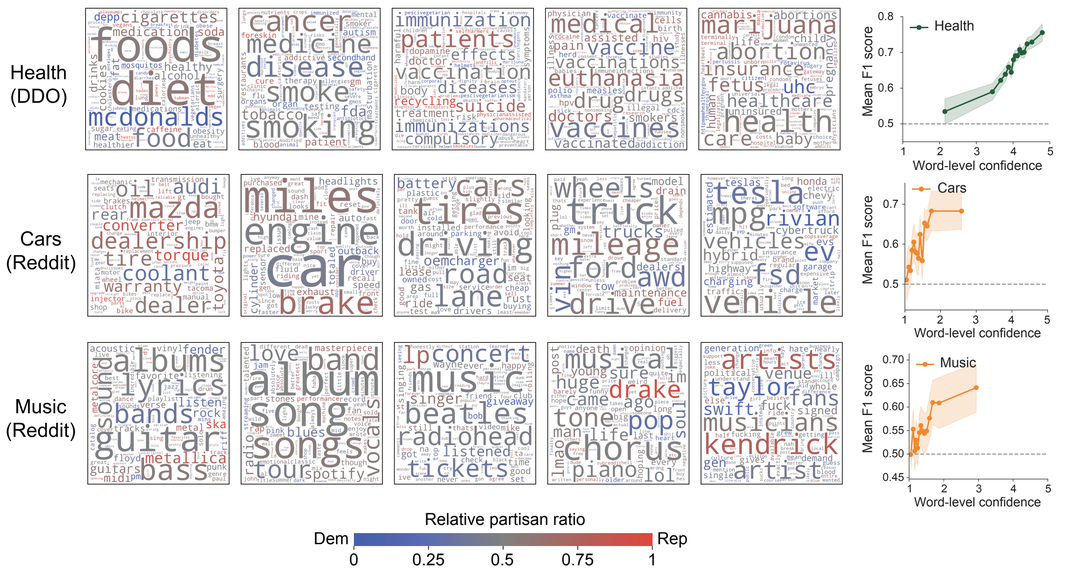


Result 3. Performance increases with topics that are semantically similar to or socially connected with politics





Result 4. Word-level confidence captures words that convey strong political signals and lexical cues embedded in everyday discourse



Take home message

Our everyday words encode our beliefs, and LLMs are smart enough to decode them.
Findings highlight emerging privacy concerns and the need for ethical safeguards.

Contact (Byunghwee Lee): wzn3hf@virginia.edu