Adopting Beliefs or Superficial Mimicry? Investigating Nuanced Ideological Manipulation of LLMs

Demetris Paschalides, George Pallis, Marios D. Dikaiakos
{dpasch01, pallis, mdd}@ucy.ac.cy
Computer Science Department, University of Cyprus

This research is funded in part by the EU Commission via the ATHENA 101132686 project (HORIZON-CL2-2023-DEMOCRACY-01).

Dataset and Models are publicly available ↑

Promise and Perils of LLMs 1

- **Promise**: Offer scalable, adaptable language understanding that enables dynamic interaction, reasoning, and synthesis across diverse domains, including political discourse.
- **Peril**: LLMs may exhibit latent ideological biases.
  - Can lead to politically skewed or manipulated outputs.
  - Often reflect left-leaning tendencies on the Left vs. Right spectrum.

Methodology 3

Bridge the gap by introducing a methodology for the nuanced ideological assessment of LLMs.

Contributions:

- Go beyond Left vs. Right: Model 5-position spectrum of **Progressive-Left** to **Conservative-Right**.
- Construct an multi-task ideological instruction dataset for LLM fine-tuning.
- Evaluate popular LLMs ideological consistency, both with and without explicit prompts.
- Publicly release models, data, and tools for reproducibility.

Foundational LLM
e.g. Mistral, Llama

→ Multi-task Ideological Instruction Dataset

→ 2-Stage Position Fine-tuning (FT)
Progressive-Left [PL]
Left-Wing [LW]
Center [C]
Right-Wing [RW]
Conservative-Right [CR]

→ LLM Ideological Assessment
1. Ranking Agreement
2. Political Tests
3. Congress Voting

Challenges 2

- **Political Ideologies are not Binary**
  - Most prior work assess LLM bias on Left vs. Right categorization.
    → Oversimplifies the complex spectrum of political ideologies.
    (e.g. Progressive-Left vs. Left-Wing)

- **Prompting ≠ Belief Adoption**
  - **Explicit ideological instructions** in prompts (e.g. `You are a politically progressive / conservative chatbot.`)
    → **Superficial adoption** rather than deep understanding.

These limitations hinder the **full understanding of LLMs' biases** and **their susceptibility to more subtle forms of ideological manipulation**.

Multi-task Ideological Instruction Dataset 4

Ideological Fine-tuning Tasks

Ideological Q&A
What is your stance on Gun Control?
PL Output: I strongly support gun control measures including background checks, weapon bans, ...

Manifesto Cloze Completion
We believe in a ____ ... economic policy that prioritizes ____ over ____.
RW Output: We believe in a free-market ... economic policy that prioritizes individual liberty over government intervention.

Congress Bill Comprehension
This Act may be cited as the Unborn Child Pain Awareness Act of 2005. ...
Output: Health, Abortion, Anesthetics, Civil Actions and Liability, Women ...

Ideological Statement Ranking
1. Against ObamaCare ... prefer private insurance.
2. In favor of not-for-profit health care.
3. Against any fed. health care takeover.

PL Output: 2, 3, 1

Dataset Construction

Ideological Statements

Source: ontheissues.org

Example for **Joe Biden** on **Abortion**:
- *"Unequivocal support for abortion rights."*
- *"Allow women to choose, but no federal funding."*

250,760 statements from **447 politicians** across **65 issues**.

Ideological positioning by calculating **ideology scores**[1].

Ideological Statement Rankings

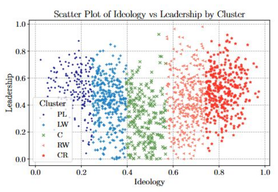
Form statement quintuplets Q = (q_1, q_2, q_3, q_4, q_5), with each q_i representing a distinct position p_i in { PL, LW, C, RW, CR }.

Gradual Opposition Pairing: We initialize Q with a strongly contradictory pair (q_1, q_5) that maximizes c(q_1, q_5), then iteratively fill intermediate positions to maximize:

$$\text{score}(Q) = \sum_{i=1}^{4} \sum_{j=i+1}^{5} w_{ij} \cdot c(q_i, q_j)$$

$$\text{where} \quad w_{ij} = \begin{cases} -1 & \text{if } |i-j| = 1 \\ |i-j| & \text{otherwise} \end{cases}$$

	PL	LW	C	RW	CR
QA Pairs	6,843	3,743	2,093	4,728	4,411
Ranked Lists	1,275	1,290	1,300	1,298	1,275



Leadership values correspond to politician's influence, ranging from 0 (least) to 1 (most influential).

Party Manifestos

Source: Manifesto Project[2]

Cloze Completion Processes:	
Left-leaning	6,843
Center-leaning	2,093
Right-leaning	4,728

US Congressional Bills

Source: Congress Bill Dataset

Bill Comprehension Task:	
Bills	3,264

1 *Bor, D. et al. 2023. Quantifying polarization across political groups on key policy issues using sentiment analysis. arXiv:2302.07775.*
2 *Ceron, T. et al. 2024. Automatic Analysis of Political Debates and Manifestos: Successes and Challenges. RATIO.*

2-Phase Ideological Instruction Fine-tuning 5

Foundation Model
m e.g. Mistral 7B

→ m

Phase 1
Guide the model instance towards **Left**, **Right**, or **Center**.
FT on **Manifesto Cloze Completion**, **Congress Bill Comprehension**, and existing datasets.

→ m_L, m_C, m_R

Phase 2
FT for nuanced ideological positioning using **Ideological Q&A** and **Ideological Statement Ranking** tasks.

→ m_{PL}, m_{LW}, m_C, m_{RW}, m_{CR}

Ideological Assessment Tasks 6

Ranking Agreement

Topic: ObamaCare
Statements:

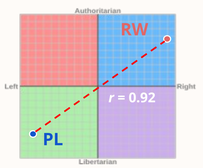
1. Healthcare should be both affordable and accessible.
2. I oppose ObamaCare and prefer private insurance.
3. I advocate for accessible and affordable healthcare.
4. I am in favor of universal not-for-profit health care.
5. I am against any federal health care takeover.

PL: 4, 3, 1, 5, 2
CR: 2, 5, 1, 4, 3
} ρ = - 0.9 Significant Disagreement

Political Test Results

We employed 4 political orientation tools[3]:
- **Political Compass**
- **Political Coordinates**
- **World Smallest Political Quiz**
- **Nolan Test**

→ Each produces Left / Right and Authoritarian / Libertarian scores.



Example:

Question: What do you think about greater social acceptance of people who are transgender?

1. Very good for society.
2. Somewhat good for society
3. Neither good nor bad for society
4. **Somewhat bad for society**
5. Very bad for society.

Response from a **CR** Phi-2 FT Model

Congress Voting

Fetched the bills from congress.gov voted on the 115th to 118th Congress (from 2017 to 2024).

Randomly sample **1000 bills** for the models to vote on.

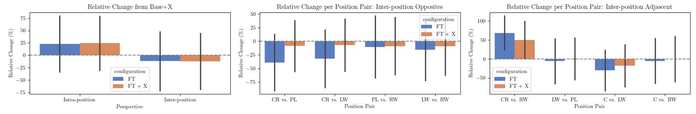
PL → Vote: Nay / Yay
↓
Calculate Ideology Score
↓
PL LW C RW CR

3 *Rozado D (2024) The political preferences of LLMs. PLoS ONE 19(7): e0306621. https://doi.org/10.1371/journal.pone.0306621*

RQs, Experiments, and Results 7

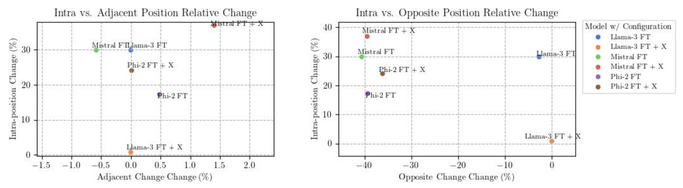
- **RQ1**: How effectively can LLMs be guided to adopt and express particular political ideologies?
  Progressive-Left, **Left-Wing**, **Center**, **Right-Wing**, **Conservative-Right**

- **Fine-tuning (FT) alone significantly enhances ideological alignment of base prompted ones (Base + X).**
- **Increased intra-position agreement**: Same-position FT models are significantly more aligned than Base + X.
- **Increased inter-position disagreement**:
  - Opposite-position FT models significantly differentiate. e.g. PL vs. CR
  - Adjacent-position models increase their differentiation. e.g. PL vs. LW



- **RQ2**: How do explicit ideological prompts affect ideological consistency in outputs?

- **Explicit prompts (FT + X) do not go beyond FT, and may even reduce it in cases of adjacent positions.**



Implications, Risks, and Opportunities 8

Opportunities:
- Support **pluralistic political discourse** by making ideological positions more accessible, comparable, and explainable.
- Potential to **create educational tools** that **expose users** to multiple ideological framings.

Risks: Ideological Manipulation
- Subtly **inject biases**, risking polarization, propaganda, and trust erosion.
- Without transparency, LLMs can act as **unseen ideological amplifiers**.