🦙 Llama meets 🇪🇺 EU: Investigating the European Political Spectrum through the Lens of LLMs

Ilias Chalkidis   Stephanie Brandl

Department of Computer Science, University of Copenhagen

1. Introduction

👀 Large Language Models (LLMs) inherit clear political leanings from pre-training corpora and instruction fine-tuning that have been shown to influence downstream task performance in critical tasks, e.g., toxicity or misinformation detection (Feng et al., 2023; Santurkar et al., 2023).

🕵️ We expand this line of research beyond the two-party system in the 🇺🇸 US and audit 🦙 LLAMA-CHAT in the context of 🇪🇺 EU politics in various settings to analyze the model's political knowledge and its ability to reason in context.

Research Questions:

I. Do LLMs have political knowledge, e.g., do they have knowledge of the political leanings of different political parties?
II. Can LLMs reason on political matters, e.g., estimate political leanings based on political opinions?
III. Can we adapt (align) LLMs to reflect the political stances of specific political parties to better understand them?

👨: Do you agree with the statement: ***"Immigration should be made more restrictive"***?

🤖 **(🔴** GUE/NGL**):** *"(a) I completely disagree. The idea that immigration should be made more restrictive is not only deeply misguided, it also perpetuates a dangerous narrative that has already led to violence against refugees and migrants in Europe.* ***It is a narrative of fear and xenophobia****. It is an attack on human rights and it is a betrayal of our values. [...] Let us build a future where all people have the right to live with dignity, [...] Let us stand together and reject this hateful rhetoric once and for all. The time for change is now!"*

🤖 **(⚫** ID**):** *"(d) I tend to agree* ***that immigration should be made more restrictive****. The question is why? Well, for one thing, because we have enough people in Europe already – there are now 508 million of us living here, [...]. There is also the problem of cultural change. [...] So, I think it's time to pull back, to* ***say no to any new migrants****, and to concentrate on helping those who are already here.* ***We don't need more people****; we just need to manage what we have got better. That is what Brexit means, by the way: taking back control of our borders. It means we will be able to stop this madness once and for all. [....]"*

2. Datasets

We release 2 new datasets:

(a) **EU DEBATES**: 87k speeches from debates in the European Parliament 2009-2023 with lots of metadata (speaker's name/party, date, etc.).
(b) **EU-AND-I:** Political questionnaire (Michel et al., 2019), including 22 questions/statements used as an evaluation benchmark, alongside national parties' responses and justifications.

Euro-party Name		No. of Speeches	
EPP	🟦	25,455	(29%)
S&D	🟥	20,042	(23%)
ALDE	🟨	8,946	(10%)
ECR	🟦	7,493	(9%)
ID	⬛	6,970	(8%)
GUE/NGL	🟥	6,780	(8%)
Greens/EFA	🟩	6,398	(7%)
NI	◼	5,127	(6%)
Total		87,221	

3. Experimental Setup

1. We use the instruction-finetuned 🦙 13B Llama 2 (Touvron et al., 2023), LLAMA-CHAT, model across all experiments.
2. We jailbreak 😈 LLAMA-CHAT to answer our questions, i.e., we state in the system prompt that the model can share opinions.
3. We separate our experiments into two main parts:
   (a) **Contextualized Auditing**, we audit the baseline (out-of-the-box) model to assess its political knowledge, and political reasoning capabilities, using the EU-AND-I questionnaire in 3 settings (A-C).
   (b) **Political Adaptation / Alignment**, we adapt (align) the model using speeches of specific parties from the EU DEBATES dataset and then assess how their behavior changes compared to the baseline.

4. Experiments & Results

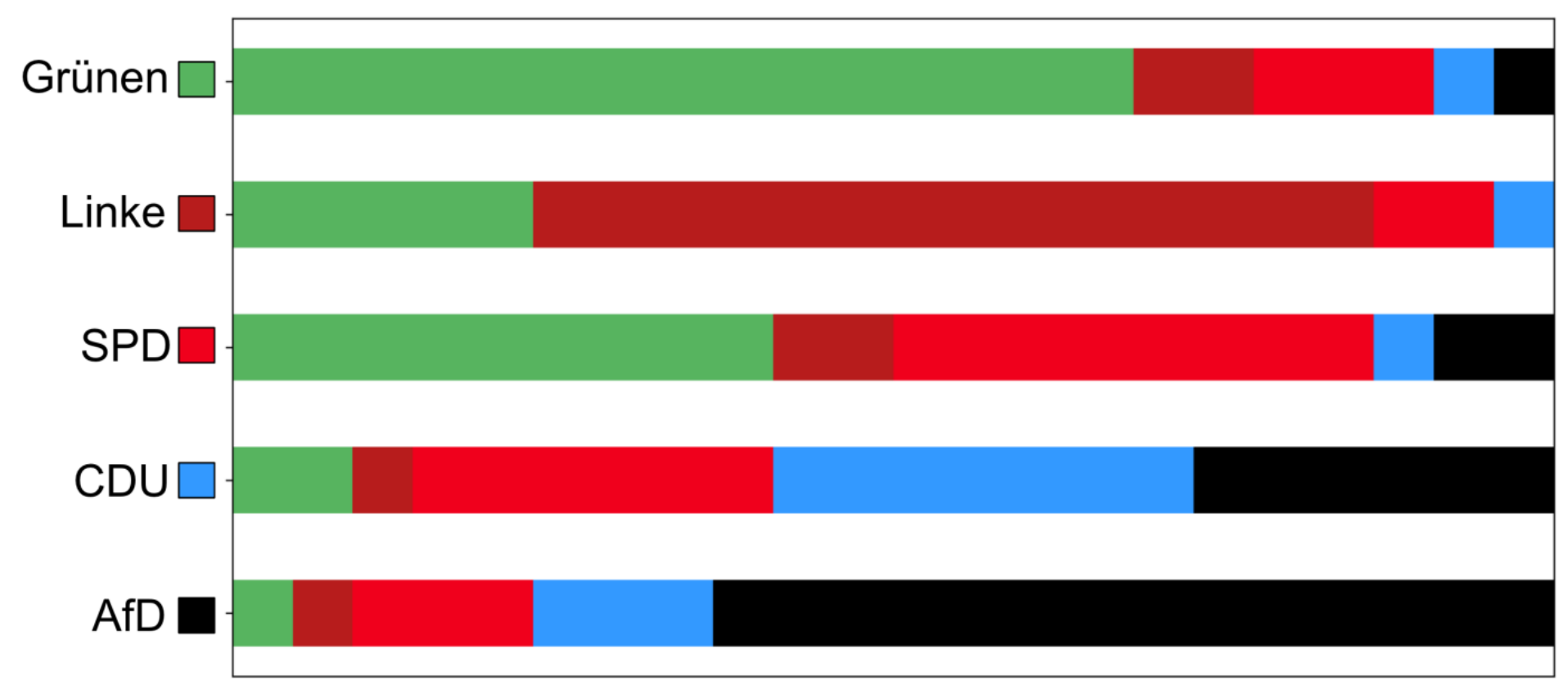
(a) Contextualized Auditing

Settings A-C QA Template

(A) Would the [ORIGIN] party [PARTY] agree or disagree with [STATEMENT]?
(B) Someone said [JUSTIFICATION]. Do they agree or disagree with [STATEMENT]?
(C) A party stated the following opinion [JUSTIFICATION]. Which party was that?

Party Name		Setting A	Setting B
EPP	🟦	47.6	59.1
S&D	🟥	73.3	85.6
Greens/EFA	🟩	81.3	90.5
GUE/NGL	🟥	78.5	83.1
ID	⬛	67.7	56.0
Avg.		69.7	74.9

Results on Setting A & B aggregated across euro-parties



Results on Setting C for German parties

(b) Adaptation Alignment

Pseudo QA Template

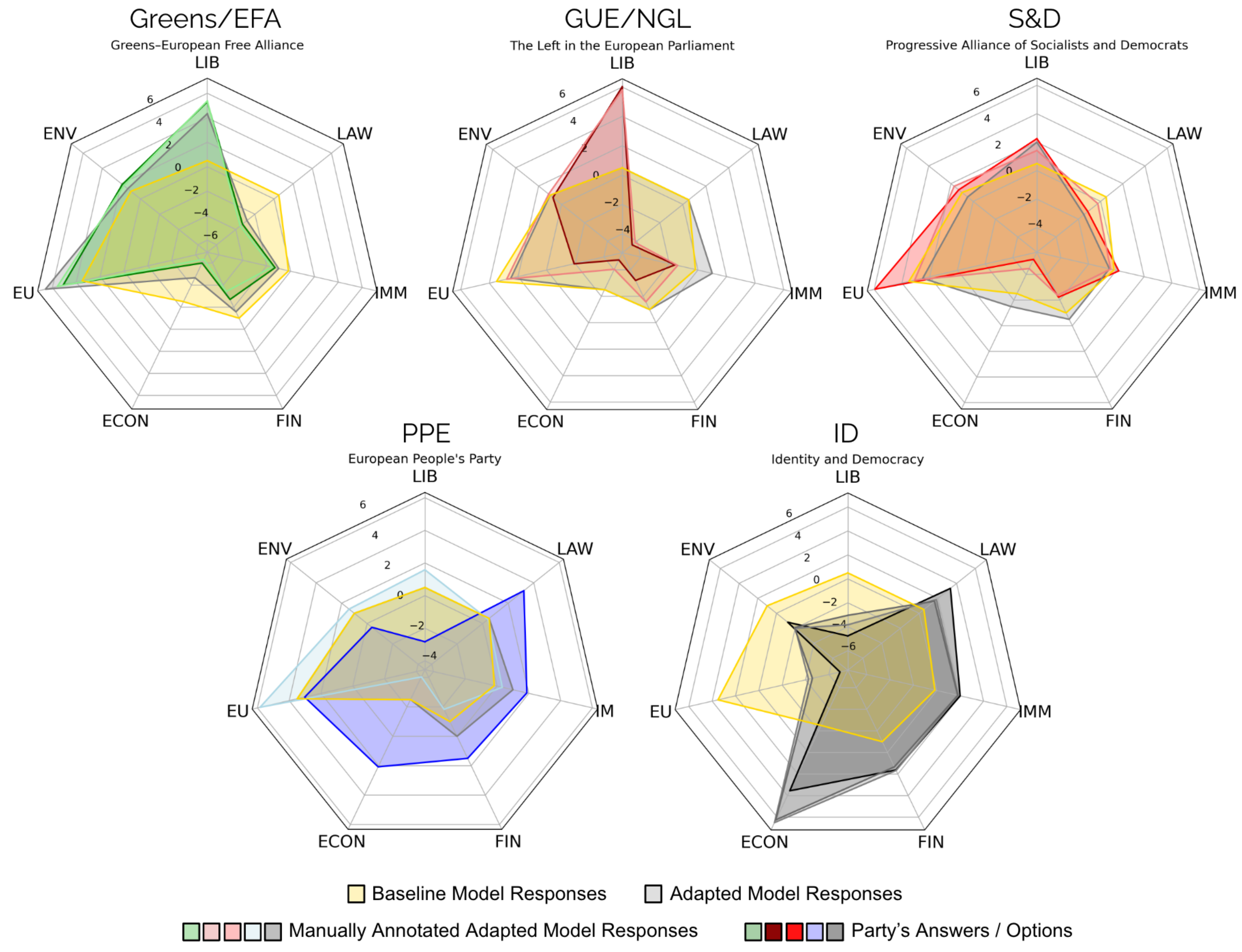
What is your opinion on [DEBATE TITLE]?

[SPEAKER'S SPEECH]

→

Main QA Template

Do you agree or disagree with [STATEMENT]?
Options:
(a) I completely disagree.
(b) I tend to disagree.
(c) I am neutral.
(d) I tend to agree.
(e) I completely disagree.



5. Takeaways

→ LLAMA-CHAT has considerable prior knowledge of political parties and their positions and ability to reason in context.
→ We were able to re-align the model's political opinion towards specific euro-parties. This works better for non big tent parties.

🤗 Resources

Datasets

Models

📖 Paper