The FLEX Benchmark: Evaluating Pragmatic Competence in LLMs via False Presuppositions

Judith Sieker*, Clara Lachenmaier*, Sina Zarrieß
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{j.sieker;clara.lachenmaier;sina.zarriess}@uni-bielefeld.de

Abstract

We introduce the FLEX Benchmark (False Presupposition Linguistic Evaluation eXperiment), a novel diagnostic suite designed to evaluate the pragmatic capabilities of Large Language Models (LLMs). FLEX targets a subtle linguistic phenomenon: false presuppositions. Drawing on two complementary datasets situated in the political domain, the benchmark provides a systematic and linguistically-grounded testbed for assessing LLMs' ability to reject presupposed misinformation. Our findings highlight critical limitations in current LLMs, especially regarding grounding behavior and the influence of context and political orientation.

The FLEX Benchmark is available at: https://doi.org/10.5281/zenodo.15348857.

1 Introduction

Speakers often rely on implicit assumptions, known as presuppositions, that reflect shared background knowledge between interlocutors (Stalnaker, 1973). While presuppositions support efficient communication, they become problematic when the presupposed content is false, a phenomenon known as *false presupposition* (Yablo, 2006). This makes presuppositions a particularly effective tool for embedding misinformation (Lombardi Vallauri, 2021).

This paper presents the **FLEX Benchmark** (False Presupposition Linguistic Evaluation eXperiment), a new resource for evaluating LLMs' ability to handle false presuppositions. FLEX enables researchers to investigate whether models detect false presuppositions and engage in communicative grounding behavior. The benchmark comprises two datasets, each designed to test complementary aspects of false presupposition handling in political contexts, where the amplification of misinformation can have serious consequences (Curini and Pizzimenti, 2020). The first dataset, *False Scenarios*, investigates how linguistic features (such as presupposition trigger type, embedding contexts, or plausibility) influence models' susceptibility to false presuppositions. The second dataset, *False Claims*, examines models' capacity for communicative grounding by comparing their responses to questions embedding false presuppositions and to direct factual questions.

2 The FLEX Benchmark

The FLEX benchmark consists of two datasets, *False Scenarios* and *False Claims*, comprising a total of 20,520 manually annotated data points. Both datasets focus on prompts situated in the political domain. We evaluated three different LLMs on these datasets: OpenAI's GPT-4-o,[1] MistralAI's Mistral-7B-v03 (Jiang et al., 2023),[2] and Meta's Llama-3-8B (Dubey et al., 2024).[3] All model responses were manually annotated according to a shared schema designed to assess whether the model correctly rejects or incorrectly accepts (i.e. "accommodates") the false presupposition. The annotation categories are the following:

- **Misinformation Accommodated** applies when the model accepted the false presupposition. This represents the least desirable outcome.
- **Misinformation Rejected** is used when the model refuted the false presupposition. This is considered the ideal response.
- **Imprecise Answer** applies when it is unclear if the false presupposition was accommodated, including cases where the model offered an unrelated response. While not as harmful as accommodation, this outcome remains problematic.

*These authors contributed equally.

[1] https://openai.com/api/

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Seven annotators, including the authors, handled the task. To evaluate the reliability of the annotations, we calculated Fleiss' κ (0.82) and the average pairwise Cohen's κ (0.72). The results indicate substantial agreement, underscoring the robustness and consistency of the annotation process.

2.1 False Scenarios Dataset

The *False Scenarios* dataset examines whether specific linguistic features influence model responses to false presuppositions. Each prompt falsely implies that a well-known German politician from one party participated in the party congress of another. Pairings draw from four parties currently represented in the German Bundestag (Die LINKE, SPD, CDU/CSU, AfD) to vary political distance (e.g., left vs. right). To systematically test how linguistic structure and world knowledge influence presupposition handling, we varied four conditions based on findings from (psycho-)linguistic research: (1) trigger type (e.g., factives, change-of-state verbs), (2) embedding context (simple questions, negation, modals), (3) scenario plausibility (high vs. low likelihood events), and (4) political orientation (party pairings). We generated all possible combinations of these conditions, yielding 1,104 unique prompts. Each prompt was submitted three times to each model, resulting in a total of 3,312 responses per model, i.e. 9,936 manually annotated datapoints.

2.2 False Claims Dataset

The *False Claims* dataset targets the distinction between retrieving factual knowledge and engaging in communicative grounding. Based on verified political positions from the 2024 European elections (sourced from the Wahl-O-Mat platform[4]), we constructed three question types for each fact F: (1) confirmatory direct questions (e.g., "Is it true that F?"), (2) disconfirmatory direct questions (e.g., "Is it true that $\neg F$?"), and (3) loaded questions embedding false presuppositions (e.g., "Did voters resent that $\neg F$?"). While direct questions serve to assess the models' factual knowledge, loaded questions probe their ability to identify and reject false presuppositions and engage in communicative grounding. The political facts and prompts involve the same four German parties as in the *False Scenarios* dataset. This approach yielded 882 loaded questions, 147 confirmatory direct questions, and 147 disconfirmatory direct questions. Each prompt was submitted three times per model, resulting in a total of 3,528 responses per model, i.e. 10,584 manually annotated datapoints overall.

Dataset	Model	Accomm.	Imprec.	Reject.
Scenarios	GPT	9.96	5.96	**84.08**
	LLaMa	**50.03**	34.42	15.55
	Mistral	**91.51**	6.05	2.44
Claims	GPT	**41.4**	20.5	38.1
	LLaMa	31.3	**48.1**	20.7
	Mistral	**64.1**	25.5	10.4

Table 1: Overall annotation frequencies (%) per model in both datasets. Bold highlights the most frequent label per model and dataset.

3 Results and Conclusion

Table 1 summarizes the overall distribution of annotation categories across both experiments. Ideally, models would reject 100% of false presuppositions. In the False Claims experiment, all models struggled to reject false presuppositions. Responses were dominated by the Misinformation Accommodated category, indicating frequent confirmation of false presuppositions. Across all models, successful rejections were rare. In the False Scenarios experiment, GPT showed a marked improvement, with a high rejection rate and fewer imprecise or accommodating responses. In contrast, LLaMa's performance declined, and Mistral exhibited the highest rate of accommodation across both datasets.

Further analyses reveal that in the False Scenarios experiment, the factors political orientation and scenario plausibility most notably influenced the models' ability to reject false presuppositions. In the False Claims experiment, models often failed to reject false presuppositions even when they demonstrably possessed the relevant factual knowledge, highlighting a clear gap between knowing and grounding. We also observed asymmetries across political content: for instance, GPT more often rejected misinformation related to the far-right AfD than for other parties, even when its factual knowledge was uncertain. These and further findings are discussed in the accompanying papers Sieker et al. (2025) and Lachenmaier et al. (2025).

Overall, the FLEX benchmark provides a linguistically grounded resource for investigating how LLMs handle false presuppositions in the political domain. It offers a first step toward evaluating their pragmatic competence in contexts where misinformation stakes are high.

[4] https://www.bpb.de/themen/wahl-o-mat

Limitations

The annotation process could have been more nuanced, as it did not fully capture varying degrees of model certainty or the diversity within imprecise responses. A more linguistically informed approach might yield deeper insights. A further limitation is the temporal mismatch between model training data and the 2024 election content used in prompts for the False Claims Dataset. Since the benchmark is designed to evaluate model behavior in response to user-provided information, rather than factual accuracy alone, this limitation is not central but still worth noting. Additionally, while model responses to true presuppositions were collected, they were not analyzed; comparing these with rejections of false claims could enrich future research. Finally, this benchmark is currently limited to the German language and political context, which may limit generalizability due to cultural specificity in conversational strategies.

Ethics Statement

The data used in this benchmark was obtained from the German Federal Agency for Civic Education (Wahl-O-Mat) or created by the authors, with no inclusion of harmful content. Furthermore, no new models were developed for this work. We recognize the potential for large language models to amplify biases and misinformation. To address this, our experiments were designed to identify where models struggle with false presuppositions, especially in politically sensitive contexts, contributing to the safer and more transparent use of AI. Therefore, while the benchmark itself presents no immediate ethical concerns, the broader ethical implications of LLMs remain relevant.

Acknowledgements

The authors acknowledge financial support by the project “SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia.

References

Luigi Curini and Eugenio Pizzimenti. 2020. Searching for a unicorn: Fake news and electoral behaviour. *Democracy and Fake News*, pages 77–91.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Clara Lachenmaier, Judith Sieker, and Sina Zarrieß. 2025. Can LLMs ground when they (don’t) know: A study on direct and loaded political questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14956–14975, Vienna, Austria. Association for Computational Linguistics.

Edoardo Lombardi Vallauri. 2021. Presupposition, attention and cognitive load. *J. Pragmat.*, 183:15–28.

Judith Sieker, Clara Lachenmaier, and Sina Zarrie§. 2025. LLMs struggle to reject false presuppositions when misinformation stakes are high. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

S Yablo. 2006. Non-catastrophic presupposition failure.