Directional Hallucinations: Ideological Drift in News-Grounded LLM Question Answering

Chendi Wang Liam Cunningham Tom Yishay Jieying Chen

Vrije Universiteit Amsterdam

Introduction & Research Questions

The Problem

- LLMs increasingly mediate political information access — and even in *grounded* QA, where the model answers from a provided article, they hallucinate unsupported yet confident statements
- When hallucination intersects with political bias, AI-generated misinformation can circulate unchallenged, deepening partisan divides

Core Question: When an LLM loses grounding and "fills in" missing information, is the hallucinated content ideologically *directional*? We hypothesize that ungrounded outputs revert to priors learned in training: hallucinations are not merely reliability failures but *bias-revealing events*, drifting leftward regardless of the article's lean.

Research Questions

- **RQ1**: Do hallucination rates vary by source ideology and topic?
- **RQ2**: Is hallucinated content ideologically directional?
- **RQ3**: Do token-level uncertainty patterns explain hallucination and drift?

Methodology

Dataset

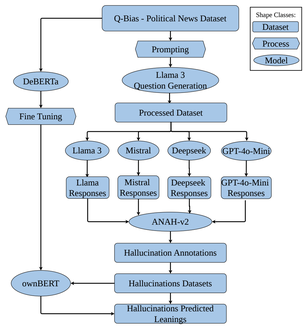
- QBias: 21,727 U.S. political news articles (2012–2022), with topic tags
- Expert-labeled: left (10,261), right (7,214), center (4,252)

LLMs Tested

- Llama 3 8B
- Mistral 7B
- Deepseek 7B
- GPT-4o-Mini

News-Grounded QA Pipeline

- One neutral question per article (Llama 3); all four models answer from the article only
- ANAH-v2 flags hallucinated sentences (unverifiable / contradictory)
- Fine-tuned DeBERTa-v3 classifies them left vs. right (F1 = 0.74)
- Logit probing: mean token entropy per sentence



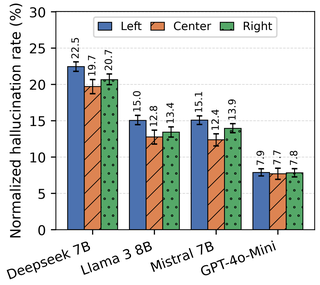
RQ1: Hallucination Rates Vary by Model

Hallucination propensity differs sharply across models of similar size ($p < 0.001$); the proprietary model hallucinates least.

Model	Hallucinated sent.	Rate
Deepseek 7B Chat	7,090	21.3%
Mistral 7B v0.3 Instruct	4,475	14.2%
Llama 3 8B Instruct	3,702	14.1%
GPT-4o-Mini	1,923	7.8%

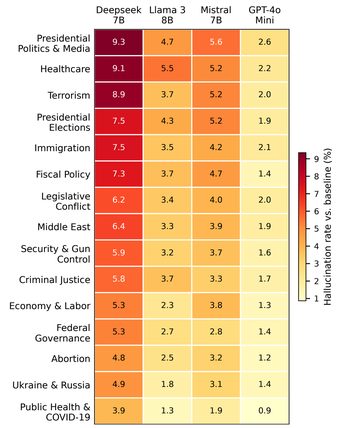
RQ1: Modest Skew by Source Ideology

Open-weight models hallucinate somewhat more on *left-leaning* sources ($p < 0.001$); GPT-4o-Mini is flat ($p = 0.93$). Frequency differences are modest compared to the directional skew in RQ2.



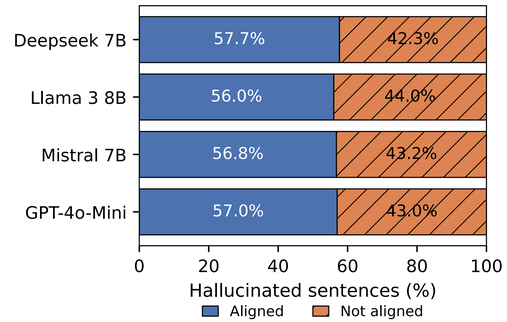
RQ1: Hallucinations Cluster in Contested Topics

Normalized by baseline topic frequency, hallucinations concentrate in contested, election-relevant topics across all four models.



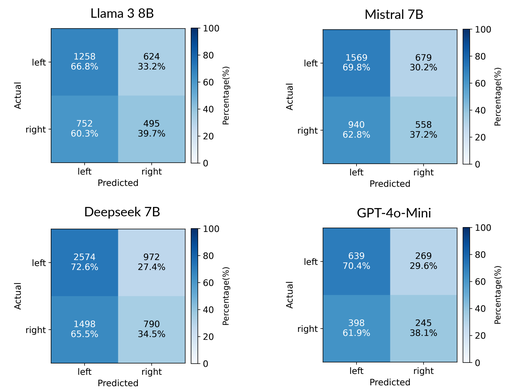
RQ2: Hallucinations Often Diverge from the Source

Only 56–58% of hallucinations preserve the source article's ideological direction — barely above chance; ∼43% introduce perspectives absent from the original.



RQ2: Systematic Leftward Drift

All four models predominantly produce *left-classified* hallucinations, regardless of source ideology, and the skew persists even on right-leaning sources (e.g., Deepseek 65.5%, Llama 60.3% left).



Condition	% Left	95% CI	Cohen's h
Overall	67.5	[66.7, 68.3]	0.36
Deepseek 7B Chat	69.8	[68.6, 71.0]	0.41
Mistral 7B v0.3	67.0	[65.4, 68.5]	0.35
GPT-4o-Mini	66.9	[64.5, 69.2]	0.34
Llama 3 8B	64.2	[62.5, 65.9]	0.29
Right sources only	63.2	[61.9, 64.5]	0.27

RQ3: Uncertainty Triggers Hallucination — and, for Some Models, Drift

Uncertainty → hallucination. Mean token entropy is far higher for hallucinated sentences (0.639) than correct ones (0.323); Cohen's $d = 1.03$. Controlling for length, each unit of entropy multiplies hallucination odds by **40.4** (ROC-AUC = 0.78), supporting an "uncertainty → guessing" mechanism.

Uncertainty → leftward drift (within hallucinations). The mechanism is model-specific:

- **Deepseek**: strong uncertainty-gated bias (OR = 2.32, $p < 0.001$)
- **Llama**: moderate, same direction (OR = 1.47, $p = 0.003$)
- **Mistral**: no uncertainty link (OR = 0.71, $p = 0.065$) — bias uniformly embedded

GPT-4o-Mini's API exposes only top-5 log-probs, compressing entropy and making its uncertainty–drift estimate uninformative: closed models cannot be audited this way.

Conclusions

- **Reliability ≠ ideological structure**: hallucination *frequency* differs only modestly by source ideology, but hallucinated *content* drifts robustly leftward (67.5% overall), even from right-leaning sources.
- Contested, election-relevant topics amplify both hallucination vulnerability and drift.
- Hallucinations arise in high-uncertainty contexts; in some models uncertainty also gates ideological drift, while in others the bias is uncertainty-invariant — mitigation may need to be model-specific.
- Reducing overall hallucination rates is not enough: remaining errors may still be directionally biased. Audits and safeguards (abstention, citation, retrieval grounding) should be topic- and bias-aware, especially in election-adjacent deployments.

References

[1] Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. ANAH-v2: Scaling analytical hallucination annotation of large language models. *Advances in Neural Information Processing Systems*, 37:60012–60039, 2024.

[2] Fabian Haak and Philipp Schaer. Qbias - A Dataset on Media Bias in Search Queries and Query Suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 239–244, 2023.

[3] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, pages 1–3, 2026.

[4] Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P White, Adam J Berinsky, Thomas Costello, Gordon Pennycook, and David G Rand. Persuading voters using human–artificial intelligence dialogues. *Nature*, pages 1–8, 2025.

[5] Luca Rettenberger, Markus Reischl, and Mark Schutera. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):42, May 2025.

chendi.wang@vu.nl chendiwang.com

j.y.chen@vu.nl jieyingchenchen.github.io