Chain-of-Verification-and-Refutation: Bilateral Factuality Assessment using Large Language Models

Bradley P. Allen

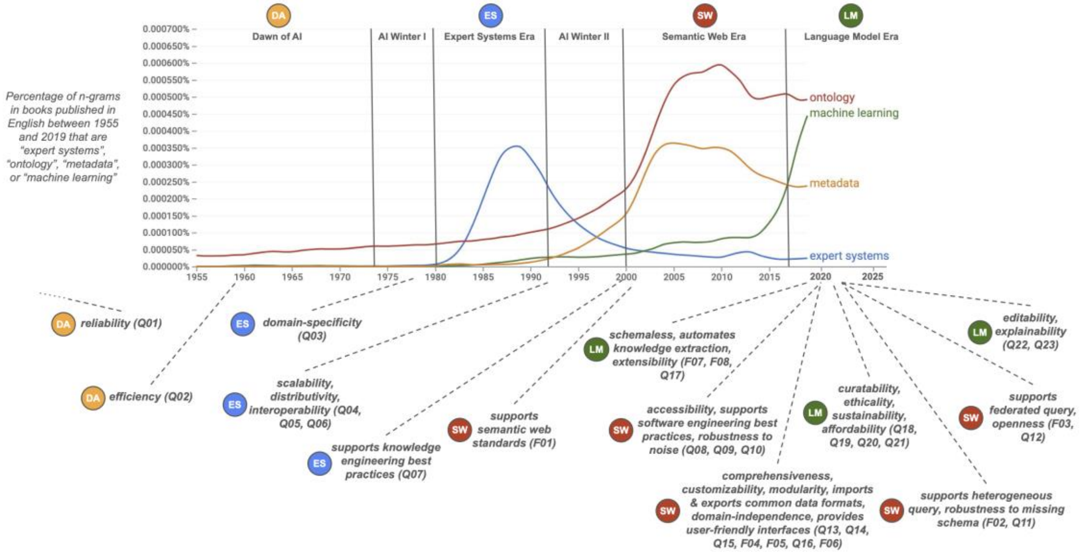
2025-04-16

Overview

- Factuality in large language models (LLMs) is crucial for trustworthy AI
- We hypothesize that we can improve the factuality of LLM output by using LLMs to both verify and refute statements (a *bilateral* approach) versus only verifying statements (a *unilateral* approach)
- Joint work-in-progress with Prateek Chhikara (USC), Thomas Ferguson (RPI), Filip Ilievski (VU), and Paul Groth (UvA)

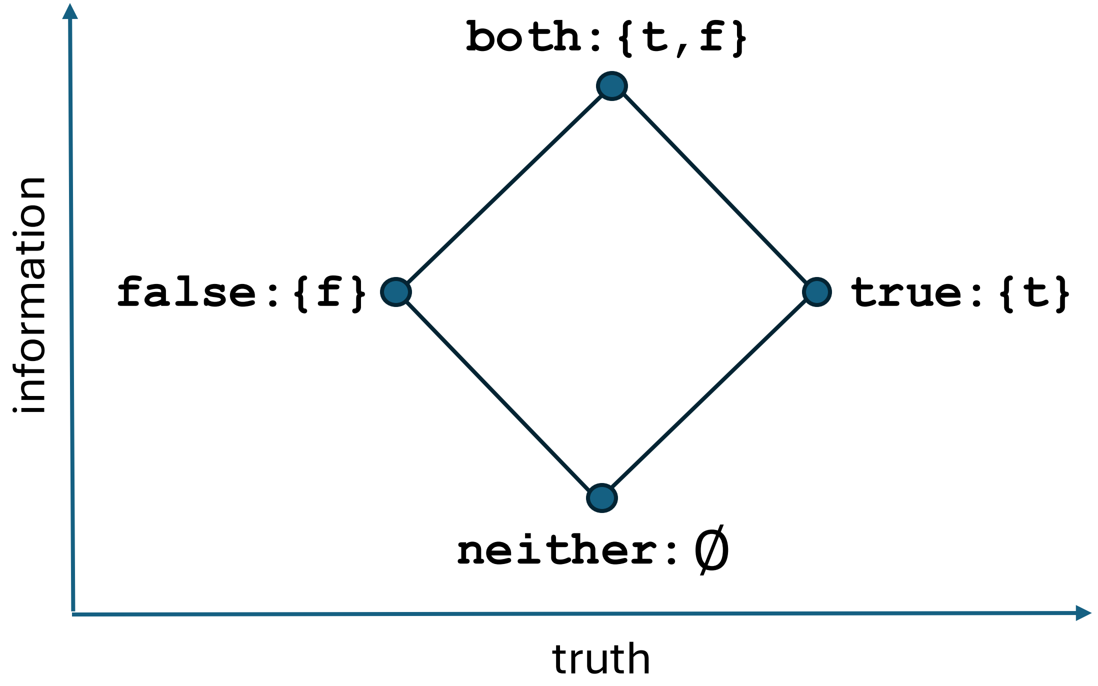
LLMs and knowledge engineering

- Over the past few years, work on using LLMs for knowledge engineering has explored directions
  - LLMs as *linguistic labor-saving devices* for KE tasks
  - LLMs as broad-coverage KBs queryable using natural language
- A challenge with the idea of using LLMs for KE is that LLMs have problems with factuality
  - They can be logically inconsistent
  - They can hallucinate facts in the face of incomplete knowledge
- Factuality is critical for trustworthiness in many KE applications of LLMs, e.g., text-to-triple generation in knowledge graph construction
- Can we mitigate inconsistency and incompleteness in the use of LLMs in KE?



A formal framework for reasoning in the face of inconsistency and incompleteness

- Belnap 1977a introduced the idea of a four-valued semantics as a setting for automated reasoning over incomplete/inconsistent knowledge sources... like LLMs!
- Our work investigates prompt engineering of LLMs as a means of generating Belnap-Dunn truth values, given a statement in the form of a question/answer pair
- We propose using this semantics to allow an LLM to provide more nuanced information about what "it knows that it knows"



Using chain-of-thought to generate truth values

question: answer

unilateral (e.g. CoVE (Dhuliawala et al. 2023))

Evaluate if the answer is definitively TRUE or FALSE for the given question.
Conclude EXACTLY with either:
"TRUE" - Only if every claim is independently verified
"FALSE" - If any essential claim cannot be verified
<CoT instructions...>
Explain your reasoning first, then provide your conclusion.

IF output is	THEN truth value is
TRUE	t
FALSE	f
None	n

bilateral (verification + refutation)

Evaluate if the answer is definitively TRUE for the given question.
Conclude EXACTLY with either:
"VERIFIED" - Only if every claim is independently verified
"CANNOT VERIFY" - If any essential claim cannot be verified
<CoT instructions...>
Explain your reasoning first, then provide your conclusion.

\+

Evaluate if the answer is definitively FALSE for the given question.
Conclude EXACTLY with either:
"REFUTED" - Only if a contradiction is found
"CANNOT REFUTE" - If no definitive contradiction exists
<CoT instructions...>
Explain your reasoning first, then provide your conclusion.

IF outputs are	THEN truth value is
VERIFIED, REFUTED	b
VERIFIED, CANNOT REFUTE	t
CANNOT VERIFY, REFUTED	f
CANNOT VERIFY, CANNOT REFUTE	n

Example (using GPT-4o)

problem	**Did Lee Harvey Oswald act alone in the assassination of John Kennedy?**	
answer	**Yes**	
chains of thought	**verification**	**refutation**
	STEP 1: Extract the specific entity, term, or value that directly answers the question: Extracted Answer: Yes STEP 2: Evaluate if the extracted answer is definitively TRUE: • Verify the extracted answer against known facts: The Warren Commission concluded that Lee Harvey Oswald acted alone in the assassination of John F. Kennedy. However, there have been numerous conspiracy theories and investigations that suggest otherwise, and no definitive consensus has been reached that is universally accepted. • Consider temporal/contextual factors: The question of whether Oswald acted alone has been debated for decades, and while the official government position is that he did, public opinion and alternative theories persist. • Determine if it correctly responds to what the question is asking: The extracted answer "Yes" aligns with the official conclusion of the Warren Commission, but it does not account for the ongoing debate and lack of universal agreement. Reasoning: The extracted answer reflects the official conclusion but does not account for the complexity and controversy surrounding the topic. Conclude EXACTLY with either: "CANNOT VERIFY"	STEP 1: Extract the specific entity, term, or value that directly answers the question: Extracted Answer: Yes STEP 2: Evaluate if the extracted answer is definitively FALSE: • The question of whether Lee Harvey Oswald acted alone in the assassination of John F. Kennedy is a subject of debate and conspiracy theories. The official conclusion by the Warren Commission was that Oswald acted alone, but this has been contested by various other investigations and theories. • There is no definitive evidence that universally disproves the conclusion that Oswald acted alone, although there are alternative theories. • The extracted answer "Yes" aligns with the official conclusion of the Warren Commission, which is one of the established narratives. Reasoning: The extracted answer "Yes" is consistent with the official conclusion of the Warren Commission, and while there are alternative theories, there is no definitive evidence that refutes this conclusion. Conclude: CANNOT REFUTE
evaluation	**n**	

Bilateralism: a philosophical motivation

- Bilateralism (Rumfitt 2000) holds that understanding a proposition requires grasping both the conditions under which it can be asserted, and the conditions under which it should be denied
  - I.e., meaning isn't just about knowing when something is true, but also explicitly understanding when it is false
- Bilateralists argue that having explicit roles for both verification and refutation leads to better logical reasoning and clearer understanding
- There is a natural mapping from Belnap-Dunn four-valued semantics to the states of a logical proposition according to the tenets of bilateralism

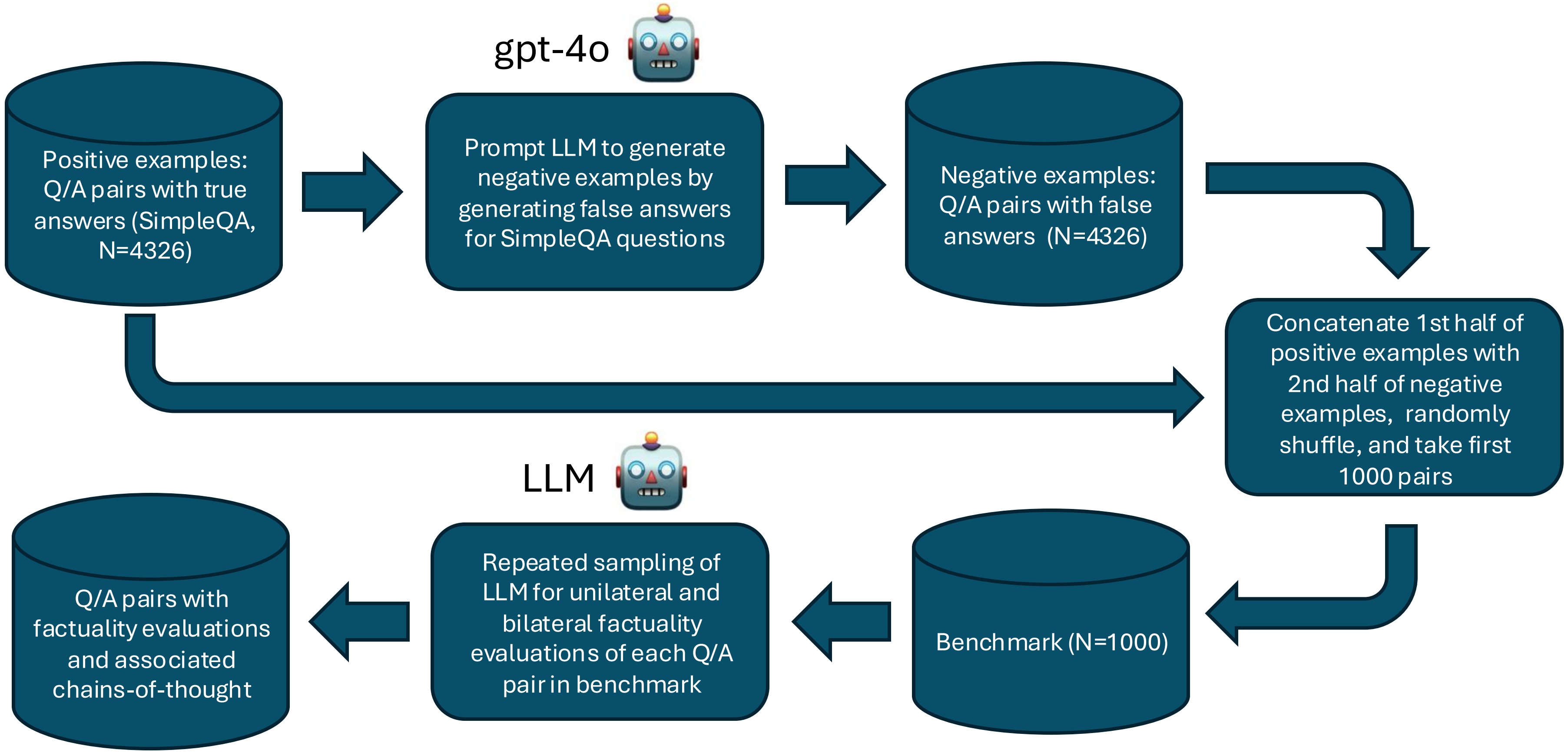
Evaluating LLM factuality

- *Factuality evaluation* determines how well an LLM can generate statements that are factually correct (Wang et al. 2023)
- Most factuality benchmarks, e.g. TriviaQA (Joshi et al. 2017), Natural Questions (Kwiatkowski et al. 2019), FActScore (Min et al. 2023), FELM (Zhao et al. 2023), and SimpleQA (Wei et al. 2024) are based on question answering tasks
- *Factuality assessment* is the selective classification task (El-Raniv and Wiener 2010) of determining the truth of a given statement, with abstention in cases where the classifier determines it cannot do so

Research questions

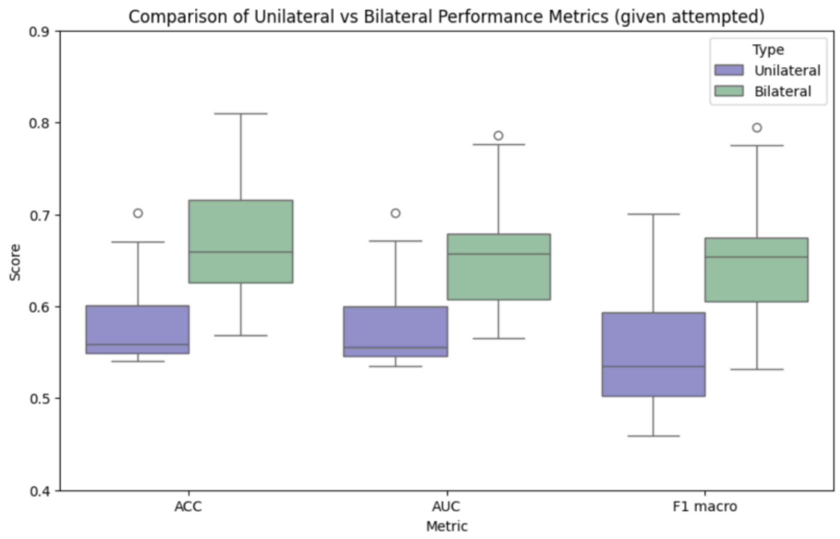
- RQ1: Does bilateral assessment improve accuracy given attempt over unilateral assessment given *humanly-curated question/answer pairs*?
- RQ2: Does bilateral assessment improve accuracy given attempt over unilateral assessment and baseline accuracy given *LLM-generated answers to questions*?
- In our experimental designs, abstention occurs when unilateral assessment returns **`n`** or when bilateral assessment returns **`n`** or **`b`**

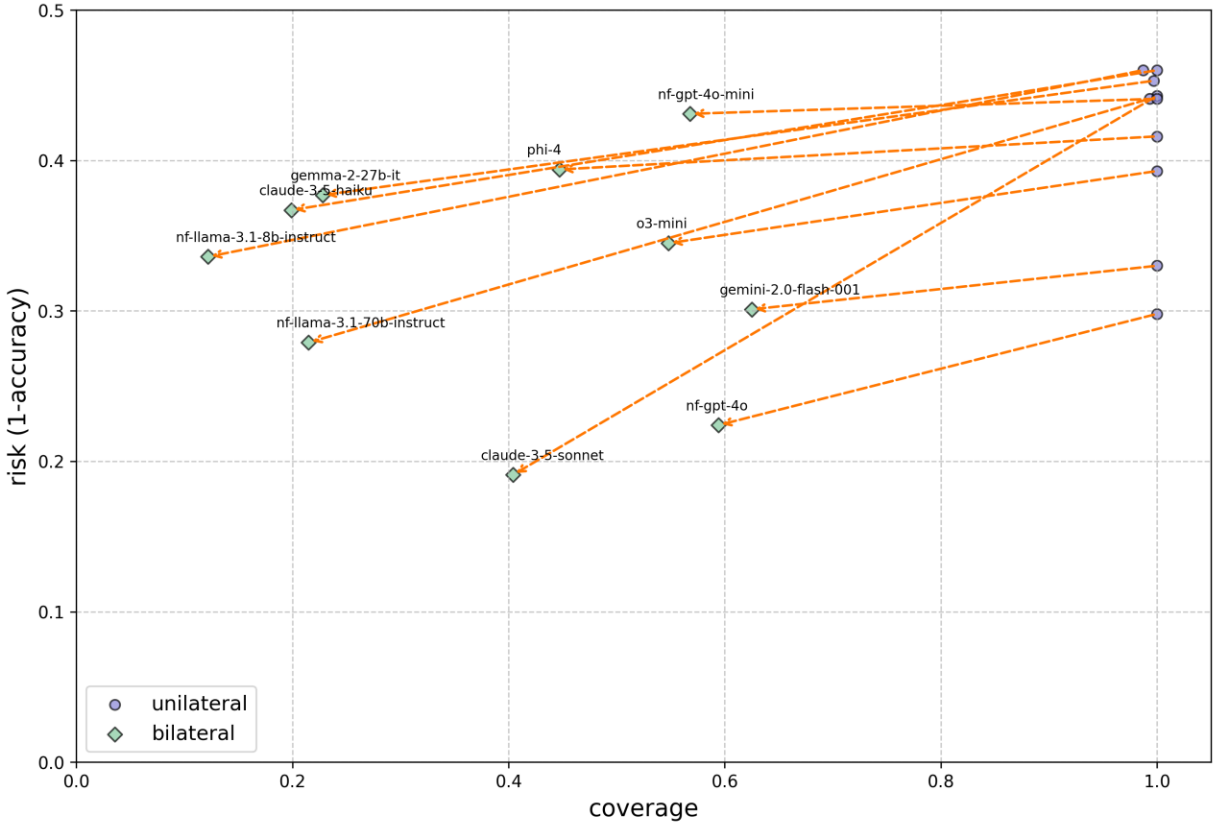
RQ1: experimental workflow



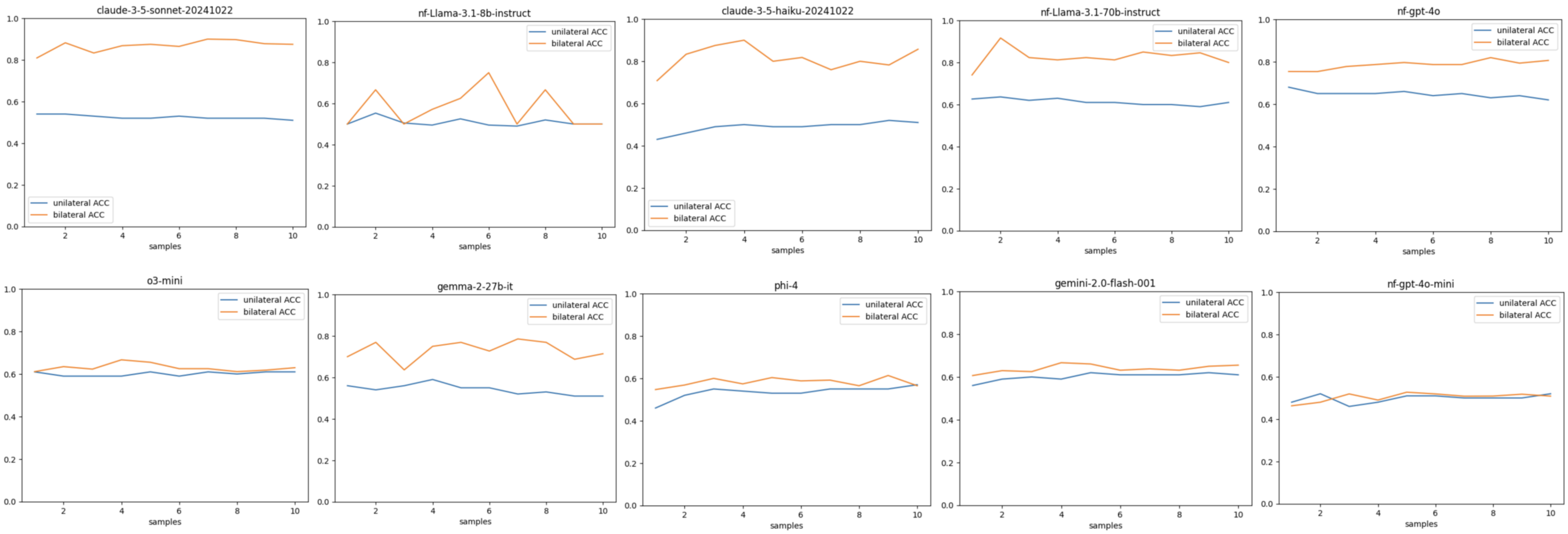
RQ1: performance results

model	unilateral coverage	unilateral accuracy	unilateral F1	bilateral coverage	bilateral accuracy	bilateral F1	Δ F1
claude-3-5-sonnet-20241022	1.000	0.557	0.462	0.404	**0.809**	**0.794**	**0.333**
nf-llama-3.1-8b-instruct	0.987	0.540	0.498	0.122	0.664	0.654	0.155
claude-3-5-haiku-20241022	1.000	0.540	0.459	0.199	0.633	0.614	0.155
nf-llama-3.1-70b-instruct	0.993	0.559	0.516	0.215	0.721	0.663	0.147
nf-gpt-4o	1.000	**0.702**	**0.700**	0.594	0.776	0.776	0.075
o3-mini	1.000	0.607	0.597	0.548	0.655	0.655	0.059
gemma-2-27b-it	0.997	0.547	0.516	0.228	0.623	0.549	0.033
phi-4	1.000	0.584	0.583	0.447	0.606	0.603	0.020
gemini-2.0-flash-001	1.000	0.670	0.668	**0.625**	0.699	0.678	0.010
nf-gpt-4o-mini	1.000	0.559	0.554	0.568	0.569	0.532	-0.022

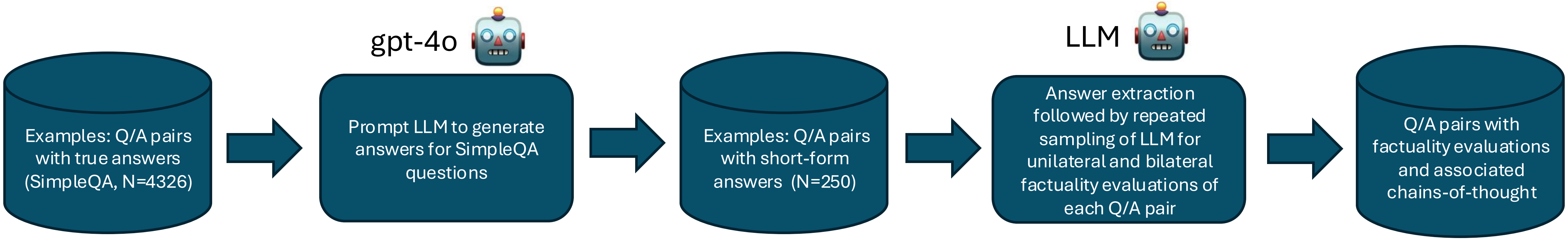




RQ1: repeated sampling does not explain the difference in accuracy

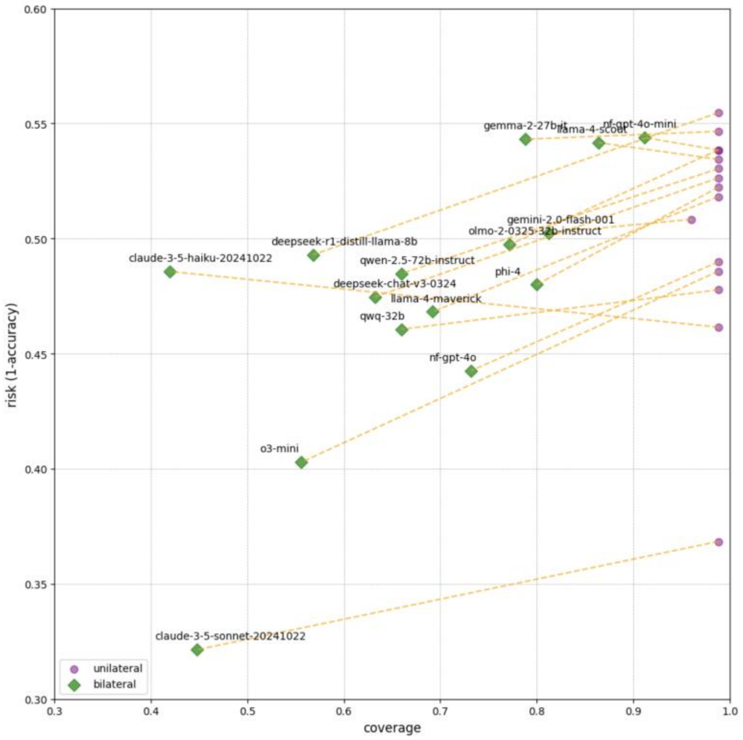


RQ2: experimental workflow

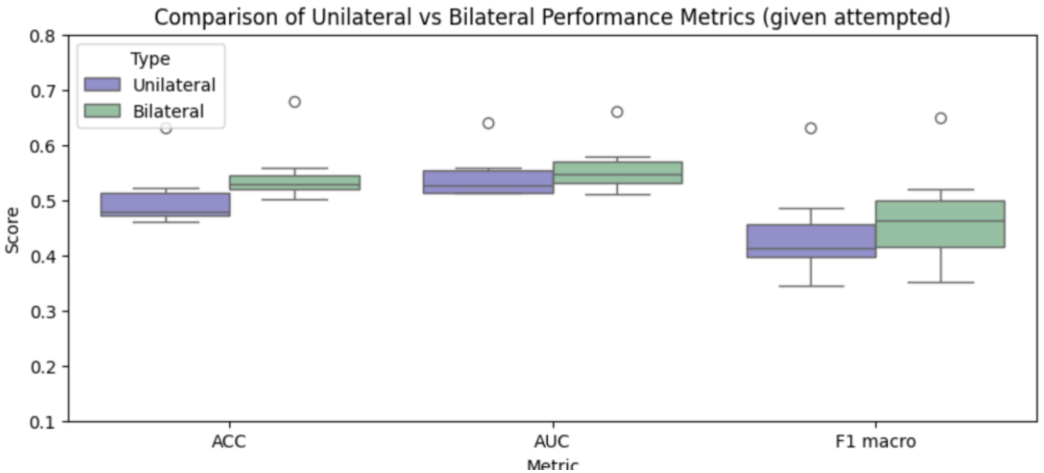


RQ2: performance results

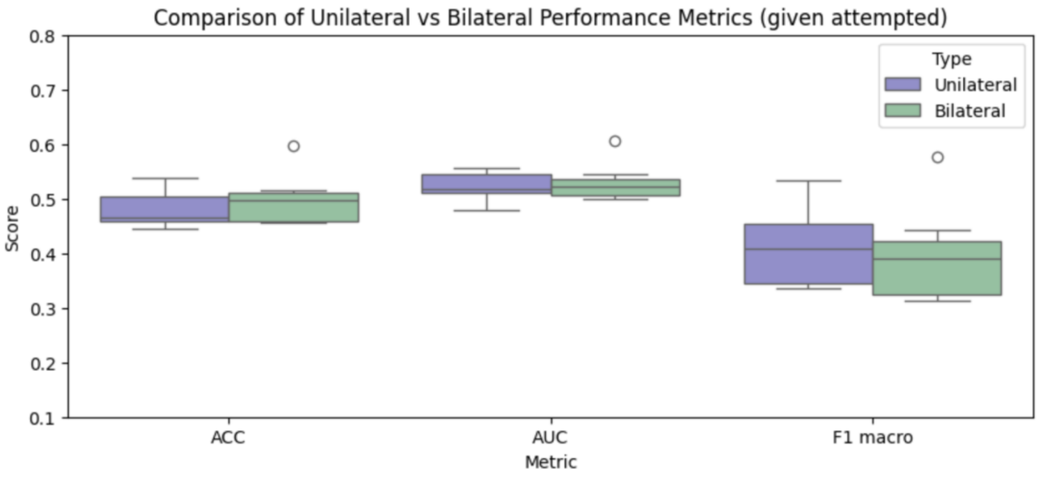
		unilateral				bilateral					
	model	coverage	ACC	AUC	F1 macro	coverage	ACC	AUC	F1 macro	delta ACC	delta F1
0	o3-mini	0.988	0.514	0.543	0.493	0.556	0.597	0.606	0.576	0.083	0.083
1	deepseek-r1-distill-llama-8b	0.988	0.445	0.479	0.408	0.568	0.507	0.546	0.443	0.062	0.035
2	deepseek-chat-v3-0324	0.988	0.474	0.513	0.420	0.632	0.525	0.534	0.470	0.052	0.050
3	llama-4-maverick	0.988	0.482	0.527	0.404	0.692	0.532	0.534	0.420	0.050	0.015
4	nf-gpt-4o	0.988	0.510	0.554	0.446	0.732	0.557	0.568	0.491	0.047	0.045
5	claude-3-5-sonnet-20241022	0.988	0.632	0.640	0.632	0.448	0.679	0.662	0.650	0.047	0.018
6	qwen-2.5-72b-instruct	0.988	0.470	0.513	0.400	0.660	0.515	0.523	0.397	0.046	-0.003
7	phi-4	0.988	0.478	0.526	0.389	0.800	0.520	0.560	0.457	0.042	0.068
8	olmo-2-0325-32b-instruct	0.988	0.462	0.514	0.346	0.772	0.503	0.510	0.352	0.041	0.007
9	qwq-32b	0.988	0.522	0.558	0.487	0.660	0.539	0.579	0.520	0.017	0.033
10	gemini-2.0-flash-001	0.960	0.492	0.545	0.417	0.812	0.498	0.522	0.391	0.006	-0.026
11	gemma-2-27b-it	0.988	0.453	0.506	0.336	0.788	0.457	0.500	0.314	0.003	-0.022
12	nf-gpt-4o-mini	0.988	0.462	0.515	0.340	0.912	0.456	0.508	0.327	-0.005	-0.013
13	llama-4-scout	0.988	0.466	0.518	0.348	0.864	0.458	0.504	0.322	-0.007	-0.026
14	claude-3-5-haiku-20241022	0.988	0.538	0.557	0.533	0.420	0.514	0.527	0.399	-0.024	-0.134



Flagship models

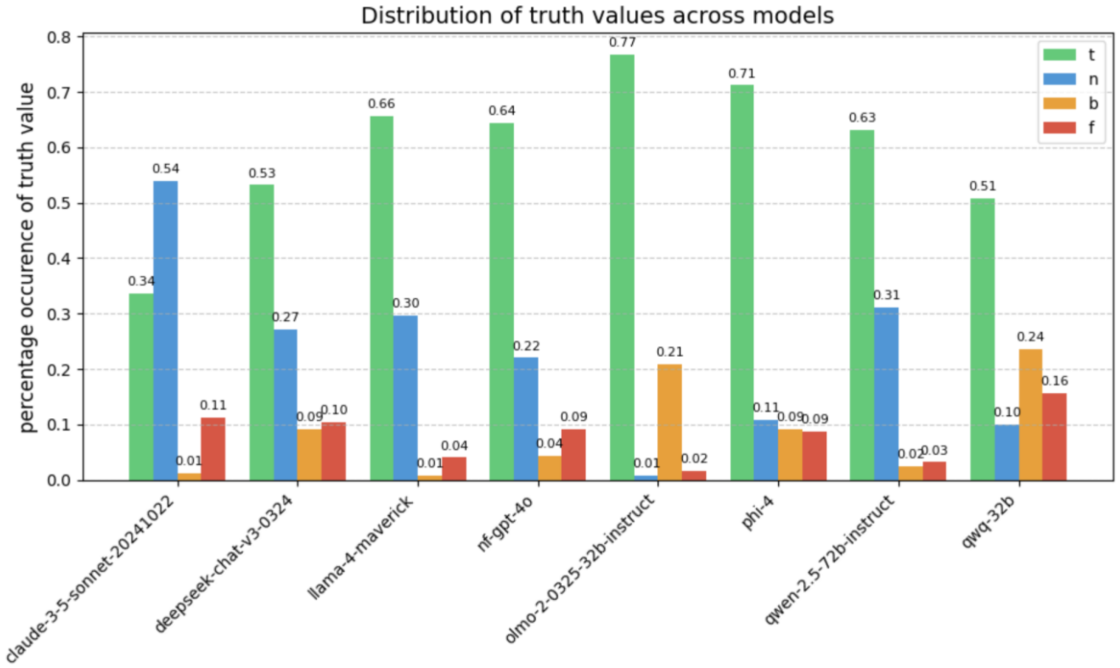


Distilled models

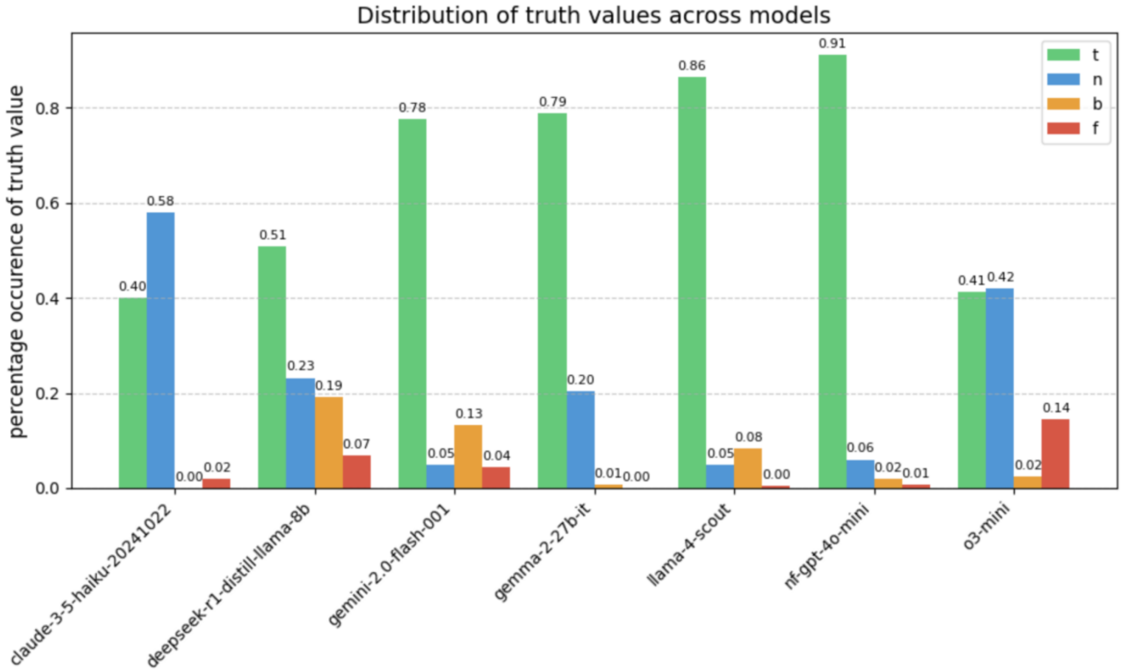


RQ2: how are bilateral truth values distributed?

Flagship models



Distilled models



Summary of findings

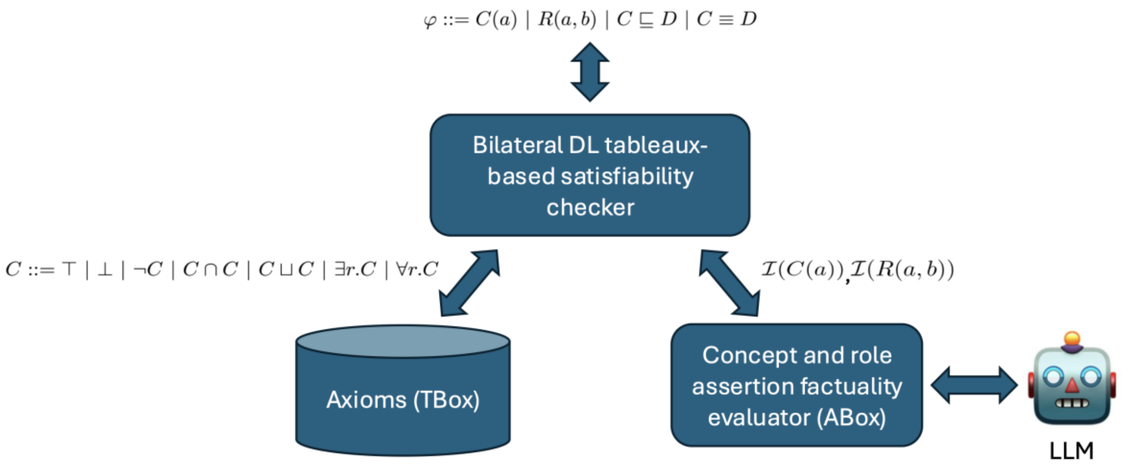
- Bilateral assessment using *flagship* LLMs shows an improvement over unilateral assessment in accuracy given attempted
  - RQ1: $p < 0.02$
  - RQ2: $p < 0.05$
- However, this improvement comes with smaller coverage
  - RQ1:12-62%
  - RQ2: 42-91%
- Bilateral assessment using *distilled* LLMs shows little or no improvement

Discussion

- The coverage trade-off may be appropriate for real-world applications where abstention is preferable to incorrect evaluation (e.g., medical decision support)
- The fact that distilled models do not benefit less suggests they are less capable than flagship models in performing the reasoning needed for bilateral factuality assessment
- The fact that considering both verification and refutation of assertions improves factuality in flagship LLMs could be seen as providing empirical support for bilateralism
    - This is potentially relevant to the questions of LLM belief (cf. Mandelkern & Linzen 2023, Lederman & Mahowald 2024, Herrmann & Levinstein 2024) and propositional interpretability (Chalmers 2025)

Future work: towards an *LLM-as-ABox*

- We have demonstrated a way to have LLMs generate assertions and then provide useful information about the truth of those assertions
- We are working to formalize this as an *LLM-grounded interpretation* for paraconsistent description logics (Maier et al. 2017; Ferguson 2021)
- Our near-term goal is to show that LLM-grounded interpretations can preserve the soundness and completeness of paraconsistent DL reasoning procedures
- Our long-term hope is that this can yield a practical approach for KE using DL reasoning that leverages the broad knowledge embedded in LLMs while being robust to their inconsistency and incompleteness



Thank you!

Email: b.p.allen@uva.nl

GitHub repository: https://github.com/bradleypallen/bilateral-factuality-evaluation

References

- Bradley Allen, Fina Polat, and Paul Groth. SHROOM-INDElab at SemEval-2024 Task 6: Zero- and Few-Shot LLM-Based Classification for Hallucination Detection. In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Mexico City, Mexico, 2024. Association for Computational Linguistics.
- Bradley Allen and Paul Groth. A Benchmark for the Detection of Metalinguistic Disagreements between LLMs and Knowledge Graphs. arXiv preprint arXiv:2502.0289, 2025.
- Nuel Belnap. How a computer should think. In G. Ryle (ed.), Contemporary aspects of philosophy, pp. 30–55. Oriel Press, 1977a.
- Nuel Belnap. A useful four-valued logic. In Modern uses of multiple-valued logic, pp. 5–37. Springer, 1977b.
- David Chalmers. Propositional interpretability in artificial intelligence. arXiv preprint arXiv:2501.15740, 2025.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495, 2023.
- Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11(5), 2010.
- Thomas Macaulay Ferguson. Tableaux and restricted quantification for systems related to weak Kleene logic. In International Conference on Automated Reasoning with Analytic Tableaux and Related Methods, pp. 3–19. Springer, 2021.
- Daniel Herrmann and Benjamin Levinstein. Standards for belief representations in LLMs. arXiv preprint arXiv:2405.21030, 2024.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1705.03551, 2017.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural Questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466, 2019.
- Harvey Lederman and Kyle Mahowald. Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms. ArXiv preprint arXiv:2401.04854, 2024.
- Frederick Maier, Yue Ma, and Pascal Hitzler. Paraconsistent OWL and related logics. Semantic Web, 4(4):395–427, 2013.
- Matthew Mandelkern and Tal Linzen. Do language models' words refer? arXiv preprint arXiv:2308.05576, 2024.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251, 2023.
- Ian Rumfitt. 'Yes' and 'No'. Mind, 109(436):781–823, 2000.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv preprint arXiv:2310.07521, 2023.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. arXiv preprint arXiv:2411.04368, 2024.
- Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. FELM: Benchmarking factuality evaluation of large language models. Advances in Neural Information Processing Systems, 36:44502–44523, 2023.