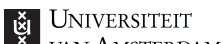
$\mathbf{B_{BL}}$: A Bilateral Modal Logic for LLM Factuality Evaluation

Bradley P. Allen

TLLM 2026 · Tsinghua University · 3 April 2026

Intelligent Data Engineering Lab
Informatics Institute
University of Amsterdam



Motivation

- LLM factuality evaluation is *the assessment of whether an LLM's assertions merit epistemic trust*
- Current approaches model LLM factuality informally, typically via ternary valuations or confidence scores
- These conflate factuality challenges with hallucination and penalize uncertainty as failure (Bang et al., 2025)
- Syntactic variations in prompts can dramatically affect factuality in LLM outputs, leading to the concept of *prompt-sensitive knowledge* (Li et al., 2025)
- As a formalism to help clarify these issues, we present B_{BL}, a bilateral modal logic for reasoning about LLM doxastic states

Object language and bilattice semantics

- Propositional atoms $p, q, \ldots$ closed under:

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \Box\varphi \mid \Diamond\varphi$$

- Following Allen et al. (2025), we define a semantics for B_{BL} based on $\mathcal{NINE} = \mathcal{V}_3 \times \mathcal{V}_3$ (Arieli et al., 1996) where $\mathcal{V}_3 = \{\mathfrak{f}, \mathfrak{e}, \mathfrak{t}\}$.
- Each value $\langle u, v \rangle$ has verification (u) and refutation (v) dimensions, with key values of $\langle \mathfrak{t}, \mathfrak{f} \rangle$ (verified), $\langle \mathfrak{f}, \mathfrak{t} \rangle$ (refuted), $\langle \mathfrak{f}, \mathfrak{f} \rangle$ (ignorance), $\langle \mathfrak{t}, \mathfrak{t} \rangle$ (inconsistency).
- Truth ordering $\leq_t$ and knowledge ordering $\leq_k$ organize these values; operations on $\mathcal{NINE}$ use strong Kleene truth functions component-wise, with the standard ordering $\mathfrak{f} < \mathfrak{e} < \mathfrak{t}$.
- Negation swaps components: $\neg\langle u, v \rangle = \langle v, u \rangle$; meet and join combine dimensions appropriately.

Models

- A model for $\mathcal{L}$ is a structure $\mathfrak{M} = \langle S, R, V \rangle$ where S is a non-empty set of situations, $R \subseteq S \times S$ is a reflexive and symmetric accessibility relation, and $V : S \times \text{Prop} \to \mathcal{V}_3 \times \mathcal{V}_3$ is a valuation function, defined recursively:

$$
\begin{aligned}
[\![p]\!]^{\mathfrak{M}}_s &= V(s,p) \\
[\![\neg\varphi]\!]^{\mathfrak{M}}_s &= \neg[\![\varphi]\!]^{\mathfrak{M}}_s \\
[\![\varphi \wedge \psi]\!]^{\mathfrak{M}}_s &= [\![\varphi]\!]^{\mathfrak{M}}_s \sqcap [\![\psi]\!]^{\mathfrak{M}}_s \\
[\![\varphi \vee \psi]\!]^{\mathfrak{M}}_s &= [\![\varphi]\!]^{\mathfrak{M}}_s \sqcup [\![\psi]\!]^{\mathfrak{M}}_s \\
[\![\Box\varphi]\!]^{\mathfrak{M}}_s &= \langle \min_{s' : sRs'} \pi_1([\![\varphi]\!]^{\mathfrak{M}}_{s'}), \max_{s' : sRs'} \pi_2([\![\varphi]\!]^{\mathfrak{M}}_{s'}) \rangle \\
[\![\Diamond\varphi]\!]^{\mathfrak{M}}_s &= \langle \max_{s' : sRs'} \pi_1([\![\varphi]\!]^{\mathfrak{M}}_{s'}), \min_{s' : sRs'} \pi_2([\![\varphi]\!]^{\mathfrak{M}}_{s'}) \rangle
\end{aligned}
$$

- Soundness established via encoding of B_{BL} valuations in classical system B (Blackburn et al., 2001; Fitting, 1991)

LLM-grounded models

- An *LLM-grounded model* $\mathfrak{M}_M = \langle S, R_M, V_M \rangle$
  - an LLM M
  - situations $s \in S$ corresponding to natural language statements δ_s
  - accessibility relation R_M such that sR_Ms' iff $\delta_s \sim \delta_{s'}$, where $\delta_s \sim \delta_{s'}$ when $\delta_{s'}$ is a meaning-preserving syntactic variant (paraphrase, reordering, reformatting) of δ_s
  - a valuation function $V_M(s, p) = \langle u, v \rangle$ where u, v are computed by binary classification with abstention (El-Yaniv et al., 2010), resolved by majority vote over repeated sampling of responses to verification/refutation prompts with δ_s and δ_p, where δ_p is a natural language statement corresponding to $p \in$ Prop and unparsable responses are mapped to abstention ($\mathfrak{e}$)
- This establishes stable valuations for p at each situation s, which the modal operators then aggregate across R_M

Verification and refutation prompts for computing $V_M(s, p)$

Verification:

```
You are tasked with determining whether the following
assertion can be verified as true based on available
evidence and knowledge.
Assertion: {δ_p}
Context: {δ_s}
Your task is to determine if this assertion can be verified.
Consider all available evidence, facts, and reliable sources
of information.
You must respond with exactly one of these two token
sequences:
- VERIFIED (if the assertion can be confirmed as true based
on evidence)
- CANNOT VERIFY (if the assertion cannot be confirmed as
true, either due to lack of evidence, uncertainty, or
because it is false)
Do not provide any explanation or additional text. Respond
with only the required token sequence.
Response:
```

$$u = \begin{cases} \mathfrak{t} & \text{if response} = \texttt{VERIFIED} \\ \mathfrak{f} & \text{if response} = \texttt{CANNOT VERIFY} \\ \mathfrak{e} & \text{otherwise} \end{cases}$$

$$v = \begin{cases} \mathfrak{t} & \text{if response} = \texttt{REFUTED} \\ \mathfrak{f} & \text{if response} = \texttt{CANNOT REFUTE} \\ \mathfrak{e} & \text{otherwise} \end{cases}$$

Analogously for refutation

A modal interpretation of prompt-agnostic and prompt-sensitive knowledge

- Designated values $D = \{\langle \mathfrak{t}, \mathfrak{f} \rangle\}$
- p is **prompt-agnostic** at s iff $[\![\Box p]\!]_s^{\mathfrak{M}} \in D$
  - The model believes p and the belief survives syntactic variation
  - This resembles Williamson's **safety condition** (Williamson, 2000): knowledge requires that the belief would not easily have been different in nearby situations
- p is **prompt-sensitive** at s if $[\![p]\!]_s^{\mathfrak{M}} \in D$ but $[\![\Box p]\!]_s^{\mathfrak{M}} \notin D$
  - The model believes p but the belief does not survive syntactic variation

Experimental setup

Model	Provider
Claude Opus 4.1	Anthropic
Llama 4 Maverick	Meta
Llama 4 Scout	Meta
Gemini 2.5 Flash	Google
DeepSeek-V3	DeepSeek
Qwen 2.5 72B	Alibaba

Benchmark	Content type
TruthfulQA	Factoids (2021)
SimpleQA	Factoids (2024)
MMLU-Pro	Expert STM knowledge
FACTScore	Biographical facts

Protocol: $N = 250$ assertions per cell (seed $= 42$, balanced 50/50 +/-); $n = 3$ paraphrase variants; bilateral samples $= 3$ (majority vote). Variants generated once per dataset using Claude Opus 4.6, cached across all models. Ternary unilateral as baseline.

Example: prompt-agnostic knowledge (Qwen 2.5 72B $\times$ SimpleQA)

The answer to "Who was the lawyer who developed matrix algebra and also worked on higher-dimensional geometry?" is **Arthur Cayley**.

Situation	Paraphrase	$[\![\Box p]\!]^{\mathfrak{M}}_{s_i}$
s_0	*Who was the lawyer who developed matrix algebra…?*	$\langle \mathrm{t}, \mathrm{f} \rangle$
s_1	*Which lawyer is known to have developed matrix algebra…?*	$\langle \mathrm{t}, \mathrm{f} \rangle$
s_2	*Who was the attorney who developed matrix algebra…?*	$\langle \mathrm{t}, \mathrm{f} \rangle$
s_3	*Can you identify the lawyer responsible for matrix algebra…?*	$\langle \mathrm{t}, \mathrm{f} \rangle$

$$[\![p]\!]^{\mathfrak{M}}_{s_0} = \langle \mathrm{t}, \mathrm{f} \rangle$$
$$[\![\Box p]\!]^{\mathfrak{M}}_{s_0} = \langle \min(\mathrm{t}, \mathrm{t}, \mathrm{t}, \mathrm{t}), \max(\mathrm{f}, \mathrm{f}, \mathrm{f}, \mathrm{f}) \rangle = \langle \mathrm{t}, \mathrm{f} \rangle$$

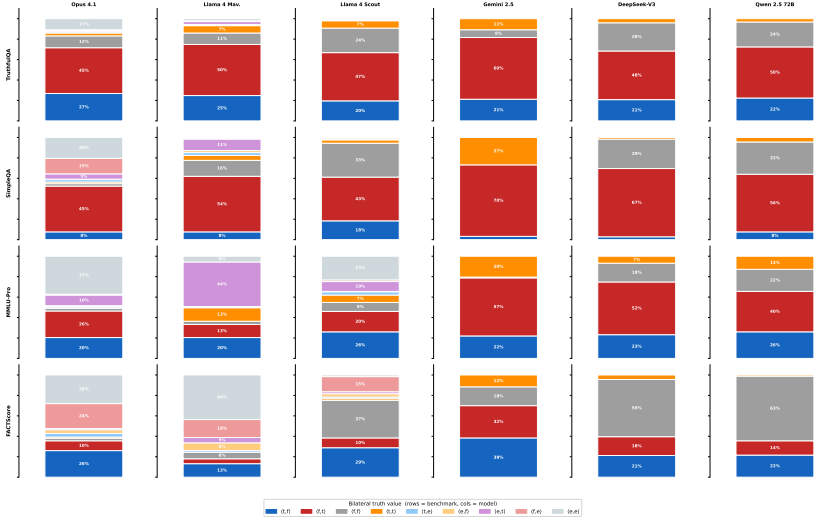
Example: prompt-sensitive knowledge (Qwen 2.5 72B × SimpleQA)

The answer to "In Severance, whom does Mark Scout replace as department head?" is **Peter "Petey" Kilmer**.

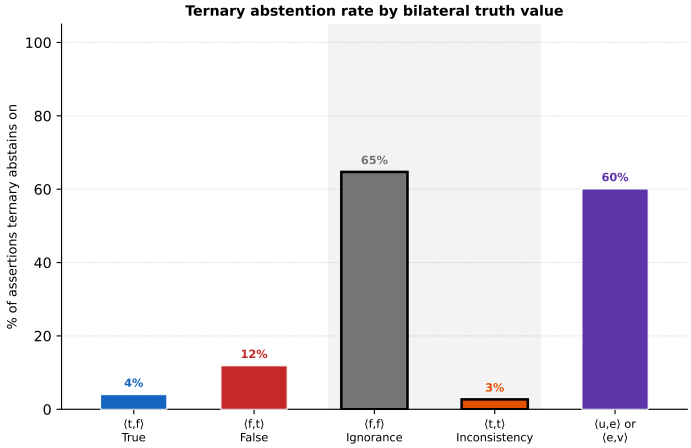
Situation	Paraphrase	$[\![\Box p]\!]^{\mathfrak{M}}_{s_i}$
s_0	*In Severance, whom does Mark Scout replace as dept. head?*	$\langle \mathfrak{t}, \mathfrak{f} \rangle$
s_1	*In Severance, who is the person that Mark Scout succeeds?*	$\langle \mathfrak{f}, \mathfrak{f} \rangle$
s_2	*Whom does Mark Scout take over from in the role of dept. head?*	$\langle \mathfrak{t}, \mathfrak{f} \rangle$
s_3	*Which individual did Mark Scout replace as dept. head?*	$\langle \mathfrak{t}, \mathfrak{f} \rangle$

$$[\![p]\!]^{\mathfrak{M}}_{s_0} = \langle \mathfrak{t}, \mathfrak{f} \rangle$$
$$[\![\Box p]\!]^{\mathfrak{M}}_{s_0} = \langle \min(\mathfrak{t}, \mathfrak{f}, \mathfrak{t}, \mathfrak{t}), \max(\mathfrak{f}, \mathfrak{f}, \mathfrak{f}, \mathfrak{f}) \rangle = \langle \mathfrak{f}, \mathfrak{f} \rangle$$

Distribution of $[\![\Box p]\!]$ reveals model- and content-specific structure



Bilateral valuation distinguishes ignorance ($\langle \mathrm{f}, \mathrm{f} \rangle$) from inconsistency ($\langle \mathrm{t}, \mathrm{t} \rangle$)



Future work

- Neighborhood semantics for $\mathrm{B_{BL}}$
  - Connection to tradition of epistemic logics for multi-agent systems
- Parameterization of $\mathrm{B_{BL}}$ on choice of designated and anti-designated values $D, D^* \subseteq V_3 \times V_3$
- De Morgan-Płonka decomposition of $\mathcal{NINE}$ (Ferguson, 2017; Paoli et al., 2026)
  - Yielding completeness of $\mathrm{B_{BL}}$ via Kamide-Shramko embedding (Kamide et al., 2017) over the $\mathrm{DM_4}$ core
  - Reading $\mathfrak{e}$ as *disengaged* as opposed to *unparsable*: hallucination as containment violation?
- Using $[\![\Box\varphi]\!]$ to guide moves by an LLM opponent in a prover-skeptic dialogue game (Allen, 2026)

Summary

- A bilateral modal logic with bilattice semantics for reasoning about LLM belief
- A modal interpretation of prompt-agnostic and prompt-sensitive knowledge
- Actionable insight into the structure of LLM belief given a subject area

Instead of getting LLMs to reason logically, use logic to reason about LLMs

Dank u wel · 谢谢 · Thank you

b.p.allen@uva.nl

References i

- Allen, Bradley P (2026). **"Elenchus: Generating Knowledge Bases from Prover-Skeptic Dialogues".** In: *arXiv preprint arXiv:2603.06974*.
- Allen, Bradley P. et al. (2025). **"Sound and Complete Neurosymbolic Reasoning with LLM-Grounded Interpretations".** In: *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*. Ed. by Leilani H. Gilpin et al. Vol. 284. Proceedings of Machine Learning Research. PMLR, pp. 392–419. url: https://proceedings.mlr.press/v284/allen25a.html.
- Arieli, Ofer and Arnon Avron (1996). **"Reasoning with logical bilattices".** In: *Journal of Logic, Language and Information* 5.1, pp. 25–63.
- Bang, Yejin et al. (2025). **"HalluLens: LLM Hallucination Benchmark".** In: *arXiv preprint arXiv:2504.17550*.
- Blackburn, Patrick, Maarten De Rijke, and Yde Venema (2001). ***Modal logic*.** Cambridge University Press.

References ii

- Ferguson, Thomas Macaulay (2017). **"Cut-Down Operations on Multilattices".** In: *Meaning and Proscription in Formal Logic: Variations on the Propositional Logic of William T. Parry*. Springer, pp. 133–161.
- Fitting, Melvin C (1991). **"Many-valued modal logics".** In: *Fundamenta informaticae* 15.3-4, pp. 235–254.
- Kamide, Norihiro and Yaroslav Shramko (2017). **"Embedding from multilattice logic into classical logic and vice versa".** In: *Journal of Logic and Computation* 27.5, pp. 1549–1575.
- Li, Moxin et al. (2025). **"Knowledge boundary of large language models: A survey".** In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5131–5157.
- Paoli, Francesco, Damian Szmuc, and Martina Zirattu (2026). **"The Algebra of Analytic Containment".** In: *Journal of Logic, Language and Information*, pp. 1–31.
- Williamson, Timothy (2000). ***Knowledge and its Limits*.** OUP UK.

 El-Yaniv, Ran and Yair Wiener (2010). **"On the Foundations of Noise-free Selective Classification".** In: *Journal of Machine Learning Research* 11.5.