Dataset	Perturbation	True to True	Not True to Not True	Epistemic Expansions $\mathcal{E}$	Epistemic Retractions $\mathcal{R}$
City Locations	Synthetic	2503 (25.0)	3301 (33.0)	567 (5.7)	**3629 (36.3)**
	Fictional	3521 (35.2)	2708 (27.1)	1160 (11.6)	2611 (26.1)
	Fictional (T)	3678 (36.8)	2529 (25.3)	1339 (13.4)	2454 (24.5)
Medical Locations	Synthetic	1539 (14.1)	5562 (51.0)	843 (7.7)	**2952 (27.1)**
	Fictional	1816 (16.7)	4489 (41.2)	1916 (17.6)	2675 (24.6)
	Fictional (T)	1724 (15.8)	4476 (41.1)	1929 (17.7)	2767 (25.4)
Word Definitions	Synthetic	1880 (19.0)	4246 (42.9)	1647 (16.7)	**2115 (21.4)**
	Fictional	2179 (22.0)	4319 (43.7)	1574 (15.9)	1816 (18.4)
	Fictional (T)	2159 (21.8)	4389 (44.4)	1504 (15.2)	1836 (18.6)

Table 4. Epistemic expansions $\mathcal{E}$ and retractions $\mathcal{R}$ under perturbed zero-shot experiments. Counts (percentages) of beliefs that remain stable or undergo expansions $\mathcal{E}$ or retractions $\mathcal{R}$ under `Synthetic` (yellow), `Fictional` (gray), and `Fictional(T)` (blue) perturbations. Despite domain variation, `Synthetic` perturbations induce the highest retraction rates in all cases, mirroring the hierarchy observed under probing in Table 3.

City Locations is highly stable, Medical Indications shows moderate instability, and Word Definitions is markedly more fragile. These results again indicate that representational stability is governed by epistemic familiarity, given that City Locations are likely richly represented in training corpora, while Medical Indications are common but specialized, and Word definitions are the most semantically flexible.

5.3 Behavioral Stability under Zero-Shot Perturbations

We next evaluate `P-StaT` behavioral stability by applying the same perturbations Θ via belief context in zero-shot prompting and measuring epistemic retractions. Table 4 reports belief changes across LLMs, with LLM-level results in Supplementary Section D.3. Overall belief changes are larger than in probing experiments, reflecting the greater flexibility of behavioral responses compared to linear reclassification over fixed representations.

Despite this increase, the perturbation hierarchy observed representationally also appears behaviorally. The `Synthetic` perturbation induces the highest epistemic retraction rate in every domain, with 36.3% in City Locations, 27.1% in Medical Indications, and 21.4% in Word Definitions. `Fictional` and `Fictional(T)` also induce substantial retractions, but consistently fewer than `Synthetic` within each dataset.

However, domain ordering differs from the probing setting. City Locations has the highest zero-shot retraction rates, while Word Definitions has the lowest. This reversal likely reflects interactions between belief stability and behavioral accuracy. Specifically, zero-shot retractions conflate belief changes with the number of baseline beliefs available to retract (i.e., a statement must be considered `True` in the baseline case in order to become a retraction). In contrast, probing isolates changes in the decision boundary under fixed representations.

Our results show that **epistemic familiarity governs stability not only in activation space, but also in LLM behavior**. Unfamiliar `Synthetic` content systematically destabilizes LLM beliefs under semantic perturbation, whereas familiar `Fictional` content produces smaller and more context-dependent effects.

6 Discussion

We provide a unified perspective on how LLMs organize and maintain truth judgments under semantic perturbation. Across LLMs, domains, and evaluation settings, `Neither` statements play a central role in belief stability: unfamiliar `Synthetic` statements consistently induce the largest epistemic retractions, while familiar `Fictional` statements are more stable. Our findings indicate that belief stability in LLMs is governed as much by epistemic familiarity as by linguistic form.

At the representational level, we observe a decoupling between linguistic similarity and latent-space organization. `Fictional` content exhibits domain-dependent linguistic variation that does not align with distance in representation space. This mismatch suggests that veracity representations encode higher-level epistemic context learned during training rather than lexical statistics alone. Through the lens of `P-StaT`, these representational differences translate into systematic differences in stability under semantic perturbation. Expanding the semantic definition of truth to include unfamiliar `Synthetic` content reliably destabilizes previously held beliefs, while

perturbations involving familiar `Fictional` content are better tolerated. These results hold in both representational and behavioral instantiations, suggesting that latent-space organization captures meaningful epistemic structure.

More broadly, `P-StaT` complements accuracy-based factuality benchmarks by shifting focus from correctness under fixed prompts to stability under principled semantic shifts. By emphasizing epistemic retractions, it targets a particularly consequential form of instability. This perspective aligns with formal accounts of rational belief change, such as Leitgeb's notion of P-stability, which requires that established beliefs be preserved under justified changes in evidential context [16]. From this viewpoint, the sensitivity of LLMs to `Synthetic` perturbations highlights a structural limitation: distributional plausibility alone does not ensure a stable internal organization of truth, falsity, and indeterminacy. Together, these findings suggest that evaluating belief stability under controlled semantic perturbations offers a principled way to probe the epistemic organization of LLMs beyond what can be inferred from accuracy alone.

7 Conclusion

We show that **epistemic familiarity is a key determinant of belief stability under semantic reframing**. By linking internal representations and behavioral responses under matched perturbations, `P-StaT` reveals that unfamiliar `Synthetic` content consistently induces the largest epistemic retractions across LLMs and domains. These results demonstrate that evaluating belief stability under controlled perturbations of `Neither` content provides a principled complement to accuracy-based factuality metrics and a step toward more epistemically reliable language models.

8 Limitations

Our perturbations focus on a specific notion of epistemic familiarity and a limited set of factual domains. While this design enables controlled comparisons across representational and behavioral settings, it does not exhaustively cover all types of semantic variation. Extending `P-StaT` to other forms of epistemic ambiguity, such as disputed claims, probabilistic beliefs, or evolving facts, would further test its generality. In addition, our analysis considers fixed LLM parameters. Although `P-StaT` isolates how semantic perturbations interact with existing internal representations, it does not address how belief stability may change in settings where representations themselves evolve over time. Finally, retraction rates depend on accuracy on `True` statements, so our notion of stability reflects both robustness to perturbation and an LLM's propensity to assert truth in the baseline case. While we emphasize epistemic retractions as a primary signal of instability, other applications may require alternative notions of stability.

9 Ethical Considerations

Our study examines the conditions under which perturbations can systematically destabilize what an LLM considers true. For example, we find that unfamiliar synthetic content induces the largest epistemic retractions, producing up to 32.7% retractions in representational evaluations and up to 36.3% in behavioral evaluations. This could inform efforts to undermine LLM reliability and has implications for trust in LLMs.

While our framework could be misused to deliberately destabilize LLM beliefs, our intent is diagnostic rather than adversarial: `P-StaT` is designed to identify epistemic vulnerabilities in order to inform more robust evaluation and LLM design. We do not propose interventions or belief manipulation techniques, and all experiments are conducted on fixed, open-source LLMs in offline settings. More broadly, our findings highlight a structural limitation of current LLMs rather than a prescription for exploiting it. `P-StaT` is intended solely for research and diagnostic evaluation of pretrained LLMs and is not designed for deployment, belief steering, or real-world decision-making systems.

9.1 Data availability

We use the `True`, `False`, and `Synthetic` statements available at https://huggingface.co/datasets/carlomarxx/trilemma-of-truth under a `CC-BY-4.0` license. `Fictional` statements are available at https://huggingface.co/datasets/samanthadies/representational_stability. The code used to generate the `Noise` activations can be found at https://github.com/samanthadies/representational_stability.

9.2 Code availability

We release the code used to generate activations, generate the `Noise` activations, and run the `P-StaT` stability experiments under an `MIT License` at https://github.com/samanthadies/representational_stability. ChatGPT and Copilot were used to help write initial versions of experiment and plotting scripts, and helped clean and comment the code. All AI-generated content was reviewed and verified by the authors.

All models used in this work are publicly available for research use under their respective licenses (`MIT License` for `sAwMIL`; `Gemma` for Gemma-7b, Gemma-7b-it, Gemma-2-9b, and Gemma-2-9b-it; `llama3.1` for Llama-3.1-8b and Llama-3.1-8b-Instruct; `llama3.2` for Llama-3.2-3b, Llama-3.2-3b-Instruct; `llama3` for Llama3-Med42-8b; `Bio-Medical-Llama-3-8b LLM License` for Bio-Medical-Llama-3-8b; `apache-2.0` for Mistral-7B-v0.3, Mistral-7B-Instruct-v0.3, Qwen2.5-7B, Qwen2.5-

7B-Instruct, Qwen2.5-14B, and Qwen2.5-14B-Instruct). Marks and Tegmark did not specify a license for the `Mean Difference` probe [7].

Acknowledgments

We thank Hannes Leitgeb and Branden Fitelson for discussions on P-stability and how it might be related to epistemic uncertainty in LLMs. We also thank Zohair Shafi and Moritz Laber for their feedback and discussions on methodological and empirical portions of this work.

Funding

This material was sponsored by the Government of the United States under Contract Number FA8702-15-D-0002. The view, opinions, and/or filings contained in this material are those of the author(s) and should not be construed as an official position, policy, or decision of the Government of the United States or Carnegie Mellon University or the Software Engineering Institute unless designated by other documentation.

Competing interests

The authors declare no competing interests.

References

1. AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M. & Ghazvininejad, M. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031* https://doi.org/10.48550/arXiv.2204.06031 (2022).

2. Han, J. *et al.* Simple factuality probes detect hallucinations in long-form natural language generation. In *Findings of the Association for Computational Linguistics (EMNLP 2025)*, 16209–16226. https://doi.org/10.18653/v1/2025.findings-emnlp.880 (2025).

3. Abbasi Yadkori, Y., Kuzborskij, I., György, A. & Szepesvari, C. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. *Adv. Neural Inf. Process. Syst.* **37**, 58077–58117 (2024).

4. Suzgun, M. *et al.* Language models cannot reliably distinguish belief from knowledge and fact. *Nat. Mach. Intell.* 1–11. https://doi.org/10.1038/s42256-025-01113-8 (2025).

5. Liu, Y. *et al.* Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*. https://doi.org/10.48550/arXiv.2308.05374 (2023).

6. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Inf. Syst.* **43**, 1–55. https://doi.org/10.1145/3703155 (2025).

7. Marks, S. & Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of True/False datasets. In *Proceedings of the 1st Conference on Language Modeling (COLM 2024)* (2024).

8. Bürger, L., Hamprecht, F. A. & Nadler, B. Truth is universal: Robust detection of lies in LLMs. *Adv. Neural Inf. Process. Syst.* **37**, 138393–138431 (2024).

9. Savcisens, G. & Eliassi-Rad, T. Trilemma of truth in large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025* (2025). https://openreview.net/forum?id=z7dLG2ycRf.

10. Turpin, M., Michael, J., Perez, E. & Bowman, S. Language models don't always say what they think: Unfaithful explanations in Chain-of-Thought prompting. *Adv. Neural Inf. Process. Syst.* **36**, 74952–74965 (2023).

11. Elazar, Y. *et al.* Measuring and improving consistency in pretrained language models. *Transactions Assoc. for Comput. Linguist.* **9**, 1012–1031. https://doi.org/10.1162/tacl_a_00410 (2021).

12. Li, Y., Miao, Y., Ding, X., Krishnan, R. & Padman, R. Firm or fickle? evaluating large language models consistency in sequential interactions. In Che, W., Nabende, J., Shutova, E. & Pilehvar, M. T. (eds.) *Findings of the Association for Computational Linguistics: ACL 2025*, 6679–6700. https://doi.org/10.18653/v1/2025.findings-acl.347 (Association for Computational Linguistics, Vienna, Austria, 2025).

13. Wei, A., Haghtalab, N. & Steinhardt, J. Jailbroken: How does LLM safety training fail? *Adv. Neural Inf. Process. Syst.* **36**, 80079–80110 (2023).

14. Harding, J. Operationalising representation in natural language processing. *Br. J. for Philos. Sci.* https://doi.org/10.1086/728685 (2023).

15. Herrmann, D. A. & Levinstein, B. A. Standards for belief representations in LLMs. *Minds Mach.* **35**, 5. https://doi.org/10.1007/s11023-024-09709-6 (2024).

16. Leitgeb, H. The stability theory of belief. *Philos. review* **123**, 131–171. https://doi.org/10.1215/00318108-2400575 (2014).

17. Conneau, A., Kruszewski, G., Lample, G., Barrault, L. & Baroni, M. What you can cram into a single