Security and Privacy of Machine Learning, 2025 Critique Hallucination – (1) Why Language Models Hallucinate (2) Learning to Reason for Hallucination Span Detection

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

Critique: (1) Why Language Models Hallucinate

Summary

Kalai et al. [1] present a learning-theoretic account of why language models (LMs) produce overconfident falsehoods. The core idea is a reduction from generative error in density estimation to misclassification error in an *Is-It-Valid* (IIV) binary classification problem. For base models trained by cross-entropy, the paper proves lower bounds linking a model's generative error rate to its IIV error, both without and with prompts, and identifies statistical drivers such as the prevalence of *arbitrary facts* (e.g., birthdays) and agnostic-learning limits for imperfect model classes. For post-training, the authors argue that binary, guess-rewarding evaluations structurally incentivize bluffing over abstention; they propose *explicit confidence targets* that penalize wrong answers relative to confidence, encouraging behavioral calibration. The work positions hallucination not as a Transformer quirk but as a statistical inevitability under current objectives and evaluation regimes, aligning with empirical surveys on hallucination causes and mitigations [2]. The account interfaces with findings on internal uncertainty signals [3] and on alignment methods such as RLHF [4], while grounding the formal view in classic learning theory [5].

Strengths

- **Unifying reduction.** The mapping from generative error to IIV misclassification is elegant and clarifies that some hallucinations are the supervised-learning analogue of unavoidable generalization errors. This frames debates about "fixing hallucination" in terms of sample complexity and hypothesis class rather than solely decoding tricks.
- **Prompt-aware formalism.** Extending the reduction to conditional generation with prompts (varying E_c, V_c) makes the results relevant to real LM usage patterns, not just unconditional modeling.
- **Actionable evaluation proposal.** Confidence-targeted scoring is simple to implement, compatible with existing benchmarks, and provides a clean decision rule (answer iff $p(\text{correct}) > t$) that aligns incentives with truthful uncertainty.
- **Clear separation of stages.** The paper disentangles pre-training (statistical calibration and inherent errors) from post-training (socio-technical reinforcement of overconfidence), avoiding the common conflation of the two.
- **Conceptual economy.** By avoiding architecture-specific assumptions, the analysis travels across LM families, RAG settings, and reasoning-augmented systems, matching observations catalogued in surveys [2].

Weaknesses

- **Idealized data assumption.** Key lower bounds assume $p(V) = 1$ (noiseless training), yet modern corpora contain nontrivial noise and contradiction. While the authors note noise would often *increase* errors, the proofs and constants hinge on this idealization, leaving open how bounds translate under realistic, heterogeneous label noise.
- **Thresholding choice and $|E|$ dependence.** The IIV classifier uses a probability threshold of $1/|E|$ (or $\min_c |E_c|$ under prompts). This dependence can be brittle: $|E_c|$ is latent, prompt-specific, and unidentifiable. Practical surrogates (e.g., top-k mass, logit margins) may yield different constants or even different qualitative behavior.
- **From calibration to behavior.** The link $\delta \approx 0$ via cross-entropy local optimality supports *probability* calibration, yet the post-training prescription relies on *behavioral* calibration (abstain below t). The paper does not empirically validate that pretrained or aligned models exhibit monotone answer/IDK switching at precise thresholds across diverse tasks.
- **Limited treatment of tools and externalization.** While the theory claims to cover RAG and tool-use at a high level, it abstracts away failure modes like retrieval *coverage*, citation faithfulness, or tool-call budget constraints, which interact with abstention incentives in practice.
- **Adoption and Goodhart risk.** Confidence-targeted scoring could itself be gamed: models may *under-answer*

to maximize score, or overfit to benchmark-specific t distributions. The paper acknowledges socio-technical barriers but does not propose concrete governance or auditing protocols to mitigate Goodharting.

Potential Improvements / Extensions

- **Empirical validation of the reduction.** Construct controlled "arbitrary-facts" datasets with tunable singleton rates and report (i) measured generative error vs. predicted lower bounds, (ii) IIV misclassification vs. generative error across model sizes and training budgets.
- **Selective prediction baselines.** Compare confidence-targeted scoring to established *selective classification* and *risk-coverage* frameworks, reporting coverage-accuracy curves, AURC/AURRA, and abstention ECE; connect to theory via abstention-aware losses.
- **Behavioral calibration audits.** Release an evaluation where each item is graded under multiple $t \in \{0.5, 0.75, 0.9\}$ simultaneously; measure whether a single decision boundary (on log odds) explains answer/IDK flips across t. Probe across math, QA, code, and RAG.
- **Noise-robust bounds.** Extend the IIV reduction under label noise models (Massart/Tsybakov) and corrupted-corpus mixtures; quantify how noise inflation competes with calibration (δ) and with $|E_c|$ estimates.
- **Tool-use and retrieval economics.** Formalize a budgeted decision: *abstain*, *search* (cost c, success prob. q), or *answer*; derive optimal policies under confidence-targeted scoring and evaluate on open-domain QA with controllable retrieval costs.
- **Human factors.** Explore user acceptance of abstentions: randomized trials where the same model answers under $t=0$ vs. $t=0.75$; measure task utility, trust, and perceived competence.

Questions for the Authors

- **Identifiability of $|E_c|$.** How sensitive are your guarantees to mis-specifying $|E_c|$? Can the reduction be reformulated using quantiles of $\hat{p}(r \mid c)$ or margin-based surrogates that avoid cardinality terms?
- **Calibration-objective alignment.** Have you experimented with abstention-aware objectives (e.g., selective NLL, coverage constraints) during post-training to *jointly* optimize behavioral calibration for a distribution over t?
- **RAG failure taxonomy.** Under confidence-targeted scoring, when retrieval fails (noisy/empty hits), do models learn to abstain reliably, or do we observe systematic over-abstention on long-tail topics? Any evidence from ablations?
- **Gameability and meta-eval.** What safeguards (held-out t, adversarial prompts, uncertainty falsification tests) do you recommend to prevent benchmark overfitting and to audit honest confidence?
- **Scope of inevitability.** Your results show inevitability for base models under density estimation. With strong tool-use and verification (e.g., program-of-thought with checkers), do the same lower bounds meaningfully constrain end-to-end *systems*, or can verification asymptotically decouple generative error from IIV?

Critique: (2) Learning to Reason for Hallucination Span Detection

Summary

The paper proposes RL4HS, a reinforcement learning (RL) framework that trains a chain-of-thought (CoT) *reasoning* model to *localize* hallucinated spans in conditional generation, directly optimizing a span-level F1 reward. The method builds on Group Relative Policy Optimization (GRPO) and introduces Class-Aware Policy Optimization (CAPO) to correct a reward imbalance (non-hallucination predictions being over-rewarded). On the RAGTruth benchmark—spanning summarization, QA, and data-to-text—RL4HS outperforms instruction-only prompting, general reasoning models, and supervised fine-tuning baselines. Crucially, the paper argues that *task-specific* reasoning trained with span-level rewards is superior to generic reasoning models for hallucination localization. The work contrasts with prior efforts on (i) binary hallucination detection and pipeline factuality methods, and (ii) span-level detection via supervised or attention-based schemes [6]–[8]. Optimization-wise, it connects to recent GRPO-style reasoning training [9].

Strengths

- **Clear problem reframing:** The paper foregrounds span-level *localization* (not just binary presence), reflecting actual user needs in RAG-style systems, where actionable feedback requires pointing to unsupported spans [6].
- **Direct, verifiable objective:** Optimizing the *evaluation* metric (span-F1) as reward is compelling and minimizes metric-mismatch. The reward is naturally verifiable and avoids heuristic proxies.
- **Reasoning for localization:** Framing span detection as multi-step claim extraction and support checking, and then *training* the reasoning policy (vs. only prompting) is novel and well motivated relative to attention-based token classifiers [7].
- **CAPO to mitigate reward hacking:** The analysis of advantage imbalance (non-hallucination inflation) and the simple, effective reweighting are insightful and practically useful for other imbalanced RL-for-NLP tasks.
- **Thorough comparisons:** Including strong SFT and reasoning baselines (and an OOD evaluation) supports the claim that *domain-specialized* reasoning is needed beyond generic CoT skills [8], [9].

Weaknesses

- **Span extraction by string matching:** Mapping predicted text segments back to indices via naive matching can be brittle (duplicates, paraphrases, tokenization drift). This may conflate localization errors with surface-form variance, slightly inflating reward variance.

- **Metric singularity:** Solely optimizing span-F1 risks overfitting to overlap-based set metrics; difficult spans (paraphrastic or compositional hallucinations) might be under-rewarded despite being semantically correct localizations.
- **Limited generality claims:** The argument that generic reasoning models underperform may partly reflect task-set/domain mismatch. Without extensive cross-benchmark tests (beyond RAGTruth), the generality to long-context and multi-hop grounding remains suggestive rather than conclusive.
- **CAPO sensitivity:** The α down-weighting for non-hallucination advantages is tuned on validation; stability under domain shift (different class priors) is not fully explored, and adaptive schemes are not compared.
- **Ablations on reasoning traces:** While a case study is provided, a larger-scale analysis of trace faithfulness (e.g., step-to-span attribution, perturbation tests) would strengthen claims that RL induces genuine, not decorative, reasoning.

Potential Improvements / Extensions

- **Robust span alignment:** Replace string matching with alignment under edit distance or semantic matching (e.g., entailment-aware alignment) to make rewards less brittle and more semantically grounded [8].
- **Multi-objective rewards:** Combine span-F1 with (i) calibration terms (penalize overconfident empty outputs), (ii) coverage/diversity constraints, and (iii) NLI-based support checks for paraphrases. A Pareto or risk-aware GRPO could further stabilize precision–recall.
- **Adaptive CAPO:** Learn α online via class-conditional advantage normalization or uncertainty-weighted rebalancing; alternatively, use a per-group conditional standardization that respects label mixture proportions.
- **Trace-grounded training signals:** Add auxiliary supervision that each reasoning step references evidence spans; use counterfactual data augmentation (remove evidence and test whether the step disappears) to encourage *faithful* traces.
- **Broader evaluation:** Test on open-domain factuality sets (e.g., long-context QA with retrieval drift), code/text tables, and multilingual data. Include human judgments of span usefulness in downstream editing workflows.

Questions to the Authors

- **Faithfulness vs. performance:** Does higher span-F1 correlate with human *trust* in the reasoning traces? Any evidence that RL induces stepwise *causal* reliance on cited context rather than post-hoc rationalization?
- **Class prior shift:** How does CAPO behave when the non-hallucination prior changes (e.g., domain with very frequent hallucinations)? Could α be learned from group statistics to avoid hand-tuning?
- **Negation & paraphrase:** How often does RL4HS miss paraphrastic hallucinations (factually unsupported but lexically distant) compared to entailment-driven detectors [8]? Would hybridizing with NLI losses help?
- **Safety trade-offs:** Can the CAPO scaling re-introduce false positives that degrade user experience in low-hallucination settings? Any user-facing thresholding or abstention mechanism evaluated?
- **General-purpose reasoning:** If generic reasoning models are adapted with a small amount of span-level RL (few-shot), how quickly do they catch up? Is the gap mainly data-type, reward-type, or pretraining-style?

References

[1] A. T. Kalai and coauthors, "Why language models hallucinate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
[2] Z. Ji, N. Lee, R. Frieske *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, 2023.
[3] S. Kadavath, E. P. Wang *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.
[4] L. Ouyang, J. Wu, X. Jiang *et al.*, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
[5] M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory*. MIT Press, 1994.
[6] Y. Wu, J. Zhu, S. Xu, K. Shum, C. Niu, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models," *arXiv preprint arXiv:2311.05200*, 2023.
[7] Y. Ogasa and Y. Arase, "Hallucinated span detection with multi-view attention features," *arXiv preprint arXiv:2504.04335*, 2025.
[8] A. Scirè, K. Ghonim, and R. Navigli, "Fenice: Factuality evaluation of summarization based on natural language inference and claim extraction," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 14 148–14 161.
[9] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024, introduces GRPO-style training for reasoning.