A Reasonable Approach to Hallucination Mitigation on HotPotQA

Columbia COMS4705 Project Milestone
Keywords: *hallucination, Q&A, reinforcement learning, decoding*

Joshua Hegstad
Department of Computer Science
Columbia University
jlh2288@columbia.edu

Ahmed Jaber
Department of Computer Science
Columbia University
amj2234@columbia.edu

Farhaan Siddiqui
Department of Computer Science
Columbia University
fs2872@columbia.edu

Abstract

Small Language Models frequently hallucinate during multi-hop reasoning; we hypothesize these errors manifest as high-entropy states reflecting epistemic uncertainty. We address this in Qwen2.5-7B-Instruct via a three-stage pipeline: reasoning distillation using Chain-of-Thought, followed by entropy-aware Reinforcement Learning that aligns model confidence with NLI-verified factuality. Preliminary results on HotpotQA indicate that this approach, complemented by Recurrent Attention-based Uncertainty Quantification (RAUQ) at inference, improves factual accuracy and effectively detects hallucinations through risk-coverage analysis.

1 Key Information to include

- TA mentor: Melody Ma
- External collaborators: No
- Sharing project: No

2 Approach

We adopt a three-stage approach grounded in the hypothesis that hallucinations in Small Language Models (SLMs) stem from weak reasoning and poor uncertainty calibration.

Reasoning Distillation via CoT We begin by distilling reasoning capabilities from a large teacher model into our SLM. Using Fine-tune-CoT [1], we generate reasoning traces on HotpotQA via Zero-shot-CoT prompting. These traces are filtered for correctness and used to LoRA fine-tune Qwen2.5-7B-Instruct. This process conditions the model to follow structured, multi-hop inference patterns. We posit that explicit reasoning steps sharpen the model's next-token probability distributions, naturally reducing output entropy compared to direct-answer generation. We wrote our own code for this method, basing it off the original paper's methods. We generated a preliminary set of 1000 reasoning traces and performed Fine-tune-CoT, but have yet to evaluate this approach due to time constraints.

Entropy-Aware Reinforcement Learning (RLVF) To directly address calibration, we apply reinforcement learning with verifier-based feedback (RLVF). We define a confidence metric based on normalized token-level entropy: $\text{conf} = 1 - (H_{\text{avg}} / \log |V|) \in [0, 1]$. To distinguish between uncertainty and hallucination, we employ a DeBERTa-v3-large NLI model to score factuality ($f \in [-1, 1]$). The

reward function multiplicatively combines factuality and confidence: $R(x, y) = f \cdot (\lambda_{\text{base}} + \lambda_{\text{conf}} \cdot \text{conf})$. This formulation is critical for calibration: while confident correct answers receive the highest reward, confident hallucinations (where f is negative and confidence is high) incur the steepest penalties. We optimize this objective via REINFORCE with a batch-mean baseline, updating the policy (using LoRA) to align high confidence with factual correctness.

Orthogonal Uncertainty Quantification via RAUQ Finally, we integrate Recurrent Attention-based Uncertainty Quantification (RAUQ) [2] to detect hallucinations that may bypass output-entropy filters. While our RL objective minimizes output uncertainty, RAUQ analyzes internal processing uncertainty. We dynamically identify "uncertainty-aware" attention heads—those maximizing attention to the preceding token—and calculate a recurrent confidence score ($\alpha = 0.2$). This allows us to detect cases where the model generates low-entropy (confident) tokens despite high internal uncertainty states, serving as a robust secondary signal for hallucination detection.

Baselines:

- Out-of-the-box Qwen2.5-7B-Instruct [3]
- Qwen2.5-7B-Instruct, supervised finetuned on HotPotQA training data (10k samples)
- Qwen2.5-7B-Instruct, Fine-tune-CoT on HotPotQA with teacher Qwen3-235B-A22B-Instruct-2507 (no RL)

We also compare our results to SoTA performance on HotPotQA, both open- and closed-book (See Table 1 in the Appendix).

3 Experiments

Data: We use HotPotQA [4] in the closed-book distractor setting. We utilized distinct subsets of the training split for SFT ($N = 10k$), reasoning trace generation ($N = 1k$), and RLVF ($N = 1k$) to prevent leakage. Evaluation is performed on the first 1,000 validation samples.

Evaluation method: We evaluate performance using Exact Match (EM) and F1 Score (F1). To evaluate the quality of our uncertainty estimation, we utilize Risk-Coverage Analysis via rejection curves. By sorting model responses by their RAUQ score [2] and iteratively rejecting the most uncertain samples, we measure if the model successfully flags its own hallucinations (indicated by monotonically increasing accuracy in the retained set).

Experimental details: We evaluated both the base Qwen2.5-7B-Instruct and a version LoRA-finetuned on 10,000 HotpotQA samples. We calculated RAUQ scores during generation with a fixed $\alpha = 0.2$, balancing token probability with attention weights [2].

Results: Quantitative results are reported in the Appendix (Table 1). Figures 1 and 2 (see Appendix A) visualize the correlation between RAUQ uncertainty and model accuracy. As shown in the rejection curves, rejecting the top $x\%$ of samples with highest RAUQ scores consistently improves the EM and F1 of the retained samples. This confirms that high RAUQ scores are strongly correlated with incorrect answers (hallucinations).

4 Future work

To refine our uncertainty quantification, we plan to implement a **novel cascading framework** that synergizes the computational efficiency of RAUQ with the robustness of Semantic Entropy (SE). Since RAUQ operates in a single pass while SE requires expensive sampling, we will use RAUQ as a lightweight "triage" filter, triggering the computationally intensive SE check only when internal uncertainty exceeds a calibrated threshold. Additionally, we will conduct ablation studies to optimize the mixing hyperparameter α through grid search, ensuring the metric is tuned specifically for the distribution of the HotpotQA dataset.

Additionally, we will scale our reasoning trace generation to the full dataset to support robust Chain-of-Thought distillation. We will proceed with RLVF training on the closed-book setting, experimenting with reward shaping to balance helpfulness and factuality. Finally, we will extend our evaluation to the full HotpotQA distractor validation set and, time permitting, assess transferability to open-book settings where external knowledge retrieval is permitted.

References

[1] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers, 2023.

[2] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms, 2025.

[3] Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[4] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018.

A Appendix: Detailed Results and Figures

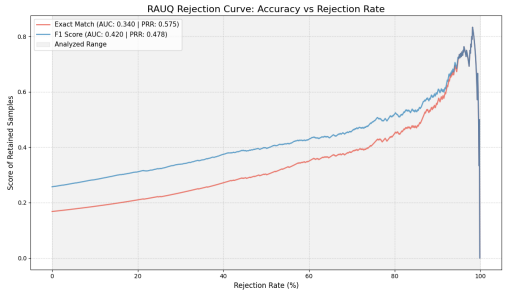

Figure 1: Vanilla Qwen2.5-7B Rejection Curve

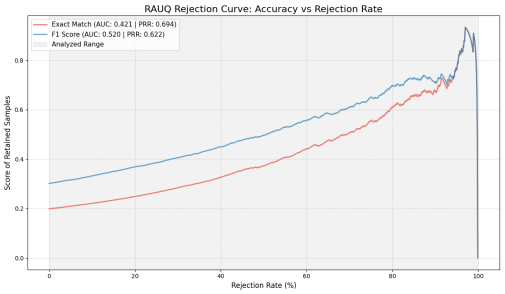

Figure 2: Finetuned Rejection Curve

Figure 3: RAUQ Rejection Curves Comparison. Both models show improved accuracy as high-uncertainty samples are rejected, with the finetuned model achieving higher overall retention quality.

Table 1: Comparison of Experimental and State-of-the-Art Results on the HotpotQA Dataset (EM and F1 scores are percentages).

Model	**Setting / Context**	**EM**	**F1**	**Avg. RAUQ**
Experimental Results ($N = 1000$ Samples)				
Qwen2.5-7B-Instruct	Closed-Book	16.80	25.73	1.39
Qwen2.5-7B-Instruct (SFT)	Closed-Book	19.90	30.15	1.74
Qwen2.5-7B-Instruct (Fine-Tune-CoT)	Closed-Book	TBD	TBD	TBD
Qwen2.5-7B-Instruct (SFT + RLVF)	Closed-Book	**20.20**	**30.02**	1.74
Qwen2.5-7B-Instruct (Fine-Tune-CoT + RLVF)	Closed-Book	TBD	TBD	TBD
SOTA Closed-Book Results (RECITE Paper)				
Codex-002	Closed-Book	37.11	48.37	N/A
PaLM-62B	Closed-Book	26.46	35.67	N/A