G4: LLM Hallucination Detection and Mitigation Leveraging Mechanistic Signals in GRPO Pipelines

Group Members: Deema Alnuhait, Krishnaveni Unnikrishnan, Rudhi Bashambu, Krishna Hota

Introduction: The emergence of Large Language Models (LLMs) has transformed NLP by enabling fluent and human-like text generation. However, these models often produce hallucinations, confidently generating incorrect or unverifiable information. Hallucinations can be intrinsic (contradicting the input) or extrinsic (introducing fabricated content). Studies by Xu et al. [12] show that hallucination results from the probabilistic nature of generation, making complete elimination unlikely. Mitigating hallucinations remains a major challenge in aligning LLMs with real-world factuality

Motivation: LLMs are often trained to prioritize fluency and human preference, but this can lead to outputs that sound natural yet lack factual accuracy. Reinforcement learning methods like GRPO [8] and RLHF [7] may exacerbate this by over-rewarding style over truth. While inference-time methods such as retrieval, reranking, and self-correction aim to reduce hallucinations, they act too late, after unreliable patterns are already learned. Our goal is to move hallucination mitigation into the training phase, using the model's own internal representations rather than external data. Recent work by Chuang et al. [1] shows that earlier transformer layers tend to encode more factual knowledge. We propose that agreement between early and final layer representations can act as a self-supervised signal for factuality.

Data and Environment: We will use standard QA and reasoning datasets like TruthfulQA [6], GSM8K [2] etc for evaluating hallucinations and factuality. For experimentation, we will use open-source LLMs such as LLaMA, Qwen and Deepseek. All experiments will be conducted within the `verl`[9] library, which we will use to develop and test our RL pipeline.

Plan of Work: At each decoding step, we compute token distributions from both an earlier layer ($p^{(k)}$) and the final layer ($p^{(L)}$). Their agreement is computed via KL divergence or token-level consistency. This signal is integrated into GRPO as a shaped reward: $R = R_{\text{task}} + \lambda C$, where C is the agreement signal. A regularization term may also be added: $\mathcal{L}_{\text{reg}} = \beta \, \mathrm{KL}(p^{(L)} \| p^{(k)})$. We will evaluate the impact of layer agreement on hallucination reduction, perform ablations over layer choice and signal type, and compare with standard GRPO.

Related Work: Hallucination mitigation in LLMs [4] has been explored across the model lifecycle. At the data stage, filtering and model editing [10] reduce unreliable patterns during pretraining. During training, RLHF [7] and GRPO [8] align models with human preferences, but flawed reward models may still reinforce fluent yet incorrect outputs. Recent mechanistic RL [11] leverages internal signals (e.g., neuron activations, reasoning depth) to guide training without human feedback. At inference, retrieval-augmented generation (RAG) [5] supplies external knowledge, while post-editing methods like Chain-of-Verification [3] improve factuality but remain costly and reactive. Of particular relevance, DoLa [1] contrasts early and late layer token distributions to suppress hallucinations, though it operates only at inference and does not impact training.

Anticipated Challenges: Key challenges include selecting the appropriate intermediate layer, quantifying agreement, and balancing task reward with the agreement signal to avoid performance degradation. We will address these through empirical tuning and ablations on (λ, β), and by testing whether the agreement term is best integrated into GRPO as part of the reward or as a regularizer (an important design choice for stable and effective training).

References

[1] Yung-Sung Chuang et al. "DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=Th6NyL07na.

[2] Karl Cobbe et al. "Training Verifiers to Solve Math Word Problems". In: *CoRR* abs/2110.14168 (2021). arXiv: 2110.14168. URL: https://arxiv.org/abs/2110.14168.

[3] Shehzaad Dhuliawala et al. "Chain-of-Verification Reduces Hallucination in Large Language Models". In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 3563–3578. DOI: 10.18653/v1/2024.findings-acl.212. URL: https://aclanthology.org/2024.findings-acl.212/.

[4] Lei Huang et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Trans. Inf. Syst.* 43.2 (Jan. 2025). ISSN: 1046-8188. DOI: 10.1145/3703155. URL: https://doi.org/10.1145/3703155.

[5] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[6] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *CoRR* abs/2109.07958 (2021). arXiv: 2109.07958. URL: https://arxiv.org/abs/2109.07958.

[7] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

[8] Zhihong Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. URL: https://arxiv.org/abs/2402.03300.

[9] Guangming Sheng et al. "HybridFlow: A Flexible and Efficient RLHF Framework". In: *arXiv preprint arXiv: 2409.19256* (2024).

[10] Anton Sinitsin et al. "Editable Neural Networks". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=HJedXaEtvS.

[11] Zhongxiang Sun et al. *Detection and Mitigation of Hallucination in Large Reasoning Models: A Mechanistic Perspective*. 2025. URL: https://arxiv.org/abs/2505.12886.

[12] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. *Hallucination is Inevitable: An Innate Limitation of Large Language Models*. 2025. arXiv: 2401.11817 [cs.CL]. URL: https://arxiv.org/abs/2401.11817.