Hallucination of LLMs: Fine-tuning methods to prevent extrinsic hallucinations

Elton Chun-Chai Li
cceli@connect.ust.hk

Abstract

Large Language Models (LLMs) like GPT and LLaMA have advanced natural language processing but often produce hallucinations—fabricated or inconsistent outputs—limiting their reliability in critical domains such as healthcare and legal services. This review examines fine-tuning methodologies to enhance LLM factual accuracy, categorizing approaches into Self-Aware and Self-Supervised Fine-Tuning. Training-based methods show promise in fostering model honesty but face single objective limitations. Building on these insights, this report proposes extending the Multi-Objective Direct Preference Optimization (MODPO) framework to address multiple factuality dimensions—such as temporal, structural, and multilingual—enabling targeted fine-tuning for enhanced accuracy, paving the way for dependable LLM deployment in high-stakes applications.

1 Introduction

Large Language Models (LLMs) such as GPT [1] and LLaMA [20] have significantly advanced natural language processing (NLP), enabling applications from digital assistance to automated content creation [8]. Despite their strengths, LLMs often produce hallucinations—outputs that are fabricated, unfaithful, or inconsistent with the provided context or their pre-training data [5, 17, 6].

Hallucinations in LLMs can be classified into two types [22]: *in-context hallucinations*, where outputs are inconsistent with the immediate context, and *extrinsic hallucinations*, where outputs are not grounded in the model's pre-training knowledge. These inaccuracies pose significant risks in critical domains such as healthcare, legal advice, and education, where factual reliability is paramount [8].

Enhancing factuality in LLMs is essential to ensure their safe and reliable deployment. Factuality awareness training aims to improve models' ability to distinguish verified facts from uncertain information and encourage acknowledgment of knowledge limitations.

This literature review focuses on fine-tuning methodologies for enhancing LLM factuality. We first review current fine-tuning techniques. Next, we explore emerging innovations in this area, and finally, we discuss the challenges and future directions for improving factual awareness through fine-tuning strategies.

2 Preliminaries

This section overviews three primary methodologies for fine-tuning Large Language Models (LLMs): Supervised Fine-Tuning (SFT), Proximal Policy Optimization (PPO), and Direct Preference Optimization (DPO).

Preprint. Under review.

2.1 Supervised Fine-Tuning (SFT)

SFT [25] represents the most straightforward approach, where a pre-trained language model is further trained on human-annotated datasets to improve response effectiveness. This method involves adjusting model weights based on labeled inputs and desired outputs, thereby enhancing the model's ability to generate accurate and contextually appropriate responses. While effective for basic alignment, SFT relies heavily on the quality and coverage of the training data. Limitations include potential overfitting to the training set and insufficient generalization to unseen scenarios, which may still result in hallucinations in novel contexts.

2.2 Reinforcement Learning with Human Feedbacks w/ Proximal Policy Optimization (PPO)

PPO operates within the Reinforcement Learning from Human Feedback (RLHF) framework, extending beyond basic supervised learning. The process involves initial SFT, followed by training a reward model based on human preferences, and finally optimizing the model's policy to maximize this reward [18]. PPO leverages policy gradients to iteratively adjust the model's parameters, balancing the trade-off between exploring new responses and exploiting known effective strategies. While PPO has shown impressive results in aligning model outputs with human values, it introduces significant computational complexity and requires careful tuning to maintain training stability. Additionally, the reliance on human feedback for reward modeling can limit scalability.

2.3 Direct Preference Optimization (DPO)

DPO [16] offers a more efficient alternative to PPO by directly incorporating human preferences into the training objective. By eliminating the need for explicit reward modeling and reinforcement learning loops, DPO significantly simplifies the training pipeline while achieving comparable or better results. This approach adjusts the likelihood of preferred responses over less desirable ones, streamlining the fine-tuning process. DPO demonstrates particular promise for practical applications where computational resources are limited or rapid iteration is necessary. However, ensuring the quality and representativeness of preference data remains a critical challenge.

These methodologies play pivotal roles in fine-tuning LLMs, each offering distinct advantages: SFT provides foundational accuracy through labeled data, while PPO and DPO incorporate human feedback more directly, with DPO offering significant advantages in terms of computational efficiency and training stability.

3 Training-Based Approaches

3.1 Self-Aware Fine-Tuning

Self-aware fine-tuning [9] aims to enhance the factual reliability of Large Language Models (LLMs) by enabling them to recognize and communicate their knowledge boundaries. This approach focuses on aligning LLMs with human values of honesty—hereafter referred to as *alignment for honesty*—and factual accuracy, ensuring that models refuse to answer when lacking sufficient knowledge, thereby mitigating the risk of generating fabricated information.

Alignment for Honesty by Yang et al. [23] introduces a framework designed to align LLMs with the human value of honesty. The primary objective is to ensure that models can distinguish between questions they can answer correctly and those they cannot, thereby enhancing user trust. The authors propose "idk responses" such as "I don't know" to express uncertainty. Their methodology includes both prompt-based approaches and supervised fine-tuning using labeled examples. The key findings demonstrate that supervised fine-tuning methods significantly improve honesty scores without adversely affecting model accuracy. Notably, the MULTISAMPLE method achieved the highest honesty score, indicating the effectiveness of training on multiple sampled responses. However, the approach relies on simplifying assumptions about knowledge boundaries and primarily focuses on acknowledging limitations without addressing other aspects of honesty.

In a similar vein, Cheng et al. [2] explore the development of truthful AI assistants that can identify and communicate their knowledge gaps. Their approach involves constructing an "I Don't Know" (Idk) dataset based on the model's performance on knowledge-intensive questions, followed by alignment techniques including supervised fine-tuning and preference-aware optimization. The results indicate a substantial reduction in hallucinations and improved user trust, with larger models exhibiting better knowledge awareness. Nonetheless, the method simplifies knowledge representation into binary classifications and depends heavily on the quality of the underlying datasets.

Zhang et al. [24] present Refusal-Aware Instruction Tuning (R-Tuning), which aims to reduce hallucinations by aligning instruction tuning data with the model's parametric knowledge. R-Tuning involves identifying knowledge gaps by categorizing data into certain and uncertain sets, augmenting them with uncertainty expressions, and fine-tuning the model accordingly. The study finds that R-Tuning significantly enhances the model's ability to refuse answering unknown questions while improving accuracy on known ones. However, R-Tuning employs binary confidence expressions and relies on accurate knowledge gap identification, which may not fully capture the model's nuanced knowledge representation.

Wan et al. [21] propose Knowledge Consistent Alignment (KCA) to address hallucinations by ensuring consistency between the training data's knowledge and the model's internal knowledge. KCA involves classifying instruction-response pairs based on knowledge requirements, generating reference knowledge snippets, and fine-tuning the model using strategies like Open-Book Tuning and Refusal Tuning. Their findings reveal a strong correlation between knowledge inconsistency and hallucination rates, with KCA effectively reducing hallucinations across various LLMs. However, the approach depends on a well-aligned LLM for knowledge verification and may incur significant computational costs, particularly for long-form generation tasks.

Comparative Insights: While all four studies emphasize the importance of enabling LLMs to recognize their knowledge limitations, they diverge in their methodological executions. Yang et al. [23] and Cheng et al. [2] focus on supervised fine-tuning with explicit refusal responses, whereas Zhang et al. [24] introduce R-Tuning to align instruction data with parametric knowledge. Wan et al. [21] extend this by ensuring knowledge consistency through Knowledge Consistent Alignment (KCA). These variations highlight a spectrum of strategies from direct response modification to comprehensive knowledge verification, each contributing uniquely to enhance factuality awareness.

Future Directions: Building on these findings, future research could explore more nuanced confidence expressions beyond binary classifications and develop advanced methods for accurately mapping knowledge boundaries. Integrating reinforcement learning-based self-aware fine-tuning could also provide deeper insights into strategic alignment of LLM responses. Furthermore, combining self-aware fine-tuning with other approaches, such as self-supervised techniques, may offer synergistic benefits in enhancing LLM factuality and reliability.

3.2 Self-Supervised Fine-Tuning

Self-supervised fine-tuning leverages the internal knowledge of Large Language Models (LLMs) to enhance their factual accuracy without relying on external annotations or human supervision. This approach capitalizes on the models' inherent capabilities to self-evaluate and optimize their responses, thereby mitigating the generation of inaccurate or "hallucinated" information.

Factuality Tuning by Tian et al. [19] introduces a method called "factuality tuning" aimed at improving the factual accuracy of LLMs in long-form text generation without necessitating human fact-checking. The methodology first employs GPT-3.5 to extract atomic claims from long-form generations and creates verification questions for each claim. Multiple answers are then sampled from the LLM to calculate consistency scores across these answers. The approach uses Direct Preference Optimization (DPO) to fine-tune models based on preference pairs ranked by their average consistency scores. The key findings demonstrate that factuality tuning significantly improves factual accuracy across benchmark datasets, outperforming traditional Reinforcement Learning from Human Feedback (RLHF) and decoding strategies.

Zhang et al. [26] present a self-alignment framework that enables LLMs to evaluate the factuality of their own responses, thereby reducing hallucinations. The framework operates by generating responses and employing self-evaluation to verify each claim, followed by DPO for optimization. The study reveals that this self-evaluation approach significantly enhances factual accuracy across various tasks. Despite its effectiveness, the framework relies on external models for claim extraction, introducing potential dependencies.

FLAME: Factuality-Aware Alignment by Lin et al. [11] takes a more straightforward approach by directly prompting LLMs to generate responses for fact-based instructions, which are then used for fine-tuning. The method involves sampling responses from the fine-tuned model and using the model itself for assessment, followed by DPO for further optimization. The findings indicate that FLAME significantly improves factual accuracy without compromising the ability to follow instructions, as evidenced by increased FActScores on benchmark datasets. However, the method's effectiveness is contingent on accurate instruction classification and may encounter challenges when scaling to larger models or diverse tasks.

Comparative Insights: The reviewed studies collectively demonstrate the potential of self-supervised fine-tuning in enhancing the factual accuracy of LLMs through internal evaluation mechanisms. Common methodologies involve generating responses, assessing their factuality through automated or self-derived metrics, and fine-tuning models using DPO. However, challenges persist, including reliance on external models for tasks like claim extraction and limitations in self-evaluation capabilities. Recent studies by Huang et al. [4] and Jiang et al. [7] have raised important concerns about the reliability of LLMs' self-evaluation abilities, suggesting that current self-supervised fine-tuning approaches may need further refinement to ensure robust factual improvements.

4 Conclusion

This literature review has explored various fine-tuning methodologies aimed at enhancing the factual accuracy of Large Language Models. Supervised techniques like SFT provide foundational improvements through labeled data, while reinforcement learning approaches such as PPO and DPO offer avenues for integrating human feedback more directly. Training-based methods, including self-aware and self-supervised fine-tuning, present innovative strategies for enabling models to recognize and mitigate their knowledge limitations.

The findings reveal that training-based approaches demonstrate considerable potential in enhancing model honesty and factual reliability, although they face inherent limitations in single-objective optimization and nuanced knowledge representation. While significant advances have been made, the challenge of achieving scalable and multi-objective factual reliability persists. Future research should prioritize three critical directions: the development of sophisticated evaluation frameworks, the exploration of hybrid methodologies that synthesize multiple training approaches, and the integration of multi-objective fine-tuning techniques. This review underscores the crucial importance of innovative fine-tuning strategies to ensure robust and reliable deployment of large language models across critical application domains. Such progress will be fundamental to establishing trust and maintaining precision in high-stakes scenarios where factual accuracy is paramount.

5 Research Plan

Existing fine-tuning approaches predominantly treat factual knowledge as a monolithic concept, potentially overlooking the nuanced nature of different types of factual information. Recent work has proposed that factual knowledge in Large Language Models (LLMs) can be categorized into distinct dimensions, like multifaceted, structural, and temporal [3]. This categorization suggests that a more targeted and specialized approach to fine-tuning could yield better results than current methods that treat all factual knowledge uniformly. By addressing these distinct facets, we aim to enhance the factuality of LLMs, thereby improving their applicability across various domains.

6 Research Objectives

Recent studies, such as Hu et al. [3], have extended the scope of factual knowledge to include multifaceted, structural, adversarial, temporal, real-world, domain-specific, and multilingual aspects, these works often treat all types as a single entity with a unified objective. This approach may not fully capture the intricacies and specific requirements of each factuality dimension, potentially limiting the effectiveness of fine-tuning methods in enhancing LLMs' factual awareness. To address this gap, this research aims to extend the Multi-Objective Direct Preference Optimization (MODPO) framework [27] to handle multiple distinct factuality categories. By fine-tuning LLMs with specialized objectives for each factuality category, we seek to improve their factual accuracy and reliability. The key objectives of this research are threefold. First, we intend to adapt the existing MODPO framework to manage multiple factuality dimensions, enabling specialized optimization tailored to each factuality type. Second, we aim to enhance the model's factual accuracy across various dimensions, including temporal facts, compositional facts, cross-domain reasoning, and region-specific knowledge, thereby ensuring robust performance in diverse applications. Additionally, comprehensive evaluations using diverse datasets will be conducted to benchmark the improvements and provide theoretical insights into the benefits and challenges of multi-objective optimization in LLM fine-tuning.

7 Proposed Methodology

This research extends the Multi-Objective Direct Preference Optimization (MODPO) framework to enhance the factual accuracy of Large Language Models (LLMs) across distinct factuality categories. By categorizing factual knowledge into different types, like temporal facts, domain specific facts, and multi-lingual fact, we perform targeted fine-tuning that addresses the unique challenges of each category.

7.1 Framework Extension and Factuality Categorization

Traditional Direct Preference Optimization (DPO) frameworks optimize a single objective, typically focusing on overall factual accuracy without distinguishing between different types of factual information. This monolithic approach can overlook the nuanced requirements of various factuality dimensions, potentially limiting the model's reliability in specialized contexts. To overcome this limitation, we extend the DPO framework to accommodate multiple objectives, each corresponding to a specific type of factual knowledge. This transition allows for simultaneous optimization across various factuality categories, enabling more targeted and effective fine-tuning.

Building upon the Pinocchio dataset [3], factual knowledge is categorized into seven distinct types: multifaceted, structural, adversarial, temporal, real-world, domain-specific, multi-lingual. The seven factual domains assess different dimensions of factual knowledge in LLMs. The Multifaceted domain examines factual accuracy when combining multiple pieces of information from various sources, testing the model's ability to maintain facts across complex information networks. The Structural domain evaluates factual precision when processing organized data like tables and databases, ensuring accuracy translates across different data formats. The Adversarial domain tests factual robustness by examining how well models maintain facts in the face of deliberately misleading information. The Temporal domain assesses factual currency, focusing on how models handle evolving facts and changing information over time. The Real-World domain measures factual verification capabilities by checking how accurately models validate claims against diverse real-world knowledge sources. The Domain-Specific area evaluates factual expertise in specialized fields, where technical and scientific accuracy is paramount. Finally, the Multi-Lingual domain examines factual consistency across languages, ensuring facts remain preserved regardless of linguistic expression. Together, these domains comprehensively assess an LLM's ability to maintain factual accuracy across different contexts and challenges.

7.2 MODPO Loss Function and Implementation

The core innovation lies in adapting the DPO loss function to handle multiple objectives. The MODPO loss function integrates distinct alignment objectives by introducing specific weights for each factuality category, allowing the model to balance and optimize diverse factuality dimensions effectively. The loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{MODPO}}(\pi_{\theta,\mathbf{w}}; \mathbf{r}_\phi, \pi_{\text{sft}}, D) = -\mathbb{E}_{(x,y_w,y_l)\in\mathcal{D}} \Bigg[\log\sigma\Bigg(\sum_{k=1}^{K} \beta w_k \log\frac{\pi_{\theta,\mathbf{w}}(y_w|x)}{\pi_{\text{sft}}(y_w|x)} \\ - \sum_{k=1}^{K} \beta w_k \log\frac{\pi_{\theta,\mathbf{w}}(y_l|x)}{\pi_{\text{sft}}(y_l|x)} - \sum_{k=1}^{K} \frac{1}{w_k}\mathbf{w}^\top_{-k}\left(\mathbf{r}_{\phi,-k}(x,y_w) - \mathbf{r}_{\phi,-k}(x,y_l)\right)\Bigg)\Bigg]\end{aligned}$$

Here, $\mathbf{w}$ denotes the weight vector for each factuality objective, and β controls the influence of the Kullback-Leibler (KL) divergence penalty. By incorporating weighted terms for each factuality category, MODPO effectively combines multiple reward signals into a unified loss function, ensuring balanced optimization across all factuality categories.

The implementation involves segmenting the Pinocchio dataset into the defined factuality categories, generating preference pairs (y_w, y_l) for each category using the FLAME SFT model [11], and assigning weights w_k based on importance or complexity. The pre-trained LLM is then fine-tuned using this MODPO loss function, employing gradient descent-based optimization techniques to minimize the loss across all objectives.

7.3 Assumptions and Expected Properties

This methodology operates under several key assumptions. Firstly, each factuality category can be treated as an independent objective without significant overlap. Secondly, the preference pairs generated accurately reflect the factual accuracy of the responses, ensuring reliable fine-tuning outcomes. Thirdly, the MODPO framework is expected to scale effectively with the number of factuality types without introducing prohibitive computational overhead, maintaining efficiency throughout the fine-tuning process.

The fine-tuned LLM is anticipated to demonstrate enhanced factual accuracy across all specified dimensions, increased robustness in handling diverse factuality challenges, and the ability to accommodate additional factuality types as needed. Balanced optimization ensures that improvements in one factuality category do not negatively impact others, resulting in a well-rounded and reliable language model. Compared to existing fine-tuning methods, MODPO offers targeted optimization by addressing each factuality type individually, leading to more precise fine-tuning. It maintains computational efficiency relative to multi-objective Reinforcement Learning from Human Feedback (RLHF) while achieving superior performance. Additionally, its flexibility allows for the easy integration of additional objectives, making it adaptable to evolving factuality requirements. Preliminary studies [27] indicate that MODPO outperforms standard DPO and multi-objective RLHF in enhancing factual accuracy.

8 Experimental Design

The proposed methodology will be validated through comprehensive experiments designed to assess the factual accuracy of the fine-tuned LLM across various datasets and factuality dimensions. Quantitative evaluations will be conducted using established metrics to compare the performance of the MODPO approach against baseline models.

8.1 Datasets

This research will utilize a diverse set of datasets to ensure a thorough evaluation of the model's factual accuracy across different contexts and domains. The Biography dataset [12], Alpaca Fact [11], FAVA [14], MedQA [15], and TruthfulQA [10] will be employed. These datasets encompass a wide range of domain, allowing for a comprehensive assessment of the model's performance across various factuality contexts and domains. However, these dataset includes both factual and non-factual questions. To ensure fairness, a SFT from FLAME will be used to extract all the factual question only, which is same as the setting at FLAME.

8.2 Evaluation Metrics

The primary metric for assessing performance will be the FActScore (FS) [13], designed to evaluate the factual accuracy of long-form text generated by LLMs. FS provides a fine-grained assessment of factual correctness across various dimensions, enabling a detailed comparison between the proposed MODPO approach and baseline methods.

8.3 Baseline Methods

The MODPO approach will be compared against several baseline methods, including FLAME [11], the standard Direct Preference Optimization (DPO) method, which treats factuality as a single objective, and Multi-objective Reinforcement Learning from Human Feedback (RLHF) [27], a multi-objective reinforcement learning approach for language model alignment. These comparisons will highlight the relative performance gains achieved by the proposed method.

8.4 Ablation Studies

To understand the contribution of each component in the proposed methodology, ablation studies will be conducted. These studies will involve fine-tuning the model under varying configurations to isolate the effects of different elements. Initially, all factuality knowledge will be treated as a single category, and the model will be fine-tuned using the original DPO approach. Subsequently, the model will be fine-tuned using subsets of factuality types, such as three and five types, respectively, to assess the impact of each additional objective. Finally, the model will undergo full multi-objective DPO fine-tuning using all seven factuality types to evaluate the cumulative effect on factual accuracy. These ablation studies will provide insights into the effectiveness of each factuality category and the overall impact of multi-objective optimization on the model's performance.

8.5 Implementation Details

The implementation of the proposed methodology will utilize machine learning frameworks such as PyTorch or TensorFlow. Llama2 70B will serve as the base for fine-tuning, and the FLAME's model will be employed for generating preference pairs. Custom scripts will be developed to handle the multi-objective DPO process, including the separation of datasets, generation of preference pairs, and integration of multiple objectives during fine-tuning. The experimental setup will also involve configuring the models to handle the distinct factuality objectives effectively, ensuring that each factuality type is appropriately addressed during the optimization process.

8.6 Key Optimization Strategies

To optimize performance and efficiency, several strategies will be employed. Learning rate scheduling will be implemented to ensure stable and efficient training by dynamically adjusting the learning rates based on the training progress. Mixed-precision training techniques will be utilized to accelerate the training process while maintaining model accuracy, leveraging lower-precision computations where appropriate. Gradient clipping will be applied to prevent exploding gradients and ensure stable optimization. Additionally, parallel processing techniques will be leveraged to expedite the fine-tuning process across multiple objectives, ensuring that the multi-objective optimization is performed efficiently without significant delays.

8.7 Expected Results and Potential Challenges

The MODPO fine-tuned model is anticipated to outperform the FLAME benchmark in factual accuracy across multiple datasets. By addressing each factuality type individually, the fine-tuned model should demonstrate enhanced reliability and versatility in generating factual information across diverse contexts. These improvements will be quantitatively validated through superior FActScore (FS) metrics.

Several challenges may arise during the research. Balancing multiple objectives is critical; ensuring that the fine-tuning process effectively balances all factuality types without compromising any single aspect is essential for achieving comprehensive improvements. Computational overhead is another concern, as the increased complexity of multi-objective fine-tuning may lead to higher computational requirements, necessitating efficient resource management strategies. Additionally, ensuring data quality and coverage is paramount; the selected datasets and generated preference pairs must comprehensively cover all factuality types to provide a robust foundation for fine-tuning.

To address these challenges, dynamic weighting schemes will be implemented to effectively balance objectives, ensuring that no single factuality type disproportionately influences the optimization process. The training pipeline will be optimized to manage computational resources efficiently, potentially through techniques such as mixed-precision training and parallel processing. Thorough data preprocessing and augmentation will be conducted to enhance data quality and coverage, ensuring that all factuality types are adequately represented and addressed during fine-tuning.

9 Expected Outcomes and Impact

The proposed MODPO fine-tuning approach is expected to significantly enhance the factual accuracy of LLMs across multiple dimensions, outperforming existing approaches such as FLAME. By addressing each factuality type individually, the fine-tuned model should demonstrate improved reliability and versatility in generating accurate information across diverse contexts. These improvements will be quantitatively validated through superior FActScore (FS) metrics, thereby establishing the efficacy of the MODPO framework in enhancing LLM factuality.

Enhancing the factual accuracy of LLMs has profound implications for various applications, including education, information dissemination, and other knowledge-dependent fields. Reliable and factually accurate models can support better decision-making, reduce misinformation, and increase trust in AI-driven systems. Furthermore, the multi-objective fine-tuning framework developed in this research could serve as a foundation for future studies aiming to improve other nuanced aspects of LLMs, such as ethical reasoning and contextual understanding, thereby contributing to the development of more robust and reliable AI systems.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners.

[2] Q. Cheng, T. Sun, X. Liu, W. Zhang, Z. Yin, S. Li, L. Li, Z. He, K. Chen, and X. Qiu. Can AI assistants know what they don't know?

[3] X. Hu, J. Chen, X. Li, Y. Guo, L. Wen, P. S. Yu, and Z. Guo. TOWARDS UNDERSTANDING FACTUAL KNOWLEDGE OF LARGE LANGUAGE MODELS.

[4] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. Large language models cannot self-correct reasoning yet.

[5] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. 55(12):248:1–248:38.

[7] D. Jiang, J. Zhang, O. Weller, N. Weir, B. V. Durme, and D. Khashabi. SELF-[IN]CORRECT: LLMs struggle with discriminating self-generated responses.

[8] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. Challenges and applications of large language models.

[9] S. Li, C. Yang, T. Wu, C. Shi, Y. Zhang, X. Zhu, Z. Cheng, D. Cai, M. Yu, L. Liu, J. Zhou, Y. Yang, N. Wong, X. Wu, and W. Lam. A survey on the honesty of large language models.

[10] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods.

[11] S.-C. Lin, L. Gao, B. Oguz, W. Xiong, J. Lin, W.-t. Yih, and X. Chen. FLAME: Factuality-aware alignment for large language models.

[12] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation.

[13] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation.

[14] A. Mishra, A. Asai, V. Balachandran, Y. Wang, G. Neubig, Y. Tsvetkov, and H. Hajishirzi. Fine-grained hallucination detection and editing for language models.

[15] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[16] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model.

[17] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S. M. T. I. Tonmoy, A. Chadha, A. P. Sheth, and A. Das. The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations.

[18] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback.

[19] K. Tian, E. Mitchell, H. Yao, C. D. Manning, and C. Finn. Fine-tuning language models for factuality.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and efficient foundation language models.

[21] F. Wan, X. Huang, L. Cui, X. Quan, W. Bi, and S. Shi. Knowledge verification to nip hallucination in the bud.

[22] L. Weng. Extrinsic hallucinations in llms. *lilianweng.github.io*, Jul 2024.

[23] Y. Yang, E. Chern, X. Qiu, G. Neubig, and P. Liu. Alignment for honesty.

[24] H. Zhang, S. Diao, Y. Lin, Y. R. Fung, Q. Lian, X. Wang, Y. Chen, H. Ji, and T. Zhang. R-tuning: Instructing large language models to say 'i don't know'.

[25] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang. Instruction tuning for large language models: A survey.

[26] X. Zhang, B. Peng, Y. Tian, J. Zhou, L. Jin, L. Song, H. Mi, and H. Meng. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965. Association for Computational Linguistics.

[27] Z. Zhou, J. Liu, J. Shao, X. Yue, C. Yang, W. Ouyang, and Y. Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization.