AI-Generated Text Detection with SVM and LoRA-Finetuned RoBERTa

Marie Yang[1*] **Harris Song**[1*]
marieyang@ucla.edu harris.song@ucla.edu

Anish Pal[1*] **Aditya Patil**[1*]
anishmpal@ucla.edu adityapatil@ucla.edu
[1]Department of Computer Science
University of California, Los Angeles
[*] *Denotes Equal Contribution*

1 Introduction

As large language models (LLMs) like ChatGPT and Claude become increasingly capable and accessible, AI-generated content is rapidly flooding the internet. While these tools have many beneficial applications, they also bring about significant issues, including the spread of misinformation and hallucinations. (Xu et al., 2024). Given the limitations of existing detectors, our goal was to build a more robust and reliable classifier. We trained a traditional Support Vector Machine as a baseline and fine-tuned RoBERTa, a widely used BERT variant, using Low-Rank Adaptation on the HC3 dataset.

2 Modeling Approach

In this section, we outline our modeling strategy for text classification. We begin with a classical baseline presented in subsection 2.1, then progressively adopt more sophisticated neural methods in subsection 2.2 and subsection 2.3; this comparative approach allows us to quantify the improvements gained from recent advancements in language modeling and parameter-efficient fine-tuning.

2.1 Support Vector Machine (SVM)

For our baseline model, we used a support vector machine (SVM), a classical machine learning technique (Cortes and Vapnik, 1995). We compared this approach with a more complex, attention-based approach to see what benefits it would yield in comparison and will discuss this in subsection 5.1.

2.2 Finetuning RoBERTa

We fine-tuned an encoder-only model, RoBERTa (Liu et al., 2020), which is a robustly optimized variant of BERT that significantly outperforms BERT on a number of metrics. For the full model architecture, refer to Figure 1. We also considered variants such as DeBERTa (He et al., 2021), which adds a decoder mask on top of BERT and a better attention mechanism.

Upon further research on existing text classifiers, we found OpenAI's RoBERTa GPT-2 generated text detector (OpenAI) that was released alongside the GPT-2 XL model and fine-tuned on GPT-2 text data. We decided against fine-tuning OAI's existing classifier, however, given that we saw no improvement in accuracy from the OAI baseline of about 87% after some initial training. Instead, we ultimately chose to train RoBERTa base on our own training data.

2.3 Low Rank Adaptation

To improve accuracy and reduce training time and compute, we used Low Rank Adaptation (LoRA) (Hu et al., 2021), shown in Figure 1 to train our model. LoRa freezes the pre-trained weights of the RoBERTa model and inserts low-rank trainable adapters between certain layers. These adapters are small matrices that allow efficient fine-tuning with far fewer parameters and training time, while maintaining performance.

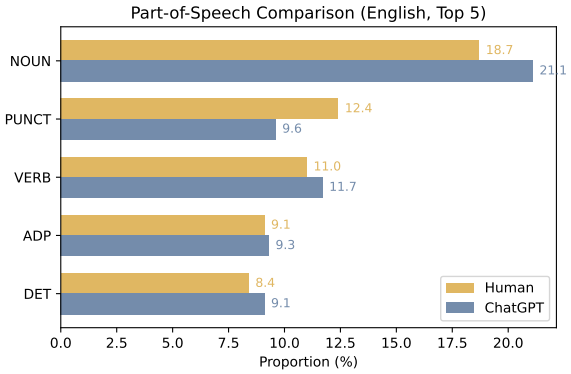

Figure 2: POS Comparison (Marcus et al., 1993) from the HC3 Dataset (Guo et al., 2023b)

3 Dataset Collection

We initially trained on ahmadreza13's 3.6 million-sample dataset on HuggingFace (ahmadreza13,

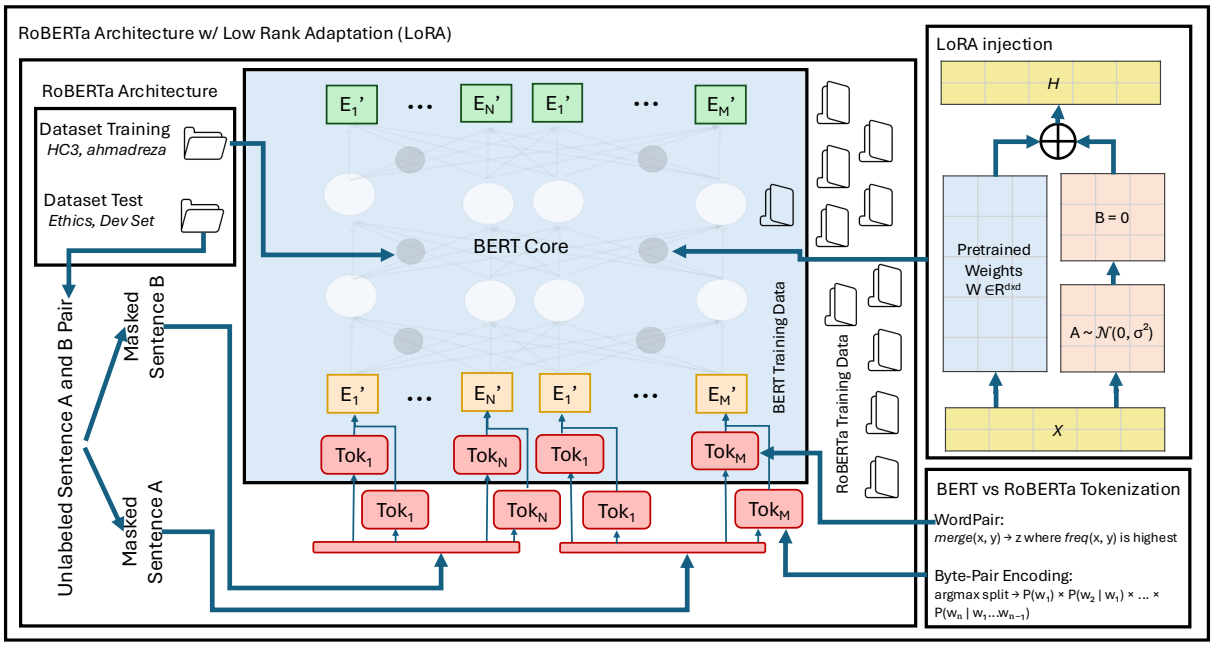

Figure 1: Our architecture w/ Training and Test dataset, including the baseline BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), and LoRA (Hu et al., 2021)

2024). Fine-tuning base RoBERTa on this dataset resulted in extremely poor performance with an accuracy of 0.457. This dataset had no explicit comparison between human and AI-generated texts for the same prompt, and we felt there were higher quality datasets available. Therefore, we chose to switch to the **HC3 dataset** (Guo et al., 2023b).

The **HC3 (Human ChatGPT Comparison Corpus)** dataset is a practical and well-structured database used for analyzing differences between human-written and ChatGPT-generated text. The English portion of the dataset contains roughly 24,300 question–answer pairs, covering five main domains: Reddit Explain Like I'm 5 (ELI5) (∼17,000 examples), finance (∼3,900), medicine (∼1,250), open-domain QA (∼1,190), and computer science/artificial intelligence wiki questions (∼840). Each example consists of a question, one or more human-written responses, and one or more responses generated by ChatGPT.

In addition to its breadth of content, the dataset also exhibits valuable structural patterns which can inform model development. For example, ChatGPT responses generally tend to be longer and more formally structured, often including introductory phrases and carefully worded explanations, while human responses are typically more concise and stylistically varied. In the finance and medicine subsets, human responses frequently reference external knowledge or personal anecdotes, while ChatGPT tends to produce more generalized and less polarizing answers. Stylistically, ChatGPT answers include redundant phrasing or cautionary disclaimers, especially in sensitive domains such as medicine and law.

4 Hyperparameters

We used the following hyperparameters for finetuning roberta-base with LoRA:

Listing 1: LoRA Configuration

```
lora_config = LoraConfig(
    r=8,
    lora_alpha=16,
    target_modules=["query", "value"],  #
    Commonly targeted layers
    lora_dropout=0.1,
    bias="none",
    task_type=TaskType.SEQ_CLS
)
```

Listing 2: Training Arguments

```
training_args = TrainingArguments(
    output_dir=output_dir,
    eval_strategy="epoch",
    save_strategy="epoch",
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    label_names=["labels"],
    weight_decay=0.01,
    logging_steps=50,
    load_best_model_at_end=True,
    metric_for_best_model="accuracy",
    save_total_limit=2,
```

```
    resume_from_checkpoint=True if os.path.
    exists(output_dir) else False)
```

5 Results

5.1 Support Vector Machine Baseline

Training an SVM on our dataset resulted in a decent performance baseline with an accuracy of ∼70%.

Table 1: SVM Classification Report

Class	Precision	Recall	F1-score	Support
AI (0)	0.69	0.70	0.70	12,000
Human (1)	0.70	0.69	0.70	12,000
Accuracy			**0.70**	24,000
Macro Avg	0.70	0.70	0.70	24,000
Weighted Avg	0.70	0.70	0.70	24,000

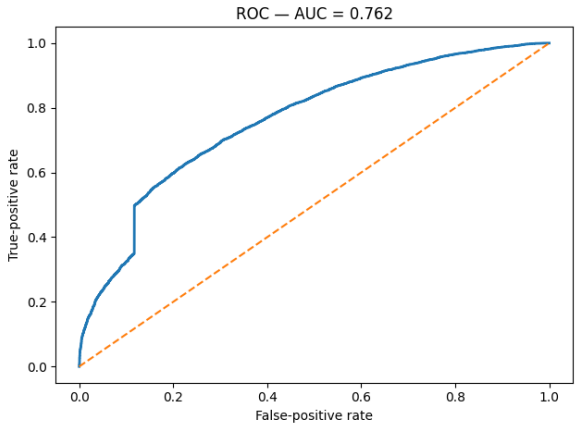

Figure 3: SVM ROC-AUC Curve

Our AUC-ROC (Hanley and McNeil, 1982) curve for our SVM at Figure 3. Area under the curve being 0.762 suggests that the SVM model has a moderate ability to discriminate between AI and human-written text.

5.2 Finetuned RoBERTa

Fine-tuning with LoRA on 48,644 data points in the HC3 dataset gave us a model accuracy of 92% over the entire development set (excluding the ethics dev set).

Table 2: Finetuned RoBERTa Classification Report for Entire Dev Set

Class	Precision	Recall	F1-score	Support
AI (0)	0.8994	0.9392	0.9189	12,000
Human (1)	0.9364	0.8950	0.9152	12,000
Accuracy			**0.9171**	24,000
Macro Avg	0.9179	0.9171	0.9170	24,000
Weighted Avg	0.9179	0.9171	0.9170	24,000

We can see from the table that our model has slightly higher precision for the Human (1) class and higher recall for the AI (0) class. This implies that when the model predicts a text to be human-generated, it is likely to be correct, but it may miss some human examples. On the other hand, the model catches most AI-generated texts, but mistakenly flags some human texts as AI in order to do this. Overall, there is a precision-recall trade-off where the model errs on the side of classifying some human text as AI in order to catch more AI examples.

This trade-off has significant implications for real-world applications of AI detection systems. On one hand, a model that incorrectly flags human work as AI-generated can lead to serious consequences, such as students being falsely accused of plagiarism. On the other hand, failing to detect AI-generated content could undermine the integrity of educational assessments and allow cheating to go undetected. Therefore, striking a careful balance is essential.

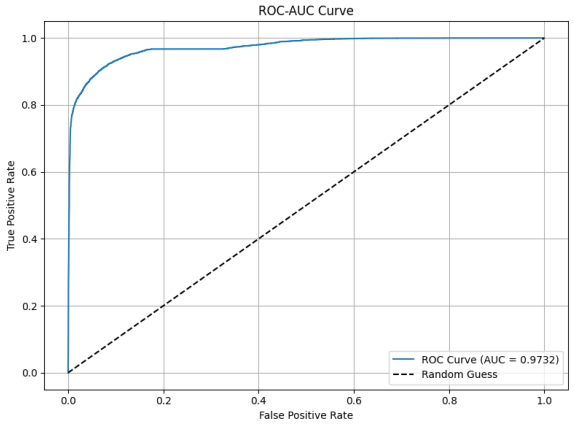

Figure 4: Finetuned RoBERTa ROC-AUC Curve

5.3 Discussion

Our model had outstanding performance on the arxiv_chatgpt dataset, achieving an accuracy of 99%, and lower, relatively similar performances of

around 89% across the other datasets (arxiv_cohere, reddit_chatgpt and reddit_cohere).

This is somewhat surprising. The HC3 dataset consists entirely of ChatGPT-generated text, so we would expect it to perform better on ChatGPT text than Cohere text, which it did. However, a significant portion of the HC3 dataset – 17.1k out of 24.3k samples, or roughly 70% – is made up of text sourced from Reddit. Despite our training dataset being dominated by Reddit text, our model still performs better on arxiv_chatgpt than reddit_chatgpt.

Our hypothesis for this phenomenon is that Reddit texts are simply inherently harder to classify than arXiv texts. While Reddit texts can vary greatly in tone, content and vocabulary, arXiv papers follow a more uniform academic writing style with consistent technical language. This uniformity may make it easier for the model to detect patterns in the arXiv dataset. Furthermore, the HC3 dataset does still include a significant amount of technical texts in the domains of medicine and computer science/artificial intelligence. Table 3, Table 4, Table 5, and Table 6 include our results.

Table 3: Report for `arxiv_chatGPT.jsonl`

Class	Precision	Recall	F1-score	Support
0	1.0000	0.9800	0.9899	3000
1	0.9804	1.0000	0.9901	3000
Accuracy			**0.9900**	6000
Macro Avg	0.9902	0.9900	0.9900	6000
Weighted Avg	0.9902	0.9900	0.9900	6000

Table 4: Report for `arxiv_cohere.jsonl`

Class	Precision	Recall	F1-score	Support
0	1.0000	0.7890	0.8821	3000
1	0.8258	1.0000	0.9046	3000
Accuracy			**0.8945**	6000
Macro Avg	0.9129	0.8945	0.8933	6000
Weighted Avg	0.9129	0.8945	0.8933	6000

Table 5: Report for `reddit_chatGPT.jsonl`

Class	Precision	Recall	F1-score	Support
0	0.8264	0.9997	0.9048	3000
1	0.9996	0.7900	0.8825	3000
Accuracy			**0.8948**	6000
Macro Avg	0.9130	0.8948	0.8937	6000
Weighted Avg	0.9130	0.8948	0.8937	6000

Table 6: Report for `reddit_cohere.jsonl`

Class	Precision	Recall	F1-score	Support
0	0.8247	0.9880	0.8990	3000
1	0.9850	0.7900	0.8768	3000
Accuracy			**0.8890**	6000
Macro Avg	0.9049	0.8890	0.8879	6000
Weighted Avg	0.9049	0.8890	0.8879	6000

5.4 Error Analysis

Our model displayed the ability to distinguish between a human-written paper abstract and a ChatGPT-generated abstract:

LoRA Abstract (Source: LoRA paper)

"An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to particular tasks or domains. As we pre-train larger models, full fine-tuning, which retrains all model parameters, becomes less feasible...."

Prediction: Human (1) **Confidence:** 1.0000

LoRA Abstract (Source: ChatGPT)

"Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning technique for large pre-trained models, particularly transformer-based architectures. Instead of updating all model weights during downstream training, LoRA injects trainable low-rank matrices ..."

Prediction: AI (0) **Confidence:** 0.9757

However, it was unable to predict some human-generated text that had similar tones and content to what you would expect to see from AI-generated text.

Source: Reddit Dev Set

"You most likely subscribe to a progressive view of history as a whole, meaning that society advances linearly (populations become more civilized, more knowledgeable, more advanced, etc, essentially superior to those in the past). You most likely believe (implicitly or explicitly) the purpose of history is to define and bolster the nation state..."

Prediction: AI (0), **True Label:** Human (1)
Confidence: 1.0000

6 Ethics

Our model had lower accuracy on the ethics dev than on the general dev set, but performed almost as well as on the original dev set as it did on the

Hewlett dataset as demonstrated at Table 7. Lower performance on the German Wikipedia dataset was expected in Table 8, as the HC3 dataset we used to train had primarily English text, making our model less likely to perform well on datasets that were in a foreign language. This explains why it performed so well on the hewlett dataset, but not the low performance on TOEFL Table 9. This could have been due to the fact that for TOEFL, people are required to speak very formally and in less of a casual tone, which the model might have associated with being more similar to AI generated text.

The performance of our model on these ethics-specific dev sets highlights the importance of textual diversity in training models. With an internet overrun by English text, it's still important to recognize that many other languages are out there, and that they're vastly underrepresented in model training; this makes it so that many models are biased towards English.

Table 7: Report for `hewlett.json`

Class	Precision	Recall	F1-score	Support
0	1.0000	0.9091	0.9524	88
1	0.0000	0.0000	0.0000	0
Accuracy			**0.9091**	88
Macro Avg	0.5000	0.4545	0.4762	88
Weighted Avg	1.0000	0.9091	0.9524	88

Table 8: Report for `german_wikipedia.jsonl`

Class	Precision	Recall	F1-score	Support
0	0.5325	0.9820	0.6906	500
1	0.8846	0.1380	0.2388	500
Accuracy			**0.5600**	1000
Macro Avg	0.7086	0.5600	0.4647	1000
Weighted Avg	0.7086	0.5600	0.4647	1000

Table 9: Report for `toefl.json`

Class	Precision	Recall	F1-score	Support
0	1.0000	0.4176	0.5891	0
1	0.0000	0.0000	0.0000	91
Accuracy			**0.4176**	91
Macro Avg	0.5000	0.2088	0.2946	91
Weighted Avg	1.0000	0.4176	0.5891	91

7 Related Work

We included related works, and further directions where time constraints and computational bottlenecks are not significant. (Gururangan et al., 2020) is an analysis on how domain-adaptive pretraining, such as pretraining within biomedical fields, or in mechanical engineering, is still useful; future ideas including domain-specific writing styles to identify patterns of authorship. (Jiao et al., 2021) introduces another interesting approach, discussing translation between languages. Fine-tuning techniques from the language translation (German, Chinese, and English) would be useful in edge cases.

Lastly, (Mitchell et al., 2023) is a paper that uses a novel approach to analyze probabilities. We are also able to graph some of the logit-probabilities through the Transformers package (Wolf et al., 2020); a benefit includes a lack of fine-tuning due to a more probabilistic approach rather than a generative approach. A lot of our approaches in subsection 5.2 and subsection 5.1, are focused on training, (Gururangan et al., 2020), (Jiao et al., 2021), and (Mitchell et al., 2023) define theory-based methods.

8 Conclusion

Our experiments demonstrate that fine-tuning RoBERTa with Low-Rank Adaptation (LoRA), as discussed in subsection 2.2, subsection 2.3, subsection 5.2, and modeled with Figure 1 on the HC3 dataset (Guo et al., 2023b) and analyzed in Figure 2 leads to a substantial improvement in detecting AI-generated text compared to traditional machine learning approaches such as SVM, which was extensively discussed in subsection 2.1 and tested in subsection 5.1. In subsection 5.4, we detail specific test cases, and then we performed our experiments on the ethics dataset in section 6. We finally discuss other related work in section 7.

A limitation of our approach is that our model is trained on primarily ChatGPT and Reddit text, clearly established in Table 5. As a result, our architecture is achieves better performances compared to text generated by other LLMs such as Gemini or Claude, or sourced elsewhere, such as arXiV as established with Table 4. We could incorporate multilingual data into our model, such as the HC3 Chinese dataset for inclusivity. (Guo et al., 2023a).

The LoRA-augmented RoBERTa model achieved an overall accuracy of over 91% on the primary evaluation set, with especially high precision in distinguishing human from AI text. In comparison, the SVM baseline reached 70% accuracy, confirming the advantage of transformer-based models for this task.

A Appendix: Contribution Statement

Attached is a contribution statement provided by each author, including the technical and non-technical research work conducted.

Marie Yang: Train RoBERTa base using HC3 dataset, model evaluations on original dev datasets, model visualizations, write Results section

Harris Song: Set up GCP Virtual machine, read papers and write Related Work section, help write and proofread report, create architecture and data diagrams, SVM training and visualizations

Anish Pal: Train OAI RoBERTa classifier using ahmadreza13 dataset, model evaluations on OAI RoBERTa classifier, analyze training datasets and write Data section

Aditya Patil: Finding and preprocessing training datasets, evaluating and analyzing model on ethics dataset, cleaning Github and writing scripts, writing Ethics section

References

ahmadreza13. 2024. human-vs-ai-generated-dataset. https://huggingface.co/datasets/ahmadreza13/human-vs-Ai-generated-dataset. 3.61M samples.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023a. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Jiaxin Guo, Shuning Chang, Ziyu Wang, and Ming Zhou. 2023b. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *Preprint*, arXiv:2301.11305.

OpenAI. Roberta base openai detector. https://huggingface.co/openai-community/roberta-base-openai-detector. Accessed: 2025-05-21.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Software available from https://github.com/huggingface/transformers.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*. Formal proof that LLM hallucination cannot be fully eliminated.