# LLMs are Frequency Pattern Learners in Natural Language Inference

Anonymous ACL submission

## Abstract

While fine-tuning LLMs on NLI corpora improves their inferential performance, the underlying mechanisms driving this improvement remain largely opaque. In this work, we conduct a series of experiments to investigate what LLMs actually learn during fine-tuning. We begin by analyzing predicate frequencies in premises and hypotheses across NLI datasets and identify a consistent **frequency bias**, where predicates in hypotheses occur more frequently than those in premises for positive instances. To assess the impact of this bias, we evaluate both standard and NLI fine-tuned LLMs on bias-consistent and bias-adversarial cases. We find that LLMs exploit frequency bias for inference and perform poorly on adversarial instances. Furthermore, fine-tuned LLMs exhibit significantly increased reliance on this bias, suggesting that they are learning these frequency patterns from datasets. Finally, we compute the frequencies of hyponyms and their corresponding hypernyms from WordNet, revealing a correlation between frequency bias and textual entailment. These findings help explain why learning frequency patterns can enhance model performance on inference tasks.

## 1 Introduction

Natural Language Inference (NLI) is a core task in language understanding, aiming to determine whether a hypothesis follows logically from a premise. The rise of LLMs has driven significant progress in NLI (He et al., 2024; Liu et al., 2024), with studies (Liu et al., 2020; Li et al., 2025) showing that training on NLI corpora improves performance on inference benchmarks. Moreover, Cheng et al. (2025) introduce counterfactual NLI datasets to train LLMs, significantly enhancing their inferential abilities while reducing hallucinations.

Despite the widespread use of NLI datasets for training, the underlying mechanisms by which LLMs acquire inferential capabilities remain unclear. Li et al. (2022) argue that the performance gains of LLMs from inference data result from overfitting to dataset artifacts. Mckenna et al. (2023a) prove that LLMs benefit from memorizing such artifacts and leveraging them as shortcuts for inference. However, these studies fail to account for the finding of Cheng et al. (2025) that training on counterfactual reasoning data can still lead to improved inference performance. To further understand the source of LLMs' performance gains from training on NLI corpora, we conduct a series of controlled experiments analyzing their inferential behavior.

*First*, we examine the frequency of predicates in premises and hypotheses separately across multiple NLI datasets to analyze their underlying distributional properties. Our experiments reveal a clear **frequency biases** in NLI datasets: predicates in hypotheses tend to occur with significantly higher frequency than those in premises in examples labeled as positive. *Second*, we evaluate the performance of both standard and NLI-tuned LLMs on inference benchmarks, presenting evidence that these models are sensitive to frequency bias and tend to rely on this bias during inference. We further prove that fine-tuning on NLI datasets amplifies this reliance on frequency bias. Furthermore, we partition the NLI test set based on whether samples are *consistent* with or *adversarial* to frequency bias. Evaluation results show that models fine-tuned on inference data perform poorly on frequency-adversarial inference, proving that training on NLI datasets leads models to learn the frequency bias from NLI datasets. *Finally*, we present experiments demonstrating a correlation between frequency bias and textual entailment, offering an explanation for why learning frequency bias may serve as a proxy for enhancing inferential capability.

The main contributions of this paper are summarized as follows:

(a) We identify frequency bias in the various NLI data sets, where hypotheses tend to occur more frequently than premises when labeled is positive.

| | Label = Entail | | | Label = No-Entail | | |
|---|---|---|---|---|---|---|
| | premise | hypotheiss | *more frequent* | premise | hypothesis | *more frequent* |
| EGs | 36.05 | 209.1 | hypothesis | 109.35 | 32.84 | premise |
| Levy/Holt | 9.97 | 24.29 | hypothesis | 8.38 | 6.53 | premise |
| RTE | 60.89 | 62.24 | hypothesis | 73.01 | 60.09 | premise |
| MNLI | 69.30 | 72.71 | hypothesis | 69.10 | 65.80 | premise |

Table 1: Average predicate frequency in hypothesis and premise across different datasets. Frequencies are computed separately for positive (Entail) and negative (No-Entail) examples.

(b) We evaluate a range of LLMs and their fine-tuned counterparts, demonstrating that these models tend to exploit frequency bias during inference, and that fine-tuning on NLI datasets further amplifies this reliance.

(c) We present the relation between frequency bias and textual entailment, providing explanations why learning frequency bias can improve inference performance.

## 2 Methods

### 2.1 Calculate average frequency of predicates

For each hypothesis and premise in NLI datsets, we compute n-gram frequencies using the *WordFreq* library (Speer, 2022), which aggregates frequency data from multiple text sources to provide reliable usage statistics. In our experiments, we focus exclusively on the frequency of *verbal* predicates, abstracting away from the specific entity arguments that accompany them. We calculate the average predicate frequency for each statement. This approach enables us to capture general usage patterns of predicates across corpora while minimizing effects introduced by different subjects or objects[1].

### 2.2 Metrics

We calculate frequency bias as the difference between the frequencies of the hypothesis and the premise, computed as follows:

$$Bias(hypo, prem) = Freq(hypo) - Freq(prem)$$

where $Freq(\cdot)$ denotes the average frequency of predicates in a given statement[2].

## 3 Experimental Setup

### 3.1 Datasets

We fine-tune LLMs on a range of widely used NLI datasets and evaluate their inferential performance. The training sets include **RTE** (Dagan et al., 2006; Wang et al., 2019), **MNLI** (Williams et al., 2018; Wang et al., 2019), and **Entailment Graphs (EGs)** (Hosseini et al., 2018, 2021; Cheng et al., 2025), which is a counterfactual yet logically valid reasoning dataset. For evaluation, we use the **Levy/Holt** (Levy and Dagan, 2016; Holt, 2019) dataset, where each premise–hypothesis pair contains a single predicate with two named entity arguments, allowing for clear analysis of predicate frequency patterns.

We adopt the same prompt templates used in prior work (Schmitt and Schütze, 2021; Mckenna et al., 2023a; Cheng et al., 2025) for both fine-tuning and inference, which format samples as binary questions to determine whether the premise entails the hypothesis. A positive label corresponds to Entail, and a negative label to No-Entail. Details of the datasets and prompt configurations are provided in Appendix A and B.

### 3.2 Fine-tune LLMs

We fine-tune several widely used LLMs on NLI datasets, including DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), Mistral-7B, LLaMA-3-8B-instruct, and LLaMA-3-70B-instruct. Fine-tuning is conducted using LoRA (Hu et al., 2022) within the PEFT framework (Ding et al., 2023), with a learning rate of $1e^{-4}$, 12 training epochs, rank 8, and a dropout rate of 0.05.

## 4 Results

### 4.1 Finding 1: Frequency bias in NLI datasets

We analyze average predicate frequency on the various NLI datasets. From Table 1, we observe a consistent pattern: for instances labeled Entail, the hypothesis typically contains predicates with higher corpus frequency. Conversely, for No-Entail instances, predicates in the hypothesis tend to be less frequent. This observation present a consistent ***frequency bias*** embedded in these NLI datasets. The presence of frequency biases in NLI datasets raise the potential possibility that LLMs may be learning and leveraging these frequency patterns during fine-tuning.

[1] Our code and used data will be released upon publication.

[2] To better illustrate the frequency differences, we scale all frequency values by a factor of 1,000.

| | Label = Entail | | | | Label = No-Entail | | | |
|---|---|---|---|---|---|---|---|---|
| | premise | hypothesis | *Bias(hypo,prem)* | *follow bias* | premise | hypothesis | *Bias(prem,hypo)* | *follow bias* |
| DeepSeek-8B ✓ | 20.27 | 65.84 | 45.57 | consistent | 88.98 | 11.75 | 77.23 | consistent |
| DeepSeek-8B ✗ | 8.27 | 52.59 | **44.32** | consistent | 40.83 | 24.15 | **16.68** | consistent |
| LLaMA-3-8B ✓ | 17.58 | 66.18 | 48.6 | consistent | 112.15 | 13.36 | 98.79 | consistent |
| LLaMA-3-8B ✗ | 26.28 | 33.63 | **7.35** | consistent | 45.95 | 20.07 | **25.88** | consistent |
| Mistral-7B ✓ | 20.06 | 66.83 | 46.77 | consistent | 92.87 | 18.31 | 74.56 | consistent |
| Mistral-7B ✗ | 11.02 | 50.71 | **39.69** | consistent | 36.39 | 18.25 | **18.14** | consistent |
| LLaMA-3-70B ✓ | 17.13 | 67.96 | 50.83 | consistent | 77.56 | 19.37 | 58.19 | consistent |
| LLaMA-3-70B ✗ | 24.66 | 39.15 | **14.49** | consistent | 44.03 | 16.75 | **27.28** | consistent |
| DeepSeek-8B$_{EG}$ ✓ | 15.78 | 79.85 | 64.07 | consistent | 95.18 | 8.26 | 86.92 | consistent |
| DeepSeek-8B$_{EG}$ ✗ | 25.64 | 15.55 | **-10.09** | *adversarial* | 26.50 | 30.06 | **-3.56** | *adversarial* |
| LLaMA-3-8B$_{EG}$ ✓ | 14.92 | 66.27 | 51.35 | consistent | 130.86 | 8.26 | 122.6 | consistent |
| LLaMA-3-8B$_{EG}$ ✗ | 69.18 | 23.45 | **-45.73** | *adversarial* | 38.01 | 22.10 | **15.91** | consistent |
| Mistral-7B$_{EG}$ ✓ | 18.23 | 75.67 | 57.44 | consistent | 96.69 | 8.02 | 88.67 | consistent |
| Mistral-7B$_{EG}$ ✗ | 18.39 | 15.66 | **-2.73** | *adversarial* | 28.49 | 29.88 | **-1.39** | *adversarial* |
| LLaMA-3-70B$_{EG}$ ✓ | 13.83 | 103.23 | 89.4 | consistent | 71.65 | 9.76 | 61.89 | consistent |
| LLaMA-3-70B$_{EG}$ ✗ | 23.60 | 16.02 | **-7.58** | *adversarial* | 19.67 | 64.88 | **-45.21** | *adversarial* |

Table 2: Frequency of premise and hypothesis when evaluating the Levy/Holt dataset using different LLMs and their EG-tuned variants ($_{EG}$). ✓ indicates correct predictions, while ✗ indicates incorrect predictions.

| | Label = Entail | | | Label = No-Entail | | |
|---|---|---|---|---|---|---|
| | premise | hypothesis | *Bias(hypo,prem)* | premise | hypothesis | *Bias(prem,hypo)* |
| DeepSeek-8B ✓ | 20.27 | 65.84 | 45.57 | 88.98 | 11.75 | 77.23 |
| DeepSeek-8B ✗ | 8.27 | 52.59 | **44.32** | 40.83 | 24.15 | **16.68** |
| DeepSeek-8B$_{EG}$ ✓ | 15.78 | 79.85 | 64.07 | 95.18 | 8.26 | 86.92 |
| DeepSeek-8B$_{EG}$ ✗ | 25.64 | 15.55 | **-10.09** | 26.50 | 30.06 | **-3.56** |
| DeepSeek-8B$_{RTE}$ ✓ | 21.17 | 75.42 | 54.25 | 73.99 | 10.86 | 63.13 |
| DeepSeek-8B$_{RTE}$ ✗ | 12.97 | 42.14 | **29.17** | 44.88 | 31.68 | **13.2** |
| DeepSeek-8B$_{MNLI}$ ✓ | 19.62 | 71.14 | 51.52 | 86.89 | 8.86 | 78.03 |
| DeepSeek-8B$_{MNLI}$ ✗ | 9.29 | 13.88 | **4.59** | 51.33 | 23.25 | **28.08** |

Table 3: Frequency of premise and hypothesis when evaluating the Levy/Holt using LLMs fine-tuned on different NLI datasets. ✓ denotes correct predictions, while ✗ indicates incorrect predictions.

### 4.2 Finding 2: LLMs are sensitive to frequency bias

We evaluate a range of LLMs and their EGs-tuned variants on Levy/Holt. Model predictions are classified as either *correct* (✓) or *incorrect* (✗), and we analyze frequency bias across these categories.

We reports the frequency biases in Table 2. The results reveal a clear trend: both standard and fine-tuned LLMs can make *correct* predictions when test samples are consistent with the frequency bias. In contrast, when this bias is reduced or adversarial, the likelihood of *incorrect* predictions increases. It shows that LLMs' performance is sensitive to the frequency bias. Furthermore, we observe that the sensitivity is especially pronounced in EG-tuned models, where the frequency bias is higher in *correct* predictions and more reduced in *incorrect* ones, compared to standard LLMs.

We also evaluate LLMs fine-tuned on different NLI datasets. As shown in Table 3, we observe that models consistently tend to make correct predictions on samples with stronger frequency bias, but are more likely to produce incorrect predictions when the bias is reduced. The phenomenon is consistently more pronounced in LLMs fine-tuned on NLI datasets compared to standard models.

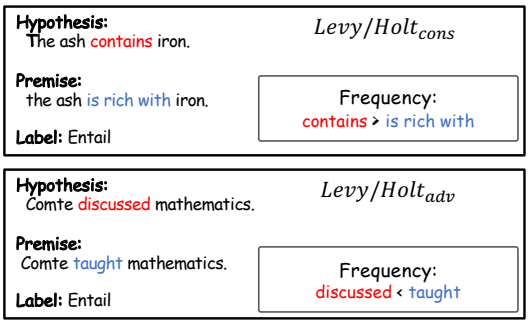


Figure 1: A sample in Levy/Holt$_{cons}$ and Levy/Holt$_{adv}$.

### 4.3 Finding 3: NLI-tuned LLMs struggle with frequency-adversarial inference

To further measure the impact of frequency bias, we divide the Levy/Holt dataset into two subsets: those *consistent* with the frequency bias (Levy/Holt$_{cons}$) and those that are *adversarial* to it (Levy/Holt$_{adv}$). As illustrated in Figure 1, in Levy/Holt$_{cons}$ samples labeled `Entail`, the predicate in the hypothesis is more frequent than that in the premise. In contrast,

| Models | Levy/Holt$_{cons}$ | Levy/Holt$_{adv}$ | Δ |
|---|---|---|---|
| LLaMA-3-8B | 74.0 | 61.74 | -12.26 |
| DeepSeek-8B | 73.51 | 64.99 | -8.52 |
| Mistral-7B | 65.31 | 57.23 | -8.08 |
| LLaMA-3-70B | 84.25 | 70.55 | -13.7 |
| LLaMA-3-8B$_{EG}$ | 85.2 | 62.5 | **-22.7** |
| DeepSeek-8B$_{EG}$ | 80.8 | 62.18 | **-18.62** |
| Mistral-7B$_{EG}$ | 83.61 | 58.79 | **-24.82** |
| LLaMA-3-70B$_{EG}$ | 85.25 | 69.67 | **-15.58** |

Table 4: AUC scores on the frequency-consistent and frequency-adversarial Levy/Holt.

Levy/Holt$_{adv}$ contains cases, where the premise is more frequent than the hypothesis.

We evaluate model performance separately on Levy/Holt$_{cons}$ and Levy/Holt$_{adv}$, and report the Area Under the Curve (AUC) scores in Table 4. Results show a substantial drop in AUC scores on Levy/Holt$_{adv}$ for both standard and fine-tuned LLMs. This performance gaps on Levy/Holt$_{cons}$ and Levy/Holt$_{adv}$ suggests that LLMs rely on frequency bias as a shortcut, performing well on reasoning from low-frequency to high-frequency statements but struggling when this pattern is reversed. Compared to standard LLMs, EG-tuned LLMs exhibit a more substantial performance gaps, indicating that training on EGs reinforces their reliance on the frequency bias in datasets. We further fine-tune LLMs on additional NLI datasets and observe consistent findings, as shown in Appendix D.

These results highlight two key findings: (1) LLMs perform well on bias-consistent cases but struggle with adversarial ones, indicating that they exploit frequency bias as a proxy for inference. (2) After fine-tuned on NLI datasets, LLMs exhibit increased reliance on frequency bias. This suggests that fine-tuning process encourages learning frequency-based patterns from these datasets, which reinforces inferences from low-frequency to high-frequency statements while diminishing the ability to reason in the opposite direction.

### 4.4 Finding 4: Frequency bias is a proxy for gradient of semantic generalization

To further investigate the relationship between predicate frequency and entailment relation, we analyze the frequency of **hyponym–hypernym pairs**[3] extracted from WordNet (Miller, 1994). These pairs represent a specific form of upward semantic entailment, namely generalization, where a more specific concept (hyponym) entails its corresponding general one (hypernym). For example, *"whisper"* is a hyponym of *"talk"*, so the statement *"X whisper to Y"* semantically entails *"X talk to Y"*.

[3]Unlike Mckenna et al. (2023b), we focus on *verb* pairs to examine generalization and specificity, as verbs often involve more abstract and dynamic semantic shifts than nouns.

| | Hyponyms | Hypernyms |
|---|---|---|
| WordNet | 4.13 | 12.18 |

Table 5: Average frequency of Hyponym-Hypernym in WordNet. The results prove that more general concepts (hypernym) have higher frequency.

Table 5 reports the average frequency of hyponyms and hypernyms separately. Hypernyms occur more frequently than their corresponding hyponyms, indicating that more general concept are more frequent in natural language. This finding suggests that frequency bias may serve as a proxy for the generalization gradient, whereby inferences from lower-frequency to higher-frequency predicates reflect a generalization from specific to more abstract concepts. In NLI datasets, we also observe that when the label is `Entail`, hypotheses often contain more hypernyms than corresponding premises, as shown in Appendix C. This suggests that most samples in NLI datasets require inference from more specific concepts to more general ones.

These findings offer an explanation for why training LLMs on NLI datasets can serve as a proxy for enhancing inferential capability: LLMs learn frequency biases from datasets during training and these biases align with a generalization gradient that supports entailment from specific to more general concepts. Although LLMs can learn frequency bias as a proxy to enhance their inferential capability, learning this bias limits model robustness in frequency-adversarial settings, as observed in §4.3.

## 5 Conclusion

In this work, we investigate what LLMs actually learn during the fine-tuning process. First, we identify significant frequency bias in various NLI datasets. Next we prove that LLMs exploit this bias for inference, and that fine-tuning further increases their reliance on such patterns. Finally, we show a strong correlation between frequency bias and a particular variety of entailment that is common in NLI datasets, namely hyponym-to-hypernym generalization. It offers an explanation for why learning these frequency patterns can enhance model inference performance. Our work also reveals that a key limitation of LLMs is their vulnerability to frequency-adversarial inference cases.

## Limitations

In this work, our findings suggest that LLMs internalize frequency biases from training data and utilize them as a proxy for inference. However, due to computational constraints, our experiments are limited to a range of smaller LLMs with 8B variants and a single extremely large-scale model, LLaMA-3-70B. This restricts our ability to draw conclusions about extremely large models. In future work, we will explore broader model scales to assess the consistency of these trends.

## References


Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2010. Global Learning of Focused Entailment Graphs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1220–1229.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global Learning of Typed Entailment Rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 610–619.

Liang Cheng, Tianyi Li, Zhaowei Wang, Tianyang Liu, and Mark Steedman. 2025. Neutralizing bias in llm reasoning using entailment graphs. *arXiv preprint arXiv:2503.11614*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499.

Xavier Holt. 2019. Probabilistic Models of Relational Implication. *arXiv:1907.12048 [cs, stat]*. ArXiv: 1907.12048.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R Holt, Shay B Cohen, Mark Johnson, and Mark Steedman. 2018. Learning Typed Entailment Graphs with Global Soft Constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Mohammad Javad Hosseini, Shay B Cohen, Mark Johnson, and Mark Steedman. 2021. Open-Domain Contextual Link Prediction and its Complementarity with Entailment Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2790–2802.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Omer Levy and Ido Dagan. 2016. Annotating Relation Inference in Context via Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 249–255, Berlin, Germany. Association for Computational Linguistics.

Dengchun Li, Naizheng Wang, Zihao Zhang, Haoyang Yin, Lei Duan, Meng Xiao, and Mingjie Tang. 2025. Dynmole: Boosting mixture of lora experts fine-tuning with a hybrid routing mechanism. *arXiv preprint arXiv:2504.00661*.

Tianyi Li, Mohammad Javad Hosseini, Sabine Weber, and Mark Steedman. 2022. Language models are poor learners of directional inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 903–921.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. 2024. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*.

Nick Mckenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023a. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774.

Nick Mckenna, Tianyi Li, Mark Johnson, and Mark Steedman. 2023b. Smoothing entailment graphs with language models. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 551–563.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Martin Schmitt and Hinrich Schütze. 2021. Language Models for Lexical Inference in Context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1267–1280, Online. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*.

# A Dataset

## A.1 Train set

**Entailment Graphs (EGs)** (Berant et al., 2010, 2011; Hosseini et al., 2018, 2021) are symbolic graphs to preserve the textual entailment in open-domain corpora. These graphs contain entailment between predicates, formatted as triples consisting of predicate pairs and their typed arguments. Cheng et al. (2025) propose a method to initialize the extracted EGs into inference datasets for training LLMs. The datasets comprise premise–hypothesis pairs, with each hypothesis being counterfactual yet logically entailed by its corresponding premise. Fine-tuning LLMs on this data has been shown to substantially enhance inferential capability while reducing hallucinations (Cheng et al., 2025).

**RTE** (Dagan et al., 2006; Wang et al., 2019) is a NLI benchmark dataset designed for evaluating models on the task of recognizing textual entailment (RTE). It consists of sentence-level premise-hypothesis pairs collected from various sources, which are collected from various sources such as news articles and information extraction tasks.

**MNLI** (Williams et al., 2018; Wang et al., 2019) is a widely used benchmark for evaluating language models on NLI, It consists of sentence-level premise-hypothesis pairs, with premises drawn from diverse sources and hypotheses manually written.In our experiments, we fine-tune LLMs using the MNLI training split.

## A.2 Test set

**Levy/Holt** (Levy and Dagan, 2016; Holt, 2019) dataset is a widely used for NLI, which comprises premise-hypothesis pairs structured in a specific task format: "Given [premise $P$], is it true that [hypothesis $H$]?". Each $P$- and $H$-statement has the property of containing one predicate with two named entity arguments, where the same entities appear in both $P$ and $H$. The Levy/Holt dataset contains inverse of all entailment pairs. In our experiments, we study the challenging directional subset, where the entailments hold in one direction but not both.

# B Prompts Used in Experiments

**Prompt templates** are widely acknowledged for their significant and sometimes decisive impact on the behavior of LLMs. In our experiments, we categorize the prompt templates into two distinct types based on their usage: prompt templates for fine-tuning LLMs and prompt templates used during inference.

## B.1 prompt template for fine-tuning

We follow the fine-tuning setup of Cheng et al. (2025), adopting the same prompt templates used in prior inference studies (Schmitt and Schütze, 2021; Mckenna et al., 2023a; Cheng et al., 2025), which follow the format outlined below:

If [PREMISE], then [HYPOTHESIS].

To make LLMs better understanding the task, we format it as Boolean questions and include indicator words such as "Question:" and "Answer:". For each option, we automatically provide explanations for every answer by adding affirmation or negation to the propositions. As a result, the NLI training data is structured as shown in Table 8. We fine-tune our models using these templates.

## B.2 prompt template for inference

Following the evaluation settings of prior works (Schmitt and Schütze, 2021; Mckenna et al., 2023a; Cheng et al., 2025), we use the same few-shot examples in our inference prompts for NLI tasks, consisting of two positive and two negative instances. The examples are shown in Table 9.

# C Hyponymn-Hypernym Pairs in NLI datasets

We analyze the distribution of hyponym–hypernym pairs in the NLI dataset by counting the occurrences of hypernyms and hyponyms in premises and hypotheses. We focus our analysis on the EG and Levy/Holt datasets because their samples contain explicit predicate structures, unlike MNLI and RTE, which are at the sentence level and lack clearly defined predicates. As shown in Table 6, we observe that for instances labeled as `Entail`, hypernyms appear more frequently in the hypothesis than in the premise. Conversely, for `No-Entail` instances, hypernyms are more frequent in the premise. These results suggest that hypotheses tend to contain more abstract concepts than premises in positive examples, aligning with the observed frequency bias and reinforcing the findings discussed in §4.4.

| | when Label = Entail | | when Label = No-Entail | |
|---|---|---|---|---|
| | Hypernmys | Hyponyms | Hypernmys | Hyponyms |
| Levy/Holt | 92 | 65 | 56 | 101 |
| EG | 93 | 64 | 31 | 61 |

Table 6: Counts of Hypernyms and Hyponyms in hypotheses and premises of Levy/Holt and EGs. We extract all hyponym–hypernym pairs from the data. For samples labeled as `Entail`, hypotheses tend to contain more hypernyms, indicating a more general statement. In contrast, for `No-Entail` samples, premises typically include more hypernyms, suggesting that the corresponding hypotheses are more specific or less general.

| Models | Levy/Holt$_{cons}$ | Levy/Holt$_{adv}$ | $\Delta$ |
|---|---|---|---|
| DeepSeek-R1-8B | 73.51 | 64.99 | -8.52 |
| DeepSeek-R1-8B$_{EG}$ | 80.8 | 62.18 | **-18.62** |
| DeepSeek-R1-8B$_{RTE}$ | 73.12 | 61.74 | -11.38 |
| DeepSeek-R1-8B$_{MNLI}$ | 74.09 | 61.85 | -12.24 |
| LLaMA-3-8B | 74.0 | 61.74 | -12.26 |
| LLaMA-3-8B$_{EG}$ | 85.2 | 62.5 | **-22.7** |
| LLaMA-3-8B$_{RTE}$ | 73.97 | 57.49 | -16.48 |
| LLaMA-3-8B$_{MNLI}$ | 72.33 | 58.91 | -13.42 |

Table 7: AUC scores on the frequency-consistent and frequency-adversarial Levy/Holt.

## D Fine-tuned LLMs Performance on Frequency-adversarial Inference

Table 7 presents the performance of LLMs fine-tuned on various NLI datasets, consistently showing that these models struggle with frequency-adversarial inference.

## E Computing Costs

We fine-tuned the LLaMA-3-70B model on the EGs dataset using four NVIDIA RTX A6000 GPUs over 21 hours. Fine-tuning on the MNLI dataset required approximately 46 hours. For inference, evaluation on the Levy/Hot datasets takes around 30 minutes.

| | |
|---|---|
| | Question: If [PREMISE], then [HYPOTHESIS]. Is that true or false?<br>(A) True; (B) false |
| label=*True* | (A) True.<br>Yes, it is true. [PREMISE] entails [HYPOTHESIS]. |
| label=*False* | (B) False.<br>No, it is false. [PREMISE] does not entail [HYPOTHESIS]. |

Table 8: The table present the prompt template using in our training steps.

| **Few-shot Examples Instantiated Prompt for Inference Task** |
|---|
| If Google bought Youtube, then Google owns Youtube. Is that true or false?<br>A) True<br>B) False<br>Answer: A) True. Owning is a consequence of buying.<br>If Google owns Youtube, then Google bought Youtube. Is that true or false?<br>A) True<br>B) False<br>Answer: B) False. Owning does not imply buying, the ownership may come from other means.<br>If John went to the mall, then John drove to the mall. Is that true or false?<br>A) True<br>B) False<br>Answer: B) False. John may have gone to the mall by other means.<br>If John drove to the mall, then John went to the mall. Is that true or false?<br>A) True<br>B) False<br>Answer: A) true. Driving is a means of going to the mall.<br>If John F. Kennedy was killed in Dallas, then John F. Kennedy died in Dallas. Is that true or false?<br>A) True<br>B) False<br>Answer: |

Table 9: Example instantiated prompts in Few-shot settings, for the sample "PREMISE: [Google bought Youtube], HYPOTHESIS: [Google owns Youtube]".