Neutralizing Bias in LLM Reasoning using Entailment Graphs

Liang Cheng[†] **Tianyi Li**[‡*] **Zhaowei Wang**[§] **Tianyang Liu**[†] **Mark Steedman**[†]
[†]University of Edinburgh [‡]Amazon Alexa AI [§]HKUST
l.cheng@ed.ac.uk tylteddy@amazon.co.uk m.steedman@ed.ac.uk

Abstract

LLMs are often claimed to be capable of Natural Language Inference (NLI), which is widely regarded as a cornerstone of more complex forms of reasoning. However, recent works show that LLMs still suffer from hallucinations in NLI due to *attestation bias*, where LLMs overly rely on propositional memory to build shortcuts. To solve the issue, we design an unsupervised framework to construct counterfactual reasoning data and fine-tune LLMs to reduce attestation bias. To measure bias reduction, we build *bias-adversarial* variants of NLI datasets with randomly replaced predicates in premises while keeping hypotheses unchanged. Extensive evaluations show that our framework can significantly reduce hallucinations from attestation bias. Then, we further evaluate LLMs fine-tuned with our framework on original NLI datasets and their bias-neutralized versions, where original entities are replaced with randomly sampled ones. Extensive results show that our framework consistently improves inferential performance on both original and bias-neutralized NLI datasets.

1 Introduction

Natural Language Inference (NLI) has long been recognized as a foundational understanding task in language understanding with various downstream applications (Cheng et al., 2023; Deng et al., 2023; Gao et al., 2023). It assesses the understanding ability of models by requiring them to determine if a given premise logically entails a hypothesis. Recently, with the rise of LLMs, the field of NLI has witnessed significant advancements (Brown et al., 2020; He et al., 2023; Liu et al., 2024b). These models, pre-trained on vast amounts of text data, have been claimed to capture inferential relations between statements enabling reasoning, positioning them as state-of-the-art tools for NLI.

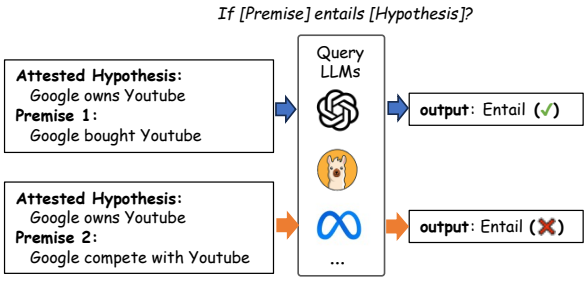

Figure 1: An example of attestation bias. LLMs tend to evaluate entailment with their memorized knowledge rather than given premise.

However, despite the apparent success of LLMs on natural NLI tasks, they continue to embody response biases, leading to the phenomenon of hallucination (Huang and Chang, 2023; Ji et al., 2023; Gallegos et al., 2024). In NLI, this issue arises because models rely more on memorization from their training corpus rather than inference from the given premises, causing false positive entailment judgments when the hypothesis is *attested* in the training data (Poliak et al., 2018; Rawte et al., 2023), the phenomenon known as **attestation bias** (Mckenna et al., 2023). The attestation bias leads to brittleness in bias-adversarial cases, and poses challenges in accurately assess the bias-free reasoning capabilities of LLMs (Mckenna et al., 2023).

In this study, we explore the method of fine-tuning LLMs to improve their robustness against the attestation bias. We propose an *unsupervised* approach to construct counterfactual but logically consistent datasets for training LLMs. Our approach begins with unsupervised extraction of textual entailment relations between predicates from large-scale open-domain corpora using semantic parsing. The extracted data is then formatted into Entailment Graphs (EGs) (Hosseini et al., 2018, 2021), which consist of typed predicate pairs. Finally, we generate counterfactual samples by randomly selecting named entities and other argu-

[*]Work completed while the author was at the University of Edinburgh.

Findings of the Association for Computational Linguistics: ACL 2025, pages 13714–13730

ments to instantiate these types.

We evaluate the effectiveness of our method along two dimensions: bias reduction and general inferential performance improvement.

First, to measure how well our training has reduced the attestation bias, we compare LLMs before and after our training on *bias-adversarial* variants of NLI datasets. Specifically, we randomly alter the predicates in the premises while keeping the hypothesis fixed. The newly generated premises are non-entailing, so any positive judgments by the LLM are false positives, arising from attestation bias relating to the hypothesis. The results demonstrate that training LLMs with our method can significantly reduce attestation bias, ensuring a more reliable evaluation of their reasoning capabilities.

Second, to evaluate the effectiveness of our training in improving inferential performance on the original NLI tasks, we conduct experiments on both the *standard* and a further *bias-neutralized* NLI datasets, where entities in the original dataset are replaced with randomly selected entities of the same type. In both cases, our approach outperforms baseline models, demonstrating its robustness and effectiveness in enhancing inferential capability.

The main contributions of this paper are summarized as follows:

(a) To reduce attestation bias while enhancing inferential capabilities, we propose an unsupervised approach using EGs to generate a logically sound inference dataset, free of the artefacts that plague human-constructed NLI datasets, for fine-tuning LLMs.

(b) We show that for 4 different LLMs, our approach reduces attestation bias and improves performance in NLI tasks.

(c) We introduce a bias-neutralized method for a more accurate evaluation of LLMs' true inferential capabilities. This approach generates bias-neutralized test sets, where our EG-enhanced models consistently achieve superior reasoning performance on both standard and these bias-neutralized inference datasets.

2 Related Work

Hallucination in Inference: Hallucination in LLMs has emerged as a significant area of concern in NLP, as these models often generate content that is either factually inaccurate or contextually inappropriate. Talman and Chatzikyriakidis (2019) report that many models struggle to generalize across different NLI datasets, even when the task format remains the same. In smaller language models, Li et al. (2022) observed a reliance on dataset artifacts when performing directional NLI on predicates. Furthermore, Poliak et al. (2018) found a range of NLI datasets containing artifacts that are memorized by supervised models trained on only sample hypotheses, causing overestimation of their inference performance. Gulati et al. (2024) proves that SOTA LLMs rely on memorized data to answer math questions, and their performance drops significantly once variable names are altered. Carlini et al. (2023) found that LLMs are capable of memorizing significantly more data compared to smaller models, raising doubts on whether their performance gains stem from advanced inferential capabilities or more memorization.

Attestation Bias: Attestation bias occurs when LLMs show a significantly higher probability of predicting `Entail` when the hypothesis is *attested*, indicating that the inference process of LLMs is heavily influenced by their reliance on memorization about hypotheses. As a result, LLMs are inherently prone to disregarding the premise and responding incorrectly by relying on memorized information about hypothesis from their training corpus, as illustrated in Figure 1.

Mckenna et al. (2023) conducted a hypothesis-only test on LLMs, revealing that when labels contradict attestation bias, LLMs can be poor or even near-random classifier. Their research demonstrates that attestation bias is the primary source of hallucination in LLMs on inference tasks.

Entailment Graphs: EGs are symbolic graphs used to preserve entailment relations between predicates (Berant et al., 2010, 2011; Hosseini et al., 2018, 2021). Unlike sentence-level inference data, EGs are formatted as sets of triples, with each triple consisting of predicate pairs and typed arguments. For example, "*(Person.X, visited, Location.Y)* $\models$ *(Person.X, went to, Location.Y)*". EGs have been utilized in open-domain question answering and knowledge inference (Cheng et al., 2023; Wang et al., 2024).

3 Methodology

We propose an *unsupervised* approach for constructing counterfactual reasoning datasets to fine-tune LLMs, enabling them to generalize beyond memorized knowledge and enhance inferential ca-

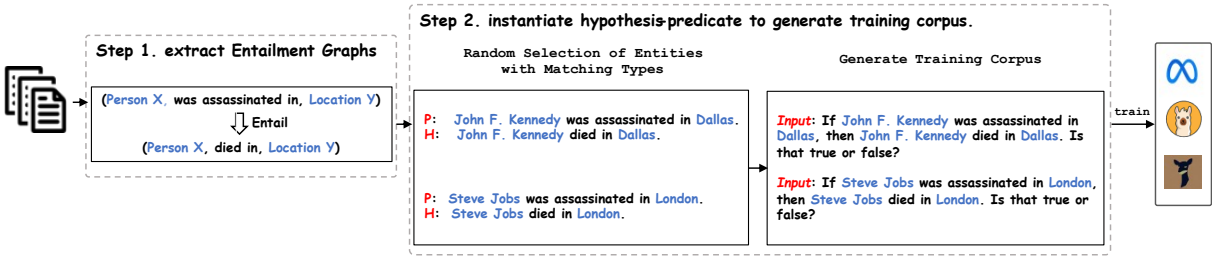

Figure 2: The pipeline of our approach: **Step 1:** Build EGs in unsupervised manner. **Step 2:** Instantiate predicates using random entities with matching types, then wrap instantiated predicates into prompts to generate training corpus.

pabilities.

3.1 Constructing Counterfactual Reasoning Datasets

As demonstrated in Figure 2, our counterfactual reasoning dataset is built in two key steps: unsupervised extraction of EGs (§3.1.1) and instantiating EG rules into NLI training sets (§3.1.2).

3.1.1 Extracting Entailment Graph

We adopt the approach proposed by Hosseini et al. (2018) for constructing EGs, which involves following main steps:

First, we employ a combinatory categorial grammar (CCG) parser (Steedman, 2000), GraphParser (Reddy et al., 2014), to extract binary relations between a predicate and its arguments from sentences. Each argument, typically a noun, is then linked to its corresponding FreeBase IDs and entity type using the Named Entity Linking tool Aidalight (Nguyen et al., 2014), formatting the extracted triple as predicates with typed arguments. Then, we compute the distributional similarity score[1] by calculating the co-occurrence of predicates associated with entities of the same `types`, assuming that predicates linked to the same entities refer to the same event or episode. We generate negative examples by replacing predicates with the same typed arguments. The entire process is *unsupervised*.

Corpus: Following Hosseini et al. (2021), we use the multiple-source **NewsSpike** (Zhang and Weld, 2013) corpus to extract the EGs. NewsSpike was deliberately built to include different articles from different sources describing identical news events. From this dataset, we extract 5,500 positive and 5,500 negative samples.

[1]We use the Weeds similarity score (Weeds and Weir, 2003) as the entailment score in our construction.

3.1.2 Instantiating Premise-Hypothesis Pairs

Entailment graphs consist of entailment rules, each rule in an EG involves a pair of predicates expecting two typed arguments. To instantiate EG rules into NLI data entries, we first replace the type arguments with specific named entities. To ensure consistent entity replacements, we categorize entities from the open-domain corpus into 48 FIGER types (Ling and Weld, 2012), such as "person" or "location", aligning them with the types in Freebase (Bollacker et al., 2008). We assign a default type "thing" in failure cases. Using the entity-to-type mapping, we randomly select entities that match the corresponding types to instantiate typed EGs into premise-hypothesis pairs.

EG rules are predicate focused and apply to any context, so the instantiation process preserves the entailment relationships within typed EGs. Figure 2 illustrates this process with an example, where the entailment "*(Person X, was assassinated in, Location Y)* $\models$ *(Person X, died in, Location Y)*" is formatted into the following NLI format: "**[Premise]**: Steve Jobs was assassinated in London. **[Hypothesis]**: Steve Jobs died in London". Our constructed premise-hypothesis pairs preserve the logical inferential relationship but are counterfactual. We manually reviewed 50 samples from the generated counterfactual Entailment Graphs and found no incorrect entailment labels. Notably, unlike previous work (Kaushik et al.; Elazar et al., 2024), our construction process is fully automatic and unsupervised.

We adopt the prompt templates from previous studies (Schmitt and Schütze, 2021; Mckenna et al., 2023), formatting the premise-hypothesis pairs as a two-way answer choice: A) True, B) False. These generated natural questions are then used to query the models. The full list of concrete prompts can be found in Appendix E.1.

3.2 Training LLMs with Instantiated EGs

We use the counterfactual reasoning data generated from EGs to fine-tune DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), Mistral-7B, LLaMA-3-8B-instruct, and LLaMA-3-70B-instruct models. These models are widely recognized for their strong reasoning capabilities and have drawn significant interest from researchers.

Following the *llama-recipes* tools published by Meta[2], we fine-tune LLMs using LoRA (Hu et al., 2022) within the PEFT (Parameter-Efficient Fine-Tuning) (Ding et al., 2023) framework. During fine-tuning, we set a fixed learning rate of $1e^{-4}$ and train for 12 epochs. The LoRA rank is set to 8, with LoRA dropout rate of 0.05.

4 Experimental Setup

To evaluate the effectiveness of our EG-enhanced models, we conduct two experiments: (1) examining whether EG-enhanced models reduce attestation bias in §5, and (2) assessing their inferential performance in NLI benchmarks in §6. Both evaluation experiments follow the same settings, including the evaluation datasets and baselines, as detailed in the following sections[3].

4.1 Evaluation Datasets

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. Following Mckenna et al. (2023), we study the challenging *directional* subset, where the entailments hold in one direction but not both.

SNLI (Bowman et al., 2015) is another widely used datasets for NLI. It consists of human-generated premise-hypothesis pairs with manually assigned labels. Unlike Levy/Holt, which contains named-entity artifacts, SNLI is composed of general sentences in both premises and hypotheses, typically without named entities.

Fomatting as two-choice questions: For evaluation, premise-hypothesis pairs in Levy/Holt and SNLI are formatted as two-choice natural questions using prompts, where choice A corresponds to `Entail` and choice B to `No-Entail`, ensuring alignment with the Levy/Holt and SNLI annotations[4]. These prompts are crafted by human experts (Schmitt and Schütze, 2021; Mckenna et al., 2023), as shown in Appendix §E.3. During evaluation, all models successfully selected either A or B for every development set question, indicating compatibility with the QA format.

4.2 Baselines

Standard LLMs with Few-Shot Setting: To evaluate the performance of EG-enhanced LLMs, we compare them against the original LLMs as baseline models. During inference, we adopt a few-shot approach, hand-annotating a minimal set of 4 examples in the style of the template. These examples are prepended before the query (see Appendix E.3 for an example). Our goal is to analyze model behavior as conditions change, rather than maximize the score on a specific dataset. Therefore, we maintain a minimal 4-example setup to evoke positive responses across LLMs.

Chain-of-Thought Reasoning: In our experiments, we incorporate manually written explanations in few-shot examples, providing step-by-step reasoning before each answer to guide LLMs. Specifically, we utilize a three-step analytical process for guidance: analyzing the premise, analyzing the hypothesis and clarifying the relationship between premise and hypothesis. These explanations, detailed in Appendix E.4, serve as Chain-of-Thought (CoT) prompts, establishing a baseline for evaluating LLM performance under CoT guidance.

5 Experiment 1: Attestation Bias Reduction

To evaluate the effectiveness of our method in reducing attestation bias, we measure attestation bias by comparing estimated probabilities of predicting `Entail` conditioned on whether the hypothesis is predicted `Attested` or not.

However, in original NLI dataset entailments may coincidentally refer to attested facts, which could lead to spurious correlation between inference and attestation scores, making it difficult to

[2] https://github.com/meta-llama/llama-recipes

[3] All code and data are provided in https://github.com/LeonChengg/EGs-tuned-LLMs.git

[4] In SNLI, contradiction and neutral labels are categorized as No-Entail.

Task	Sample Query: [premise] ⇒ [hypothesis]	Dataset Label
I	George Bush was the governor of Texas ⇒ George Bush is a politician from Texas	`Entail`
RPI	George Bush lives in Texas ⇒ George Bush is a politician from Texas	`No-Entail`

Table 1: A sample of generate *RPI* from original inference task I.

determine whether LLMs rely on memory or reasoning to generate predictions. To address this issue, we adopt the **Random Premise Inference Task** proposed by Mckenna et al. (2023), which serves as a *bias-adversarial* benchmark for accurately quantifying attestation bias.

The Random Premise Task (*RPI*) modifies the original NLI dataset by replacing the original premise predicate with a randomly selected predicate while keeping the hypothesis fixed and maintaining the same entity arguments. As a result, this process creates non-entailing premise-hypothesis pairs, as the example illustrated in Table 1. This transformation produces a dataset in which all samples are labeled as negatives (`No-Entail`), as two randomly paired predicates are highly unlikely to form entailment relations[5]. We determine whether LLMs can correctly identify these samples as negatives or if they rely solely on attested hypotheses to make false positive predictions. An ideal model should predict zero `Entail`. The *RPI* task effectively tests the model's reliance on propositional memory, as it prevents true entailments while maintaining the attestedness of the conclusions (hypotheses).

To calculate attestation biases, we first determine the attestedness of each hypothesis by prompting the LLM to classify it as true, false, or unknown, following the same prompts[6] used in prior studies (Poliak et al., 2018; Mckenna et al., 2023). Then, we categorize all samples into two groups: *attested* and *non-attested*, depending on whether the LLM identifies the hypothesis as true. We calculate the proportion of `Entail` predictions in the *attested* set and compare it to the *non-attested* set, providing an effective measure of attestation bias by highlighting differences in prediction behavior between the two sets.

We adopt the Levy/Holt dataset for this attestation bias measurement experiment because this dataset includes artifacts containing named entities, allowing us to assess whether these artifacts in hypothesis are attested by LLMs.

Model	AttBias	$\Delta_{AttBias}$
DeepSeek-R1-Llama-8B	26.04	-
DeepSeek-R1-Llama-8B$_{CoT}$	15.10	-10.94
DeepSeek-R1-Llama-8B$_{EG}$	**7.58**	**-18.46**
Mistral-7B	32.98	-
Mistral-7B$_{CoT}$	22.64	-10.34
Mistral-7B$_{EG}$	**13.0**	**-19.98**
Llama-3-8B	23.12	-
Llama-3-8B$_{CoT}$	13.94	-9.18
Llama-3-8B$_{EG}$	**5.99**	**-17.13**
Llama-3-70B	19.20	-
Llama-3-70B$_{CoT}$	15.96	-3.24
Llama-3-70B$_{EG}$	**8.34**	**-10.86**

Table 2: AttBias scores on the *RPI* Levy/Holt dataset across various models (lower is better). The subscript *EG* denotes LLMs trained on our constructed EGs, while *CoT* refers to models employing chain-of-thought reasoning.

5.1 Scoring Attestation Bias

Attestation bias reflects a significantly higher likelihood of predicting `Entail` for *attested* hypotheses compared to *non-attested* hypotheses. To quantify this bias, we define the **Attestation Bias score** as follows:

$$AttBias = P(tok = \texttt{Entail}|Att(hypo)) - P(tok = \texttt{Entail}|\neg Att(hypo))$$

where $P(tok = \texttt{Entail} \mid Att(hypo))$ represents the estimated conditional probability of predicting `Entail` when the hypothesis is attested. When the models achieve the same accuracy rate[7], we calculate the proportion of `Entail` predictions that fall within the *attested* hypothesis set compared to *non-attested* set.

Lower AttBias scores representing reduced influence from attested memorization, indicating less impact of attestation bias.

5.2 Results of Experiment I: Degree of Bias Reduction

Table 2 presents the Attestation Bias scores for the *RPI* task. The results demonstrate that our EG-

[5] Mckenna et al. (2023) manually inspect the generated random premise entries for the Levy/Holt dataset and found only 9.6% to be true entailment.

[6] The attestation prompt is provided in Appendix E.2.

[7] In our experiments, accuracy rate is set to 0.5.

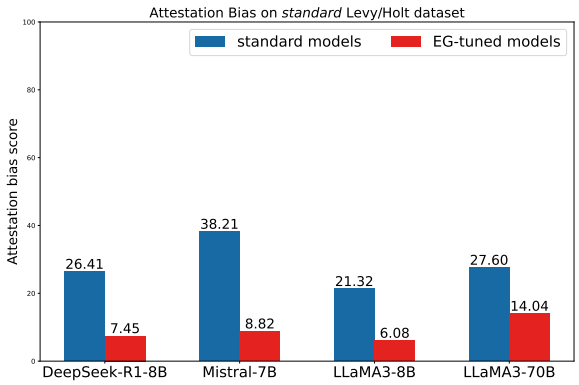

Figure 3: The Attestation Bias scores for the original Levy/Holt, demonstrating a consistent attestation bias reduction after fine-tuning with EGs.

enhanced LLMs significantly reduce attestation bias, addressing the challenges of hallucination in NLI. Additionally, we report the reduction of false positives on *attested* hypotheses in Appendix C, indicating reduced reliance on the model's memorization of the training corpus.

We observe that the reduction in attestation bias is more pronounced in smaller models, compared to larger models, such as LLaMA-3-70B. This suggests that the fine-tuning with our EGs is more effective for smaller-sized models, as larger models require more extensive EGs data for fine-tuning. To validate this, we fine-tune LLMs with increasing amounts of training data generated from EGs. Our results indicate that attestation bias in LLaMA-3-70B continues to decline as training data increases, reaching levels comparable to those observed in smaller LLMs (detailed in Appendix D).

In addition to the *RPI* task, which is explicitly designed to be bias-adversarial, we also present the AttBias scores on the original Levy/Holt dataset in Figure 3, which demonstrate a consistent reduction in attestation bias following fine-tuning with EGs.

The significant reduction in attestation bias can be attributed to the counterfactual nature of our EG-based training corpus. During fine-tuning, the LLMs learn entailment between predicates while incorporating counterfactual knowledge, thereby reducing their reliance on memorized artifacts for inference. Our method minimizes hallucinations from attestation bias and enhances overall robustness.

6 Experiment 2: Evaluate Capability of Inference

EG-enhanced LLMs exhibit a notable reduction in attestation bias. To further investigate how this attestation bias reduction affects their capability of inference, we evaluate their general inferential performance on original NLI tasks.

However, attestation biases and reported data leakage (Gururangan et al., 2020; Balloccu et al., 2024; Ravaut et al., 2024) often lead to an overestimation of LLMs performance on original NLI dataset, where LLMs rely on memorization to form reasoning shortcuts rather than genuinely reasoning, making it challenging to accurately evaluating their *true* inference capabilities.

To address these challenges, we introduce a method to generate bias-neutralized datasets from original NLI datasets for evaluation. We neutralize biases by replacing the original entities with others of the same type, generating new inference data that includes counterfactual statements while preserving original entailment labels. We then evaluate our EG-enhanced LLMs on both ***original*** and ***bias-neutralized*** NLI test sets, enabling a more accurate assessment of their *true* inferential capabilities.

6.1 Neutralizing Biases

To neutralize the biases arising from memorization, we generate counterfactual premise-hypothesis pairs from the original dataset by randomly replacing the original entities with other entities of the same type. Since the entities are randomly selected, the newly generated pairs are logically sound but are likely not to exist in original training corpora. This method generates novel statements without changing their entailment labels, enabling an unbiased assessment of the model's true inferential capability without interference of memorization.

We use entity type constraints here to ensure polysemous predicates maintain the same sense. For instance, the verb "run" has different meanings in "[person] runs [organization]" versus "[person] runs [software]", but when substituting entities of the same type, the sense remains consistent. Therefore, the specific entity names do not affect the entailment labels (Yarowsky, 1993). Notably, this approach ensures that the dataset remains distinct from the original inference data and can allow for building new datasets for every experiment, thereby reducing potential overestimation caused by evaluation data leakage.

Task	Sample Query: [premise] ⇒ [hypothesis]	Dataset Label
LH	**George Bush** was the Governor of **Texas** ⇒ **George Bush** is a politician from **Texas**	`Entail`
$LH_{rpArg\downarrow}$	**Jan Hus** was the Governor of **Svaneti** ⇒ **Jan Hus** is a politician from **Svaneti**	`Entail`
$LH_{rpArg\uparrow}$	**Elon Musk** was the Governor of **Paris** ⇒ **Elon Musk** is a politician from **Paris**	`Entail`

Table 3: An exapmle of generating $LH_{rpArg\downarrow}$ and $LH_{rpArg\uparrow}$ from original Levy/Holt data (*LH*).

6.2 Bias-neutralized Test Sets

We apply this method to construct bias-neutralized test sets from the original NLI datasets.

Replaced Argument LevyHolt (LH_{rpArg}): All original entities in Levy/Holt are named entities. We replace them with other real entities of the same type, which are extracted using the named entity linking tool Aidalight (Nguyen et al., 2014) from NewsCrawl (Barrault et al., 2019), a decade-long span of multi-source news text, in which entities are typed into FIGER types (Ling and Weld, 2012).

Pre-trained LLMs are likely to contain more memorized knowledge about high-frequent named entities. Thus these generated counterfactual samples involving high-frequency entities have higher probability to conflict with the model's memorized knowledge, exacerbating attestation biases. To further analyze the effect of entities frequency, we sample new entities uniform randomly from the 5% least common entities in NewsCrawl ($LH_{rpArg\downarrow}$), and the 5% most common ($LH_{rpArg\uparrow}$), separately. We insert the sampled entities while preserving the rest of each statement. Examples are shown in Table 3.

Replaced Argument SNLI ($SNLI_{rpArg}$): Since the SNLI dataset typically consists of general sentences without named entities, making it impractical to directly extract real named entities from NewsCrawl to replace these general entities.

To address this, we adopt the approach proposed by Liu et al. (2024a), using ChatGPT to identify entities with their types that co-occur in both the hypothesis and premise (these are general entities, not limited to named entities) and then generate new entities that match the specified types to replace them. For instance, an original SNLI sample such as "*John gives Mary an apple ⇒ Mary receives an apple from John*", will be modified to "*John gives Mary a book ⇒ Mary receives a book from John*".

We manually check 50 samples from LH_{rpArg} and $SNLI_{rpArg}$ separately, confirming that all of them are logically sound. Additionally, we also examine the attestedness of hypotheses in LH_{rpArg} and $SNLI_{rpArg}$ and find only 0.89% are attested, ensuring that these bias-neutralized test sets are logically consistent and bias-free.

Models	Levy/Holt	SNLI
DeepSeek-R1-Llama-8B	69.25	85.06
DeepSeek-R1-Llama-8B$_{CoT}$	65.65	85.26
DeepSeek-R1-Llama-8B$_{EG}$	**71.49**	**85.80**
Mistral-7B	69.78	85.47
Mistral-7B$_{CoT}$	65.14	83.63
Mistral-7B$_{EG}$	**72.82**	**85.64**
LLaMA-3-8B	66.87	**87.49**
LLaMA-3-8B$_{CoT}$	62.40	85.63
LLaMA-3-8B$_{EG}$	**73.69**	86.62
LLaMA-3-70B	77.40	**90.01**
LLaMA-3-70B$_{CoT}$	76.53	89.03
LLaMA-3-70B$_{EG}$	**77.46**	89.85

Table 4: AUC scores of original, EG-enhanced ($_{EG}$) and with chain-of-thought prompts ($_{CoT}$) LLMs versions on original Levy/Holt and SNLI.

6.3 Scoring Inferential Capability

Following Mckenna et al. (2023), we analyze model performance across varying confidence thresholds by converting letter choices into probabilities using the following mapping:

$$S_{\text{ent}} = 0.5 + 0.5 * \mathbb{I}[\text{tok} = \mathbf{A}] * S_{\text{tok}} - 0.5 * \mathbb{I}[\text{tok} = \mathbf{B}] * S_{\text{tok}}$$

Where $\mathbb{I}$ is the indicator function, and S_{ent} estimates the probability of `Entail` from a textual output ($0 \leq S_{\text{ent}} \leq 1$) with token probability S_{tok}. The linear transformation preserves the ordering of model confidences, which is sufficient for calculating a precision-recall curve and **Area Under the Curve (AUC)** score.

6.4 Results of Experiment II: Performance in Inference Tasks

Table 4 presents the AUC scores for original NLI datasets, demonstrating that EG-enhanced LLMs consistently outperform original LLMs across various model families on the Levy/Holt dataset. Notably, smaller models exhibit significant performance gains from EGs tuning, surpassing the improvements observed in larger models. On the original SNLI dataset, our EG-enhanced models yield

Models	Tasks					
	$LH_{rpArg\downarrow}$		$LH_{rpArg\uparrow}$		$SNLI_{rpArg}$	
	AUC	Δ_{AUC}	AUC	Δ_{AUC}	AUC	Δ_{AUC}
DeepSeek-R1-Llama-8B	66.17	-	61.04	-	76.13	-
DeepSeek-R1-Llama-8B$_{CoT}$	59.57	-6.6	56.65	-4.39	69.66	-6.47
DeepSeek-R1-Llama-8B$_{EG}$	**67.37**	+1.2	**68.92**	+7.87	**77.87**	+1.74
LLaMA-3-8B	61.80	-	59.05	-	78.31	-
LLaMA-3-8B$_{CoT}$	53.69	-8.11	54.95	-4.1	70.11	-8.20
LLaMA-3-8B$_{EG}$	**71.27**	+9.47	**70.96**	+11.91	**80.03**	+1.72
Mistral-7B	61.27	-	59.96	-	78.17	-
Mistral-7B$_{CoT}$	59.78	-1.49	57.52	-2.44	75.62	-2.55
Mistral-7B$_{EG}$	**71.20**	+9.93	**72.27**	+12.31	**80.43**	+2.26
LLaMA-3-70B	71.99	-	69.55	-	80.26	-
LLaMA-3-70B$_{CoT}$	70.65	-1.34	67.34	-2.21	80.10	-0.16
LLaMA-3-70B$_{EG}$	**76.14**	+4.15	**76.71**	+7.16	**83.29**	+3.03

Table 5: AUC scores of LLMs, LLMs with chain-of-thought prompt ($_{CoT}$) and their EG-enhanced versions ($_{EG}$) on bias-neutralized inference datasets.

only modest improvements. One possible explanation for this limited improvement is data leakage (Balloccu et al., 2024), as the original SNLI dataset may contain more memorized instances from pretrained LLMs, leading to an overestimation of inferential capacities.

To assess the true capability of inference, we evaluate bias-neutralized inference test sets and present the AUC scores in Table 5. The results highlight consistent improvements achieved by our EG-enhanced models across all bias-neutralized datasets. Additionally, the Precision-Recall curve in Appendix B further illustrates that EG-enhanced LLMs outperform the original models, with particularly notable gains in smaller LLMs. Notably, across all bias-neutralized inference datasets, the EG-enhanced smaller models (DeepSeek-8B$_{EG}$, LLaMA-3-8B$_{EG}$ and Mistral-7B$_{EG}$) achieve performance on the same level of LLaMA-3-70B. These results suggest that after training on our EGs, smaller LLMs achieve inferential capabilities comparable to the standard extreme large models. We further examine the impact of prompt templates in Appendix A, proving that LLMs are genuinely learning textual entailment during fine-tuning rather than merely memorizing the prompts.

In Table 5, we also observe that standard LLMs are limited in $LH_{rpArg\uparrow}$ and the improvement achieved by our method on $LH_{rpArg\uparrow}$ is more pronounced compared to $LH_{rpArg\downarrow}$. This suggests that standard LLMs face greater challenges when processing counterfactual information involving high-frequency entities. The observation indicates that LLMs overly rely on memorization to build reasoning shortcuts, ultimately undermining their reasoning performance on high-frequent counterfactual tasks, which have higher probability to conflict with the model's memorized knowledge. On the other hand, the improvement of our EG-enhanced models on $LH_{rpArg\uparrow}$ highlights their enhanced inferential capability and robustness, particularly in handling conflicts between given premise and memorization.

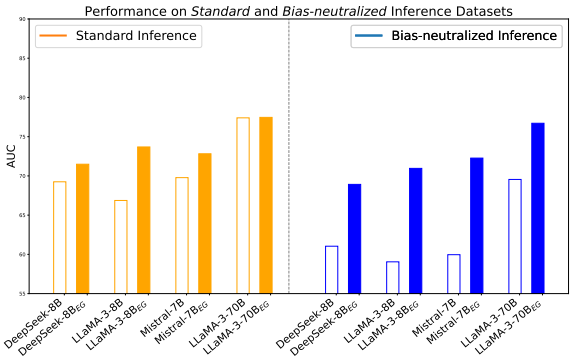

Figure 4: AUC scores of baseline (outline) and EG-trained (solid) LLMs on original (orange) and our bias-neutralized (blue) Levy/Holt.

Figure 4 presents that EG-enhanced model exhibit a more pronounced improvement on bias-neutralized test set compared to original test sets. The reduction of attestation bias explains these findings, as EG-enhanced models increasingly rely on their enhanced inferential capabilities to process NLI tasks rather than retrieving memorized information from training corpus.

7 Errors Analysis

We have shown that training LLMs on our counterfactual entailment graphs can significantly mitigate

Predictions	Models	Label = positives		
		Prem	Hypo	more frequent
Correct	LLaMA-8B$_{EG}$	14.92	66.27	Hypo
	DeepSeek-8B$_{EG}$	15.78	79.85	Hypo
	Mistral-7B$_{EG}$	18.23	75.67	Hypo
	LLaMA-3-70B$_{EG}$	13.83	103.23	Hypo
Incorrect	LLaMA-8B$_{EG}$	69.18	23.45	Prem
	DeepSeek-8B$_{EG}$	25.64	15.55	Prem
	Mistral-7B$_{EG}$	18.39	15.66	Prem
	LLaMA-3-70B$_{EG}$	23.60	16.02	Prem

Table 6: Predicate frequencies in Premise and Hypothesis. EG-tuned models tend to make incorrect predictions when the premise contains higher-frequency predicates than the hypothesis.

hallucinations while improving their inference capabilities. To further understand model behavior, we conduct a series of experiments analyzing the errors of EG-tuned models. Specifically, we examine predicate frequencies in each statement, separately analyzing cases where the model produces correct versus incorrect predictions.

As shown in Table 6, when EG-tuned LLMs produce incorrect predictions, the premise tends to contain higher-frequency predicates than the hypothesis. This suggests that EG-tuned models perform well on low-to-high frequency generalizations, but struggle in the reverse direction.

8 Conclusion

In NLI tasks, attestation bias in LLMs leads to false positives, as models rely on memorization rather than reasoning from the given premise, ultimately undermining their robustness and accuracy in inference. To address this, we propose an unsupervised method for constructing logically consistent counterfactual EGs to fine-tune LLMs, enhancing their robustness against attestation bias. Experimental results show that our method reduces attestation bias and enhances inference by learning counterfactual EGs, enabling models to learn predicate entailment while without introducing artifacts, thereby minimizing dependency on memorized patterns. Our method improves robustness and effectiveness of models in practical applications while providing a more objective evaluation of inference capability across LLMs.

9 Limitation

In this paper, we propose an unsupervised approach to construct counterfactual reasoning data by instantiating entailment graphs and demonstrate its effectiveness in reducing hallucinations, enhancing the reasoning capabilities of LLMs.

A limitation of our current work is that the counterfactual reasoning data is used to fine-tune large LLMs for a specific task, natural language inference, rather than across a broader range of tasks. Although inference is foundational to many NLP tasks, it remains uncertain whether our approach will generalize effectively to other tasks. In future work, we plan to integrate counterfactual reasoning data into instruction-tuning frameworks to evaluate its performance across a wider variety of NLP tasks.

10 Acknowledgments

This research was supported the Universit y of Edinburgh Huawei Laboratory.

References

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages

632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Liang Cheng, Mohammad Javad Hosseini, and Mark Steedman. 2023. Complementary roles of inference and language models in QA. *EMNLP 2023*, 550(2023):75.

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.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Ying Shen, and Wai Lam. 2023. Nonfactoid question answering as query-focused summarization with graph-enhanced multihop inference. *IEEE Transactions on Neural Networks and Learning Systems*.

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.

Yanai Elazar, Bhargavi Paranjape, Hao Peng, Sarah Wiegreffe, Khyathi Chandu, Vivek Srikumar, Sameer Singh, and Noah A Smith. 2024. Measuring and improving attentiveness to partial inputs with counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3603–3623.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-AXIOM: A functional and static benchmark for measuring higher level mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

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.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2023. Using natural language explanations to improve robustness of in-context learning for natural language inference. *arXiv preprint arXiv:2311.07556*.

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*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. 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.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 94–100. AAAI Press.

Tianyang Liu, Tianyi Li, Liang Cheng, and Mark Steedman. 2024a. Explicit inductive inference using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15779–15786.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. 2024b. 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. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774.

Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. AIDA-light: High-throughput Named-entity Disambiguation. In *Workshop on Linked Data on the Web 2014*, pages 1–10. CEUR-WS. org.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *NAACL HLT 2018*, page 180.

Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. 2024. How much are llms contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*.

Vipula Rawte, A. Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *ArXiv*, abs/2309.05922.

Siva Reddy, Mirella Lapata, and Mark Steedman. 2014. Large-scale Semantic Parsing without Question-Answer Pairs. *Transactions of the Association for Computational Linguistics*, 2:377–392.

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.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, MA.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the Generalization Power of Neural Network Models across NLI Benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y Wong, and Simon See. 2024. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. *arXiv preprint arXiv:2402.10646*.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

David Yarowsky. 1993. One sense per collocation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Congle Zhang and Daniel S Weld. 2013. Harvesting Parallel News Streams to Generate Paraphrases of Event Relations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786.

A Analyzing the effects of prompt templates

One concern is whether the observed improvements in LLMs performance stem from learning the predicate entailment in EGs or simply memorizing the structure of query prompt templates. To determine this, we conducted a series of controlled experiments where the predicates in the EGs were randomly shuffled while preserving the original prompt structure during fine-tuning. These controlled datasets contain incorrect entailment relations between predicates but maintain the same prompt template. For instance, a shuffled example might present: "If Steve Jobs *was assassinated in* London, then Steve Jobs *was born in* London. Is that true or false?". As shown in Table 7, the results show a significant performance drop (almost random predict) after fine-tuning on the randomly shuffled EGs. This confirms that the LLMs are indeed learning textual entailment during fine-tuning, rather than simply memorizing prompt templates.

Models	Levy/Holt	SNLI
DeepSeek-R1-Llama-8B	69.25	85.06
DeepSeek-R1-Llama-8B$_{EG}$	**71.49**	**85.80**
DeepSeek-R1-Llama-8B$_{randEG}$	52.0	51.76
Mistral-7B	69.78	85.47
Mistral-7B$_{EG}$	**72.82**	**85.64**
Mistral-7B$_{randEG}$	52.77	50.34
LLaMA-3-8B	66.87	**87.49**
LLaMA-3-8B$_{EG}$	**73.69**	86.62
LLaMA-3-8B$_{randEG}$	51.03	49.07
LLaMA-3-70B	77.40	**90.01**
LLaMA-3-70B$_{EG}$	**77.46**	89.85
LLaMA-3-70B$_{randEG}$	56.25	52.91

Table 7: The AUC score of original, EG-enhanced ($_{EG}$), with chain-of-thought prompts ($_{CoT}$) and the random-shuffled-EG-enhanced ($_{randEG}$) LLMs versions on original Levy/Holt and SNLI dataset.

B Precision-Recall curve

To reduce biases arising from memorization, we evaluate bias-neutralized inference dataset and present the Precision-Recall curve for LH_{rpArg} in Figure 6. The results show that our EG-enhanced LLMs outperform original models, with particularly significant improvements observed in smaller LLMs. Additionally, Figure 7 shows the Precision-Recall curve for the original LH dataset, further demonstrating the consistent benefits of fine-tuning with EGs.

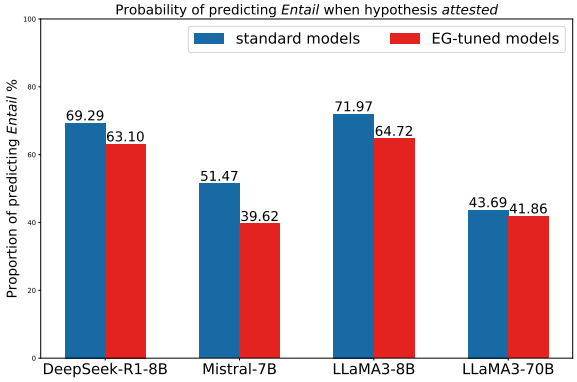

Figure 5: The probability of predicting `Entail` for *RPI* LevyHolt, conditioned on the LLMs' attestation of the hypothesis. Since predicting `Entail` in this context represents a false positive hallucination, a lower probability is better. The image clearly shows that hallucination decrease significantly after fine-tuning with EGs.

Sizes of EGs	Model	
	LLaMA-3-8B$_{EG}$	LLaMA-3-70B$_{EG}$
0 samples	23.12	19.20
2k samples	14.03	18.63
5k samples	12.30	11.93
11k samples	5.99	8.34
18k samples	4.87	5.81

Table 8: The AttBias scores of various LLMs fine-tuned with different sizes of EG datasets.

C Proportion of False Positives on *PRI*

Figure 5 presents the estimated probability of predicting *Entail* when the hypothesis is attested, highlighting a consistent decrease after fine-tuning with EGs across all LLMs. This decrease presents fewer false positives on *attested* hypotheses, indicating reduced reliance on the model's memorization of the training corpus.

D Fine-tuning LLMs with different sizes EGs

To evaluate the performance of LLMs trained on varying sizes of generated counterfactual data, we present the attestation bias scores in Table 8. The results indicate that smaller LLMs are more sensitive to fine-tuning, while extremely large LLMs require larger amounts of generated data during the fine-tuning process.

E Prompt Format Selection

Prompt templates are widely acknowledged for their significant and sometimes decisive impact on the behavior of LLMs. In our experiments, we

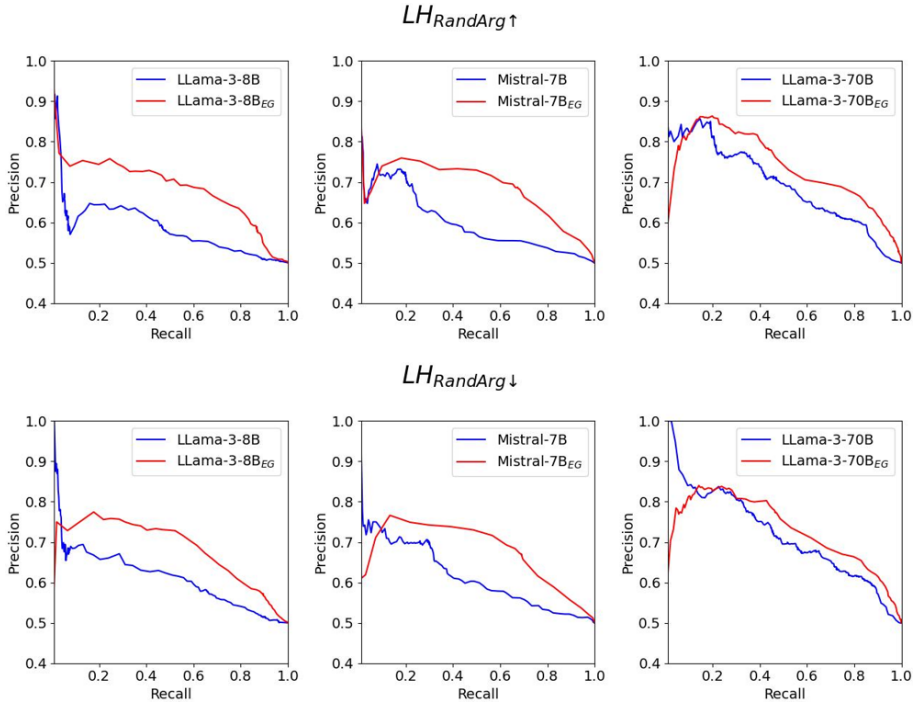

Figure 6: The Precision-Recall curve of our method compared to standard LLMs on $LH_{rpArg\downarrow}$ and $LH_{rpArg\uparrow}$ datasets.

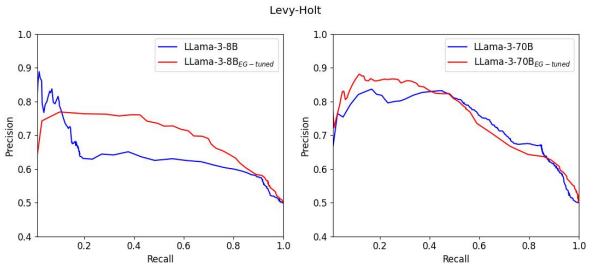

Figure 7: The Precision-Recall Curve of Levy/Holt across LLMs and the EG-tuned version.

categorize the prompt templates into three distinct types based on their usage: (a) prompt templates for fine-tuning LLMs, (b) prompt templates for hypothesis attestation, and (c) prompt templates used during inference.

E.1 prompt template for fine-tuning

We adopt the manually crafted prompts utilized in prior inference studies (Schmitt and Schütze, 2021; Mckenna et al., 2023), 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 choice, we automatically provide explanations for every answer by adding affirmation or negation to the propositions. As a result, the data in our counterfactual reasoning training corpus will be structured as Table 9 shown. We fine-tune our method on these templates.

E.2 prompt template for attesting hypothesis

In attesting hypothesis process, we using the same prompt discussed in §E.1, but mask the premise. Since the model may not be able to definitively determine whether the hypothesis is true or false, we use a question with three choices to evaluate the hypothesis, as (A) True, (B) Unknown, and (C) False.

In-context examples have been widely used for interacting with LLMs since Brown et al. (2020). Moreover, Wei et al. (2022) demonstrated that incorporating chain-of-thought reasoning, step-by-step explanations, into in-context examples enhances LLMs’ understanding of tasks. To improve the model’s comprehension of the attestation task, we include a minimal set of three examples with explanations in few-shot settings, ensuring that each choice is represented by a corresponding example. We present the manual-crafted examples in Table 10.

E.3 prompt template for inference

Similar to the hypothesis attestation experiments (§E.2), we employ the few-shot settings in our prompts for inference tasks. Specifically, we manually created four examples for the inference tasks, consisting of two positive and two negative instances. These examples are presented in Table 10.

E.4 Chain-of-Thought Prompts

Similar to the prompts used for inference (§E.3), we employ the few-shot settings in our prompts for inference tasks. Consistently, we manually created four examples for the inference tasks, each with step-by-step explanations to guide LLMs through reasoning. These examples are presented in Table 11.

F Computing Costs

In our experiments, the extraction and learning of entailment graphs from the NewsSpike corpus required approximately 220G of CPU resources over a span of 20 hours. For the fine-tuning process, we employed four NVIDIA RTX A6000 GPUs to fine-tune the LLaMA-3-70B models, a process that took 21 hours to complete. This setup ensured efficient resource utilization while achieving optimal performance for our large-scale model fine-tuning.

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

Table 9: The table present the prompt template using in our training steps.

A. Few-shot Examples Instantiated Prompt for Inference Task
If Google bought Youtube, then Google owns Youtube. Is that true or false? A) True B) False Answer: A) True. Owning is a consequence of buying. If Google owns Youtube, then Google bought Youtube. Is that true or false? A) True B) False Answer: B) False. Owning does not imply buying, the ownership may come from other means. If John went to the mall, then John drove to the mall. Is that true or false? A) True B) False Answer: B) False. John may have gone to the mall by other means. If John drove to the mall, then John went to the mall. Is that true or false? A) True B) False Answer: A) true. Driving is a means of going to the mall. If John F. Kennedy was killed in Dallas, then John F. Kennedy died in Dallas. Is that true or false? A) True B) False Answer:
B. Few-shot Examples Instantiated Prompt for Attesting Hypothesis
Google bought Youtube. Is that true or false? A) True B) Unknown C) False Answer: A) True. Yoshua Bengio likes oak trees. Is that true or false? A) True B) Unknown C) False Answer: B) Unknown. The sun rises from the west. Is that true or false? A) True B) Unknown C) False Answer: C) False. Answer:

Table 10: Example instantiated prompts in Few-shot settings, for the sample "PREMISE: [Google bought Youtube], HYPOTHESIS: [Google owns Youtube]". The few-shot prompts in part B are used throughout the main experiments in this paper. We also present an example of the prompts we use for the hypothesis-only measure as described in §E.2.

C. Few-shot Chain-of-Thought Prompts
If Google bought YouTube, then Google owns YouTube. Is that true or false? A) True B) False Explanation: 1. Analyze Premise: the premise describe Google bought YouTube. 2. Analyze Hypothesis: the hypothesis state that Google owns the YouTube. 3. Reasoning: company A bought company B, it means that the company B belongs to company A now. So the premise entails hypothesis. Answer: A) True. If Google owns YouTube, then Google bought YouTube. Is that true or false? A) True B) False Explanation: 1. Analyze Premise: the premise state that Google owns the YouTube now. 2. Analyze Hypothesis: the hypothesis describe Google bought YouTube. 3. Reasoning: owning does not imply buying, the ownership may come from other means. So the premise does not entail hypothesis. Answer: B) False. If John went to the mall, then John drove to the mall. Is that true or false? A) True B) False Explanation: 1. Analyze Premise: the premise state that a person go to someplace. 2. Analyze Hypothesis: the hypothesis describe a person went to someplace by driving car. 3. Reasoning: A person may have gone to the place by other ways, so the premise does not entail hypothesis. Answer: B) False. If John drove to the mall, then John went to the mall. Is that true or false? A) True B) False Explanation: 1. Analyze Premise: the premise describe a person went to someplace by driving car. 2. Analyze Hypothesis: the hypothesis state that a person go to someplace. 3. Reasoning: Driving present the way to went to the place, so the premise entails hypothesis. Answer: A) true. If John F. Kennedy was killed in Dallas, then John F. Kennedy died in Dallas. Is that true or false? A) True B) False Answer:

Table 11: Examples of CoT prompts, contains 3 steps: (1) analyze premise. (2) analyze hypothesis. (3) finding the relation between hypothesis and premise.