Sentence-Level Soft Attestation Bias of LLMs

Anonymous ACL submission

Abstract

While many large language models (LLMs) have demonstrated evolving reasoning ability, they are reported to hold attestation bias in inference tasks. Instead of focusing on entailment signals between a premise and a hypothesis, LLMs are easily misled by whether the hypothesis is factual in the models' knowledge. However, previous study on attestation bias requires the factuality of input sentences to be determinable, which is often not true for inference tasks. In this paper, we propose soft attestation, a measurement compatible with all kinds of NLI datasets. Then we implement a sentence-level explicit inductive inference pipeline. By reporting its performance against attestation bias on three NLI datasets with four mainstream LLMs, we demonstrate that the attestation bias persists as a severe problem on sentence-level inference, yet it can also be exploited to improve LLMs' inference performance.[1]

1 Introduction

While many contemporary large language models (LLMs) claim to have strong reasoning ability, recent studies have shown that they are subject to severe attestation bias in fundamental natural language inference tasks (McKenna et al., 2023). When a model is asked to predict whether a premise entails a hypothesis, instead of focusing on the inference signal, an LLM is easily distracted by the hypothesis's out-of-context factuality. As a result, LLMs usually perform worse when the entailment label between premise and hypothesis disagrees with the hypothesis's attestation (factuality) label.

Previous work has proposed the idea of explicit inductive inference as a solution. By doing inference on alternative entailment inquiries created by LLMs themselves, the attestation bias can be utilized to mitigate itself, and therefore improve LLMs' performance on triple-level inference tasks (Liu et al., 2024). However, in most mainstream natural language inference (NLI) challenges and downstream applications, entailment inquiries are often not presented as pairs of well-structured triples. An inference model is usually expected to pick up the entailment structure between sentences on its own, which limits the scenario where this pipeline can be used.

Furthermore, sentences in many NLI datasets do not have a well-defined factuality label. Previous discussion on attestation bias restrict the inputs to be either factual or counter-factual by locating named entities, which excludes a lot of NLI datasets and therefore limits its perspective. This encourages us to generalize the process of detecting attestation bias, and further examine on more datasets if the attestation bias is still harmful to the latest LLMs.

Based on these motivations, this paper first defines the soft attestation measurement. Then we probe the attestation bias that exist in latest LLMs, and apply the idea of explicit inductive inference to sentence-level datasets to build a pipeline as a solution. The contributions in this paper can be summarized as follows:

1. We discuss a flaw in previous studies of attestation bias, and propose the definiton of soft attestation as a better measurement that is compatible with more general NLI datasets.

2. By applying this measurement, we expand the investigation of attestation bias to more various NLI datasets with several SOTA LLMs, and show that the attestation bias still remains a severe problem.

3. We implement a sentence-level inductive inference pipeline. We demonstrate that this pipeline can improve LLMs' overall performance on NLI task. Under certain circumstances, we show that it can also mitigate the attestation bias by exploiting

[1] Codes and data on this paper will be released upon publication.

the bias itself.

2 Related work

It is widely observed that LLMs accumulate a bias towards facts that they memorized from a vast amount of pre-training corpus (Roberts et al., 2020; Carlini et al., 2022; Yan et al., 2022). In specific reasoning scenarios like solving math problems, this bias causes LLMs to perform worse even when only variable names are changed (Gulati et al., 2024).

For inference tasks, McKenna et al. (2023) designed a list of experiments specifically focused on the attestation bias, and showed that it causes LLMs to expose a significant performance drop at inference time. The following works have also confirmed that this bias keeps affecting the inference performance of newer LLMs (Liu et al., 2024).

Other works proposed different ways to alleviate the negative effect of the attestation bias by presenting models with counterfactual examples (Wang et al., 2022; Zhou et al., 2023; Wang et al., 2023) or using type labels to mask the entities (Zhou et al., 2024).

For LLMs, Liu et al. (2024) proposed a novel explicit inductive inference pipeline that can utilize the attestation bias to mitigate itself. On triple-level inference tasks, this work argues that if the premise can be controlled to be attested, the attestation label of the hypothesis will then statistically align with the entailment label between the premise and the hypothesis, which makes the attestation bias unharmful. Although this idea brings an intriguing possibility, whether it can be applied to more complex circumstances remains an open question, which we will try to answer in this paper.

3 Measuring attestation bias

The attestation bias is clearly defined by the relation between the hypothesis's attestation label and the premise-hypothesis pair's entailment label. The attestation label itself, on the other hand, can sometimes be hard to determine. The factuality label of a sentence describing real-world knowledge like "Steve Jobs was born in San Francisco" can be supported or denied by evidence or the model's knowledge. However, it is nearly impossible to determine the factuality of sentences like "Tom likes Mary" without context.

Previous work has carefully avoided them by selecting only the datasets derived from corpora where factuality can be determined, e.g. news articles. Yet, those general sentences appear in many mainstream NLI datasets, which urge us to include them into the discussion of attestation bias.

In this paper, we try to improve the measurement of attestation. When assigning attestation (factuality) labels to sentences, besides "attested" (**A**) and "non-attested" (**NA**), we now allow another label, "factuality can not be determined" (**ND**), to be an option. In previous work, the **ND** label either does not exist or only exists in prompting but is then merged into the **NA** label in later analysis. We note that definition of attestation as the **hard attestation**.

In contrast, we propose **soft attestation bias**. We argue that it is counter-intuitive to force the factuality label to be binary. While the boundary between **A** and **ND** can be vague, a model usually has strong evidence to support an **NA** claim. Therefore, we define the soft attestation by observing whether the probability $P(\mathbf{NA})$ is higher than both $P(\mathbf{A})$ and $P(\mathbf{ND})$, but not necessarily $P(\mathbf{A})+P(\mathbf{ND})$ which is the case where the **A** category is merged with the **ND** category.

Formally, we call a sentence soft attested when $P(\mathbf{NA}) > P(\mathbf{A}) \wedge P(\mathbf{NA}) > P(\mathbf{ND})$. Moreover, we define a pair of premise and hypothesis soft attestation-consistent if their entailment label agrees with the soft attestation label of the hypothesis. Otherwise, we note them as soft attestation-adversarial.

In all the following experiments, we always display results under both the hard and the soft attestation measurement.

4 Sentence-Level Explicit Inductive Inference

In this section, we explain the design of the **S**entence-**L**evel **E**xplicit **I**nductive **I**nference (**SLEII**) pipeline. Similar to the triple-level explicit inductive inference pipeline, the methodology of the SLEII pipeline is to explicitly help an LLM to expand a single entailment inquiry to a list of alternatives, and inductively draw a more reliable conclusion from all the similar cases. However, unlike triples which each possess one object and one subject, sentences include more complex semantic components. It is therefore necessary to design modules that can handle any kind of input sentence.

In the SLEII pipeline, each pair of one premise and one hypothesis sequentially goes through four modules. To better present the function of each module, we take the following pair of sentences as an example to illustrate how this pipeline works:

> Premise: John is eating chocolate after meeting Mary.
>
> Hypothesis: He received some chocolate as a gift.

Now we introduce the four modules one by one by process this sentence pair as an example.

4.1 Alignment

This module decides which part of the sentences should be substituted to create alternative sentences. By creating variations of the original sentence pair, we aim to alter the meaning of the sentences within a reasonable range while keeping the entailment label between the sentences unchanged. To achieve this, we only replace those entities that appear in both sentences. This module locates these entities and tags them with type labels.

Type labels are necessary here to prevent ambiguity problems. While (X kills Y) entails (X is a cure of Y) when X is medicine and Y is a disease, in most of the other cases their meaning is entirely opposite. To avoid this kind of undesirable substitution, the alignment module is asked to provide a contextualized type label for each entity. Here, different mentions of the same entity can have different type labels.

We encourage this module to do reference resolution between the premise and the hypothesis. A reference counted as an entity may be tagged with the type labels like "pronoun". For instance, after going through this module, the example will become:

> Premise: [entity#1: person] is eating [entity#2: food] after meeting Mary.
>
> Hypothesis: [entity#1: pronoun] received some [entity#2: food] as a gift.

4.2 Premise variation

Once we obtain the tagged premise and hypothesis, we independently instantiate the premise into alternative variations. We encourage the LLM to write factual sentences if possible. These new premises are instantiated with entities that are more familiar to the LLM, and therefore these premises are more likely to be attested.

For each premise, we create k different new alternatives. Now our example looks like this:

> Premise$'_1$: Jane is eating an apple after meeting Mary.
>
> Premise$'_2$: Steve is eating popcorn after meeting Mary.
>
> ......
>
> Premise$'_k$: Tom is eating chips after meeting Mary.

An explanation of the mapping relations between tags and entities will be generated and passed to the next module.

4.3 Hypothesis instantiation

This module then derives the alternative hypotheses, based on the mapping information received from the previous module on how entities in the corresponding premise are replaced. Now we have a list of k variations of new sentence pairs. The example will look like this:

> Premise$'_1$: Jane is eating an apple after meeting Mary.
>
> Hypothesis$'_1$: She received some apples as a gift.
>
> ......
>
> Premise$'_k$: Tom is eating chips after meeting Mary.
>
> Hypothesis$'_k$: He received some chips as a gift.

4.4 Prediction

Finally, based on each alternative pair of (Premise$'_k$, Hypothesis$'_k$), this module queries the LLM to get an alternative score s_k. Note the score of the original entry as s_0, the final weighted SLEII score S_w is calculated with a weight parameter w:

$$S_w = (1 - w)s_0 + w \sum_{i=0}^{k} s_i \tag{1}$$

5 Experimental setup

5.1 Dataset

To evaluate our pipeline on sentence-level inference tasks, we test the soft attestation measurement and the SLEII pipeline on three NLI datasets.

SNLI The Stanford Natural Language Inference dataset (Bowman et al., 2015) is a classic NLI dataset that provides three-way classification (entailment/contradiction/neutral) entailment inquiries. It contains 570k human-written sentence pairs with crowd-sourced labels, and is widely used for NLI evaluation. Results are reported on the test set.

RTE The Recognizing Textual Entailment dataset from GLUE (Wang et al., 2019) integrate the data from the RTE challenge series. The texts are derived from real-world corpus like news and scientific articles, which makes them a suitable source of potential attestation bias. In this dataset, the "neutral" and "contradiction" labels are merged into the "not_entailment" label. Since the labels of the test set are not accessible for RTE, we report our results on the validation set.

MNLI The Multi-Genre Natural Language Inference dataset (Williams et al., 2018) is also a widely used dataset formatted the same as SNLI but with sentences from more diverse corpus like transcribed speech, fiction, and government reports. The test set's golden labels are also not directly available. We report results on the "dev_matched" development set.

For all datasets, we select the best value of the weight parameter w between 0 to 1 on the training sets, and apply them to the test sets to report results.

5.2 Large language models

We aim to cover the latest state-of-the-art LLMs from various sources. In this paper, our pipeline is tested with four mainstream LLMs that are claimed to have robust reasoning abilities.

GPT-4o mini (OpenAI, 2024) is a cost-efficient variant of OpenAI's GPT-4 architecture. It claims to have powerful reasoning abilities over many natural language understanding benchmarks. The version that we use is "gpt-4o-mini-2024-07-18".

LLama 3 (Meta, 2024) is an open-source LLM published by Meta. In our experiment, we choose the "Llama3-8B-Instruct" version, a smaller version with 8 billion parameters.

Gemini 2.0 Flash (Google, 2024) is an enhanced comprehensive model published by Google. The version we used in our experiments is "gemini-2.0-flash".

Claude 3.5 Haiku (Anthropic, 2024) is an updated version of Anthropic's fastest LLM. The model version that we use is "claude-3-5-haiku-20241022" .

When we need to collect the probability of choices, if the LLM provides token probability access, we use the probability at the output choice mark token (A, B, or C) to represent the choice probability. When the token probabilities are not accessible, we assign 1 to the returned choice and 0 to the others.

To guarantee replicable results, whenever we query one of the LLMs, either the temperature parameter is set to 0, or the "do_sample" flag is turned off.

5.3 Prompts

In each module, the content of the prompts may affect the final results. We tried different prompt variations in our pilot studies on the training sets, and fixed the prompts that we use before doing final experiments on the test sets.

Following Liu et al. (2024), we set the number of variations k to 10 in our experiments. For the alignment, premise variation, and hypothesis instantiation module, we give few-shot examples to guarantee expected results. For the prediction and factuality determination module, we use zero-shot prompts to avoid instability from prompt engineering. More description of the prompts used in our experiments are included in appendix C.

6 Results and discussion

In this section, we first discuss the effect of applying soft attestation measurement. Then we show how the SOTA LLMs perform against both the hard and the soft attestation bias as a baseline. Following that, we report the performance of the SLEII pipeline and its impact on attestation bias.

6.1 Hard attestation vs. soft attestation

As discussed in section 3, we first compare the hard and the soft attestation bias. To better understand how each LLM classifies the datasets, we inspect the data proportion when entries are classified according to their hypotheses' attestation label.

The results in table 1 show how each LLM sets divergent standards for determining factuality. When responding to the same prompt asking about the factuality of a hypothesis, it can be seen that the four LLMs we tested give very different responses. This also demonstrates the distinction between "factuality" and "attestation", as factuality is an objective property, while attestation represents the sentence's faithfulness within an LLM.

Model	**Att. measure**	**Dataset**								
		SNLI			RTE			MNLI		
		A	NA	ND	A	NA	ND	A	NA	ND
		cons.		***adv.***	***cons.***		***adv.***	***cons.***		***adv.***
Llama3-8B		83%	16%	0%	91%	8%	0%	56%	42%	1%
	hard	*42%*		*56%*	*59%*		*41%*	*55%*		*42%*
	soft	*42%*		*56%*	*59%*		*41%*	*56%*		*42%*
GPT4o-mini		5%	90%	4%	17%	58%	25%	39%	42%	18%
	hard	*56%*		*42%*	*43%*		*56%*	*44%*		*53%*
	soft	*56%*		*43%*	*44%*		*55%*	*43%*		*56%*
Gemini-2.0		20%	1%	80%	65%	22%	12%	31%	13%	54%
	hard	*8%*		*8%*	*49%*		*29%*	*27%*		*12%*
	soft	*36%*		*63%*	*58%*		*42%*	*47%*		*51%*
Claude-3.5		47%	1%	51%	61%	13%	36%	24%	4%	72%
	hard	*23%*		*32%*	*43%*		*25%*	*19%*		*9%*
	soft	*36%*		*62%*	*56%*		*43%*	*51%*		*47%*

Table 1: Percentages of the data size when the hypothesis of an entry is determined by the LLM to be attested (A), not attested (NA), or its factuality can not be determined (ND). ***cons.*** and ***adv.*** note the size of the attestation-consistent subset and the attestation-adversarial subset.

While Llama3-8B barely assigns **NA** labels to any hypothesis, Gemini-2.0 and Claude-3.5 consider the majority of hypotheses in the SNLI and MNLI datasets as "their factuality can not be determined". Yet, the size ratios between ***cons***. and ***adv.*** are mostly close to fifty-fifty splits, indicating that attestation-adversarial inference is an essential sub-task in various NLI scenarios.

The size percentage of the attestation-consistent (***cons.***) subsets and the attestation-adversarial (***adv.***) subsets are relative to the entire datasets. For Gemini-2.0 and Claude-3.5, we observed a negative correlation between the number of **NA** labels a model assigns to hypotheses and the total data coverage under hard attestation setting, as the ratios of ***cons.*** and ***adv.*** do not add up to one. This happens because the logit scores of output tokens in these two models are not available. By default, we assign 0, 0, 1 to the probability of the **A**, **NA**, **ND** labels when the model's choice is **ND**. As a result, this kind of entry falls in neither of ***cons.*** nor ***adv.*** under hard attestation setting.

6.2 Attestation bias persists

In contrast to the non-consistent results in the previous section, we now display LLMs' universal vulnerability against attestation bias. Following McKenna et al. (2023) and Liu et al. (2024), we report LLMs' inference performance in the normalized area under the precision-recall curve in table 2. The ***diff.*** columns mark the performance difference between *cons.* and *adv.* settings.

Apart from the exception of GPT4o-mini, the other three models suffer from a severe performance drop from *cons.* to *adv.* under both attestation measures. While a slight performance drop can be evidence of an LLM's real inference ability, the reversed attestation bias effect that GPT4o-mini exposes is beyond our expectation. Since the training data and training method of GPT4o-mini are not publicly available, we are not able to provide a reasonable explanation. Nevertheless, we can draw an overall conclusion that through years of LLM evolution, attestation bias is still a significant obstacle limiting many models from doing contextualized logical inference.

It is also worth noting that the results yielded by the hard and soft attestation measures are very close under most settings. The results in section 6.1 show that the soft attestation measurement guarantees full coverage of every data point, compared to the **ND** hypotheses lost in the hard attestation

Model	**Att. measure**	**Dataset**								
		SNLI			RTE			MNLI		
		cons.	*adv.*	***diff.***	*cons.*	*adv.*	***diff.***	*cons.*	*adv.*	***diff.***
Llama3-8B	*hard*	96.9	23.6	**-73.3**	94.1	39.8	**-54.3**	83.1	30.0	**-53.1**
	soft	96.6	23.8	**-72.8**	94.1	39.8	**-54.3**	83.2	29.8	**-53.4**
GPT4o-mini	*hard*	65.9	97.6	+31.7	75.8	96.9	+21.1	77.7	88.8	+11.1
	soft	90.1	93.0	+2.9	80.6	96.8	+16.2	82.2	86.5	+4.3
Gemini-2.0	*hard*	99.1	51.7	**-47.4**	85.6	67.0	**-18.6**	78.9	65.9	**-13.0**
	soft	99.6	52.5	**-47.1**	86.1	65.1	**-21.0**	86.4	58.5	**-27.9**
Claude-3.5	*hard*	95.3	50.4	**-44.9**	95.3	40.0	**-55.3**	92.1	68.2	**-23.9**
	soft	92.6	51.5	**-41.1**	88.6	42.6	**-46.0**	85.2	48.2	**-37.0**

Table 2: AUC$_{norm}$ (%) scores on attestation-consistent (*cons.*) and attestation-adversarial (*adv.*) subsets under hard and soft attestation measurement. The ***diff.*** columns mark the performance difference from cons. to adv.

measurement. On this basis, we advocate for future works to use soft attestation measurement as well as the hard one, as it suits a larger variety of NLI datasets while fundamentally reflecting the same property of the attestation bias.

6.3 Overall performance of SLEII pipeline

Now we discuss whether sentence-level attestation bias can also be exploited by applying explicit inductive inference. Table 3 shows the overall performance of the SLEII pipeline. Although the attestation-consistent and attestation-adversarial subsets are not involved in this section, we still divide the results under each setting into two columns marked with *hard* and *soft*. Here, the *hard* columns only take into account those entries where the hypotheses have **A** or **NA** label, omitting entries with **ND** hypotheses as the hard attestation measurement does. The soft columns consider every entry, no matter the attestation label of a hypothesis is **A**, **NA**, or **ND**.

The SLEII$_{bw}$ rows mark the results when the weight parameter w in equation 1 is set to the best value selected over the same experiments on training sets. The pure SLEII scores are calculated by setting w to 1, which means it does not look at the original entailment inquiry at all. A table of all w values is included in appendix B.

It can be observed that the pure SLEII pipeline sometimes performs worse than baseline, while SLEII$_{bw}$ reaches best performance under most combinations of LLM and dataset. After looking into the alternative premises and hypotheses generated by the models, we believe this is caused by the improved robustness of inferring over multiple examples. While predicting inference only on generated alternatives without looking at the original entry is risky, when using the alternatives as backup evidence, it can correct cases where the model gets the original inference wrong, without harming ones that are already answered correctly by the baseline.

6.4 Mitigate the attestation bias

Liu et al. (2024) proved with the triple-level explicit inductive inference pipeline that, with careful design, the attestation bias of LLMs can be exploited to mitigate itself. In this section, we present a similar analysis to discuss whether the same effect can be found in sentence-level inference.

Table 4 and 5 show the AUC$_{norm}$ results under the hard and soft attestation measurement respectively. Similar to table 2, the ***diff.*** columns report the performance difference between *cons.* and *adv.* subsets. The ***diff.*** results closest to zero are highlighted as an indicator of more robust inference ability against the attestation bias.

For hard attestation measurement, the pure SLEII pipeline always increases the ***diff.*** value over baseline, contributing to the performance of SLEII$_{bw}$ in many cases. This effect indeed mitigates the negative effect of attestation bias for Llama3-8B, Gemini-2.0, and Claude-3.5, although it further escalates the situation for GPT40-mini due to its reverse attestation bias.

In contrast, the SLEII pipeline's performance under soft attestation measurement is less ideal. In a few cases, the pure SLEII pipeline presents an attestation bias even worse than the baseline, causing

Model	**Pipeline**	**Dataset**					
		SNLI		RTE		MNLI	
		hard	*soft*	*hard*	*soft*	*hard*	*soft*
Llama3-8B	-	78.5	78.4	78.8	78.8	64.0	63.9
	SLEII	51.5	51.5	66.2	66.2	57.3	57.3
	SLEII$_{bw}$	**80.0**	**79.5**	**79.8**	**79.8**	64.2	64.2
GPT4o-mini	-	96.0	93.9	90.7	91.5	85.1	84.5
	SLEII	75.0	84.0	77.4	81.7	71.9	81.9
	SLEII$_{bw}$	96.0	94.4	90.7	91.5	85.1	84.8
Gemini-2.0	-	77.9	84.4	77.0	75.8	74.3	74.3
	SLEII	84.7	87.7	82.4	78.7	77.8	77.5
	SLEII$_{bw}$	**85.0**	**88.8**	**84.1**	**80.6**	**78.1**	**78.0**
Claude-3.5	-	78.7	79.4	73.7	73.5	86.1	72.2
	SLEII	80.1	80.7	76.5	73.2	79.3	77.3
	SLEII$_{bw}$	**83.7**	**84.4**	**79.8**	**76.1**	86.6	**78.6**

Table 3: The overall pipeline performance under each setup (AUC$_{norm}$). SLEII$_{bw}$ marks the results using the best weight w learned from the training set under the same setting.

Model	**Pipeline**	**Dataset**								
		SNLI			RTE			MNLI		
		cons.	*adv.*	***diff.***	*cons.*	*adv.*	***diff.***	*cons.*	*adv.*	***diff.***
Llama3-8B	-	96.9	23.6	-73.3	94.1	39.8	-54.3	83.1	30.0	-53.1
	SLEII	70.6	11.9	**-58.7**	78.5	31.8	**-46.7**	76.0	25.9	**-50.1**
	SLEII$_{bw}$	95.5	29.2	-66.3	92.8	46.0	-46.8	82.0	31.2	-50.8
GPT4o-mini	-	65.9	97.6	+31.7	75.8	96.9	**+21.1**	77.7	88.8	+11.1
	SLEII	35.7	92.2	+56.5	54.1	93.3	+39.2	74.5	86.4	+11.9
	SLEII$_{bw}$	71.8	97.5	**+25.7**	75.8	96.9	**+21.1**	78.4	88.9	**+10.5**
Gemini-2.0	-	99.1	51.7	-47.4	85.6	67.0	-18.6	78.9	65.9	-13.0
	SLEII	98.9	54.1	-44.8	82.4	81.9	**-0.5**	80.8	71.2	-9.6
	SLEII$_{bw}$	99.0	54.6	**-44.4**	85.1	82.0	-3.1	80.8	71.7	**-9.1**
Claude-3.5	-	95.3	50.4	-44.9	95.3	40.0	-55.3	92.1	68.2	-23.9
	SLEII	91.1	48.2	**-42.9**	89.8	46.5	**-43.3**	87.8	73.6	**-14.2**
	SLEII$_{bw}$	94.5	50.1	-44.4	94.4	43.6	-50.8	93.4	72.4	-21.0

Table 4: AUC$_{norm}$ (%) scores on attestation-consistent (*cons.*) and attestation-adversarial (*adv.*) subsets under **hard** attestation measurement. The ***diff.*** column marks the difference from *cons.* to *adv.* The ***diff.*** value closeset to zero under each setting is highlighted.

Model	**Pipeline**	**Dataset**								
		SNLI			RTE			MNLI		
		cons.	*adv.*	***diff.***	*cons.*	*adv.*	***diff.***	*cons.*	*adv.*	***diff.***
Llama3-8B	-	96.6	23.8	-72.8	94.1	39.8	-54.3	83.2	29.8	-53.4
	SLEII	70.5	12.3	**-58.2**	78.5	31.9	**-46.6**	76.2	25.7	**-50.5**
	SLEII$_{bw}$	95.4	29.6	-65.8	92.9	46.1	-46.8	82.1	31.0	-51.1
GPT4o-mini	-	90.1	93.0	**+2.9**	80.6	96.8	**+16.2**	82.2	86.5	+4.3
	SLEII	43.8	88.1	+44.3	56.4	93.6	+37.2	69.9	73.0	**+3.1**
	SLEII$_{bw}$	88.0	94.9	+6.9	80.6	97.3	+16.7	83.0	86.6	+3.6
Gemini-2.0	-	99.6	52.5	**-47.1**	86.1	65.1	-21.0	86.4	58.5	**-27.9**
	SLEII	99.0	33.2	-65.8	84.0	69.7	**-14.3**	87.3	57.4	-29.9
	SLEII$_{bw}$	99.7	33.6	-66.1	86.1	70.5	-15.6	87.7	57.9	-29.8
Claude-3.5	-	92.6	51.5	**-41.1**	88.6	42.6	-46.0	85.2	48.2	**-37.0**
	SLEII	88.7	46.8	-41.9	82.5	44.6	**-37.9**	91.1	47.0	-44.1
	SLEII$_{bw}$	91.6	49.0	-42.6	87.1	45.1	-42.0	91.6	49.5	-42.1

Table 5: AUC$_{norm}$ (%) scores on attestation-consistent (*cons.*) and attestation-adversarial (*adv.*) subsets under **soft** attestation measurement. The ***diff.*** column marks the difference from *cons.* to *adv.* The ***diff.*** value closeset to zero under each setting is highlighted.

the SLEII$_{bw}$ to perform poorly. This comparison suggests that although the idea of exploiting the attestation bias with explicit inductive reasoning is innovative, it may not work effectively when a lot of ND sentences are included.

This proposition can be further supported by focusing on the performance difference between datasets. Under both attestation measurements, SLEII pipeline shows best improvement on RTE and least on SNLI, which is correlative to the nature of the three datasets. RTE is derived from news corpora and therefore includes abundant factual statements, SNLI consists of made-up sentences with nearly no named entities, and MNLI is a mixture between the former two. This further proved that the more sentences in the datasets of which factuality can be determined, the better the explicit inductive inference pipeline works.

To summarize, we argue that mitigating the soft attestation bias is an even more challenging task that requires a model to handle sentences when their factuality can not be determined.

7 Conclusion

In this paper, we first propose the concept of soft attestation, and prove with experiments that it is a better measurement than hard attestation, as it allows more NLI datasets to be included in the research of attestation bias. At the same time, it captures the same attestation proxy as the hard measurement does. We advocate the soft attestation measurement to be used besides the hard one in future studies of attestation bias.

Then we report the severe attestation bias existing in three out of four SOTA LLMs. Under both attestation measurements, the LLMs struggle to achieve average performance on attestation-adversarial subsets. We conclude that the attestation bias is still a severe problem impairing SOTA LLMs' inference ability.

Finally, by showing results of the SLEII pipeline under various experimental settings, we argue that the explicit inductive inference method can be used to mitigate attestation bias on sentence level. Still, it is limited to scenarios where the factuality of input sentences can be determined. For more general inference tasks, in remains a difficult challenge to resolve the attestation bias without relying on factuality of sentences.

Limitations

For many modules in this paper, it is possible that prompt engineering may substantially affect the outcome. For instance, adding and editing few-shot examples are common ways to control the model's response. Further prompt engineering work may be required into these experiments.

Using the SLEII pipeline to infer on k extra variations results in k times of extra resource spent, indicating that this method may not be suitable for computationally intensive cases where a model needs to process a large number of entries.

References

Anthropic. 2024. Claude 3.5.

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.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Google. 2024. Gemini 2.0.

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*.

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

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, Singapore. Association for Computational Linguistics.

Meta. 2024. Llama3.

OpenAI. 2024. Gpt 4o-mini.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

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.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184, Singapore. Association for Computational Linguistics.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. Conceptual and unbiased reasoning in language models. *arXiv preprint arXiv:2404.00205*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Computational cost

Our experiments on Llama3-8B-Instruct are applied on one A6000 GPU. For every 1,000 entries (10 variations for each entry), the entire pipeline runs on average for 6 hours.

Other experiments are executed with online APIs. The typical time consumed for every 1,000 entries is 9 hours for GPT 4o-mini, 6 hours for Gemini 2.0 Flash, and 21 hours for Claude 3.5 Haiku.

B Learned overall weights for SLEII pipeline

Table 6 reports the weight parameter w learned from training sets for SLEII pipeline under each setting.

C Prompts

We list all of the zero-shot prompts and instructions for prompting in our experiments. For few-shot prompts that we used, please check the files associated with this paper.

C.1 Alignment

Two sentences are given. Find the entities that appears in both sentences, and replace them with unique tags together with type labels. In this task, pronouns are also considered as entities. Only tagged those mentions that refer to an exact same entities in both sentences. If two mentions of one entity are semantically different in the two sentences, they should not be tagged. If there is no entity that appears in both sentences, leave the sentence as it is. Always output the two tagged sentences first, and then provide an explanation.

C.2 Premise variation

Rewrite the given tagged sentence into a factual sentence by replacing the type tags with named entities. For each rewritten version, start with "Rewritten version X", then output the rewritten sentence, and then provide an explanation.

C.3 Hypothesis instantiation

Given two tagged sentences, the Sentence 1 is already rewritten by replacing the type tags with actual entities. Explanation on how the tags are replaced is provided. Now rewrite Sentence 2, so that the tags that appears in both sentences with the same '#' tag number are replaced with the same entity.

C.4 Prediction

Given a premise and a hypothesis, predict whether the premise entails the hypothesis. Return only one choice mark 'A', 'B', or 'C' to answer the question, and then explain your choice.

C.5 Determine factuality

Instruction:

Given a statement, determine whether the statement is factual. Return only one choice mark 'A', 'B', or 'C' to answer the question.

Prompt:

Statement: statement
Question: Is the statement factual?
Choices:
A. Yes, factual
B. No, not factual
C. Can not be determined

Model	**Dataset**					
	SNLI		RTE		MNLI	
	soft	*hard*	*soft*	*hard*	*soft*	*hard*
Llama3-8B	0.03	0.12	0.01	0.01	0.23	0.23
GPT4o-mini	0	0.04	0	0	0.24	0.36
Gemini-2.0	0.01	0.28	0.54	0.54	0.52	0.46
Claude-3.5	0.10	0.44	0.53	0.61	0.41	0.57

Table 6