Which Words Matter? Understanding How Large Language Models Comprehend Arguments

Anonymous ACL submission

Abstract

Pioneering developments in large-scale language models (LLMs) have marked a substantial stride in their ability to comprehend multifaceted debate topics and to construct argumentative narratives. Despite this progress, there remains a notable lack of scholarly understanding of the processes by which LLMs engage with and analyze computational arguments. Classical studies have delved into the linguistic frameworks of arguments, encapsulating their essence within the realms of structural organization and logical coherence. Yet, it remains unclear whether LLMs utilize these recognized frameworks in addressing argument-related tasks. In an effort to illuminate this research void, our study introduces three hypotheses centered on the dynamics of claim, evidence and stance identification in argument mining tasks: 1) Omitting specific logical connectors in an argument does not change the implicit logical relationship, and LLMs can learn it from the modified context. 2) The importance of words or phrases in an argument is determined by the extent of implicit information they encapsulate, regardless of their individual components within the structure of the argument. 3) Removing crucial words or phrases from an argument alters the implicit logical relationship, making it impossible for LLMs to learn the original logic from the modified text.Through comprehensive assessments on the standard IAM dataset, it is revealed that information contained in the phrases within the argument has a greater impact on the understanding of the argument by large models, and the experiment results validate our hypothesis.

1 Introduction

Argumentation is a fundamental aspect of communication, pervading diverse facets of daily life (Wachsmuth et al., 2016; El Baff et al., 2020). It manifests in everyday discourse (Swanson et al., 2015; Misra et al., 2016; Lugini and Litman, 2018), legal deliberations (Rinott et al., 2015; Šavelka and Ashley, 2016; Poudyal et al., 2020), and scientific inquiry (Lauscher et al., 2018b,a; Al Khatib et al., 2021). Argumentation not only enhances mutual understanding by revealing varied perspectives and rationales but also strengthens the articulation and persuasiveness of opinions. With the increasing interest in computational argumentation within natural language processing, scholars have embarked on exploring various argument mining tasks. These include identifying argument components (Levy et al., 2014; Rinott et al., 2015; Lippi and Torroni, 2016), extracting argument pairs (Cabrio and Villata, 2012; Cheng et al., 2021, 2020), and assessing argument quality (Wachsmuth et al., 2017; Toledo et al., 2019; Lauscher et al., 2020). Meanwhile, large language models (LLMs) have exhibited a sophisticated understanding of language semantics, capably navigating complex text comprehension and generation tasks (Maynez et al., 2023; Cheng et al., 2023a; Yuan et al., 2023; Cheng et al., 2023b; Wu et al., 2023), while also functioning as adept social agents in human and artificial interactions (Park et al., 2023; Andreas, 2022).

Current research on computational argument largely falls into two categories: The first focuses on extracting argument components based on structure, such as claims and evidence (Sardianos et al., 2015; Goudas et al., 2014; Li et al., 2021, 2019), while the second is centered on argument generation, including counter-arguments (Schiller et al., 2021; Hua et al., 2019; Lin et al., 2023; Alshomary and Wachsmuth, 2023). Studies like (Chen et al., 2023) evaluate LLMs in computational argument tasks, revealing their strengths and highlighting evaluation challenges within this domain. As a crucial direction of computational argumentation, argument mining is centered on comprehending unstructured texts and automatically extracting diverse argumentative elements which require models to discern logical relationships between sentences, a capability we aim to assess in LLMs through various tasks. This enables us to objectively quantify the language model's ability to understand arguments through argument mining tasks. Drawing on insights from psychology and cognitive science, which suggest humans can comprehend texts even with missing words or disrupted word order (Grainger and Whitney, 2004; Perfetti and Bolger, 2018), similar phenomena have been observed in LLMs. For instance, (Li et al., 2023) demonstrate prompt compression by eliminating less informative words or phrases, thus reducing computational costs.

This study undertakes a novel examination of LLMs' argument comprehension across three argument mining tasks: claim extraction, evidence extraction, and stance classification. We test LLMs in various settings, including fine-tuned and zero-shot configurations, to validate their logical comprehension abilities in the claim-evidence context. We hypothesize that: 1) Omitting logical connectors does not obscure implicit logical relationships, with LLMs capable of inferring them from the modified context. 2) The relevance of words or phrases in an argument is constrained to the implicit information they hold, independent of their structural role. Eliminating words or phrases with critical information distorts the implicit logic, preventing LLMs from grasping the original reasoning. These hypotheses are tested using the IAM dataset (Cheng et al., 2022).

The major contributions are threefold: 1) It is the inaugural study to utilize LLMs to perform claim extraction, evidence extraction, and stance classification on a large-scale dataset, employing both zero-shot and fine-tuning approaches. 2) We demonstrate that the omission of certain logical connectors does not necessarily change the underlying logical relationship, which can still be inferred by LLMs from the altered context. 3) We establish that removing words or phrases that contain critical information disrupts the implicit logical relationship, challenging the LLM's ability to derive the original logical connections.

2 Background

2.1 Argument mining tasks

Argument mining focuses on extracting components of an argument from text. In tasks like claim and evidence extraction, the aim is to automatically retrieve relevant claims or supporting/opposing evidence based on a given topic or claim. Unlike other NLP tasks, argument mining not only requires relevance to the query but also emphasizes identifying persuasive elements in the text. This distinctive feature sets argument mining apart from tasks concentrating on different patterns or information types.

For a long time, it has been believed that argument structure plays a crucial role in the process of understanding arguments. For instance, many efforts have focused on leveraging structural information such as syntactic and discourse structures to solve argument mining tasks (Ye and Teufel, 2021; Peldszus and Stede, 2015; Huber et al., 2019). Nevertheless, in everyday spoken language and online forums, some expressions are considered to deviate from the strict paradigm of argumentation yet still possess persuasive power. Many studies also posit that implicit reasoning plays a significant role in understanding arguments (Habernal et al., 2018; Singh et al., 2021). This implicit reasoning lies behind specific vocabulary, aiding us in comprehending and establishing logical connections between different argument components. Based on this phenomenon, we propose our first hypothesis: Omitting certain logical connectors in an argument does not alter the implicit logical relationship within the context.

2.2 Prompt compression

While LLMs demonstrate remarkable comprehension and generation capabilities across various natural language processing tasks, their demanding computational resource requirements remain a significant obstacle. The cost of the API service is also a factor to consider when utilizing closed-source LLMs, which is often associated with the input sequence length. Recently, many efforts have been focused on compressing prompts while attempting to maintain the performance of the model (Chevalier et al., 2023; Jiang et al., 2023a,b). (Li et al., 2023) achieve prompt compression by retaining high-entropy words and eliminating other low-entropy words to preserve model performance. Although the compressed prompt ensures the performance of large model inference, the compressed prompt becomes unreadable for humans. This compression disrupts the original prompt text's argumentative structure but is still comprehensible to large models. We posit that the reason high-entropy words can ensure model performance is that these words contain crucial clues for implicit reasoning. Large models can leverage these clues to comprehend the text. Therefore, we propose our second and third hypothesis: The importance of words or phrases in an argument may be unrelated to their specific components within the structure of the argument, and is determined by the extent of implicit information they convey. Removing words or phrases containing crucial information from the argument alters the implicit logical relationship within the context.

3 Methodology

To examine the comprehension abilities of large-scale language models (LLMs), we conduct experiments on three different tasks, namely claim extraction, evidence extraction, and stance classification. The claim extraction and evidence extraction tasks assess the model's understanding of logical aspects. Stance classification evaluates the model's ability to recognize emotional differences.

3.1 Word Removal

In general, we attempt to disrupt the implicit reasoning of arguments through two means. The first involves removing specific conjunctions such as "because,", "and," and so on. The second opts for the removal of informative words, specifically high-information entropy words. As seen in Table 1, it provides examples of both removal ways, and the resulting texts become unreadable for humans after applying either removal to the argument text, especially the removal of informative words.

3.1.1 Removal of Connectives

In the initial works solving tasks related to identifying and extracting various parts of arguments, many previous methods rely on structural information (Nguyen and Litman, 2015; Aker et al., 2017; Morio and Fujita, 2019). Connectives are often considered key cues for revealing the logical relationships for discourses within a sentence, aiding our understanding of contextual relationships during reading.

Specifically, PDTB-2.0 (Prasad et al., 2008) is leveraged to find out the connectives, wherein discourse relations are annotated on the one million Wall Street Journal (WSJ) corpus. As seen in Table A.1, we choose phrases (or words) like “instead" or subordinating conjunctions like “because," which are considered to have logical meanings to form our connective library $\mathcal{C}$.

$$\mathcal{S} = \{P_1, P_2, \cdots, P_n\} \tag{1}$$

$$\overline{\mathcal{S}} = \{P_i\}_{P_i \notin \mathcal{C}} \tag{2}$$

Given a sentence $\mathcal{S}$, which is derived from a claim, evidence or topic text, we check all phrases P_i within it, if P_i is found in our connective library $\mathcal{C}$, we remove it. This results in a modified sentence $\overline{\mathcal{S}}$ devoid of these connectives.

3.1.2 Removal of Informative Words

(Li et al., 2023) remove phrases with low information entropy to compress the prompt while maintaining performance. In this work, to test if removing crucial phrases alters the logical relationship between sentences, we take a different approach to eliminate the top-N high information entropy tokens, where the information entropy is denoted as $\mathcal{I}$.

Specifically, we choose GPT-2 (Radford et al., 2019) as our base model to compute the information entropy of each token w_i within sentence $\mathcal{S}$. As shown in Eq. 4, we obtain the entropy values by applying a softmax function to the logits of each token w_i derived from encoding sentence $\mathcal{S}$ with GPT-2, namely, the generation probability of each token is used as the information entropy of this token.

$$\mathcal{S} = \{w_1, w_2, \cdots, w_n\} \tag{3}$$

$$\mathcal{I}(w_i) = \text{softmax}(\text{logits}_{\mathcal{V}})_{w_i} \tag{4}$$

wherein $\mathcal{V}$ is the vocabulary.

After obtaining the information entropy of all tokens, we remove the top three tokens in sentences. For sentences with a length of less than three tokens, we only discard one token with the most information entropy from the argument.

3.2 LLMs for Argument Mining

All experiments are conducted in two settings: zero-shot and fine-tuning settings. Our prompt design is shown in figure 1, we use the evidence extraction task as an example: initially, we provide a brief text describing the task requirements and introducing, followed by an evidence candidate sentence *S* and target claim sentence *T*. Considering the different formatting requirements for the T5 model and the Llama model, there are slight differences in the prompt format design for both models. For the Llama model, we prefix each component with ### to indicate differentiation. For the training data required for fine-tuning, we directly append a ground truth label to the response.

Setting	Text
Original Sentence	Not only is saving confined to money but also to time.
Removal of Connectives	is saving confined to money to time.
Removal of Informative Words	Not only is saving to but also to .

Table 1: The remove by connectives method eliminated the connective phrase "not only ... but also," while the remove by entropy method removed the three words: "confined," "money," and "time."

For label selection, we prefer choosing meaningful labels such as we have opted to use "evidence" and "not evidence" as our labels for experimentation in evidence extraction task, rather than simple "yes" or "no." In a zero-shot setting, we assess the inherent ability of the LLMs to comprehend arguments. Under the conditions of fine-tuning, we examine whether the LLMs can acquire the capability to understand arguments across datasets with different word removal configurations.

4 Experiments

4.1 Experimental Setup

Tasks and datasets We experiment on the claim extraction, evidence extraction, and stance classification tasks from the IAM dataset, where the training, validation, and test sets are directly adopted from the original splits with a ratio of 8:1:1 (Cheng et al., 2022). The claim extraction and evidence extraction processes can assess the logical comprehension capabilities of LLMs under different scenarios, while stance classification can evaluate the emotional recognition abilities of LLMs. IAM Datasets (Cheng et al., 2022) contains 123 debating topics with a diverse range sourced from online forums, containing 69,666 sentences extracted from these articles.

The goal of claim extraction is to automatically identify and extract the claims from articles associated with a particular debating topic. This task is crucial in the field of argument mining, as claims play a pivotal role in constructing and supporting arguments. On the other hand, evidence extraction involves the automatic identification of relevant evidence within documents associated with a specific topic and its related claim. The model is tasked with extracting pertinent evidence to support or refute the given claim. Stance classification is defined as the process of determining, for each claim associated with a given topic, whether it aligns with or contradicts the overall stance on the topic. This task involves evaluating the relationship between claims and the overarching theme to understand the position they take in relation to the given topic.

Large language models We conduct experiments using two open-source LLMs, LLama2-7b (Touvron et al., 2023) and Flan-T5-XL (Chung et al., 2022), as well as a nonopen-source LLM, ChatGPT-3.5-Turbo (OpenAI). For the open-source LLMs LLama2-7b and Flan-T5-XL, we conducted zero-shot and fine-tuning experiments under three settings: original text, text with removed conjunctions, and text with removed informative words. For the two settings involving the removal of conjunctions and informative words, we applied the removal to both training and testing data. During the fine-tuning process, we employed the LORA fine-tuning method (Hu et al., 2021). During the fine-tuning stage, we uniformly set the number of training epochs to 5.

For the zero-shot setting experiments, we incorporate the corresponding *S* and *T* from the respective tasks of the IAM test datasets into the prompt, requesting the LLM to provide answers. In the fine-tuning configuration, the training is based on two different LLMs, Flan-T5-XL and Llama2-7b, on the training sets of the three argument mining tasks. For the word removal, we apply the same word removal process to both the training and test sets of data.

Metrics In order to assess the impact of removing different words from argument context in argument mining tasks, we employ both accuracy and Macro F1 score as performance metrics. These metrics remain consistent with previous work (Cheng et al., 2022; Chen et al., 2023).

At the same time, we observed that the occurrence of hallucinations in LLMs changes when certain words are removed using different methods. We define instances where the labels returned by the LLMs are inconsistent with the labels provided in the prompt as hallucination. The presence of hallucinated responses significantly affects the per-

```
Identify whether the given sentence is a evidence
towards the given target. Choose from ’evidence’
or ’not a evidence’. \n
Sentence: S \n

Target: T \n

Label:
```

(a) Prompt for flan-T5 model

```
###Instruction: Identify whether the given
sentence is a evidence towards the given target.
Choose from ’evidence’ or ’not a evidence’. \n
###Sentence: S \n

###Target: T \n

###Label:
```

(b) Prompt for llama models

Figure 1: The evidence extraction prompts for Flan-T5 model and Llama model, where *S* represents the evidence candidate sentence, *T* represents the target claim sentence

formance of the LLMs in argument mining tasks. Therefore, we also record the frequency of hallucinated responses in the model under different conditions.

4.2 Results and Discussions

Tables 2 & 3 present the results obtained for the argument mining task and the hallucination frequency of flan-T5-xl model on claim extraction task and evidence extraction task, respectively. For cases where the response results from LLMs do not match the labels provided in the instructions, such as responses like ’not enough information’, we consider them uniformly as instances of hallucination. Due to our fine-tuning and testing on the all data of three tasks in the IAM dataset, the final results show some discrepancies compared to the findings in the (Chen et al., 2023).

4.2.1 Overall Results

Overall, Flan-T5-xl demonstrates the best performance in claim extraction, evidence extraction, and stance classification. Additionally, we observe that for all models, there is a decrease in performance across these three tasks when connectives and informative words are removed from the argument text. Among these, the removal of informative words leads to the most significant performance drop.

We will further analyze the impact of word removal on ChatGPT-3.5, Flan-T5-xl, and LLama2-7b in claim extraction, evidence extraction, and stance classification tasks in the following paragraphs.

In the *claim extraction* task, after word removal, all models have a significant performance drop under zero-shot setting or fine-tuning setting, with the impact of removing informative words surpassing that of removing connectives. We observe that fine-tuned Flan-T5-xl achieves the best performance with an accuracy of 0.923 and a macro F1 of 0.811. Following closely in zero-shot settings, Chat-GPT3.5 demonstrates a performance of 0.673 accuracy and 0.672 macro F1. LLama2-7b performs the least favorably, even after fine-tuning, with the best accuracy and macro F1 reaching only 0.659 and 0.712, respectively. Meanwhile, we observe a significant reduction in the frequency of hallucination in LLMs after fine-tuning, as shown in Table 3.

The Flan-T5-xl model, after fine-tuning, demonstrates a noticeable decrease in the occurrence of hallucination in responses, regardless of whether word removal is applied or not.

After implementing word removal, the performance of Chat-GPT-3.5, Flan-T5-xl, and Llama2-7b in claim extraction shows a decrease, with the impact of removing informative words being greater than that of removing connectives. In the case of removing connectives, ChatGPT-3.5-turbo and Flan-T5-xl experience approximately a 0.005 drop in accuracy and a decrease of 0.025 and 0.038 in macro F1, respectively, under zero-shot settings. Llama2-7b, on the other hand, exhibits a decrease of 0.025 in accuracy and 0.047 in macro F1. In contrast, when removing informative words, ChatGPT-3.5-turbo and Flan-T5-xl, under zero-shot settings, witness a decrease in accuracy ranging from 0.1 to 0.2, with macro F1 decreasing by approximately 0.2 and nearly 0.15, respectively. The impact of removing informative words is more substantial on the already poorly performing Llama2-7b model, with a decrease of nearly 0.11 in accuracy and 0.3 in macro F1.

After fine-tuning, both Flan-T5-xl and Llama2-7b show improvements in performance on claim extraction. Flan-T5-xl exhibits an increase of 0.03 in accuracy and 0.06 in macro F1. Llama2-7b, on the other hand, sees an improvement from an accuracy below 0.5 and a macro F1 of 0.601 before fine-tuning to an accuracy of 0.659 and a macro F1 of 0.712 after fine-tuning. The performance improvement after fine-tuning is mainly attributed to

Task	Setting	GPT-3.5-Turbo		Flan-T5-XL		LLama2-7b	
		Accuracy	F1	Accuracy	F1	Accuracy	F1
Claim Detection	Zero-shot	0.673	0.672	0.893	0.759	0.477	0.601
	- remove connectives	0.669	0.647	0.886	0.721	0.452	0.554
	- remove info words	0.578	0.479	0.818	0.623	0.343	0.279
	Fine-tune	-	-	0.923	0.811	0.659	0.712
	- remove connectives	-	-	0.914	0.801	0.622	0.675
	- remove info words	-	-	0.834	0.638	0.523	0.396
Evidence Detection	Zero-shot	0.477	0.475	0.793	0.747	0.341	0.387
	- remove connectives	0.435	0.411	0.782	0.714	0.345	0.357
	- remove info words	0.381	0.357	0.719	0.570	0.317	0.282
	Fine-tune	-	-	0.864	0.795	0.585	0.618
	- remove connectives	-	-	0.843	0.745	0.550	0.561
	- remove info words	-	-	0.828	0.527	0.492	0.453
stance classification	Zero-shot	0.603	0.594	0.535	0.476	0.342	0.384
	- remove connectives	0.598	0.592	0.537	0.478	0.361	0.333
	- remove info words	0.421	0.403	0.482	0.379	0.312	0.322
	Fine-tune	-	-	0.583	0.471	0.486	0.355
	- remove connectives	-	-	0.569	0.466	0.481	0.341
	- remove info words	-	-	0.517	0.413	0.447	0.325

Table 2: Results of GPT-3.5-Turbo, Flan-T5-XL and llama2-7b on claim detection, evidence detection and stance classification tasks. For open source model Flan-T5-XL and llama2-7b, we test their performance under both zero-shot and fine-tuning setting. For ChatGPT-3.5-Turbo, we only test their performance under zero shot setting.

a significant reduction in hallucination. As shown in Table 3, Flan-T5-xl experiences a decrease in the number of hallucinations from 223 to 54 after fine-tuning without word removal. This indicates that through LORA fine-tuning, LLMs can learn the content of instructions.

For Flan-T5-xl, fine-tuning with connectives removed yields an accuracy and macro F1 of 0.914 and 0.801. However, removing informative words results in a drop to an accuracy of 0.834 and a macro F1 of 0.638. In the case of Llama2-7b, fine-tuning with connectives removed leads to a decrease in accuracy and macro F1 from 0.659 to 0.622 and 0.712 to 0.675, respectively. Removing informative words has a more substantial impact on performance, with Flan-T5-xl dropping to an accuracy of 0.834 and a macro F1 of 0.638, and Llama2-7b declining to an accuracy of 0.523 and a macro F1 of 0.396. Interestingly, fine-tuning after removing informative words does not enhance macro F1 compared to models without fine-tuning, indicating the challenge LLMs face in capturing implicit logical relationships when informative words are excluded.

In the claim extraction experiments under the zero-shot setting, we observed that ChatGPT-3.5-turbo experienced the least impact after removing connectives. The Llama2-7b model consistently performed the worst across various settings, with its performance being particularly affected by word removal, especially the removal of informative words. After fine-tuning, there was a significant reduction in hallucination instances, and the large models were able to essentially return the labels as required by the prompt, contributing to the overall performance improvement. However, in the case of removal of informative words, hallucinations stemming from refusal to answer still persisted.

For *evidence extraction*, the effect of eliminating informative words is more pronounced compared to the impact of removing connectives. Similar to claim extraction, both in zero-shot and fine-tuning settings, all models experience a notable decline in performance. Under zero-shot settings, Flan-T5-xl achieves an accuracy of 0.782 and a macro F1 of 0.747, outperforming ChatGPT3.5-turbo with an accuracy of 0.477 and a macro F1 of 0.475, as well as LLama2-7b with an accuracy of 0.341 and a macro F1 of 0.387. Due to the increased demand for logical understanding in evi-

Setting	Hallucination Frequency	
	Claim Extraction	Evidence Extraction
Zero-shot Flan-T5-XL	223	302
w/ remove connectives	298	371
w/ remove info words	772	1148
Fine-tuned Flan-T5-XL	54	88
w/ remove connectives	81	96
w/ remove info words	516	705

Table 3: The hallucination frequency of Flan-T5-XL model under different settings.

dence extraction tasks for LLMs, the overall performance decreases compared to the claim extraction task.

After applying word removal, evidence extraction performance diminishes for Chat-GPT-3.5, Flan-T5-xl, and LLama2-7b, with the removal of informative words having a more substantial impact than removing connectives. LLama2-7b, surprisingly, shows a 0.04 improvement in accuracy after removing connectives. Specifically, under zero-shot settings, ChatGPT-3.5-turbo and Flan-T5-xl witness a decline in accuracy from 0.477 and 0.793 to 0.435 and 0.782, respectively. Their macro F1 values also decrease from 0.475 and 0.747 to 0.411 and 0.714. When removing informative words, ChatGPT-3.5-turbo and Flan-T5-xl experience a decrease in accuracy to 0.381 and 0.719. LLama2-7b, after removing informative words, sees a drop in accuracy and macro F1 to 0.317 and 0.282, respectively. Similar to the claim extraction task, evidence extraction in zero-shot settings demonstrates an increase in hallucination after word removal. Results in Table 3 show that removing connectives leads to a 69 hallucination increase, while removing informative words results in a substantial increase of 777 hallucinations.

After fine-tuning, both Flan-T5-xl and LLama2-7b show a significant improvement in performance in evidence extraction. Flan-T5-xl exhibits an increase of around 0.07 in accuracy and 0.05 in macro F1. LLama2-7b, similarly, sees an improvement to an accuracy of 0.585 and a macro F1 of 0.618 after fine-tuning. Similar to the claim extraction task, the performance improvement after fine-tuning is mainly contributed to a significant reduction in hallucination.

The Flan-T5-xl model, following fine-tuning with connectives excluded, achieves an accuracy of 0.864 and a macro F1 of 0.795. However, when informative words are omitted, its performance declines to an accuracy of 0.828 and a macro F1 of 0.527. As for the Llama2-7b model, fine-tuning with connectives removed results in a decrease in accuracy from 0.550 to 0.492 and a decrease in macro F1 from 0.561 to 0.453 compared to fine-tuning without word removal. It's noteworthy that the impact of word removal on model performance varies between Flan-T5-xl and Llama2-7b.

For *stance classification* Removing words significantly affects performance, aligning with findings from previous tasks. The impact of eliminating informative words surpasses that of removing connectives.Unlike the claim extraction and evidence extraction tasks, the stance classification task places more emphasis on emotion recognition and understanding capabilities rather than logical reasoning.The performance of ChatGPT-3.5-turbo excels in a zero-shot setting, contrasting with the optimal performance of Flan-t5-xl in the previous two tasks achieved under fine-tuning settings. In this task, ChatGPT-3.5-turbo achieved the best performance under zero-shot settings, reaching an accuracy of 0.603 and a macro F1 of 0.594. Meanwhile, Flan-T5-xl achieved an accuracy of 0.535 and a macro F1 of 0.476. On the other hand, LLama2-7b performed the poorest, with only 0.342 accuracy and 0.384 macro F1.

For ChatGPT3.5-turbo, after removing connectives, the accuracy and macro F1 are 0.598 and 0.592, respectively. However, when informative words are removed, ChatGPT3.5-turbo's performance drops to an accuracy of 0.421 and a macro F1 of 0.403. The performance of ChatGPT-3.5-turbo experiences a significant decline when informative words are removed, compared to the performance in the first two tasks A.2. For the LLama2-7b model and Flan-T5-xl, their performance in the stance classification task is relatively poor. When informative words are removed, their accuracy and macro F1 reach the lowest levels.

4.3 Discussions

Statistics of word removal We employ the method outlined in the previous section 3 to remove words from arguments. To ensure the fairness of the experiment, we aim to maintain consistent argument lengths across different methods after word removal. Therefore, we choose to eliminate the top-3 words with the highest information entropy in the "remove by information entropy" method3.1.2, ultimately achieving a comparable length for arguments in both methods. Table 4 shows the average length after word removal.

	Original	Remove Connector	Remove Info
All sentences	21.05	18.97	18.57
Claim	23.44	20.81	20.62
Evidence	25.09	22.73	22.16

Table 4: The average length after word removal

Analysis of hallucination In this study, we classify instances where the LLMs provide a label different from those specified in the prompt as hallucination. Any deviation from the required label in the prompt by the LLMs' output is considered as contributing to the hallucination frequency. We observed that the occurrence of hallucination in LLMs after word removal is different from the situation before word removal. Especially after word removal, particularly when informative words are removed, as shown in Table 3, the frequency of hallucination significantly increases. Moreover, most hallucinations after removing informative words consist of sentences such as 'I don't know what you are talking about' and 'not enough proof.' In contrast, when informative words are not removed, most hallucinations involve repeating prompts or labels that do not strictly adhere to the prompt requirements. These occurrences are substantially reduced after fine-tuning, while the former type does not decrease significantly. This also demonstrates that after removing informative words and undergoing fine-tuning, LLMs cannot learn the intrinsic meaning of arguments.

Analysis of removed words As shown in Figure 2, we create a word cloud for the informative words removed from the training sets of claim extraction, evidence extraction, and stance classification.

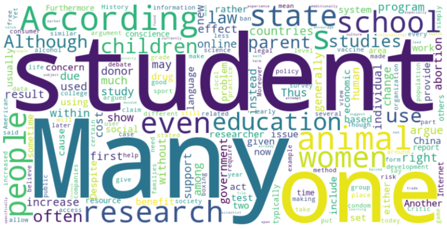


Figure 2: The word cloud of removed informative words.

From Figure 2, we can observe that the removed informative words mainly consist of nouns commonly used as subjects or objects in sentences, such as 'student,' 'system,' and 'education,' among others. Additionally, there are adverbs like 'many' that can imply the potential emotional meaning of a sentence. Some connectives, such as 'according' and 'although,' also appear in the word cloud. These words can help us identify the key parts of sentences. This also indicates the reasons for the success of past methods, indeed, some conjunctions contain rich information.

We observe that these words, in the stance classification task, assist in understanding the emotional inclination of arguments. In claim extraction and evidence extraction tasks, they help us establish the underlying connections between different argument components

5 Conclusion

Our experiments on claim extraction, evidence extraction, and stance classification affirm three initial hypotheses: 1. The absence of specific logical connectors in an argument doesn't alter the implicit logical relationship; language models can learn it from the modified context. 2. The importance of words or phrases in an argument is tied to the implicit information they convey, unrelated to their structural components. 3. Removing words or phrases with crucial information from an argument changes the implicit logical relationship, making it challenging for language models to learn the original relationship from the modified text.

This work, through testing the importance of different types of words in the argument context, helps us understand which words carry more crucial argumentative information in the comprehension process of LLMs. It will aid LLMs in understanding and reconstructing implicit meanings in the future.

Limitations

Our study conducted experiments on three argument mining tasks: claim extraction, evidence extraction, and stance classification, validating three hypotheses. However, experiments were not conducted on more complex tasks such as argument generation. Additionally, due to limitations in computational resources, fine-tuning experiments were only performed on flan-t5-xl and llama2-7b, without conducting global fine-tuning experiments. Challenges still exist in researching the effectiveness of LLMs in the field of argumentation.

References

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96, Copenhagen, Denmark. Association for Computational Linguistics.

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2023. Conclusion-based counter-argument generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 957–967, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Andreas. 2022. Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023a. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

Liying Cheng, Tianyu Wu, Lidong Bing, and Luo Si. 2021. Argument pair extraction via attention-guided multi-layer multi-cross encoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6341–6353, Online. Association for Computational Linguistics.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023b. Lift yourself up: Retrieval-augmented text generation with self memory. *arXiv preprint arXiv:2305.02437*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846, Singapore. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings 8*, pages 287–299. Springer.

Jonathan Grainger and Carol Whitney. 2004. Does the huamn mnid raed wrods as a wlohe? *Trends in cognitive sciences*, 8(2):58–59.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

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

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Laurine Huber, Yannick Toussaint, Charlotte Roze, Mathilde Dargnat, and Chloé Braud. 2019. Aligning discourse and argumentation structures using subtrees and redescription mining. In *Proceedings of the 6th Workshop on Argument Mining*, pages 35–40, Florence, Italy. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

Anne Lauscher, Goran Glavaš, and Kai Eckert. 2018a. ArguminSci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018b. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Xiangci Li, Gully Burns, and Nanyun Peng. 2019. Scientific discourse tagging for evidence extraction. *arXiv preprint arXiv:1909.04758*.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. Scientific discourse tagging for evidence extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2550–2562, Online. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16705–16720, Singapore. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Luca Lugini and Diane Litman. 2018. Argument component classification for classroom discussions. In *Proceedings of the 5th Workshop on Argument Mining*, pages 57–67, Brussels, Belgium. Association for Computational Linguistics.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles. Association for Computational Linguistics.

Gaku Morio and Katsuhide Fujita. 2019. On the role of syntactic graph convolutions for identifying and classifying argument components. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9997–9998.

Huy Nguyen and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO. Association for Computational Linguistics.

OpenAI. Introducing chatgpt.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Charles A Perfetti and Donald J Bolger. 2018. The brain might read that way. In *The Cognitive Neuroscience of Reading*, pages 293–304. Routledge.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66.

Jaromír Šavelka and Kevin D. Ashley. 2016. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 50–59, Berlin, Germany. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Keshav Singh, Farjana Sultana Mim, Naoya Inoue, Shoichi Naito, and Kentaro Inui. 2021. Exploring methodologies for collecting high-quality implicit reasoning in arguments. In *Proceedings of the 8th Workshop on Argument Mining*, pages 57–66, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Prague, Czech Republic. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678, Online. Association for Computational Linguistics.

Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. 2023. Evaluating instruction-tuned large language models on code comprehension and generation. *arXiv preprint arXiv:2308.01240*.

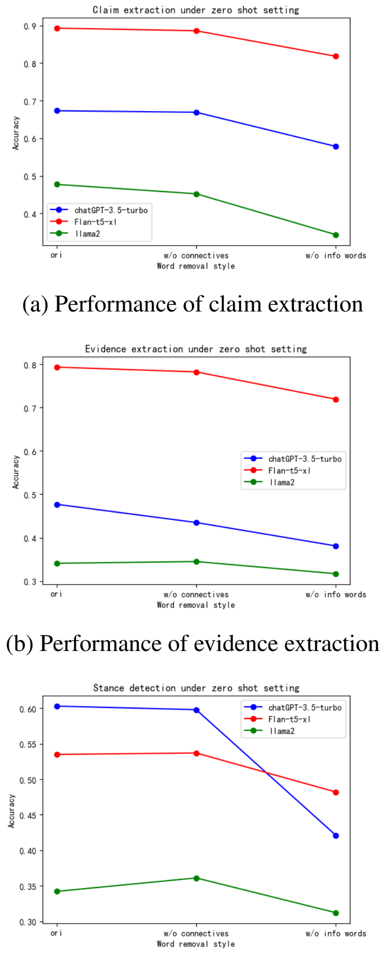
A Appendix

A.1 Connectives word and phrase

Connective Words
"once","although","though","but","because", "nevertheless","before","until","if", "previously","when","and","so","then", "while","however","also","after", "separately","still","or","moreover", "instead","as","nonetheless","unless", "meanwhile","yet","since","rather", "indeed","later","ultimately", "therefore","thus","further", "afterward","next","similarly", "besides","nor","alternatively", "whereas","overall","till", "finally","otherwise","thereby", "additionally","meantime","likewise", "regardless","thereafter","earlier", "except","furthermore","lest","specifically", "conversely","consequently","plus","And", "hence","accordingly","simultaneously", "for","else"
Connective Phrase
"as long as", "so that", "in addition", "on the other hand", "for instance", "in fact", "as a result","either or", "in turn","in particular","not only", "if and when","by comparison","in contrast", "as if","now that","before and after", "by contrast","as though", "on the one hand on the other hand", "insofar as", "as an alternative", "in the end","if then","in other words", "but also","as soon as","in short", "neither nor","as well","much as", "by then","on the contrary","in sum", "when and if","for example"

Table 5: The connectives library

A.2 Performance figure



(a) Performance of claim extraction

(b) Performance of evidence extraction

(c) Performance of stance classification

Figure 3: The performance of claim extraction, evidence extraction and stance classification