LLMs as annotators of argumentation

Anna Lindahl
Språkbanken
University of Gothenburg
Sweden
anna.lindahl@svenska.gu.se

Abstract

Annotated data is essential for most NLP tasks, but creating it can be time-consuming and challenging. Argumentation annotation is especially complex, often resulting in moderate human agreement. While large language models (LLMs) have excelled in increasingly complex tasks, their application to argumentation annotation has been limited. This paper investigates how well GPT-4o and Claude can annotate three types of argumentation in Swedish data compared to human annotators. Using full annotation guidelines, we evaluate the models on argumentation schemes, argumentative spans, and attitude annotation. Both models perform similarly to humans across all tasks, with Claude showing better agreement with humans than GPT-4o. Agreement between models is higher than human agreement in argumentation scheme and span annotation.

1 Introduction

Annotated data is essential in most natural language processing (NLP) tasks, including semantic and pragmatic analysis. While pretrained large language models (LLMs) have reduced the need for large amounts of annotated training data, labeled data remains crucial for evaluation. However, creating high-quality annotated data can be time-consuming and expensive, especially when faced with the complex aspects of linguistic meaning involved in annotating a phenomenon like argumentation.

Annotating argumentation is a challenging task, as it involves not only identifying opinions but how they are argued for. Argumentation in itself can often be implicit and context-dependent, and sometimes even subjective, which can lead to differing opinions among annotators. In NLP, the study of argumentation is usually done within the field of argumentation mining, which aims to automatically retrieve and analyze argumentation (Stede and Schneider, 2018; Lawrence and Reed, 2020). Because of the complexity of the task, argumentation annotated datasets used in this field often report lower agreement than annotation of other phenomena in NLP (Lytos et al., 2019; Lindahl and Borin, 2024). These challenges make argumentation especially suitable for investigation.

In recent years, LLMs have excelled at different complex tasks, often outperforming previous baselines (Brown et al., 2020; Chowdhery et al., 2022). Often, these models are not fine-tuned on data, but instead instructed through prompts to perform various tasks, such as classification. This way of prompting is more similar to annotation of data, rather than training and then classifying. This similarity between prompting and annotation has given rise to several studies comparing how well LLMs annotate (Pavlovic and Poesio, 2024), with potential advantages in speed and cost (Ding et al., 2023). Recently, models are able to handle longer inputs (compare OpenAI's GPT-4's context window of 8102 tokens to GPT-4o's 128k), making it possible to prompt models with actual annotation guidelines rather than shortened instructions.

Despite the capabilities of LLMs, not many studies have yet compared argumentation annotation of humans to that of models. Because of this, in this study, three Swedish datasets annotated with different types of argumentation are used to evaluate how well two LLMs, GPT-4o and Claude Sonnet, can annotate argumentation. The datasets are annotated for argumentation schemes, spans of argumentation and attitude, respectively. Because these tasks are complex and human annotators often disagree, we are also interested to see how the models annotate in cases of disagreement. For example, is there some annotators the models align with more? More specifically, this study asks:

1. How well can LLMs annotate argumentation, given annotation guidelines?

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 242–252

2. How do models annotate when humans disagree?

In answering these questions, this paper contributes a novel analysis of the models' capabilities as annotators of argumentation in three different tasks. As far as the author is aware, there are no other studies which analyse argumentation annotation of these kinds. We find that both models can annotate similarly to humans in all argumentation annotation tasks, with Claude showing better agreement with humans than GPT-4o. In the argumentation scheme and argumentation span task the agreement between the two models is higher than between human annotators.

2 Related work

In many tasks, it has been shown that LLMs can perform comparably to human annotators. For example, Gilardi et al. (2023) use ChatGPT for four annotation tasks (stance, topic, relevance, and frames) and find that the model performs similarly to, or better than, human annotators in these tasks, compared to an aggregated gold standard. Similarly, in Aldeen et al. (2023) GPT's performance in several classification tasks is presented. They find that the model overall performs well but struggles with sarcasm and emotion detection. Other areas where LLMs have been used for annotation are stance classification (Liyanage et al., 2024) and grammatical annotation (Morin and Marttinen Larsson, 2025). However, not all studies find that LLMs perform well. LLMs seem to struggle with more complex tasks, for example Wei et al. (2024) find that LLMs under-perform in the task of event extraction.

As mentioned in the previous section, the difference between annotation and classification in these kinds of studies is not always clear. In some studies (for example in Liyanage et al. (2024) models are not given instructions similar to that of what a human annotator would receive, but instead a shortened version. As pointed out by Pavlovic and Poesio (2024), most studies compare model output to a curated gold standard, without direct comparison to the other annotators.

An exception to this is Rønningstad et al. (2024), who annotate entity-level sentiment in Norwegian texts by prompting ChatGPT. They compare the model's and five human annotators' annotations to a curated dataset. Accuracy and Cohen's κ is lower for ChatGPT, with the exception of an outlier annotator. They also find that ChatGPT's errors deviate from the other annotators' labels more than the human annotators' labels. Another example is the study by Li and Conrad (2024), who annotate stance using open source LLMs. They find that LLMs show lower agreement with human annotators in cases where human annotators themselves disagree. They also find that in these cases, the stance is less explicit than in other examples.

In argumentation mining, studies have shown that LLMs can perform well on different tasks. For example, Abkenar et al. (2024) perform argument discourse unit classification and relation classification using open source models. Chen et al. (2024) explore the argumentation mining tasks of detecting claims, stance and evidence types. They find that GPT-3.5-Turbo performs best on complex tasks. Cabessa et al. (2025) fine-tune open source models and achieve state of the art results for component classification, relation classification and identification. Gorur et al. (2025) find that open source LLMs can outperform the baseline in identifying argumentative relations. There are also examples of LLMs being used to generate argumentation (Rocha et al., 2023). However, LLMs have not been successful at all tasks. Ruiz-Dolz and Lawrence (2023) find that GPT-4 performs below other models in their fallacy detection and classification task.

When looking specifically at annotation of argumentation, there are few studies. Mirzakhmedova et al. (2024) examine the annotation of argument quality. In this task, inter-annotator agreement (IAA) among human annotators is between 0.37–0.40 (expert and novice annotators). They calculate IAA across several runs with the same model, treating each run as a new annotator, and it is significantly higher (between 0.73–0.98). They also compare model annotations to human annotations in cases where there is perfect agreement between human annotators, and find that there is moderate agreement between the models and humans for most categories. Schaefer (2025) investigate how LLMs can aid in annotation of sematic argument type and find that the models can perform the task but with similar performance to a BERT model. They did not compare their results to individual human annotators.

3 Datasets

These datasets represent different annotation approaches in argumentation mining, with varying complexity, document length and genre. The datasets are all in Swedish. All datasets were annotated by annotators with a background in linguistics and with Swedish as their native language. Letters are used to represent the annotators below, for example 'annotator A'. Note, however, that the annotators differ between the annotation tasks. The annotations of these datasets are not available online, thus the risk of the model being informed by the annotations beforehand is minimal. Dataset statistics are seen in Table 1 below. Note that in the following sections all examples are translated from Swedish to English.

Annotation	Documents	Tokens	Avg. doc. size
Arg. schemes	30 editorials	20561	685
Arg. spans	9 threads	28465	3162
Attitude	500 tweets	15510	31

Table 1: Dataset statistics

3.1 Argumentation schemes

This dataset consists of 30 Swedish editorials from (Lindahl et al., 2019), with topic such as energy politics and unemployment. The editorials are annotated with Walton's argumentation schemes by two annotators. An argumentation scheme describes how inferences are being made in an argument, for example "Argument from popular opinion". The annotation task consisted of finding arguments, made up by a conclusion and one or more premises. This argument was then labeled with an argumentation scheme. Components were annotated as spans, and a span could have multiple roles (e.g. conclusion in one argument and premise in another) and be used several times. An example of an annotated argumentation scheme is given below.

Premise: 'But against this, one must weigh the obvious risks that an expansion of nuclear power entails.'
Conclusion: 'The waste must be stored for hundreds of years.'
Scheme: ARGUMENT FROM CORRELATION TO CAUSE

In this task, the annotators were instructed to use the book by Walton et al. (2008), which introduces and describes argumentation schemes, as guidelines. Although the book covers many different argumentation schemes, the number of scheme types available for annotation was restricted to 30. Because it was not feasible to provide the whole book to the model, a list of descriptions of these 30 schemes was used instead. Below is an example of a scheme description.

ARGUMENT FROM SIGN:
Premise: A is true in this situation.
Premise: Event B is generally indicated as true when its sign, A, is true in this kind of situation.
Conclusion: B is true in this situation.

Descriptions of the scheme types mentioned in this paper are found in appendix B. For a more detailed description of the annotation process, see Lindahl et al. (2019).

3.2 Argumentative spans

This dataset consists of 9 threads from two Swedish online discussion forums (Lindahl, 2020). They are annotated with argumentative spans by 8 annotators. The guidelines are approximately 2,800 tokens, with examples and diagnostic tests. The guidelines also provide a definition of argumentation and a discussion of what is to be considered argumentative. An annotated example, agreed upon by most annotators is seen below. Bold indicates argumentation.

"**I think we should eliminate home economics in schools. I consider it degrading to women.** The 1800s called and wants the school's view of women back. What do you others think?"

For a more detailed description of this annotation process and annotation disagreement, see Lindahl (2020) and Lindahl (2024).

3.3 Attitude

This dataset consists of 4280 tweets from Swedish political parties and party leaders (Lindahl, 2024, 2025). The aim of the annotation was to identify what the tweet author expressed an attitude or stance towards. This was done by marking spans of text that represented what the author expressed negative or positive attitudes about. For example, see below where bold indicates negative attitude:

"Now every penny needs to go towards counteracting **the municipal cri-**

> **sis.** Therefore, we say no to **increased Swedish EU fees.** The EU bureaucrats will have to cut their coat according to their cloth."

The guidelines describe the task and provide examples. This annotation study employed four annotators, and each tweet was annotated by a combination of three annotators. A subset of the tweets was annotated by all four annotators. From the tweets annotated by all four annotators, 500 were chosen for annotation in this study. For a more detailed description of the annotation process and annotation disagreement, see Lindahl (2025) and Lindahl (2024).

4 Method

For this study, the models GPT-4o (`gpt-4o-2024-08-06`) (OpenAI et al., 2024) and Claude Sonnet (`claude-3-7-sonnet-20250219`)[1] were chosen. For both models, the temperature was set to 0. This was to make the results more consistent and deterministic. GPT-4o was chosen because it is one of the most prominent and well-known models, and one of the most commonly used ones in annotation studies. It was also is one of the most cost effective models. Claude was chosen to provide a comparison with a different model architecture and it also performed well in preliminary experiments. Initially, in order to compare the results to an open source model, experiments were run with Llama 3.3-8B.[2] However, this model did not perform well enough to be included in this study, possibly due to the model size. A larger Llama model was not used due to computational constraints.

For each annotation task, the model was given a prompt which consisted of the original annotation guidelines together with short supplemental instructions. The prompt also included an example to be annotated. Both the prompt and examples were all in Swedish. The length of the guidelines is seen in Table 2.

	Editorials	Online forum	Tweets
Tokens	3426	2786	2349

Table 2: Length of the guidelines, tokens

[1] `https://docs.anthropic.com/en/docs/about-claude/models/overview`

[2] `https://www.llama.com/models/llama-3/`

The guidelines were not changed from the original annotation task, with the exception of removal of tool-specific instructions. The supplementary instructions introduced the task and the guidelines, and told the model in which format it should return the annotations.

As previous studies have reported (Rønningstad et al., 2024; Atreja et al., 2024), care must to be taken when crafting instructions. When writing instructions in this study, variations in wording were found to influence the annotations, and developing instructions that produced the correct output took some trial and error. In this process, both models were included, in order to develop instructions that worked for both of them. For example, mentioning the guidelines early in the instructions would increase the number of annotations returned in the correct format. An example of a prompt can be seen below, see appendix A for the other.

> "Your task is to annotate text spans that you consider to contain argumentation. Here are the annotation guidelines [guidelines]. The text is from an online forum, where each post is marked with "==". Divide the annotations per post. Make the annotations in json format, as a single object. Here is the text: [document]."

The prompt was sent for each document to be annotated. In the case of the argumentation schemes and the argumentative spans, a document would be an editorial or a thread, respectively. For the attitude annotation, ten tweets were sent at a time.

The guidelines for the argumentative spans and attitude tasks included annotated examples. In the argumentation scheme task, the guidelines did not include annotated examples. This was done in order to keep the annotation task as similar as possible to the original annotation setup.

5 Results

In general, the models struggled with keeping within the instructions and often added extra knowledge. For example, they would add a motivation for the annotation despite being explicitly asked not to. The models would also often correct the spelling of misspelled words, which made it challenging to automatically compare annotations. Some examples were returned without annotation or annotated in the wrong format. Due to these issues, around 100

tweets could not be included in the evaluation. In the editorials and the online threads, all examples were included.

5.1 Argumentation schemes

This task concerned the annotation of argumentation schemes in editorials. Comparing the annotated editorials, we found that Claude annotated more tokens than the annotators, while GPT-4o annotated fewer tokens, as seen in Table 3. GPT-4o annotated about half as many tokens as the two annotators. However, for both models the number of annotated tokens varied between editorials, sometimes matching the same number of tokens as the human annotators. Annotator A and Claude annotated a similar number of arguments, and likewise did annotator B and GPT-4o.

	A	B	GPT-4o	Claude
Anno. tokens	62%	59%	34%	71%
Arguments	345	195	187	372
Avg. nr. premises	1.26	2.03	1.15	1.33

Table 3: Annotation statistics

Looking at types of argumentation schemes, the models annotated many of the same schemes as the humans, but not necessarily for the same arguments. The most common scheme for both models were 'Argument from Consequences' (31% of the schemes for GPT-4o and 23% for Claude). This is the second most common scheme for both human annotators. The second most common scheme for GPT-4o, 'Argument from Example', was not used at all by the other annotators. The second most common scheme for Claude, 'Argument from cause to effect', which is the fourth most common scheme for both annotators.

In this annotation, an annotated span representing a component can be used in more than one argument, and take on both the role of premise and conclusion (in separate arguments). For this reason, agreement is first compared on token-level and then on individual arguments. Table 4 shows agreement in Krippendorff's α (Krippendorff, 2018) between annotators and models on whether a token is annotated as being part of an argument component, as well as agreement on the separate component types. Overall, Claude agrees more with the human annotators than GPT-4o does. Adding Claude's annotation to the humans annotations increases the agreement. GPT-4o agrees more on conclusions than on premises, while the agreement between Claude an humans is similar for both components. Interestingly, the highest agreement is between the two models, without including the human annotators. It should be noted that agreement in general is quite low, reflecting the complexity of the task.

	Arg. vs. non-arg	Conclusion	Premise
H	0.15	0.19	0.22
GPT+H	0.03	0.12	0.074
CL+H	0.16	0.16	0.18
CL+GPT	0.26	0.23	0.21
All	0.13	0.15	0.13

Table 4: α for the schemes. H = Humans, GPT = GPT-4o, CL = Claude

Agreement between humans and models varied between editorials, between 0.26– -0.26 for GPT-4o and between 0.38– -0.05 for Claude. One editorial had lowest agreement between both models and humans, as well as low agreement between the human annotators. Inspecting this editorial the annotators (both models and humans) have found different arguments. However, as seen in the example below, one conclusion was annotated by all, but together with different premises and schemes.

Conclusion All: 'It is not difficult to understand that the Social Democrats talk more about how many people are employed than about how many lack employment.'
Premise A: 'If there are many who are unemployed, it is bad to talk about how many lack employment.'
Premise A: 'Many are unemployed'
Premise B: 'The Social Democrats have, as is well known, replaced their old slogan "full employment´´ with 'work for all."
Premise B: 'When it comes to 'work for all,' the Social Democrats are more vague about the goal.'
Premise Claude: 'When it comes to 'work for all,' the Social Democrats are more vague about the goal.'
Premise GPT-4o: 'At his press conference in Malmö on Friday, the Prime Minister, apparently deliberately, downplayed the significance.'

Figure 2 shows pair-wise agreement between annotators. Both models agree more with annotator A, and Claude agrees the most with all.

Looking instead at the individual arguments and their components, there are 25 arguments between all with the same conclusion. Out of these 25, 6 share at least one premise. In these arguments,

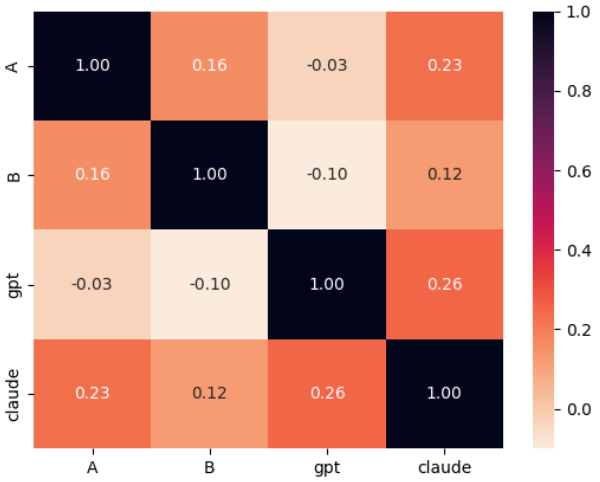

Figure 1: Pair-wise α on token-level, for the schemes

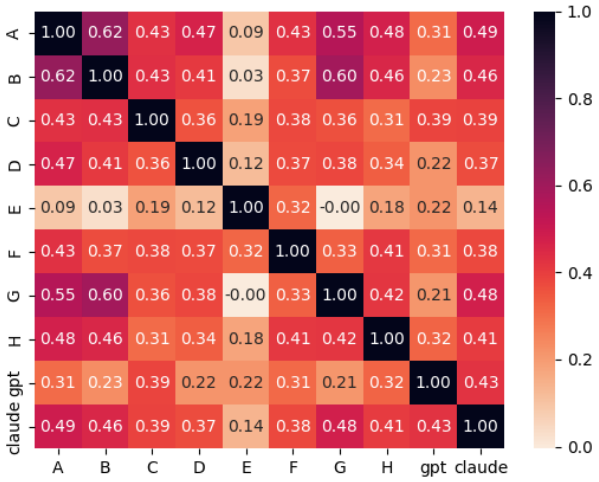

Figure 2: Pair-wise α for the online forum

the models often agreed on the scheme, while the humans more often disagreed.

Manually inspecting the annotated arguments further, more examples of annotators (both models and humans) choosing the same conclusion but different schemes and premises is found. Likewise, there are examples of components being used as both premise and conclusion, as in the example below.

Premise GPT-4o& A, Conclusion Claude: 'On election day, the individual voter is sovereign.'

Premise Claude & A, Conclusion GPT: 'This is the foundation of democracy.'

Conclusion A: 'Therefore, our appeal to our readers is this: take the opportunity to decide how our country should be governed over the next three-year period.'

Scheme A: ARGUMENT FROM CONSEQUENCES

Scheme GPT-4o: ARGUMENT FROM POPULAR OPINION

Scheme Claude: ARGUMENT FROM POSITION TO KNOW

5.2 Argumentation spans

In this task, the models were asked to annotate threads from online forums with spans that they considered to be argumentation. Similar to the previous task, GPT-4o annotated fewer tokens than most of the other annotators in the online forums. The model annotated 17.4% of the tokens as argumentative, as compared to between 20–44% for the other annotators. There is however an outlier annotator who has annotated even fewer tokens. Claude annotated 33.6% as argumentative. Agreement on token level is seen in Table 5. Agreement between the two models and humans, both together and separately is very close to the agreement between humans (0.39). However, agreement between only the models is slightly higher (0.43).

	H	GPT+H	CL+H	GPT+CL	All
α	0.39	0.37	0.39	0.43	0.38

Table 5: α on token level, argumentation spans. H = Humans, GPT = GPT-4o, CL = Claude

In Figure 2, pairwise agreement is shown. GPT-4o had among the lowest inter-annotator agreement scores, while Claude achieved higher agreement with human annotators. GPT-4o agrees the most with annotator C and H and Claude agrees the most with A and G. Interestingly, GPT-4o agrees more with the outlier annotator E than most of the other annotators do.

The pairwise token overlap between the annotators was also compared, as seen in Figure 3. For 4 of the annotators (A,B,G and F) more than 50% of GPT-4o's and Claude's annotations overlap with their annotations. In manual inspection of the threads, examples of overlap and partial overlap were also found. In the examples with high agreement, it was found that the human annotators and Claude often annotated longer spans, while GPT-4o did not. For example, below all annotators (including the models) annotated the first part of the text as argumentation, and while Claude and four of the human annotators also annotated the part in italics.

> "On the contrary. We need more home economics in schools, and more subjects need to be integrated into home economics. Home economics is the subject

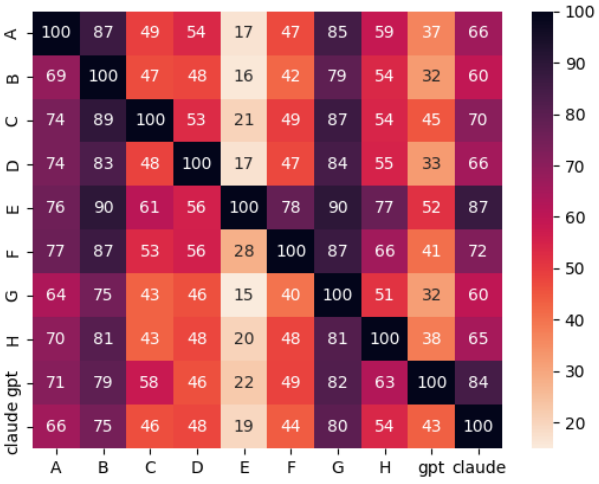

Figure 3: Pairwise percentage overlap. The figure is not symmetrical, for example are 71% GPT-4o's annotations found in A's, but 37% of A's found in GPT-4o's.

> that truly has the potential to teach practical, real-world knowledge that young people need in order to manage on their own. *And don't come and say that children learn this kind of thing at home anyway, because that's actually far from certain.*"

When inspecting spans with low agreement, the models sometimes annotated spans that were probably not intended as argumentation. For example, the span below was only annotated by one of the human annotators and is more narrative than argumentative, but the models both annotated it as argumentative.

> "After struggling for years to improve the situation without success, I have decided to leave."

5.3 Attitude

In this task, tweets were annotated for object of negative or positive attitude. The number of tweets to be annotated was originally 500, but 97 of them were excluded either due to wrong annotation format or missing annotations. As seen in Table 6, out of the remaining tweets, GPT-4o annotated the fewest tokens (10%), while Claude annotated similarly to the human annotators. Comparing label distribution, the models annotated a similar amount of negative and positive tags, while the human annotators annotated more positive tokens than negative ones.

	A	B	C	D	GPT	CL
POS	0.67	0.63	0.71	0.62	0.52	0.56
NEG	0.33	0.37	0.29	0.38	0.47	0.43
Tok.	52%	23%	25%	25%	10%	33%

Table 6: Label distributions in annotations. Tok. = annotated tokens. GPT = GPT-4o, CL = Claude

Agreement on token level is shown in Table 7. Unlike the previous tasks, agreement between the two models is lower than other annotator combinations. Agreement within humans and humans + Claude is similar, while GPT-4o has lower agreement with humans.

	H	GPT+H	CL+H	GPT+CL	All
α	0.35	0.27	0.35	0.25	0.30

Table 7: α on token level, argumentation spans. H = Humans, GPT = GPT-4o, CL = Claude

Pair-wise agreement is shown in Figure 4. It is lower between GPT-4o and the other annotators, while Claude's agreement scores are more similar to the human annotators.

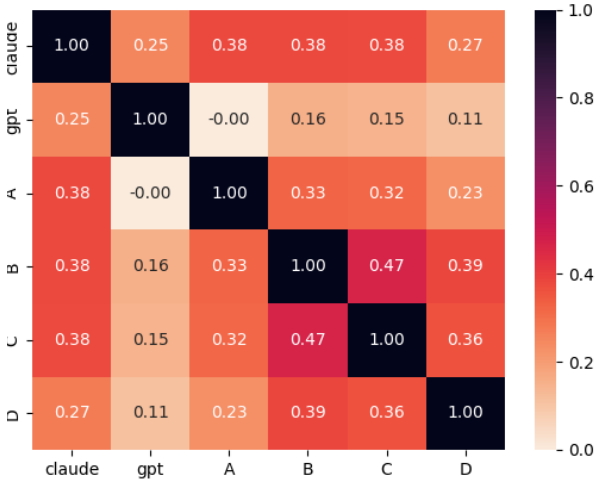

Figure 4: Pair-wise α for the tweets

For investigating labels further, Krippendorff's unitized ${}_u\alpha$ is employed (Krippendorff et al., 2016). This measure is suitable for comparing spans, but it can also show agreement on only labeled spans (ignoring label combinations such as NEG,O). As seen in Table 8, agreement is high for all annotator combinations. This tells us that the annotators and models agree substantially on the label, in instances when they have annotated the same span.

Annotators	$_{posu}\alpha$	cover	$_{negu}\alpha$	cover
H	0.82	53%	0.82	53%
GPT+H	0.83	44%	0.83	45%
CL+H	0.83	53%	0.83	54%
GPT+CL	0.97	36%	0.97	39%
All	0.84	47%	0.84	48%

Table 8: $_{u}\alpha$ for labels. Cover = coverage, % of annotated spans. H = Humans, GPT = GPT-4o, CL = Claude

Manual inspection of examples with low and high agreement revealed that low agreement often resulted from one annotator considering something argumentative that the others did not. However, it was rarely the case that the models annotated something which had not been annotated by at least one other annotator. In cases of disagreement between annotators, there was no annotator who the models with aligned more. In general, there were examples both where the models seemed to be better at following the guidelines, and cases where they annotated strangely. The models also often annotated shorter spans than the human annotators.

An example of this is shown below, where **bold** signifies positive spans, *italics* negative. In the example, "aid is needed" is not an obvious object. On the other hand, in the first sentence the models have annotated "Humanitarian superpower", which is more in line with the guidelines which instructed the annotators to keep the spans as short as possible. The other annotators have instead annotated the full sentence.

- GPT-4o: Sweden should continue to be a **humanitarian superpower.** Our **aid is needed. Humanitarian organizations** are *shamefully underfunded.*
- Claude: Sweden should continue to be a **humanitarian superpower. Our aid is needed. Humanitarian organizations** are *shamefully underfunded.*
- A & C: **Sweden should continue to be a humanitarian superpower. Our aid** is needed.*Humanitarian organizations are shamefully underfunded.*
- B: Sweden should continue to be a humanitarian superpower. **Our aid** is needed. Humanitarian organizations are shamefully underfunded.
- D: **Sweden should continue to be a humanitarian superpower.** Our aid is needed. Humanitarian organizations are shamefully underfunded.

What is included in a span can also affect the label, as seen in the example below.

- A B, & D: Sweden now has **a government that will not introduce market rents.** Tenants are today's big winners.
- GPT-4o, Claude & C: Sweden now has a government that will not introduce *market rents.* Tenants are today's big winners.

6 Conclusions

This paper first examined how well the models could perform annotation tasks when provided with guidelines. In this study we have shown that both models exhibit similar annotation patterns and agreement to that of humans, which leads us to conclude that the models can follow the guidelines and perform the task reasonably well. As these are tasks where humans often disagree, reaching comparable levels of agreement to humans could be a sign that the models, especially Claude, 'understands'[3] the task.

In all tasks, GPT-4o annotated fewer tokens than the human annotators, while Claude annotated a similar amount of tokens. This might be because GPT-4o only annotates when highly confident or due to differences in how the models approach the task. While models might respond differently to prompts, GPT-4o consistently annotated fewer tokens even during the prompt design phase.

Comparing agreement with the annotators, GPT-4o agrees less with the annotators than Claude. Claude exhibits agreement similar to that of humans, as well as higher agreement in the argumentation scheme task. When manually inspecting annotated examples, Claude would often annotate more similarly to humans, while GPT-4o shows similar patterns as another, slightly worse, human annotator. However, both models' annotations were often valid. As there can be cases where multiple interpretations are correct, for example in choosing a component as a premise or conclusion, one can not always conclude that models are wrong even if they choose to annotate differently than humans.

[3]The discussion of what understanding in this context means is left for another study.

However, there were some cases where the models' annotations did not make sense.

Because of the complexity of these tasks, and the fact that human annotators often disagree in them, the second question asked how models annotate when humans disagree. There were instances where models aligned with specific annotators. In the argumentation scheme task, both models agreed more with one of the annotators. Likewise, in the argumentation spans and attitude task, there was higher agreement with some annotators. However, in these tasks, the agreement is also different within the human annotators themselves. Most prominently, however, agreement was higher between the models than the humans in the argumentation scheme and spans task, possibly suggesting similarities in their reasoning. In the former task the models also agreed more on argumentation schemes types.

For future research, there are several promising directions. First, evaluating how other LLMs, particularly open-source models, annotate these datasets would help determine whether these findings generalize across different models and versions. Second, testing these models in zero-shot settings would reveal the extent of their inherent knowledge about argumentation without explicit guidelines. Finally, expanding the analysis to include other argumentation datasets, especially those in English, would provide broader insights into model performance across diverse argumentative contexts. In general, there are many questions to answer regarding how to use LLMs as annotators. For example, should each separate run be treated as a new annotator? In that case, should a failed run be considered as an annotator failing to perform the task?

Limitations

This study explores how two versions of the GPT and Claude models annotate, but the results might not hold for updated version versions of these models. Likewise, new models and new versions of existing models are released with increasing speed. This leads to difficulties reproducing results, not only in this study but for most studies employing LLMs.

The experiments in this study were carried out in the Swedish language, with instructions, guidelines and datasets in Swedish. This might limit the cross-linguistic generalizability of the results.

References

Mohammad Yeghaneh Abkenar, Weixing Wang, Hendrik Graupner, and Manfred Stede. 2024. Assessing open-source large language models on argumentation mining subtasks. *Preprint*, arXiv:2411.05639.

Mohammed Aldeen, Joshua Luo, Ashley Lian, Venus Zheng, Allen Hong, Preethika Yetukuri, and Long Cheng. 2023. Chatgpt vs. human annotators: A comprehensive analysis of chatgpt for text annotation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 602–609. IEEE.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *Preprint*, arXiv:2406.11980.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can large language models perform relation-based argument mining? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, Abu Dhabi, UAE. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50:2347–2364.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Mao Li and Frederick Conrad. 2024. Advancing annotation of stance in social media posts: A comparative analysis of large language models and crowd sourcing. *Preprint*, arXiv:2406.07483.

Anna Lindahl. 2020. Annotating argumentation in Swedish social media. In *Proceedings of the 7th Workshop on Argument Mining*, pages 100–105, Online. ACL.

Anna Lindahl. 2024. Disagreement in argumentation annotation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 56–66, Torino, Italia. ELRA and ICCL.

Anna Lindahl. 2025. Annotating attitude in Swedish political tweets. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 106–110, Tallinn, Estonia. University of Tartu Library, Estonia.

Anna Lindahl and Lars Borin. 2024. Annotation for computational argumentation analysis: Issues and perspectives. *Language and Linguistics Compass*, 18(1):e12505.

Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence. ACL.

Chandreen R Liyanage, Ravi Gokani, and Vijay Mago. 2024. Gpt-4 as an x data annotator: Unraveling its performance on a stance classification task. *PloS one*, 19(8):e0307741.

Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Conference on Advances in Robust Argumentation Machines*, pages 129–146. Springer.

Cameron Morin and Matti Marttinen Larsson. 2025. Large corpora and large language models: a replicable method for automating grammatical annotation. *Linguistics Vanguard*, (0).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Victor Hugo Nascimento Rocha, Igor Cataneo Silveira, Paulo Pirozelli, Denis Deratani Mauá, and Fabio Gagliardi Cozman. 2023. Assessing good, bad and ugly arguments generated by chatgpt: a new dataset, its methodology and associated tasks. In *EPIA Conference on Artificial Intelligence*, pages 428–440. Springer.

Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2024. A GPT among annotators: LLM-based entity-level sentiment annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 133–139, St. Julians, Malta. Association for Computational Linguistics.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.

Robin Schaefer. 2025. On integrating LLMs into an argument annotation workflow. In *Proceedings of the 12th Argument mining Workshop*, pages 87–99, Vienna, Austria. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Morgan & Claypool, San Rafael.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Kangda Wei, Aayush Gautam, and Ruihong Huang. 2024. Are LLMs good annotators for discourse-level event relation extraction? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1–19, Miami, Florida, USA. Association for Computational Linguistics.

A Prompts

Translated prompts for the different tasks.

A.1 Argumentation schemes

"Your task is to annotate Walton's argumentation schemes. Here is a description of these schemes: [guidelines]
Based on these descriptions, I want you to annotate schemes. Do this by marking out what is the conclusion and which premises belong to it. Then you mark which argumentation scheme is used. I want you to mark out exact text spans. Mark out all schemes you can find. Be thorough and don't stop until you can't find more schemes. Return the annotations as a single json file, in this format: {"scheme": "annotated scheme name", "components":[{"role":"conclusion","span": "the annotated conclusion"},{"role": "premise", "span": "the annotated premise"}]} Include the entire text span in the "span" field, don't abbreviate and don't correct any spelling errors. Annotate this text: [example] "

A.2 Argumentation spans

"Your task is to annotate text spans that you consider to contain argumentation. Here are the annotation guidelines [guidelines]. The text is from an online forum, where each post is marked with "==". Divide the annotations per post. Make the annotations in json format, as a single object. Here is the text: [example]. "

A.3 Attitude

"Your task is to annotate tweets. Here are the annotation guidelines . Make the annotations by marking which words are included in positive or negative spans in json format like this: [{"tweet id": 0 , "annotated tweet": Now_O needs_O every_O penny_O needs_O to_O go_O to_O counteract_O the_NEG municipal_NEG crisis_NEG ._O Therefore_O we_O say_O no_O to_O increased_NEG Swedish_NEG EU-fee_NEG ._O },{"tweet": ...}] Make sure it is valid json. Be careful to annotate what the attitude is expressed towards, not generally negative or positive words. Remember that both words, phrases and whole sentences can be annotated. Annotate these tweets: [10 examples] "

B Scheme descriptions

ARGUMENT FROM SIGN:
Premise: A is true in this situation.
Premise: Event B is generally indicated as true when its sign, A, is true in this kind of situation.
Conclusion: B is true in this situation.

ARGUMENT FROM CONSEQUENCES:
Premise: If A is brought about, then good (bad) consequences will (may plausibly) occur.
Conclusion: A should (not) be brought about.

ARGUMENT FROM EVIDENCE TO A HYPOTHESIS:
Premise: If hypothesis A is true, then a proposition B, reporting an event, will be observed to be true.
Premise: B has been observed to be true in a given instance .
Conclusion: A is true.

ARGUMENT FROM CORRELATION TO CAUSE:
Premise: There is a positive correlation between A and B.
Conclusion: A causes B.

ARGUMENT FROM POPULAR PRACTICE:
Premise: If a large majority (everyone, nearly everyone, etc.) does A, or acts as though A is the right (or an acceptable) thing to do, then A is a prudent course of action.
Premise: A large majority acts as though A is the right thing to do.
Conclusion: A is a prudent course of action.

ARGUMENT FROM EXAMPLE:
Premise: In this case, the individual a has property F and also property G
Premise: a is typical of things that have F and may or may not have G
Conclusion: Generally, if x has property F then (usually, probably, typically) x also has property G

ARGUMENT FROM POSITION TO KNOW:
Premise: a is in a position to know whether A is true (false)
Premise: a asserts that A is true (false)
Conclusion: A is true (false)