Can Large Language Models Understand Argument Schemes?

Elfia Bezou-Vrakatseli, Oana Cocarascu and Sanjay Modgil
King's College London
{elfia.bezou_vrakatseli, oana.cocarascu, sanjay.modgil}@kcl.ac.uk

Abstract

Argument schemes represent stereotypical patterns of reasoning that occur in everyday arguments. However, despite their usefulness, argument scheme classification, that is classifying natural language arguments according to the schemes they are instances of, is an under-explored task in NLP. In this paper we present a systematic evaluation of large language models (LLMs) for classifying argument schemes based on Walton's taxonomy. We experiment with seven LLMs in zero-shot, few-shot, and chain-of-thought prompting, and explore two strategies to enhance task instructions: employing formal definitions and LLM-generated descriptions. Our analysis on both manually annotated and automatically generated arguments, including enthymemes, indicates that while larger models exhibit satisfactory performance in identifying argument schemes, challenges remain for smaller models. Our work offers the first comprehensive assessment of LLMs in identifying argument schemes, and provides insights for advancing reasoning capabilities in computational argumentation.

1 Introduction

Computational argumentation is an area of research that engages natural language processing (NLP) tasks including argument mining, assessment, reasoning, and generation (Lauscher et al., 2022). Despite extensive research efforts, one task that has received limited attention is argument scheme classification. Argument schemes provide structured templates capturing stereotypical forms of arguments consisting of inferences from premise(s) to a conclusion. These schemes have traditionally been used in formal logic-based argumentation to support reasoning and deliberation (Modgil, 2017), with particular emphasis on supporting dialogues for value alignment (Bezou-Vrakatseli et al., 2024b), but have also been explored within the NLP field.

Classifying argument schemes (i.e. classifying natural language arguments according to the schemes they are instances of) enhances understanding of the reasoning process behind arguments and is an important step towards supporting human and AI uses of argument in reasoning and deliberation. Additionally, scheme classification helps in assessing the validity of arguments (including identifying fallacious reasoning) and in strengthening their quality and scope. Classifying argument schemes can enhance decision-making in AI systems, particularly in domains requiring complex reasoning, such as legal and medical applications. Incorporating argument schemes into decision-making processes promotes transparency and explainability, which are essential in high-stakes applications, such as healthcare, finance, and governance.

Various taxonomies have been proposed to systematically categorise these schemes; the most prevalent being the taxonomy developed by Walton et al. (2008). However, classifying argument schemes is a particularly challenging task (Walton and Macagno, 2015; Anthony and Kim, 2015; Lindahl et al., 2019; Bezou-Vrakatseli et al., 2024a). Indeed, the cognitive load and resources associated with this task is higher compared to other tasks such as identifying distinct components of arguments and their stance towards a topic or other arguments (Musi et al., 2016).

Large language models (LLMs) are increasingly being used to automate resource intensive cognitively demanding tasks. In particular, they have shown impressive in-context learning (ICL) ability, whereby the model learns to perform tasks from prompts consisting of input-output examples, significantly advancing the NLP field. Given their remarkable results in a variety of tasks, recent works have begun evaluating the performance of LLMs on several computational argumentation tasks, including argument component classification (Chen

Findings of the Association for Computational Linguistics: ACL 2025, pages 13666–13681
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024; Al Zubaer et al., 2023; Ren et al., 2024), fallacy classification (Pan et al., 2024), argument validity and novelty (Van Der Meer et al., 2022), argument quality (Mirzakhmedova et al., 2024), stance detection (Chen et al., 2024; Gorur et al., 2025), and argument generation (Chen et al., 2024; Verma et al., 2024). Despite these advances, use of LLMs for reasoning with and about arguments, in particular via use of argument schemes, remains relatively understudied (Gupta et al., 2024; Castagna et al., 2024; Saadat-Yazdi and Kökciyan, 2024; Trajano et al., 2024).

In this paper, we systematically evaluate whether LLMs are capable of classifying argument schemes based on the taxonomy developed by Walton et al. (2008). We examine the performance of LLM models for argument scheme classification in zero-shot, few-shot, and chain-of-thought prompting. Moreover, we explore two approaches for enhancing the task instructions: using formal definitions and LLM-generated descriptions of argument schemes.

We evaluate seven open-source and proprietary models that vary in size: Llama-3.2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Claude-3.5 Sonnet (Anthropic, 2024), GPT-4 (Achiam et al., 2023), GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Team et al., 2024), and Grok-2 (X.AI, 2024). Furthermore, we analyse which argument schemes pose a challenge to the models. Our experiments on both manually annotated and automatically generated arguments, as well as enthymemes (i.e. natural language arguments in which premises and/or the conclusion of the argument is left implicit) reveal that Claude-3.5 performs consistently better compared to the other models, although there is room for improvement in this task.

In summary, our **contributions** are as follows.

- To the best of our knowledge, our work is the first to systematically evaluate the capabilities of LLMs in classifying argument schemes.
- We evaluate the performance of seven LLMs in various setups: zero-shot, few-shot, chain-of-thought prompting, and explore two methods for enhancing the task instruction: normative information and LLM-generated descriptions.
- We examine whether there are differences in how LLMs handle manually annotated human-made arguments versus automatically generated arguments.
- Additionally, we investigate the impact of enthymemes on LLMs and evaluate the effectiveness of the models in handling the implicit content found in everyday natural language arguments.
- Our comprehensive evaluation reveals that larger models can identify argument schemes satisfactorily while smaller models struggle.

2 Background

Argument Schemes (AS) represent stereotypical patterns of reasoning that occur in everyday arguments (Walton and Godden, 2005). Their origin traces back to Aristotle's topics, and have become pivotal in modern-day argumentation theory (Macagno et al., 2017). Taxonomies for systematically categorising AS include the pragma-dialectical classification (Van Eemeren et al., 2004), the periodic table of arguments (Wagemans, 2016), and Walton's taxonomy (Walton et al., 2008). In this paper we focus on the latter, which proposes over 60 argument schemes each of which relate premises to a conclusion, and each of which have associated critical questions that identify how to challenge arguments that are instances of the scheme. Table 1 illustrates commonly used schemes and their definitions.

Enthymemes In natural language arguments, either some of the premises and/or the conclusion are left implicit. Such arguments are referred to as *enthymemes* (Walton, 2009). For example, consider the following enthymeme taken from a debate on the topic *Should the death penalty be abolished?*:

> A: "The death penalty negatively affects both the families of the victims as well as the defendant(s)."

The argument's conclusion, "therefore the death penalty should be abolished", is left implicit. It is assumed that a reader will infer the intended conclusion from *context* (i.e. the debate topic).

3 Related Work

3.1 Argument schemes

NLP research on argument scheme classification has mainly focused on developing scheme classification guidelines (Walton, 2012; Walton and Macagno, 2015; Musi et al., 2016; Lawrence et al., 2019; Visser et al., 2021) or developing datasets of

Argument from Consequences
P: If A is brought about, good/bad consequences occur.
C: Therefore, A should (not) be brought about.

Argument from Example
P: In this particular case, a has properties F and G.
C: So, generally, if x has property F, then it also has G.

Argument from Values
P1: Value V is positive/negative as judged by agent a.
P2: If V is positive it is a reason to commit to goal G.
P2': If V is negative it is a reason to not commit to goal G.
C: V is a reason for agent a to (not) commit to goal G.

Argument from Cause to Effect
P1: Generally, if A occurs, then B will (might) occur.
P2: In this case, A occurs (might occur).
C: Therefore, in this case, B will (might) occur.

Argument from Expert Opinion
P1: Source E is an expert in subject S.
P2: E asserts that proposition A of subject S is true/false.
C: A is true/false.

Argument from Alternatives
P1: Either X or Y can be the case.
P2: X is plausibly not the case.
C: Y is plausibly the case.

Argument from Analogy
P1: Generally, case $C1$ is similar to case $C2$.
P2: A is true (false) in case $C1$.
C: A is true (false) in case $C2$

Table 1: Commonly used argument schemes in discourse along with their definition, in the form of premise (P) and conclusion (C).

arguments annotated as instances of schemes. The majority of existing datasets tend to be based on Walton's taxonomy (Feng and Hirst, 2011; Hansen and Walton, 2013; Schneider et al., 2013; Lawrence and Reed, 2015; Bezou-Vrakatseli et al., 2024a), while some works have created categories based on argument schemes (Song et al., 2014; Green, 2018; Reisert et al., 2018; Jo et al., 2020; Kondo et al., 2021).

Most of the available corpora that use Walton's taxonomy are either small in size (Feng and Hirst, 2011; Lawrence and Reed, 2015), have a large number of schemes (Hansen and Walton, 2013; Schneider et al., 2013), or concentrate on a very small subset (e.g. 2) of schemes (Jo et al., 2021) in order to avoid the complexity of handling numerous schemes. Among existing datasets, *EthiX* (Bezou-Vrakatseli et al., 2024a) covers a wide range of commonly used argumentative structures, achieving a good balance in terms of size and number of classes (i.e. argument schemes), as well as a moderate inter-annotator agreement (a level typical of agreement in argument mining tasks (Lawrence and Reed, 2020)).

There have also been attempts to automatically generate arguments using argument schemes. Ruiz-Dolz et al. (2024) generated arguments from Walton's taxonomy using LLMs (GPT-3.5-Turbo and GPT-4), whilst Saha and Srihari (2023) used RoBERTa to annotate scheme-based control codes derived from Walton's argument schemes.

Other uses of argument schemes include their application in educational frameworks to reconstruct students' arguments (Macagno and Konstantinidou, 2013) and mapping schemes to discourse relation categories (Cabrio et al., 2013).

3.2 LLMs for Computational Argumentation

Given the exceptional performance of LLMs in understanding and generating natural language, recent works have started evaluating the capabilities of LLMs in several computational argumentation tasks such as argument component classification (Chen et al., 2024; Al Zubaer et al., 2023; Ren et al., 2024), fallacy classification (Pan et al., 2024), argument validity and novelty (Van Der Meer et al., 2022), argument quality (Mirzakhmedova et al., 2024), stance detection (Chen et al., 2024; Gorur et al., 2025), and argument generation (Chen et al., 2024; Verma et al., 2024).

Chen et al. (2024) assessed the performance of various LLMs (GPT-3.5-turbo, Flan and Llama models) in zero-shot and few-shot settings in argument mining (in particular claim, evidence, and stance detection, as well as evidence type classification) and argument generation. Their findings indicate promising performance across most datasets used in evaluation. Gorur et al. (2025) investigated the effectiveness of several LLMs (Llama, Mistral, Mixtral, and GPT-3.5-Turbo) for relation-based argument mining, a task similar to stance detection which involves identifying support and attack relations between arguments. Pan et al. (2024) evaluated various LLMs on the task of classifying logical fallacies. They proposed single-round and multi-round prompting strategies, which incorporated task instructions such as extraction, summarization, and chain-of-thought reasoning, and showed that LLMs have potential as zero-shot fallacy classifiers.

Other studies have explored GPT-3.5 and GPT-4 for argument component classification in the legal domain (Al Zubaer et al., 2023) as well as argument mining and argument pair extraction (de Wynter and Yuan, 2024), showing significant variation in the performance of these models. Mao et al. (2024) proposed two prompt frameworks for argument pair extraction from dialogues to improve the comprehension of complex argumentative dialogues.

Some studies focus on the argument quality assessment task. Mirzakhmedova et al. (2024) tested the capabilities of GPT-3 and PaLM 2 in evaluating the quality of arguments and found that LLMs exhibit greater consistency compared to humans. Van Der Meer et al. (2022) compared GPT-3 to transformer models trained with different paradigms, multi-task, contrastive, and intermediate-task learning, showing that GPT-3 works best for predicting argument validity, while a model trained using all three paradigms is better at predicting argument novelty.

There are few works that have explored the use of LLMs for argument reasoning. Gupta et al. (2024) introduced the argument explication task based on Toulmin's model of argumentation (Toulmin, 2003) and proposed a two-stage framework where LLMs are prompted to automatically extract claims, reasons, and warrants from informal arguments. Castagna et al. (2024) also used Toulmin's model as well as critical questions to enhance the logical and mathematical reasoning capabilities of LLMs by incorporating them into a prompting method inspired by chain-of-thought. Saadat-Yazdi and Kökciyan (2024) used LLMs to classify arguments according to their canonical form following Wagemans (2016)'s taxonomy of argument schemes. Finally, Trajano et al. (2024) used the argument schemes defined by Walton et al. (2008) to provide references to LLMs for translating natural language to computational arguments.

4 LLMs for Argument Scheme Classification

4.1 Datasets

We use two datasets annotated using Walton's taxonomy (Walton et al., 2008) in our evaluation: *EthiX* (Bezou-Vrakatseli et al., 2024a) and the English version of *NLAS-multi* (Ruiz-Dolz et al., 2024) that we will refer to as *NLAS*.

EthiX is a manually annotated dataset with 686 arguments extracted from Kialo,[1] a user-generated platform with structured and moderated debates. Each instance in the dataset contains an argument and the central question of the debate from which they were extracted. Kialo has been extensively used in argumentation research due to its structured (arguments are pro/con a topic or other arguments) yet naturalistic representation of argumentative dialogue (arguments are user-generated). In our experiments, we combined *argument from positive consequences* and *argument from negative consequences* into a single class named *argument from consequences*. This change aligns with the *argument from values* scheme, which does not distinguish between positive and negative values.

NLAS (Ruiz-Dolz et al., 2024) comprises 1,893 arguments generated automatically using GPT-3.5-Turbo and GPT-4, spanning 50 topics and instantiating 20 argument schemes. As the number of classes is large and in order to allow a direct comparison with *EthiX*, we select the argument schemes that appear in both datasets, namely *argument from example*, *argument from cause to effect*, *argument from expert opinion*, and *argument from analogy*.

We also use a subset of *EthiX* consisting of enthymemes only (see Section 2), allowing us to test the effectiveness of the models in handling implicit content. Table 2 shows statistics for the datasets used in our evaluation and Table 3 shows examples from each dataset.

Scheme	**EthiX**	**EthiX$_e$**	**EthiX$_4$**	**NLAS$_4$**
Arg from consequences[2]	174	109	-	-
Arg from example	120	72	120	97
Arg from values	117	55	-	-
Arg from cause to effect	88	23	88	98
Arg from expert opinion	81	33	81	100
Arg from alternatives	56	24	-	-
Arg from analogy	50	23	50	99

Table 2: Distribution of argument schemes across datasets. *EthiX$_e$* ony contains enthymemes. *EthiX$_4$* and *NLAS$_4$* contain the four classes that appear in both datasets.

4.2 Models

To assess the extent to which LLMs understand argument schemes, we evaluate seven open-source and proprietary models that vary in size: Llama-3.2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Claude-3.5 Sonnet (Anthropic, 2024), GPT-4 (Achiam et al., 2023), GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Team et al., 2024), and Grok-2 (X.AI, 2024). In our experiments, we set the temperature parameter to zero. We evaluate model performance using accuracy and macro F_1.

4.3 Experimental setup

To evaluate the capabilities of LLMs in classifying argument schemes, we divide our experiments into two primary setups: zero-shot and few-shot. We

---

[1] `https://www.kialo.com`

[2] Positive consequences: 96; Negative consequences: 78.

Scheme	NLAS$_4$	EthiX$_4$
Argument from Example	**A**: "*Premise*: In this particular case, John has terminal cancer and also the right to die with dignity through euthanasia. *Conclusion*: Therefore, generally, if any individual has terminal cancer, then they also have the right to die with dignity through euthanasia." **D**: Euthanasia	**A**: "In some places in the US, health-care workers are required to get vaccinated against the seasonal flu."* **D**: Should Covid-19 vaccines be mandatory?
Argument from Cause to Effect	**A**: "*Major Premise*: Generally, if a person has a favorable physical appearance, then they will be perceived more positively by others. *Minor Premise*: In this case, the person has made efforts to improve their physical appearance. *Conclusion*: Therefore, in this case, the person is likely to be perceived more positively by others, which can potentially lead to personal success." **D**: Physical appearance for personal success	**A**: "The cost of palliative care is extremely high, thus many people who opt for euthanasia are more likely to be poor." **D**: Should euthanasia be legalised?
Argument from Expert Opinion	**A**: "*Major Premise*: Dr. Nora Volkow is an expert in neuroscience and drug abuse research. *Minor Premise*: Dr. Volkow asserts that legalizing cannabis can lead to negative public health outcomes, including addiction, cognitive impairment, and mental health issues. *Conclusion*: Legalising cannabis is not advisable as it can have negative public health outcomes." **D:** Legalisation of cannabis.	**A**: "Masks reduce Covid transmission according to WHO."* **D**: Do people have a right to not wear a mask in public spaces during the COVID-19 pandemic?
Argument from Analogy	**A**: "*Similarity Premise*: Generally, killing another human being is considered morally wrong. *Base Premise*: Abortion involves the deliberate termination of a developing human being *Conclusion*: Abortion should be considered morally wrong." **D:** Abortion	**A**: "The right to reproductive freedom already includes assisted reproductive technologies such as in-vitro fertilisation. Cloning humans could be seen as another assisted reproductive technology." **D**: Is cloning humans ethical?

Table 3: Example of arguments (A) for the four schemes in NLAS$_4$ and EthiX$_4$, along with the debate topic and question (D), respectively. Arguments marked with an asterisk in EthiX$_4$ also appear in EthiX$_e$.

aim to assess the models' capabilities in identifying argument schemes by evaluating their performance across different setups. To achieve this, in addition to the standard zero- and few-shot prompting, we experiment with chain-of-thought (CoT) and explore whether enhancing the task instruction with normative information or LLM-generated descriptions of argument schemes improves the performance of the models tested in recognising these schemes.

Zero-shot Prompting We provide task instructions, clarifying the expected input (the argument and the debate topic the argument is about) and the required answer (classification into one of the predefined classes). The models classify a test instance without prior specific examples.[3] In all setups, we enrich the prompt with the phrase "According to [Walton's taxonomy]", which is shown to enhance the model's ability to ground the information against previously observed text (Weller et al., 2024).

Few-shot Prompting In the few-shot setting, the model is given a small set of labeled examples to guide its response, leveraging in-context learning (Dong et al., 2024). We provide task instructions and two examples from each class to include in the prompt as follows: first, we give one example for each scheme, and then we give a second example for each scheme in the same sequence. We observed that this scheme-by-scheme progression led to better performance, compared to grouping examples by scheme.

Chain-of-Thought Prompting CoT breaks down a task into intermediate reasoning steps. For the zero-shot setting, we use the approach by Kojima et al. (2022), which requires low human engineering to craft the prompt, and add 'Let's think step by step' at the end of the prompt. For the few-shot setting, following Wei et al. (2022), we augment the standard few-shot prompting with the reasoning steps that lead to the answer.[4]

4.3.1 Enhanced task instructions

We experiment with two approaches for enhancing the task instructions in the prompts. The first utilises the formal definitions of the schemes as found in Walton's taxonomy (Walton et al., 2008), whereas the second relies on descriptions generated using an LLM. This comparison will provide

---

[3] The prompts are provided in Appendix B.

[4] We also experimented with more structured CoT approaches, including explicitly encoding Visser's scheme classification criteria (i.e., the Argument Scheme Key steps (Visser et al., 2021)) as intermediate reasoning steps within the prompts. However, these structured approaches did not yield meaningful improvements.

further insights into the utility of structured templates for prompting relative to the LLMs' ability to 'intuitively' generate the most helpful directives.

Enriching prompts with Formal Definitions In this setup, we include the formal definitions of the schemes (see Table 1 for the definitions) in the task instruction to evaluate whether this structured approach helps ground the LLM's responses.

Enriching prompts with LLM-generated descriptions We use ChatGPT-4 to generate descriptions of the schemes, which we include in the task instruction instead of the formal representations. Our aim was to test whether LLMs inherently "know best" what types of descriptions facilitate their own reasoning, as they match the representations naturally recognised by the models themselves (analogously to how human experts often construct such descriptions for human annotators, including informal explanations, appealing to human intuitions to effectively grasp schemes). Table 4 shows the prompt and the generated descriptions.

Prompt

Provide concise descriptions of the following argument schemes to aid an LLM in accurately classifying arguments according to these categories: argument from consequences, argument from example, argument from values, argument from cause to effect, argument from expert opinion, argument from alternatives, argument from analogy.

Output

- Argument from Consequences: This argument predicts positive or negative outcomes from a proposed action or policy and uses these consequences as a reason to support or oppose the action.
- Argument from Example: This scheme uses one or more specific cases or examples to support a general conclusion.
- Argument from Values: This argument scheme appeals to the values or principles that are held by a community or an individual. It supports a conclusion based on how well it aligns with these values.
- Argument from Cause to Effect: This argument predicts that a specific cause will lead to a specific effect, using the causal connection to support the conclusion.
- Argument from Expert Opinion: This scheme supports a conclusion by citing an authority or expert in the field relevant to the argument.
- Argument from Alternatives: This scheme presents different options or courses of action and arguing for one over the others based on certain criteria or the relative advantages.
- Argument from Analogy: This type of argument claims that because two things are similar in certain respects, they are likely similar in other respects as well.

Table 4: Descriptions of argument schemes generated using ChatGPT-4.

5 Results & Discussion

5.1 Which LLM performs the best?

Table 5 shows the model performances across different experimental setups, with the highest scores highlighted in bold. A notable trend is the consistently superior performance of Claude-3.5 across all variants of *EthiX*, especially prominent in few-shot settings. Claude-3.5 achieves the highest macro F_1 in the few-shot with descriptions setup for *EthiX* and $EthiX_e$. For $EthiX_4$, Claude-3.5 achieves the highest macro F_1 in two setups: standard few-shot and few-shot with descriptions. For $NLAS_4$, two models achieve a perfect macro F_1: GPT-4o in standard few-shot and few-shot with definitions, and Gemini-1.5 in few-shot with definitions setup.

Large vs Smaller models The larger models (Claude-3.5, GPT-4 variants, Gemini-1.5, and Grok-2) yield better results accross all datasets and setups and consistently outperform smaller models. Mistral and Llama-3.2 tend to score significantly lower, sometimes by 40-50% difference in F_1. This indicates that on a challenging dataset, the larger models deliver substantially better performance.

5.2 Which setup has the best results?

Zero-shot vs Few-shot Few-shot setups consistently outperform their zero-shot counterparts. The best performance on the *EthiX* dataset is observed in the few-shot with descriptions setting, where the difference in F_1 compared to zero-shot is 11%. This trend is also evident in the $NLAS_4$ dataset, where few-shot approaches secure perfect or near-perfect scores.

Moreover, while enriching the prompt with formal definitions or descriptions might lead to better results, the overall gains remain lower compared to the improvement obtained by including task demonstrations. This suggests that the models are able to generalise from examples.

Formal definitions vs. LLM-generated descriptions In the zero-shot setting, using descriptions as opposed to definitions results in improvements for Llama-3.2 and Claude-3.5 on the three variants of *EthiX*. With few-shot, Claude-3.5 yields the best performance across the *EthiX* variants using descriptions. However, in this setup, the differences between using descriptions and definitions are overall smaller than in zero-shot. Indeed, formal definitions tend to use technical language that

	Models	Zero-shot								Few-shot							
		Standard		CoT		Definitions		Descriptions		Standard		CoT		Definitions		Descriptions	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
EthiX	Mistral	0.31	0.38	0.34	0.39	0.38	0.45	0.34	0.44	0.35	0.34	0.34	0.38	0.38	0.45	0.40	0.46
	Llama-3.2	0.18	0.20	0.15	0.19	0.19	0.28	0.26	0.35	0.24	0.31	0.24	0.31	0.24	0.32	0.26	0.36
	Claude-3.5	0.56	0.58	0.50	0.49	0.52	0.54	0.54	0.55	0.61	0.60	0.56	0.55	0.60	0.61	**0.65**	**0.65**
	GPT-4o	0.49	0.53	0.50	0.54	0.49	0.51	0.48	0.50	0.57	0.58	0.56	0.58	0.53	0.53	0.54	0.56
	GPT-4	0.45	0.47	0.48	0.49	0.44	0.48	0.47	0.51	0.55	0.57	0.56	0.56	0.57	0.58	0.57	0.57
	Grok-2	0.49	0.54	0.48	0.53	0.49	0.53	0.50	0.53	0.57	0.58	0.53	0.55	0.57	0.58	0.55	0.56
	Gemini-1.5	0.45	0.52	0.45	0.51	0.43	0.50	0.47	0.52	0.56	0.58	0.55	0.57	0.55	0.58	0.56	0.58
EthiX$_e$	Mistral	0.29	0.37	0.34	0.39	0.38	0.50	0.37	0.50	0.29	0.37	0.36	0.37	0.38	0.48	0.36	0.45
	Llama-3.2	0.17	0.22	0.15	0.19	0.23	0. 32	0.29	0.41	0.26	0.33	0.27	0.35	0.29	0.38	0.28	0.36
	Claude-3.5	0.57	0.62	0.49	0.46	0.50	0.55	0.53	0.55	0.57	0.61	0.54	0.51	0.57	0.59	**0.62**	**0.62**
	GPT-4o	0.48	0.56	0.50	0.55	0.47	0.50	0.45	0.49	0.58	0.60	0.57	0.60	0.51	0.53	0.54	0.58
	GPT-4	0.42	0.46	0.44	0.49	0.43	0.48	0.45	0.52	0.51	0.58	0.51	0.59	0.52	0.58	0.53	0.57
	Grok-2	0.51	0.58	0.48	0.56	0.49	0.55	0.49	0.54	0.55	0.58	0.52	0.54	0.55	0.57	0.53	0.56
	Gemini-1.5	0.48	0.58	0.47	0.57	0.47	0.57	0.47	0.56	0.57	0.61	0.56	0.57	0.56	0.63	0.55	0.60
EthiX$_4$	Mistral	0.26	0.43	0.41	0.44	0.44	0.44	0.42	0.47	0.52	0.53	0.60	0.60	0.49	0.50	0.55	0.56
	Llama-3.2	0.17	0.17	0.14	0.16	0.15	0.18	0.19	0.26	0.36	0.38	0.34	0.34	0.28	0.36	0.28	0.37
	Claude-3.5	0.67	0.67	0.61	0.62	0.63	0.64	0.67	0.67	**0.72**	**0.72**	0.71	0.71	0.70	0.70	**0.72**	**0.72**
	GPT-4o	0.57	0.58	0.56	0.57	0.54	0.53	0.57	0.58	0.67	0.66	0.68	0.68	0.66	0.66	0.69	0.69
	GPT-4	0.51	0.53	0.55	0.55	0.55	0.56	0.53	0.54	0.68	0.66	0.69	0.68	0.68	0.69	0.69	0.69
	Grok-2	0.53	0.57	0.54	0.57	0.57	0.60	0.53	0.56	0.65	0.66	0.68	0.68	0.63	0.64	0.64	0.65
	Gemini-1.5	0.57	0.60	0.59	0.62	0.61	0.64	0.63	0.63	0.66	0.66	0.66	0.67	0.67	0.67	0.67	0.67
NLAS$_4$	Mistral	0.86	0.86	0.86	0.86	0.94	0.94	0.87	0.87	0.87	0.87	0.90	0.91	0.91	0.91	0.88	0.88
	Llama-3.2	0.45	0.50	0.32	0.39	0.77	0.77	0.73	0.76	0.47	0.50	0.79	0.80	0.63	0.64	0.46	0.54
	Claude-3.5	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	GPT-4o	0.98	0.98	0.98	0.98	0.99	0.99	0.98	0.98	**1.00**	**1.00**	0.99	0.99	**1.00**	**1.00**	0.99	0.99
	GPT-4	0.94	0.94	0.95	0.95	0.98	0.98	0.95	0.95	0.97	0.97	0.98	0.98	0.99	0.99	0.98	0.98
	Grok-2	0.91	0.91	0.93	0.93	0.99	0.99	0.96	0.96	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Gemini-1.5	0.94	0.94	0.94	0.94	0.99	0.99	0.96	0.96	0.99	0.99	0.99	0.99	**1.00**	**1.00**	0.99	0.99

Table 5: Model performance across different setups and datasets. The horizontal dashed line distinguishes large models from smaller ones.

may not align well with the training data used to develop these models, which is typically more diverse and includes various forms of natural language. As a result, when prompted with their own descriptions, presumably more reflective of the language and styles they encountered during training, LLMs appear to engage more effectively with the task.

In *NLAS*$_4$, however, definitions prove to be a more effective prompt. This can be attributed to the structured format of arguments in *NLAS*$_4$, which closely mirrors the explicit form of the definitions used in the prompts. This alignment naturally enhances the relevance and efficacy of definition-based prompting in this context.

Our findings suggest there is potential value in using LLM-generated explanations or definitions either as complementary resources alongside expert-authored descriptions or as practical alternatives, particularly in contexts where formal, expert-generated resources are scarce or costly to produce.

Does CoT improve results? Using CoT leads to similar or even worse results for the majority of the models. The highest drop in performance when using CoT can be seen with Claude-3.5 on the three variants of the *EthiX* dataset, with up to 8% difference in F_1 compared to standard zero-shot. In contrast, Mistral showcases a stronger ability to distinguish between argument schemes using CoT on the *EthiX* dataset variants, with up to 15% improvement in F_1. In the few-shot setting, using CoT for Llama-3.2 on *NLAS*$_4$ increases F_1 from 0.47 in the standard setup to 0.79, suggesting that for smaller models, the structured reasoning provided by CoT can be particularly beneficial.

5.3 Do LLMs perform differently on manually annotated arguments compared to generated arguments?

Comparing *EthiX*$_4$ with *NLAS*$_4$ reveals a stark difference in how LLMs handle manually annotated human-made arguments versus those generated automatically. The results on *EthiX*$_4$ performance

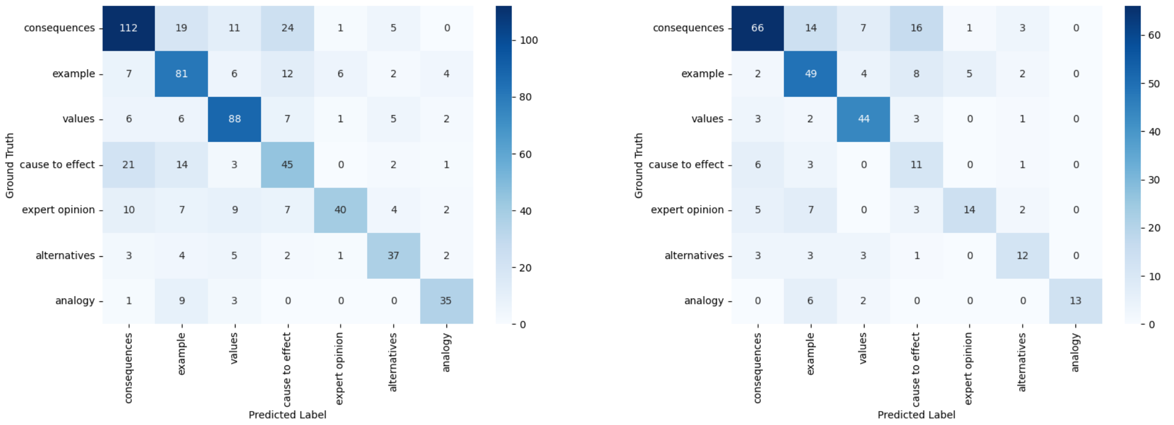

Figure 1: Confusion matrices for the best performing model, Claude-3.5 with *few-shot descriptions*, for *EthiX* (left figure) and *EthiX*$_e$ (right figure).

scores are significantly lower across all setups and models. On *NLAS*$_4$, models achieve near-perfect scores, with even the smaller models, although more variable, showing substantially higher scores.

This discrepancy suggests that LLMs are much more adept at handling automatically generated arguments. This is not surprising, as the generated arguments in *NLAS*$_4$ are more structured and consistent, with parts of the scheme name often appearing in the argument (see Table 3, e.g. “Dr. Nora Volkow is an expert in neuroscience”). In contrast, a manually annotated dataset with real-world arguments exhibits greater variability and subtlety in argument structure and language, posing a greater challenge for the models.

It is also interesting to note that for *EthiX*$_4$, performance differences among setups in the few-shot setting are minimal. This finding could suggest that, for datasets with fewer classes, simpler prompting methods that only provide a few examples might already suffice, with more elaborate prompts (such as few-shot definitions or few-shot CoT) not yielding substantial additional benefits.

5.4 Which argument schemes are more challenging?

Across all datasets, *argument from cause to effect* was the most misclassified scheme, often confused with *argument from consequences* or *argument from example*. This frequent misclassification may stem from the inherent nature of the scheme; its distinctive feature, causality, is also a common element in other schemes, leading to overlaps in classification.

The *argument from expert opinion* was the second most misclassified scheme in *EthiX*, mostly misclassified as *argument from example* or *argument from consequences* as well. This misclassification suggests a challenge with *source-based arguments*, which embed other arguments. For example, the argument “Masks reduce Covid transmission according to WHO” in Table 3 can be interpreted as ‘We should do something because it is good’, indirectly communicated by a source of expertise. Thus, the more complex, source-based *argument from expert opinion* (which presupposes the unexpressed premise that the fact that the claim is made by an expert implies its truth and acceptability) can be perceived to embed the format of another scheme, *argument from consequences*. Figure 1 illustrates the confusion matrices of Claude-3.5 in the few-shot descriptions setting, for *Ethix* and *EthiX*$_e$.

Lastly, in *NLAS*$_4$, *argument from example* and *argument from analogy* were often misclassified as one another, reflecting the fine line between generalising from a specific case and comparing case to case. This similarity highlights the subtle distinctions that pose challenges in classification tasks.

5.5 How do models perform on enthymemes?

The best performance on *EthiX*$_e$, which contains only enthymemes, is obtained with Claude-3.5 in the few-shot setting including the descriptions of the argument schemes in the prompt. Overall, we can observe the same pattern in model performance on *EthiX*$_e$ as on the entire version of *EthiX* in most experiments. For example, Claude-3.5 and GPT-4o show similar F_1 on *EthiX* and *EthiX*$_e$ in the standard zero-shot setting. In standard few-shot, Claude-3.5’s performance drops from 0.61 on *EthiX* to 0.57 on *EthiX*$_e$, while Mistral decreases from 0.35 to 0.29. These differences may indicate that the implicit nature of enthymemes introduces a small additional challenge for some models.

It is important to note that the enthymemes in *EthiX$_e$* primarily involve implicit conclusions, which can often be inferred or assumed using the debate topic. However, we did not instruct the models to reconstruct the enthymemes (only to take the debate topic into account), demonstrating how effectively LLMs leverage context independently. Regarding Gemini's improvement, a potential explanation could be the pre-training dataset used, which suggests that Gemini could be particularly adept at identifying or internally reconstructing implicit textual elements, potentially revealing an enhanced capability for context-driven reasoning.

5.6 Do LLMs perform better than pre-trained language models?

Compared to pre-trained language models (PLMs) (Bezou-Vrakatseli et al., 2024a), LLMs generally perform better. For example, BERT achieves F_1 of 0.53, with other PLMs not exceeding 0.55. In contrast, in the few-shot setting, nearly all larger LLMs achieve F_1 scores above 0.55. The best PLM reported on *EthiX*, ERNIE, yields F_1 of 0.63, while the top-performing LLM, Claude, surpasses this with an F_1 score of 0.65.

6 Conclusion

In this paper, we evaluated the ability of large language models to identify argument schemes. Our comprehensive evaluation covered zero-shot, few-shot, and chain-of-thought prompting methods across seven open-source and proprietary models on manually annotated human-made arguments, including enthymemes, as well as automatically generated arguments. Furthermore, we explored two approaches for enhancing the task instructions: normative information (i.e. the formal definitions of argument schemes as per Walton's taxonomy) and LLM-generated descriptions of argument schemes. Our analysis revealed that larger models can identify argument schemes satisfactorily in few-shot settings when given *descriptions* of argument schemes, in contrast to smaller models which struggle more. Given the inherent complexity of recognising argument schemes, a task that poses challenges even for human annotators (Walton and Macagno, 2015; Anthony and Kim, 2015; Lindahl et al., 2019; Bezou-Vrakatseli et al., 2024a), our findings are promising.

In future work, we plan to extend our analysis to other schemes and taxonomies for a more comprehensive evaluation. Furthermore, we plan to investigate argument generation using argument schemes and their associated critical questions as well as the potential of large language models for appropriate use of these schemes in dialogues.

Limitations

The presented work has certain limitations that future investigations could address. Specifically, our current work is focused exclusively on the English language. Although we have used widely recognised argument schemes from Walton's taxonomy to evaluate the classification capabilities of LLMs, argument schemes and their understanding by LLMs can vary significantly across different languages. Further research is needed to extend argument scheme classification beyond English.

Moreover, in our experiments, we only focus on Walton's taxonomy of schemes. Alternative taxonomies, such as the one proposed by Wagemans, might provide different insights into the capabilities of LLMs. Furthermore, our focus is restricted to specific argument schemes. While these schemes are the most prevalent in discourse, the exclusion of less common but equally significant schemes may skew our understanding of a model's overall argumentation proficiency. The schemes we included are primarily selected based on the available datasets, which represent schemes frequently encountered in argumentation literature and practice. Classifying argument schemes is a complex task, further compounded by the extensive list of argument schemes, which often leads to low annotator agreement (Hansen and Walton, 2013; Schneider et al., 2013; Lawrence and Reed, 2015). Although these schemes may appear clearly distinct at first glance, there remains meaningful overlap and potential ambiguity in distinguishing them (Bezou-Vrakatseli et al., 2024a). Understanding the ability of LLMs to distinguish amongst commonly encountered schemes serves as a valuable baseline for current applications and lays the groundwork for future research focused on distinguishing more closely related schemes. These limitations suggest a potential avenue for future research to explore other schemes and taxonomies.

Finally, whilst we explored different prompting strategies, our results show that LLMs are sensitive to the specifics of prompt design. This indicates that the performance variability could stem more from how the tasks are framed rather than from the models' capabilities, pointing towards the need for developing more robust evaluation methods that minimise prompt dependence.

Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [grant number EP/S023356/1].

References

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024. Medec: A benchmark for medical error detection and correction in clinical notes. *arXiv preprint arXiv:2412.19260*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. 2023. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6:1278796.

Robert Anthony and Mijung Kim. 2015. Challenges and remedies for identifying and classifying argumentation schemes. *Argumentation*, 29:81–113.

Anthropic. 2024. Claude 3.5 Sonnet.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2024a. EthiX: A Dataset for Argument Scheme Classification in Ethical Debates. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, volume 392, pages 3628–3635.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2024b. Towards Dialogues for Joint Human-AI Reasoning and Value Alignment. *arXiv preprint arXiv:2405.18073*.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *Computational Logic in Multi-Agent Systems*, pages 1–17.

Federico Castagna, Isabel Sassoon, and Simon Parsons. 2024. Critical-Questions-of-Thought: Steering LLM reasoning with Argumentative Querying. *arXiv preprint arXiv:2412.15177*.

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.

Adrian de Wynter and Tangming Yuan. 2024. "I'd Like to Have an Argument, Please": Argumentative Reasoning in Large Language Models. In *Computational Models of Argument*, pages 73–84. IOS Press.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *ACL*, pages 987–996.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can Large Language Models perform Relation-based Argument Mining? pages 8518–8534.

Nancy L Green. 2018. Towards mining scientific discourse using argumentation schemes. *Argument & Computation*, 9(2):121–135.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing Toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276.

Hans V Hansen and Douglas N Walton. 2013. Argument kinds and argument roles in the Ontario provincial election, 2011. *Journal of Argumentation in Context*, 2(2):226–258.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *TACL*, 9:721–739.

Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *LREC*, pages 1008–1018.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. Bayesian Argumentation-Scheme Networks: A Probabilistic Model of Argument Validity Facilitated by Argumentation Schemes. In *The 8th Workshop on Argument Mining*, pages 112–124.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia Potentia Est—On the Role of Knowledge in Computational Argumentation. *TACL*, 10:1392–1422.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *The 2nd Workshop on Argumentation Mining*, pages 127–136.

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

John Lawrence, Jacky Visser, and Chris Reed. 2019. An online annotation assistant for argument schemes. In *The 13th Linguistic Annotation Workshop*, pages 100–107.

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

Fabrizio Macagno and Aikaterini Konstantinidou. 2013. What students' arguments can tell us: Using argumentation schemes in science education. *Argumentation*, 27:225–243.

Fabrizio Macagno, Douglas Walton, and Chris Reed. 2017. Argumentation Schemes. History, Classifications, and Computational Applications. *IfCoLog Journal of Logics and Their Applications*, 8(4):2493–2556.

Tiezheng Mao, Jialing Fu, and Osamu Yoshie. 2024. Unearthing the Efficacy of ChatGPT in Argumentation Analysis: Performance, Potentials and Limitations . In *2024 4th Asia Conference on Information Engineering (ACIE)*, pages 157–162, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Sanjay Modgil. 2017. Dialogical Scaffolding for Human and Artificial Agent Reasoning. In *5th International Workshop on Artificial Intelligence and Cognition*, pages 58–71.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2016. Towards feasible guidelines for the annotation of argument schemes. In *The 3rd Workshop on Argument Mining*, pages 82–93.

OpenAI. 2024. Hello GPT-4o.

Fengjun Pan, Xiaobao Wu, Zongrui Li, and Luu Anh Tuan. 2024. Are llms good zero-shot fallacy classifiers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14338–14364.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *The 5th Workshop on Argument Mining*, pages 79–89.

Yupei Ren, Hongyi Wu, Zhaoguang Long, Shangqing Zhao, Xinyi Zhou, Zheqin Yin, Xinlin Zhuang, Xiaopeng Bai, and Man Lan. 2024. CEAMC: Corpus and Empirical Study of Argument Analysis in Education via LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6949–6966.

Ramon Ruiz-Dolz, Joaquín Taverner, John Lawrence, and Chris Reed. 2024. NLAS-multi: A Multilingual Corpus of Automatically Generated Natural Language Argumentation Schemes. *CoRR*, abs/2402.14458.

Ameer Saadat-Yazdi and Nadin Kökciyan. 2024. Beyond Recognising Entailment: Formalising Natural Language Inference from an Argumentative Perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9620–9636.

Sougata Saha and Rohini K Srihari. 2023. ArgU: A Controllable Factual Argument Generator. In *ACL*, pages 8373–8388.

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Computer Supported Cooperative Work (CSCW)*, pages 1069–1080.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *The 1st Workshop on Argumentation Mining*, pages 69–78.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Guilherme Trajano, Débora C Engelmann, Rafel H Bordini, Stefan Sarkadi, Jack Mumford, and Alison R Panisson. 2024. Translating natural language arguments to computational arguments using LLMs. In *Computational Models of Argument*, pages 289–300. IOS Press.

Michiel Van Der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Báez Santamaría. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103.

Frans Van Eemeren, Rob Grootendorst, and Frans H van Eemeren. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Preetika Verma, Kokil Jaidka, and Svetlana Churina. 2024. Auditing Counterfire: Evaluating Advanced Counterargument Generation with Evidence and Style. *arXiv preprint arXiv:2402.08498*.

Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. 2021. Annotating argument schemes. *Argumentation*, 35(1):101–139.

Jean Wagemans. 2016. Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: The 11th international conference of the Ontario Society for the Study of Argumentation*, pages 1–12.

Douglas Walton. 2009. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer.

Douglas Walton. 2012. Using argumentation schemes for argument extraction: A bottom-up method. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 6(3):33–61.

Douglas Walton and David M. Godden. 2005. The Nature and Status of Critical Questions in Argumentation Schemes.

Douglas Walton and Fabrizio Macagno. 2015. A classification system for argumentation schemes. *Argument & Computation*, 6(3):219–245.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. “According to . . . ”: Prompting Language Models Improves Quoting from Pre-Training Data. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2288–2301, St. Julian’s, Malta. Association for Computational Linguistics.

X.AI. 2024. Grok-2 Beta Release.

A Implementation Details

The publicly available information about the models’ parameters is the following: Llama-3.2 has 3B parameters and Mistral has 7B. The exact number of parameters of the bigger models has not been disclosed yet. Based on estimates (e.g. Abacha et al. (2024)), the larger models are calculated to have hundreds of billions of parameters.

All computations were performed on a Macbook Pro. Llama-3.2 and Mistral were run locally, using `ollama`. All other models were implemented without a dedicated GPU. GPT-4o, GPT-4, and Grok-2 were implemented using `openai`, Claude `anthropic`, and Gemini `google.generativeai`.

For Llama-3.2, we use temperature: 0, top p: 0.90, and top k: 50. For *zero-shot CoT* and *few-shot CoT*, we set max output tokens: 1000; for all other setups we use max output tokens: 50. For Mistral, GPT-4o, and GPT-4, we use temperature: 0, top p: 1.00, frequency penalty: 0, and presence penalty: 0. *zero-shot CoT* and *few-shot CoT*, we set max output tokens: 1000; for all other setups we use max output tokens: 50. For Gemini, we use temperature: 0, top p: 0.95, and top k: 40. *zero-shot CoT* and *few-shot CoT*, we set max output tokens: 1000; for all other setups we use max output tokens: 50. For Claude-3.5 and Grok-2, we use temperature: 0, top p: 1.00, and top k: 40. *zero-shot CoT* and *few-shot CoT*, we set max output tokens: 1000; for all other setups we use max output tokens: 50.

B Prompts

Prompt
Your task is to classify the Argument into one of the following argument schemes, given its Debate question, according to Walton's taxonomy. The input is given in the form: [Argument]: argument. [Debate]: debate question. You need to give an answer for the [Scheme] slot.
There are four schemes that you can choose to fill it and only these: argument from example, argument from cause to effect, argument from expert opinion, argument from analogy.
Below are the formal definitions of the schemes in the form [Definition]: Premise, Conclusion.
[Scheme]: argument from example. [Definition]: Premise: 'In this particular case, the individual a has property F and also property G.', Conclusion: 'Therefore, generally, if x has property F, then it also has property G.'
[Scheme]: argument from cause to effect. [Definition]: Major Premise: 'Generally, if A occurs, then B will (might) occur.', Minor Premise: 'In this case, A occurs (might occur).', Conclusion: 'Therefore, in this case, B will (might occur).'
[Scheme]: argument from expert opinion. [Definition]: Major Premise: 'Source E is an expert in subject domain S containing proposition A.', Minor Premise: 'E asserts that proposition A is true (false).', Conclusion: 'A is true (false).'
[Scheme]: argument from analogy. [Definition]: Similarity Premise: 'Generally, case C1 is similar to case C2.', Base Premise: 'A is true (false) in case C1.', Conclusion: 'A is true (false) in case C2.'
Below are the descriptions of the schemes:
[Scheme]: argument from example. [Description]: This scheme uses one or more specific cases or examples to support a general conclusion.
[Scheme]: argument from cause to effect. [Description]: This argument predicts that a specific cause will lead to a specific effect, using the causal connection to support the conclusion.
[Scheme]: argument from expert opinion. [Description]: This scheme supports a conclusion by citing an authority or expert in the field relevant to the argument.
[Scheme]: argument from analogy. [Description]: This type of argument claims that because two things are similar in certain respects, they are likely similar in other respects as well.
Some examples:
[Argument]: Some humans in non-developed areas rely on food in nature to survive, including meat from local animals. [Debate]: Is cannibalism ethically permissible?. [Scheme]: argument from example
[Argument]: Creating something in a controlled environment limits the potential for misuse, so developing AGI in a control environment will limit the potential of harm. [Debate]: Should an artificial general intelligence be created?. [Scheme]: argument from cause to effect
[Argument]: According to many philosophers, the fate of the universe tends towards Goodness. [Debate]: What's the right solution to the trolley problem?. [Scheme]: argument from expert opinion
[Argument]: The right to reproductive freedom already includes assisted reproductive technologies such as in-vitro fertilization. Cloning humans could be seen as another assisted reproductive technology. [Debate]: Is cloning humans ethical?. [Scheme]: argument from analogy
[Argument]: Exposure to online harassment - stressor to contend with for some resulting in some youths committing suicide -remember Ask.fm?. [Debate]: Has social media been good for humanity?. [Scheme]: argument from example
[Argument]: Nature causes irreparable damage to itself as well as constant harm, pain, and suffering to other living organisms without any help from humans. None of this can be avoided without humans. [Debate]: Would the world be a better place without humans?. [Scheme]: argument from cause to effect
[Argument]: This 2006 survey showed 49% of state prisoners, 40% of federal prisoners and 60% of all inmates in the USA have a symptom of a mental disorder. [Debate]: Should individuals sentenced to life in prison be allowed to choose death instead?. [Scheme]: argument from expert opinion
[Argument]: Airlines sometimes require 'fitness to fly' certificate from passengers flying with them. This could be extended to include proof of immunization from COVID-19. [Debate]: Should Covid-19 vaccines be mandatory?. [Scheme]: argument from analogy
[Argument]: $argument$. [Debate]: $debate\_question$. [Scheme]: Return one and only one of these 4 schemes, all in lowercase letters and written exactly as they are here. No other schemes. Do not include any additional text, repetition of the argument, or explanations!
Input
Argument: While certain attributes of the virus are still being investigated, the overwhelming consensus of scientists is that the wearing of masks will limit the spread of the virus. *Debate*: Do people have a right to not wear a mask in public spaces during the Covid-19 pandemic?
Expected Output
argument from expert opinion

Table 6: Experimental setups for *EthiX*$_4$. The *standard zero-shot* prompt is described in black text. For the *standard few-shot* setup, the red text is included. The *zero-shot definitions* setup includes the standard prompt as well as the definitions in blue, while the *few-shot definitions* setup includes the standard prompt, the definitions in blue, and the red text. Similarly, *zero-shot descriptions* includes the standard prompt as well as the descriptions in green, while *few-shot descriptions* includes the standard prompt, the descriptions in green, and the red text.

Prompt1
Your task is to classify the Argument into one of the following argument schemes, given its Debate question, according to Walton's taxonomy. Choose the scheme carefully and provide reasoning for your decision. The input is given in the form: [Argument]: argument. [Debate]: debate question. You need to give an answer for the [Scheme] slot. There are four schemes that you can choose to fill it and only these: argument from example, argument from cause to effect, argument from expert opinion, argument from analogy. Some examples: [Argument]: Some humans in non-developed areas rely on food in nature to survive, including meat from local animals. [Debate]: Is cannibalism ethically permissible?. [Reasoning]: This argument is best classified as an "argument from example" because it uses the specific instances of people in non-developed areas relying on naturally available meat for survival to support the broader claim about the permissibility of cannibalism in extreme survival situations. By stating this example, the arguer aims to provide a concrete instance that illustrates when eating meat for survival is necessary and extrapolates this necessity to justify the possibility of cannibalism under dire conditions. [Scheme]: argument from example [Argument]: Creating something in a controlled environment limits the potential for misuse, so developing AGI in a control environment will limit the potential of harm. [Debate]: Should an artificial general intelligence be created?. [Reasoning]: This is an argument from cause to effect because it establishes a causal link between controlled development (cause) and limited potential for harm (effect). The arguer is reasoning that because controlled environments generally cause limited misuse (causal principle), therefore controlling AGI development will cause limited harm (specific application). The argument's core structure relies on this causal relationship rather than examples, values, expert opinions, alternatives, or analogies. [Scheme]: argument from cause to effect [Argument]: According to many philosophers, the fate of the universe tends towards Goodness. [Reasoning]: This argument clearly relies on the authority of philosophers as experts in moral and philosophical matters. The structure follows the classic expert opinion pattern where the argument's strength comes from citing what experts (in this case, "many philosophers") believe about the nature of the universe and goodness. The argument doesn't provide examples, draw analogies, or discuss consequences - instead, it bases its persuasive power on the expertise and credibility of philosophers as authorities on moral and metaphysical questions. [Debate]: What's the right solution to the trolley problem?. [Scheme]: argument from expert opinion [Argument]: The right to reproductive freedom already includes assisted reproductive technologies such as in-vitro fertilization. Cloning humans could be seen as another assisted reproductive technology. [Debate]: Is cloning humans ethical?. [Reasoning]: This is clearly an argument from analogy because: 1. It compares human cloning to an already accepted practice (IVF); 2. It suggests that because both share the characteristic of being "assisted reproductive technologies," they should be treated similarly; 3. The argument tries to transfer the acceptance of one practice (IVF) to another (cloning) based on their shared properties; 4. The structure follows the classic form of analogical reasoning: "A is like B in relevant ways, B is accepted, therefore A should be accepted too". [Scheme]: argument from analogy [Argument]: Exposure to online harassment - stressor to contend with for some resulting in some youths committing suicide -remember Ask.fm?. [Reasoning]: This is best classified as an argument from example because it uses the specific historical case of Ask.fm and its associated suicides as an example to support the broader claim about social media's negative impact. The arguer is saying "Remember Ask.fm?" as a way to provide a concrete instance of how social media harassment has led to tragic outcomes. The argument relies on this specific example to make its broader point about social media's harmful effects on humanity. [Debate]: Has social media been good for humanity?. [Scheme]: argument from example [Argument]: Nature causes irreparable damage to itself as well as constant harm, pain, and suffering to other living organisms without any help from humans. None of this can be avoided without humans. [Debate]: Would the world be a better place without humans?. [Reasoning]: This argument is best classified as an "argument from cause to effect" because it identifies a cause (natural processes and phenomena) and directly links it to specific effects (damage, harm, pain, and suffering to itself and other living organisms). [Scheme]: argument from cause to effect [Argument]: This 2006 survey showed 49% of state prisoners, 40% of federal prisoners and 60% of all inmates in the USA have a symptom of a mental disorder. [Debate]: Should individuals sentenced to life in prison be allowed to choose death instead?. [Reasoning]: This argument is classified as an "argument from expert opinion" because it relies on the findings of a survey—presumably conducted by researchers or specialists in the field—as the basis for asserting a factual claim about the prevalence of mental health issues among prisoners. By referencing statistical data from a survey conducted in 2006, the arguer leverages the authority and credibility of the survey's authors (experts in criminology, psychology, or a related field) to substantiate the claim. [Scheme]: argument from expert opinion [Argument]: Airlines sometimes require 'fitness to fly' certificate from passengers flying with them. This could be extended to include proof of immunization from COVID-19. [Debate]: Should Covid-19 vaccines be mandatory?. [Reasoning]: This is an argument from analogy because it: 1. Draws a parallel between two similar situations (fitness certificates and vaccination proof); 2. Uses an existing practice to justify a similar proposed practice; 3.Suggests that since one health requirement is acceptable, another similar health requirement would also be acceptable; 4. Relies on the similarity between two cases to transfer the acceptability from one to the other. The argument doesn't rely on consequences, expert opinion, values, examples, or cause-effect relationships. Instead, it fundamentally works by drawing an analogy between existing airline health requirements and proposed vaccination requirements. [Scheme]: argument from analogy [Argument]: $argument$. [Debate]: $debate\_question$. [Scheme]: Return only 1 of these 4 schemes. [Reasoning]: Give the reason you chose the scheme you chose. Let's think step by step!
Prompt 2
Your task is to classify the Argument into one of the following argument schemes, given its Debate question, according to Walton's taxonomy. You are provided with the Reasoning for the classification. The input is given in the form: [Argument]: argument. [Debate]: debate question. [Reasoning]: reasoning. You need to give an answer for the [Scheme] slot. There are four schemes that you can choose to fill it and only these: argument from example, argument from cause to effect, argument from expert opinion, argument from analogy. [Argument]: $argument$. [Debate]: $debate\_question$. [Reasoning]: $reasoning$. [Scheme]: Return only 1 of these 4 schemes, all in lowercase letters and written exactly as they are here. No other schemes. Do not include any additional text, repetition of the argument, or explanations!

Table 7: Chain-of-thought setups for *EthiX$_4$*. CoT is a 2-stage setup, we prompt with Prompt1 and give the answer to the model via Prompt2. The *zero-shot CoT* prompt is described in black text. For the *few-shot CoT* setup, the red text is included.

Prompt

Your task is to classify the Argument into one of the following argument schemes, given its Debate topic, according to Walton’s taxonomy. The input is given in the form: [Argument]: argument. [Topic]: debate topic. You need to give an answer for the [Scheme] slot.
There are four schemes that you can choose to fill it and only these: argument from example, argument from cause to effect, argument from expert opinion, argument from analogy.
Below are the formal definitions of the schemes in the form [Definition]: Premise, Conclusion.
[Scheme]: argument from example. [Definition]: Premise: ‘In this particular case, the individual a has property F and also property G.’, Conclusion: ‘Therefore, generally, if x has property F, then it also has property G.’
[Scheme]: argument from cause to effect. [Definition]: Major Premise: ‘Generally, if A occurs, then B will (might) occur.’, Minor Premise: ‘In this case, A occurs (might occur).’, Conclusion: ‘Therefore, in this case, B will (might occur).’
[Scheme]: argument from expert opinion. [Definition]: Major Premise: ‘Source E is an expert in subject domain S containing proposition A.’, Minor Premise: ‘E asserts that proposition A is true (false).’, Conclusion: ‘A is true (false).’
[Scheme]: argument from analogy. [Definition]: Similarity Premise: ‘Generally, case C1 is similar to case C2.’, Base Premise: ‘A is true (false) in case C1.’, Conclusion: ‘A is true (false) in case C2.’
Below are the descriptions of the schemes:
[Scheme]: argument from example. [Description]: This scheme uses one or more specific cases or examples to support a general conclusion.
[Scheme]: argument from cause to effect. [Description]: This argument predicts that a specific cause will lead to a specific effect, using the causal connection to support the conclusion.
[Scheme]: argument from expert opinion. [Description]: This scheme supports a conclusion by citing an authority or expert in the field relevant to the argument.
[Scheme]: argument from analogy. [Description]: This type of argument claims that because two things are similar in certain respects, they are likely similar in other respects as well.
Some examples:
[Argument]: "premise": "In this particular case, John has terminal cancer and also the right to die with dignity through euthanasia", "conclusion": "Therefore, generally, if any individual has terminal cancer, then they also have the right to die with dignity through euthanasia". [Topic]: Euthanasia. [Scheme]: argument from example
[Argument]: "major premise": "Generally, if a person has a favorable physical appearance, then they will be perceived more positively by others.", "minor premise": "In this case, the person has made efforts to improve their physical appearance.", "conclusion": "Therefore, in this case, the person is likely to be perceived more positively by others, which can potentially lead to personal success.". [Topic]: Physical appearance for personal success. [Scheme]: argument from cause to effect
[Argument]:"major premise": "Dr. Nora Volkow is an expert in neuroscience and drug abuse research", "minor premise": "Dr. Volkow asserts that legalizing cannabis can lead to negative public health outcomes, including addiction, cognitive impairment, and mental health issues", "conclusion": "Legalizing cannabis is not advisable as it can have negative public health outcomes". [Topic]: Legalisation of cannabis. [Scheme]: argument from expert opinion
[Argument]: "similarity premise": "Generally, killing another human being is considered morally wrong", "base premise": "Abortion involves the deliberate termination of a developing human being", "conclusion": "Abortion should be considered morally wrong". [Topic]: Abortion. [Scheme]: argument from analogy
[Argument]: "premise": "In this particular case, individuals who are vaccinated against COVID-19 have a significantly lower risk of contracting and spreading the virus than those who are unvaccinated.", "conclusion": "Therefore, generally, if individuals in a population are vaccinated against COVID-19, the overall risk of community transmission will be reduced and the population will be better protected from the virus.”. [Topic]: Mandatory vaccination in pandemic. [Scheme]: argument from example
[Argument]: "major premise": "Generally, if a person commits a heinous crime, then they deserve a severe punishment.", "minor premise": "In this case, a person has committed a heinous crime.", "conclusion": "Therefore, in this case, the person deserves a severe punishment, which could include capital punishment.". [Topic]: Capital punishment. [Scheme]: argument from cause to effect
[Argument]: "major premise": "Legal scholar X is an expert in constitutional law", "minor premise": "Legal scholar X asserts that freedom of speech should be limited in cases where it leads to hate speech and incitement to violence", "conclusion": "Limiting freedom of speech in cases of hate speech and incitement to violence is necessary to protect individuals from harm and maintain a civil society". [Topic]: Freedom of speech. [Scheme]: argument from expert opinion
[Argument]: "similarity premise": "Generally, creating identical copies of an object or entity is similar to creating identical copies of an animal or human being through cloning technology", "base premise": "Cloning technology has been successfully used to create identical copies of animals, such as Dolly the sheep, and has the potential to create identical copies of humans", "conclusion": "Therefore, if creating identical copies of objects or entities is accepted practice (such as in the production of computer chips or mass-produced products), then cloning animals and humans should also be accepted and permissible.". [Topic]: Animal/human cloning. [Scheme]: argument from analogy
[Argument]: $argument$. [Topic]: $debate\_topic$. [Scheme]: Return one and only one of these 4 schemes, all in lowercase letters and written exactly as they are here. No other schemes. Do not include any additional text, repetition of the argument, or explanations!

Input

Argument: While certain attributes of the virus are still being investigated, the overwhelming consensus of scientists is that the wearing of masks will limit the spread of the virus.
Debate: Do people have a right to not wear a mask in public spaces during the Covid-19 pandemic?

Expected Output

argument from expert opinion

Table 8: Experimental setups for *$NLAS_4$*. The *standard zero-shot* prompt is described in black text. For the *standard few-shot* setup, the red text is included. The *zero-shot definitions* setup includes the standard prompt as well as the definitions in blue, while the *few-shot definitions* setup includes the standard prompt, the definitions in blue, and the red text. Similarly, *zero-shot descriptions* includes the standard prompt as well as the descriptions in green, while *few-shot descriptions* includes the standard prompt, the descriptions in green, and the red text.

Prompt1

Your task is to classify the Argument into one of the following argument schemes, given its Debate topic, according to Walton's taxonomy. Choose the scheme carefully and provide reasoning for your decision.

The input is given in the form: [Argument]: argument. [Topic]: debate topic. You need to give an answer for the [Scheme] slot. There are four schemes that you can choose to fill it and only these: argument from example, argument from cause to effect, argument from expert opinion, argument from analogy.

Some examples:

[Argument]: "premise": "In this particular case, John has terminal cancer and also the right to die with dignity through euthanasia", "conclusion": "Therefore, generally, if any individual has terminal cancer, then they also have the right to die with dignity through euthanasia". [Topic]: Euthanasia. [Scheme]: argument from example. [Reasoning]: This argument takes a specific instance (John's case) and uses it to make a generalization about all similar cases. It moves from a particular case to a universal claim, which is the defining characteristic of argument from example. The arguer uses John's specific situation with terminal cancer to justify a broader conclusion about all individuals with terminal cancer having the same right.

[Argument]: "major premise": "Generally, if a person has a favorable physical appearance, then they will be perceived more positively by others.", "minor premise": "In this case, the person has made efforts to improve their physical appearance.", "conclusion": "Therefore, in this case, the person is likely to be perceived more positively by others, which can potentially lead to personal success.". [Topic]: Physical appearance for personal success. [Scheme]: argument from cause to effect. [Reasoning]: This argument follows a causal chain where physical appearance (cause) leads to positive perception (effect), which in turn leads to personal success. The major premise establishes a general causal relationship between appearance and perception, while the minor premise and conclusion follow this causal chain to its outcome. The argument structure clearly shows how one event (improving appearance) is expected to lead to another (positive perception) and ultimately to a final effect (personal success).

[Argument]: "major premise": "Dr. Nora Volkow is an expert in neuroscience and drug abuse research", "minor premise": "Dr. Volkow asserts that legalizing cannabis can lead to negative public health outcomes, including addiction, cognitive impairment, and mental health issues", "conclusion": "Legalizing cannabis is not advisable as it can have negative public health outcomes". [Topic]: Legalisation of cannabis. [Scheme]: argument from expert opinion. [Reasoning]: This argument relies on the authority and expertise of Dr. Nora Volkow in neuroscience and drug abuse research. The structure follows the classic expert opinion pattern where the expert's credentials are established (major premise), their statement on the topic is presented (minor premise), and the conclusion is drawn based on accepting their expert testimony. The argument's strength comes from Dr. Volkow's recognized expertise in the relevant field of study.

[Argument]: "similarity premise": "Generally, killing another human being is considered morally wrong", "base premise": "Abortion involves the deliberate termination of a developing human being", "conclusion": "Abortion should be considered morally wrong". [Topic]: Abortion. [Scheme]: argument from analogy. [Reasoning]: This argument draws a parallel between two situations: general killing of humans and abortion. It establishes a similarity between these cases (both involve ending human life) and transfers the moral evaluation from one case to the other. The argument relies on the comparison of properties between these two situations to draw its conclusion, which is characteristic of analogical reasoning.

[Argument]: "premise": "In this particular case, individuals who are vaccinated against COVID-19 have a significantly lower risk of contracting and spreading the virus than those who are unvaccinated.", "conclusion": "Therefore, generally, if individuals in a population are vaccinated against COVID-19, the overall risk of community transmission will be reduced and the population will be better protected from the virus.". [Topic]: Mandatory vaccination in pandemic. [Scheme]: argument from example. [Reasoning]: This argument uses a specific observed instance (particular case of vaccinated individuals showing lower transmission rates) to draw a broader generalization about the effects of vaccination on population-level transmission. It moves from a specific example or case to a general conclusion, which is characteristic of argument from example.

[Argument]: "major premise": "Generally, if a person commits a heinous crime, then they deserve a severe punishment.", "minor premise": "In this case, a person has committed a heinous crime.", "conclusion": "Therefore, in this case, the person deserves a severe punishment, which could include capital punishment.". [Topic]: Capital punishment. [Scheme]: argument from cause to effect. [Reasoning]: This argument follows a causal pattern where the act of committing a heinous crime (cause) leads to deserving severe punishment (effect). The argument establishes a general causal relationship in the major premise and applies it to a specific case to reach the conclusion about deserved punishment. It's not based on examples, expert opinions, or analogies, but rather on a direct causal relationship between actions and their consequences.

[Argument]: "major premise": "Legal scholar X is an expert in constitutional law", "minor premise": "Legal scholar X asserts that freedom of speech should be limited in cases where it leads to hate speech and incitement to violence", "conclusion": "Limiting freedom of speech in cases of hate speech and incitement to violence is necessary to protect individuals from harm and maintain a civil society". [Topic]: Freedom of speech. [Scheme]: argument from expert opinion. [Reasoning]: This argument relies on the authority and expertise of a legal scholar in constitutional law to support the conclusion about limiting freedom of speech. The structure follows the classic expert opinion pattern where the expert's credentials are established (major premise), followed by their assertion (minor premise), which is then used to support the conclusion. The argument's strength comes from the expert's credibility in the relevant field rather than from examples, causal relationships, or analogies.

[Argument]: "similarity premise": "Generally, creating identical copies of an object or entity is similar to creating identical copies of an animal or human being through cloning technology", "base premise": "Cloning technology has been successfully used to create identical copies of animals, such as Dolly the sheep, and has the potential to create identical copies of humans", "conclusion": "Therefore, if creating identical copies of objects or entities is accepted practice (such as in the production of computer chips or mass-produced products), then cloning animals and humans should also be accepted and permissible.". [Topic]: Animal/human cloning. [Scheme]: argument from analogy. [Reasoning]: This argument draws a parallel between two situations: creating copies of objects and creating clones of living beings. It attempts to transfer the acceptability of one practice (copying objects) to another practice (cloning animals/humans) based on their shared characteristic of creating identical copies. The argument relies on the similarity between these two situations to justify its conclusion, which is a classic structure of analogical reasoning.

[Argument]: $argument$. [Debate]: $debate\_topic$. [Scheme]: Return only 1 of these 4 schemes. [Reasoning]: Give the reason you chose the scheme you chose. Let's think step by step!

Prompt 2

Your task is to classify the Argument into one of the following argument schemes, given its Debate topic, according to Walton's taxonomy. You are provided with the Reasoning for the classification.

The input is given in the form: [Argument]: argument. [Topic]: debate topic. [Reasoning]: reasoning. You need to give an answer for the [Scheme] slot. There are four schemes that you can choose to fill it and only these: argument from example, argument from cause to effect, argument from expert opinion, argument from analogy.

[Argument]: $argument$. [Topic]: $debate\_topic$. [Reasoning]: $reasoning$. [Scheme]: Return only 1 of these 4 schemes, all in lowercase letters and written exactly as they are here. No other schemes. Do not include any additional text, repetition of the argument, or explanations!

Table 9: Chain-of-thought setups for *NLAS*$_4$. CoT is a 2-stage setup, we prompt with Prompt1 and give the answer to the model via Prompt2. The *zero-shot CoT* prompt is described in black text. For the *few-shot CoT* setup, the red text is included.