Harnessing Toulmin's theory for zero-shot argument explication

Ankita Gupta Ethan Zuckerman Brendan O'Connor
University of Massachusetts Amherst
{ankitagupta,brenocon}@cs.umass.edu, ethanz@umass.edu

Abstract

To better analyze informal arguments on public forums, we propose the task of *argument explication*, which makes explicit an argument's structure and implicit reasoning by outputting triples of propositions ⟨*claim*, *reason*, *warrant*⟩. The three slots, or argument components, are derived from the widely known Toulmin (1958) model of argumentation. While prior research applies Toulmin or related theories to annotate datasets and train supervised models, we develop an effective method to prompt generative large language models (LMs) to output explicitly named argument components proposed by Toulmin. In particular, we prompt a language model with the theory name (e.g., '`According to Toulmin model`'). We evaluate the outputs' validity through a human study and automatic evaluation based on prior argumentation datasets, and perform robustness checks over alternative LMs, prompts, and argumentation theories. Finally, we conduct a proof-of-concept case study to extract an interpretable argumentation (hyper)graph from a large corpus of critical public comments on whether to allow the COVID-19 vaccine for children, suggesting future directions for corpus analysis and argument visualization.[1]

1 Introduction

Advances in computational methods for analyzing arguments have benefited various applications spanning debating technologies (Aharoni et al., 2014; Rinott et al., 2015), policymaking (Sardianos et al., 2015), information retrieval (Carstens and Toni, 2015), essay writing support (Stab and Gurevych, 2017) and legal decision making (Palau and Moens, 2009). However, unlike these domains with well-written arguments, web discourse on social media and public forums may feature arguments from inexperienced writers, often consisting of unclear argumentative structures and reasoning, making argument analysis quite challenging. Manual interpretation of such arguments is especially problematic in eRulemaking, where government officials are required to make sense of large amounts of public feedback (Lawrence et al., 2017).

To help automate the analysis of such informal arguments, we propose the task of *argument explication*, which involves making the *structure* and *implicit reasoning* of an argument explicit. In particular, we decompose a natural language argument into its claim and reasons. We further elucidate its reasoning by explicitly stating an implicit warrant that logically links a reason to the claim.

Argument explication can be useful for many applications. For instance, as shown in Figure 1, it can help lay out the reasoning involved in public comments, enabling quick comprehension of arguments being made. It could help identify fallacious arguments by clearly laying out an argument's logical structure (Deshpande et al., 2023), or aid theme discovery by improving text representation with implicit content (Viswanathan et al., 2023; Hoyle et al., 2023). It can also assist other NLP tasks (e.g., question-answering), where the explicated output could serve as intermediate reasoning, a method that has been demonstrated to improve downstream LM performance (Wei et al., 2022).

Traditionally, several argumentation theories (e.g., Toulmin, 1958; Freeman, 1991; Walton, 1996) have been proposed to analyze arguments, guiding the development of training datasets and supervised models trained on them (Habernal and Gurevych, 2017; Stab and Gurevych, 2017; Skeppstedt et al., 2018). Recent advances in NLP, driven by large language models (Brown et al., 2020; Touvron et al., 2023), have led to a new modeling approach using specific keywords or phrases as prompts to guide model responses (Wei et al., 2022; Kojima et al., 2022), with little or no training data. This approach is especially promising for argument

[1] All resources accompanying this project are added to: https://github.com/slanglab/argument_explication

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10259–10276
August 11-16, 2024

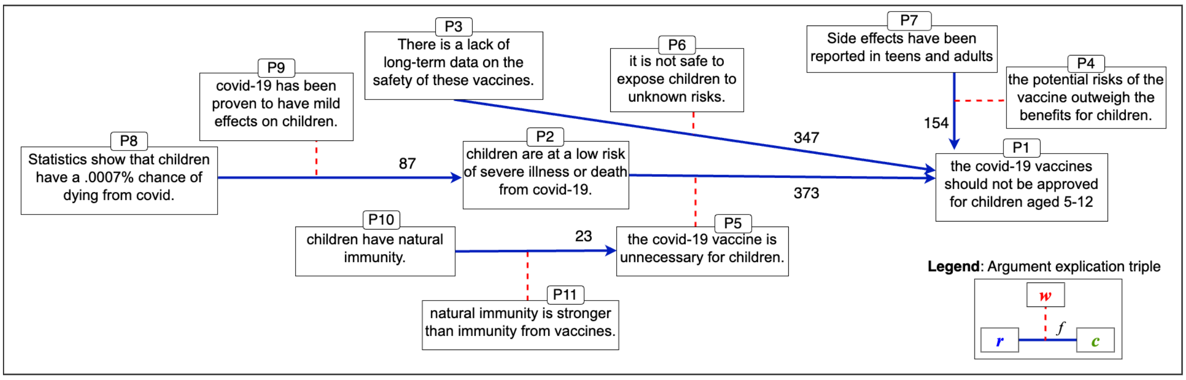

Figure 1: A portion of the corpus-level argument hypergraph[2] we automatically extract from `regulations.gov` public comments on whether to approve a COVID-19 vaccine for children. Each node is a cluster of propositions extracted from comments. An argument is a triple of nodes, $\langle$*(c)*laim, *(r)*eason, *(w)*arrant$\rangle$, visualized as *solid blue* and *dotted red* arrows connecting the explicit and implicit supporting propositions (r, w) to the claim (c). f is the triple's corpus frequency. Further details in §6.

analysis, traditionally dependent on bespoke and smaller datasets (Morio et al., 2022) compared to other NLP tasks, as it could enable the analysis of unstructured argumentative texts without requiring extensive domain-specific annotations. However, the pathway from explicit argumentation theories to prompting-based model design in the era of LMs is less well-defined.

In this work, we harness Toulmin's model of argumentation for zero-shot argument explication. Toulmin's theory proposes a schema to analyze arguments and has been commonly used to annotate real-world arguments (Habernal and Gurevych, 2017), suggesting its practical utility. This theory also has a substantial scholarly impact; for example, Google Scholar citations for Toulmin's theory (21,177) are close to Chomsky, 1957 (31,647). Beyond academic communities, this theory is also widely popular for its pedagogical use (Ellis, 2015). For instance, in a random sample of 100 documents from C4[3] (Dodge et al., 2021) mentioning *Toulmin*, we find that 21% contain worked-out examples of Toulmin-style argument breakdown, potentially serving as supervised training data in LMs' pretraining corpora. Motivated by these observations, we investigate the use of Toulmin's theory for the zero-shot argument explication task.

Our major contributions include:

- We propose the argument explication task and provide a two-stage framework to explicate arguments: identifying the claim and reasons, and then generating a warrant for each claim-reason pair. For each stage, we prompt an LM with references to Toulmin's theory (e.g., `'According to Toulmin model,'`) which elicits a theory-compliant response with correct mentions of Toulmin's terminology (§5.3) and generates reasonable values for each of these terms (§5.4).
- We further validate our results via prompt sensitivity analysis (§5.5) and comparison with other argumentation theories (§5.6). Our analysis shows that prompting with references to Toulmin's theory consistently yields better performance than other theories.
- Finally, to illustrate the usefulness of our approach and argument explication task more broadly, we apply it to a corpus of public comments related to COVID-19 vaccine approval for children (§6), visualizing them as a corpus-level argument hypergraph (Figure 1), which could be useful in drawing insights and help inform civic decision-making.

2 Related Work

Our work is related to several areas:

Argument mining involves claim-reason identification from an input argument and thus focuses on analyzing explicit content (Stab and Gurevych, 2014, 2017; Bentahar et al., 2010), while our task requires generating implicit information as well.

Argument reasoning only focuses on generating implicit information, while assuming a prior knowledge of claim and reason (Habernal and Gurevych, 2017; Becker et al., 2020b; Chakrabarty et al., 2021; Boltužić and Šnajder, 2016), which

[2]Unlike a graph, a hypergraph edge—here, an argument triple $\langle c, r, w\rangle$—can connect more than two nodes.

[3]A corpus often used to pre-train LMs (Raffel et al., 2020).

is not available when analyzing real-world arguments. In contrast, our task requires identifying claim-reason pairs before generating implicit information. Hulpus et al. (2019) investigate the end-to-end task of identifying the structure and reasoning of an argument, however only theoretically. Becker et al. (2020a) address some relevant subtasks proposed by Hulpus et al. (2019), however, they also assume pre-identified claim-reason pairs.

Argument synthesis involves generating an argument from scratch (El Baff et al., 2019; Wachsmuth et al., 2018; Gretz et al., 2020), while our task involves generating output conditioned on an input argument.

Argument mapping: While prior work has also explored visualizing arguments as maps, they have mainly focused on visualizing individual arguments (Reed, 2001) or supporting online collaborative tools, where members of a community work together to manually build an argument map (Klein, 2012). In contrast, we aim to automate the construction of a corpus-level argument hypergraph by analyzing arguments within an existing corpus.

LMs for computational argumentation have just started being explored. Chen et al. (2023) treat argument mining as a classification task, and do not consider generating implicit information. Rocha et al. (2023) consider augmenting an argument with implicit information using LMs, though only focus on explicating discourse markers.

LMs for reasoning: Several prompting frameworks, such as chain of thought (CoT) prompting (Kojima et al., 2022; Wei et al., 2022) and ReAct prompting (Yao et al., 2023), have been proposed to elicit multi-step reasoning chains from LMs by using specific phrases like 'Let's think step by step.' Our work is analogous in that we also propose prompts (e.g., 'According to Toulmin model') that elicit the reasoning involved in an argument. However, one significant difference is that the generated Toulmin breakdown of an argument has a fixed number of components, in contrast to CoT or ReAct that generate multi-step reasoning chains of varying lengths and forms.

3 Argument Explication Task

Explicating an argument involves a) identifying its *structure*: determining its claim and reasons, and b) explaining its *reasoning*: making explicit any implicit information connecting the reason to the claim. Following several argumentation theories,

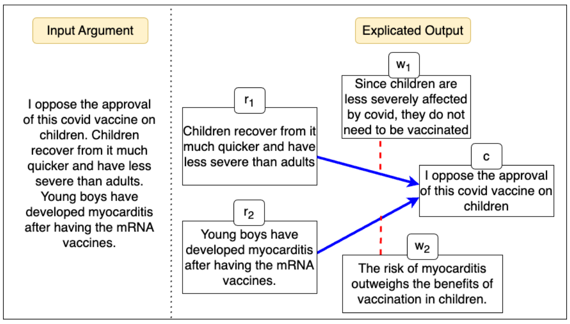

Figure 2: Illustrative example of an input argument decomposed into two explication triples of a claim (c), explicit reasons (r_i) and implicit warrants (w_i).

we propose decomposing an argument into three core components:

The ***claim*** (c) is a normative assertion or point of view put forward by the author for general acceptance. It is also known as *conclusion* (Toulmin, 1958; Walton, 1996; Freeman, 1991).

A ***reason*** (r_i) is a proposition provided by the author to convince the audience why they should accept the claim. Toulmin (1958) refers to it as *data*, and later *grounds* in Toulmin et al. (1984); others use the term *premise* (Walton, 1996; Freeman, 1991). As explained by Toulmin, '*the data represent what we have to go on.*'

The ***warrant*** (w_i) provides a logical link between reason and claim, encoding the author's current presupposed world knowledge that explains why claim (c) follows from the provided reason (r_i). A warrant is a missing piece of information that is taken for granted by the author and is assumed common knowledge, yet if it fails to hold, c cannot be inferred from the r_i. As per Toulmin (1958): '*data are appealed to explicitly, warrants implicitly.*' It is also similar to Walton (1996)'s *major premise*.

The above three core components are conceptually supported across multiple theories, including Toulmin (1958), Freeman (1991), and Walton et al. (2008), though use different terminology.[4] We tend to use terminology closer to Toulmin's version.

We consider singled-authored arguments proposing a single claim, in line with public comments, where the majority of them express support or objection to a specific policy issue. The author may give one or more reasons to support this claim. For

[4]The cited theories also propose additional components beyond the core components. We briefly review each theory in Appendix A.

every claim-reason pair, there exists a corresponding warrant. In cases where a reason sufficiently supports the claim, a trivial warrant of the form 'if reason then claim' may suffice.

Formally, the task input is a textual argument T, and its output is a collection of explication triples, $E = \{\langle c, r_i, w_i \rangle\}$ $\forall$ i=1 to N, with the same claim appearing in all triples. Figure 2 illustrates two connected triples for an input argument.

4 Method: A two-stage framework

We explicate an argument in two stages. In stage 1, we generate the claim (c) and all the reasons (r_i) from the input argument. In stage 2, we generate the warrant (w_i) between each claim-reason pair identified in stage 1.[5] Stage 2 facilitates the generation of warrants for every claim-reason pair, enabling us to fully explicate the argument. In both stages, we prompt LMs in a zero-shot setting, prompting with references to Toulmin's theory, as elaborated next.

4.1 Prompting with references to Toulmin's theory for zero-shot argument explication

Background on Toulmin's theory: Toulmin's model breaks down an argument into three core components—*claim*, *data/grounds*, and *warrant*—which directly map to the components of the argument explication task. It also has three optional components including *backing* (additional support for warrant), *rebuttal* (a view-point opposing the claim), and *qualifier* (the degree of certainty).

Prompt details: To steer an LM's response as per Toulmin's theory, we utilize the `'According to'` prompt, which has also been shown to improve an LM's ability to ground the information in an external knowledge source (Weller et al., 2023). More specifically, we empirically observe that prompting LMs with `'According to Toulmin model'` ('Toulmin prompt' for brevity) elicits a response that correctly mentions *terms* from Toulmin's theory (e.g., claim, grounds) and generates plausible *values* (propositions) for each term (Figure 3).

Obtaining explication triples from LM's response: We use the Toulmin prompt in both stages of argument explication. In stage 1, we provide a natural language argument (T) as the input. To obtain the argument's claim (c) and reasons (r_i $\forall i$), we extract the values corresponding to the term *claim* and *grounds* (or *data*), respectively from the LM response.[6] For each r_i, we construct a new argument of the form '$\{r_i\}$. Therefore, $\{c\}$', which we use as input argument in stage 2. Finally, we obtain an explication triple, $\langle c, r_i, w_i \rangle$, by extracting the values corresponding to the terms *claim*, *grounds* (or *data*), and *warrant* from the LM's response obtained in stage 2.[7]

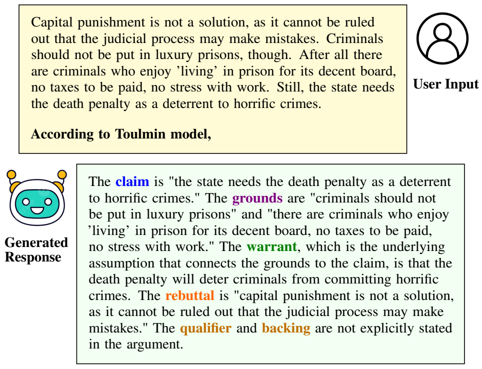

Figure 3: An input argument from MCT and an example response obtained by prompting `GPT-4` with the `'According to Toulmin model'`. The response correctly mentions *terms* from Toulmin's theory and generates plausible *values* for each of these terms.

5 Results

5.1 Experimental Details

Evaluation Datasets: We recast the following two datasets to evaluate our method.

ARCT (Habernal and Gurevych, 2017): has 445 claim-reason pairs (test split) sourced from news comments. Each pair has a correct and an incorrect warrant and the goal is to choose the correct one. For our task, we use concatenated claim-reason as the input argument and claim, reason, and correct warrant as the gold explication triple.

Microtext Corpus (MCT; Peldszus and Stede, 2015): has 112 paragraph-length arguments, each annotated with a claim and multiple reasons,

[5] We also considered providing the input argument as the context in stage 2 but found no difference in the generated warrant with or without context. Thus, we omitted context in stage 2 to reduce prompt tokens and hence cost.

[6] We post-process the LM's response into a Python dictionary (with keys as terms and values as the propositions), avoiding complex regex-based information extraction from LM's original response. We use a simple LLM-based post-processor, prompting GPT3.5 with `'Format the above text in a Python dictionary with values as a list of bullet points.'` We manually validated that in over 95% of cases, this step does not introduce errors.

[7] While the claim and reason generated in stage 1 can also be used in final explication triples, empirically we find that they are highly similar to those generated in stage 2.

based on Freeman (1991)'s theory. *MCT* was later augmented with human-written warrants (Becker et al., 2020b), allowing us to examine the model's ability to generate warrants. Even though this dataset is small in size, it has several advantages: 1) *More complex evaluation dataset*: contains multiple reasons and has argumentative relations between two non-adjacent text spans mimicking real-life arguments. 2) *Not affected by data leakage issues* (Dodge et al., 2021): While the input argument may be present in the pre-training data, the explicit Toulmin-style annotations aren't publicly available online, instead, we interpret them from the original Freeman-style annotations. The original annotations are also in XML, instead of text-to-text format typically used for pre-training LMs.

Language Models: We evaluate LMs of different sizes (40B-175B parameters), including proprietary OpenAI models and open-weight models with publicly available weights. We experiment with all the models in a zero-shot setting, without any fine-tuning. Among OpenAI models, we consider `GPT-3` (`text-davinci-003`; Brown et al., 2020), `GPT-4` (`gpt4-0613`; OpenAI, 2023). Among open-weight models, we consider `Llama-2-70B` (Touvron et al., 2023) and `Falcon-40B` (Almazrouei et al., 2023) models.[8] Following Kojima et al. (2022); Wei et al. (2022); Wang et al. (2022), we use greedy decoding for all the models.[9]

5.2 Prompting without referring to the Toulmin's theory

Our approach (§4.1) assumes a specific schema to analyze an argument and uses a reference to Toulmin's theory (as prompt) to obtain the argument components. We investigate two baselines without making these assumptions.

Baseline 1 (Without assuming task definition): In this baseline, we abandon both assumptions and use generic prompts to explain an argument without being guided by a specific task definition. We prompt `GPT-4` with the three prompts on arguments from MCT: a) '`Explain the logical steps in this argument.`' b) '`Explain this argument in a systematic way.`' and c) '`Explain this argument in an academic way.`' A useful response should analyze and explain argument components, perhaps using any terminology. Only 38.39%, 24.10%, and 21.42% responses obtained by the three prompts respectively include any discussion relevant to the three core argument components (See Appendix C for details). Qualitatively, we found many responses were only paraphrases of the input argument. Although the first prompt elicits some responses with bullet points, for most responses by all three prompts, the model generates open-ended lengthy responses, which are difficult to evaluate, with challenges like hallucinations, difficult human evaluation, an active area of research (Karpinska et al., 2021; Chang et al., 2024).

Baseline 2 (Directly prompt LM to generate argument components): In the second baseline, we assume task definition but still omit reference to Toulmin's theory in the prompt. Instead, we directly ask the LM to generate components defined in the task. Thus, in stage 1, we prompt `GPT-4` and `LLAMA-2` with '`What is the claim of this argument?`' followed by '`What are the reasons provided to support this claim?`'. We observe that while these prompts identify the correct component, the responses additionally contain a lot of irrelevant information. For instance, `GPT-4` generates reasons in addition to the claim when asked to only generate the claim. `Llama-2-70B` generates additional questions as continuations in the response, such as 'Is this argument valid?' and provides answers to these questions in the response. We observe the same issue despite limiting the maximum tokens (to generate) to the average component length in a dataset. See §5.4 for a detailed comparison with our approach. Given the low performance in stage 1, we did not investigate this baseline for warrant generation in stage 2.

5.3 Using references to Toulmin's theory

In contrast to generic prompts, the Toulmin prompt generates semi-structured responses, with mentions of theory-relevant *terms* and their *values*. Thus, this prompt offers a consistent output format and the response correctness is straightforward to assess as it is supposed to obey theory definition. We next examine the performance of this prompt in detail.

How often does the Toulmin prompt generate a theory-compliant breakdown? We compute success rate, which measures the fraction of arguments for which the LM responses contain all three core terms from Toulmin's theory: *claim*, *grounds* (or *data*), and *warrant*. As shown in

[8] We use `https://together.ai/` API for inference.

[9] See Appendix B.1 for analysis with varying temperatures.

Datasets	Success Rate (%)			
	GPT4	GPT3	Llama-2-70B	Falcon-40B
ARCT	99.0	94.6	75.2	35.0
MCT	100.0	95.4	90.2	42.5

Table 1: Fraction of responses correctly mentioning all three core terms from Toulmin's theory, across LMs and datasets, via '`According to Toulmin model`' prompt.

Table 1, a large fraction of `GPT-4` and `GPT-3` responses contain all the core terms suggesting that the model's responses are theory-compliant with high likelihood. Open-weight models also generate theory-compliant breakdowns, although at a lower frequency, with `Llama-2-70B` performing much better than `Falcon-40B`. On further analysis of responses generated by the best-performing proprietary model (`GPT-4`) and open-weight model (`Llama-2-70B`), we find that many of the responses also contain all six terms from Toulmin's theory (`GPT-4`: 96.84%, `Llama-2-70B`: 68.92%). With a low frequency (less than 5%), terms from the other argumentation theories are present in `Llama-2-70B`, but never appear in `GPT-4` responses (see Appendix D for more details), suggesting that the LM's responses conform to Toulmin's theory.

5.4 Examining the quality of explication triples obtained from LM's response

We next examine the quality of triples, $\langle c, r_i, w_i \rangle$, extracted from the LM response (§4.1). Since LMs are known to hallucinate (Maynez et al., 2020; Cao et al., 2022; Ji et al., 2023), it is imperative to examine the correctness of triples before using them for any downstream applications. We examine each of the three components in triples obtained via `GPT4` and `Llama-2-70B`, the best-performing proprietary and open-weight models in terms of success rate.

Automatic evaluation of claim and reasons: We compare generated claims and reasons with gold annotations from ARCT and MCT datasets.

Claim (c): We measure semantic similarity between generated and gold claim, using ROUGE-L (Lin, 2004, n-gram overlap) and BERTScore (Zhang et al., 2020, token-level similarity via contextualized word embeddings). In Table 2, as expected, on ARCT, `GPT-4`-generated claims exhibit near-perfect scores as it only involves the identification of the claim from two propositions (claim and reason). On MCT with longer arguments, scores are slightly lower, yet the LM responses are correct since the LM resolves coreferences in the generated claims, unlike the gold claims which are spans of the input argument. `Llama-2-70B` performs reasonably, though the similarity scores are lower than `GPT-4`. In contrast, when asking both the models to directly generate the claim, the precision drops considerably, suggesting that the LMs additionally generate a lot of irrelevant information.

Reasons (r_i, $\forall i$): Evaluation of reasons is challenging since the number of gold and generated reasons may differ and the generated reasons may not be strict spans of the input argument but light paraphrases. Thus, one-to-one mapping between generated and gold reasons is unknown. To mitigate this issue, we adopt FactScore (Min et al., 2023), which measures whether a proposition is supported by a given context. We use FactScore to measure precision (number of generated reasons supported by the gold reasons) and recall (number of gold reasons supported by generated reasons).[10]

Table 3 shows a high recall and precision on both datasets for `GPT-4`, suggesting that it can identify all relevant reasons without generating irrelevant information. `Llama-2-70B` performs reasonably, though the scores are lower than `GPT-4`. In particular, `Llama-2-70B` achieves better recall than precision on MCT, implying it identifies all relevant reasons but occasionally generates irrelevant information. In contrast, when both models are asked to generate reasons directly, their precision drops, especially for longer arguments from MCT, indicating that they generate a lot of irrelevant information in addition to the relevant reasons.

In contrast, when both models are asked to generate reasons directly, with the exception of `GPT-4` on ARCT (with simple arguments), the precision drops, especially for longer arguments from MCT, indicating that LMs generate a lot of irrelevant information in addition to the relevant reasons. Overall, directly prompting the LM to generate argument components (i.e., baseline 2) results in subpar performance when aggregated over both claim and reason components.

Human evaluation of warrants (w_i $\forall i$): Previous studies (Becker et al., 2020b; Boltužić and Šnajder, 2016) have noted variability in collecting gold warrants owing to differing annotator intuitions on what needs to be explicit or what can be taken as granted. This subjectivity results in multiple valid

[10]Aggregating pairwise similarity scores between gold and generated reasons can also be used, but we find precision and recall scores more interpretable than an aggregated score.

Prompt	Model	Dataset	BERTScore		Rouge-L	
			Recall	Precision	Recall	Precision
According to Toulmin model,	GPT4	ARCT	0.99±0.01	0.98±0.01	1.00±0.01	0.98±0.01
		MCT	0.78±0.04	0.79±0.04	0.79±0.05	0.77±0.05
	Llama-2	ARCT	0.64±0.03	0.58±0.03	0.66±0.04	0.52±0.04
		MCT	0.58±0.06	0.58±0.07	0.50±0.08	0.50±0.08
What is the claim of this argument?	GPT4	ARCT	0.95±0.01	0.91±0.02	0.99±0.01	0.90±0.02
		MCT	0.72±0.03	0.58±0.05	0.69±0.05	0.52±0.06
	Llama-2	ARCT	0.50±0.01	0.21±0.02	0.92±0.01	0.08±0.01
		MCT	0.57±0.03	0.17±0.04	0.70±0.04	0.18±0.03

Table 2: Automatic evaluation of the generated claims.

Prompt	Model	Dataset	Recall	Precision
According to Toulmin model,	GPT4	ARCT	0.88±0.03	0.87±0.03
		MCT	0.83±0.05	0.86±0.05
	Llama-2	ARCT	0.60±0.04	0.59±0.05
		MCT	0.69±0.09	0.74±0.08
What are the reasons provided to support this claim?	GPT4	ARCT	0.91±0.03	0.93±0.02
		MCT	0.82±0.07	0.75±0.05
	Llama-2	ARCT	0.74±0.04	0.43±0.04
		MCT	0.91±0.05	0.60±0.07

Table 3: Automatic evaluation of the generated reasons.

warrants per claim-reason pair, and thus a model-generated warrant could be acceptable even if it differs from gold.[11] Hence, we conduct a human evaluation to assess the quality of warrants.

Given a gold claim-reason pair, we collect acceptability judgments for gold and model-generated warrants. We consider a warrant acceptable if it is: a) relevant and fully explains the link between the claim-reason pair, b) not trivial (of the form 'if reason then claim', since each gold claim-reason pair has been annotated with a non-trivial warrant in the original datasets), and c) must hold for the claim to be inferred from the reason, even if it does not align with the annotator's or reader's personal beliefs. We hired two freelancers on Upwork[12] with graduate-level expertise in English composition and rhetoric, who were shown a claim-reason pair and three warrants (gold, `GPT4` and `Llama-2`-generated; in random order), and were asked to mark all the warrants they consider acceptable. We collected judgments for 150 pairs, with 75 random pairs from ARCT and MCT each. Appendix G provides more details.

Out of 300 judgments for each warrant type, we find that gold warrants are acceptable in 45.7%, `GPT-4`-generated in 61.7%, and `Llama-2-70B` in 26.3% cases, suggesting a preference for `GPT-4`-generated warrants, surpassing gold warrants.[13] Annotators marked a gold warrant unacceptable when it restated the claim, had incorrect wording, was irrelevant to the claim-reason pair, or failed to explain the link between the pair (examples in Appendix G). `GPT-4` warrants were mostly considered unacceptable when they repeated the reason, claim, or were of the form 'if reason then claim.' Finally, `Llama-2`-generated warrants were often repetition of reason and were acceptable in only a few cases, suggesting that `Llama-2` struggles to generate warrants, requiring further research. Nevertheless, open-weight models exhibit potential, generating Toulmin-style argument breakdown and achieving reasonable claim and reason identification.

5.5 Prompt sensitivity analysis: Can other name references to Toulmin's theory improve performance?

Toulmin's theory can be referenced in various ways (e.g., Toulmin's model/Toulmin's method). Given the prompt sensitivity of language models, we examine the performance across different references.

Extraction of alternative name references: We extract most frequent name references to Toulmin's theory, $N_t = \{n_t^1, n_t^2..n_t^k\}$, from C4 (Raffel et al., 2020), often used for pre-training LMs. Prior efforts have also studied pretraining datasets to measure data contamination (Dodge et al., 2021; Elazar et al., 2023) and its influence on model performance (Magar and Schwartz, 2022; Longpre et al., 2023), we analyze C4 for prompt design. We retrieve documents containing the word *Toulmin*[14] and identify sentences mentioning the same surname. From each sentence, we extract simple noun phrases containing common terms describing a construct (e.g., model, method, schema). After a manual review for relevance to the theorist, we compile a list of name references with their n-gram counts in C4 (Table 4); See Appendix E for more details.

The 'Toulmin model' reference performs best, though other references give comparable performance: Table 4 shows success rates obtained by prompting `GPT-4` and `Llama-2-70B` with different name references. 'Toulmin model' gives the highest success rate, while other references,

[11] Similarity between gold and generated warrants, measured using BERTScore, is low (for `GPT-4`, `ARCT`: 0.3±0.01, `MCT`: 0.4±0.04; for `Llama-2-70B`, `ARCT`: 0.2±0.01, `MCT`: 0.2±0.01), indicating the model-generated warrants differ from the gold warrants.

[12] https://www.upwork.com/

[13] We observe slight agreement among the two annotators on the gold warrants (Cohen's kappa=0.18), reflecting the inherent subjectivity of the warrant evaluation task. In comparison, we observe a fair agreement on GPT-4 generated warrants (kappa=0.24) and Llama-2 generated warrants (kappa=0.31).

[14] We use Dodge et al. (2021)'s C4 search engine at https://c4-search.apps.allenai.org/ for retrieval.

Name Reference (n_t^i)	C4 Corpus Frequency (n-gram counts)	Success Rate (%)	
		`GPT-4`	`Llama-2-70B`
(the) Toulmin model	2415	97.42	67.27
(the) Toulmin method	531	95.20	61.38
Toulmin's model	162	96.00	67.12
(the) Toulmin('s) Schema	137	96.13	61.48
(the) Toulmin('s) approach	87	95.50	56.75
Toulmin argument strategies	41	95.60	27.47
Toulmin's argument(ation) model	28	96.88	61.48

Table 4: Success rate of LMs on the ARCT dataset when prompted with '`According to` n_t^i`,`' where n_t^i is a name reference to the Toulmin's theory.

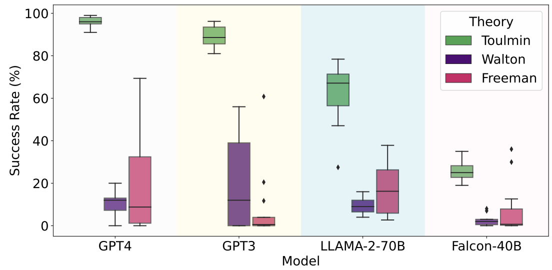

Figure 4: Across all LMs, prompting with references to Toulmin's theory results in highest success rate.

both moderate-frequency (e.g., Toulmin's model) and low-frequency (e.g., Toulmin argument model), yield comparable results. Table 4 and Figure 4 also show that `GPT4`'s performance varies less across name references, while `Llama-2-70B` exhibits greater variability, suggesting that `GPT4` is more robust to prompt variations. Finally, `Llama-2-70B` exhibits a moderate correlation between the success rate and frequency of a name reference (Spearman's correlation, ρ=0.56, though statistically non-significant with p=0.2), while GPT4 shows near zero correlation (ρ=0.04), indicating that open-weight models could benefit from optimizing prompt based on occurrence frequency of a name reference.

5.6 Can name references to alternative theories improve performance?

We investigate whether the prevalence of a theory (aggregate frequency of all name references to a theory) in LM's pretraining data correlates with its performance on the task. We examine two alternative theories, namely, Walton's argumentation schemes (Walton et al., 2008) and Freeman's theory of argument structure (Freeman, 1991), which are less frequently mentioned on the web[15] but often used in computational research (Habernal and Gurevych, 2017) and for annotation purposes (e.g., MCT used in this work is annotated according to Freeman's theory). Both theories are also relevant to the argument explication task, as they have similar core components as Toulmin, though use different terminology.

Figure 4 shows success rate distribution, when prompted with name references to the three theories,[16] across four LMs on the ARCT dataset. References to Toulmin's theory consistently yield higher success rates than references to other theories across all models, validating our hypothesis. Overall, our findings suggest that the aggregate frequency of references to a theory/concept could be an interesting factor to consider when designing prompts, an interesting avenue for future research.

In summary, the results discussed above (§5.3-§5.6) demonstrate the utility of explicitly prompting on the basis of Toulmin's theory, for the zero-shot argument explication task. For further performance improvements, the prompt could be enhanced beyond our simple '`According to Toulmin model`' by explicitly mentioning component definitions from Toulmin's theory. We leave the exploration of this approach for future work.

6 Case Study: Making sense of public opinion via argument explication

We now illustrate the use of argument explication by analyzing public comments to the FDA on COVID-19 vaccine approval for children. Prior studies used clustering (Hoyle et al., 2023) and topic modeling (Pacheco et al., 2022) to identify the main beliefs (or propositions) held by the public. However, comments are often argumentative, with inferential relations among propositions. Knowing how propositions are interconnected in the broader debate can identify not only what people believe, but also why. For example, if the public health policy has to reduce vaccine hesitancy, officials must know how propositions interconnect in a broader discussion to knock down fallacious arguments.

Method: We use Hoyle et al. (2023)'s corpus of 10,000 public comments sourced from regulations.gov, exhibiting a general vaccine hesitancy. We generate explication triples, $\langle c, r_i, w_i \rangle$, from all

[15] In C4, Toulmin's theory is referenced 3401 times (n-gram counts), Walton's theory 975 times, Freeman's theory 66 times. Appendix F provides counts from other sources.

[16] We use the method from §5.5 to extract name references for Walton and Freeman's theory, details in Appendix E. When prompting with references to each theory, we use the respective terminology for evaluating LMs. For instance, when prompting with Walton's theory, we evaluate the model's ability to identify terms as per the Walton's terminology.

comments, via the method outlined in §4.1 using GPT-4, excluding single-sentence comments which are often non-argumentative (refinement of this step left for future work). We cluster embeddings[17] of all propositions from the triples, irrespective of their role in triple, using DP-means clustering (Dinari and Freifeld, 2022; Kulis and Jordan, 2012), which automatically determines the number of clusters based on a Euclidean distance threshold. We use a threshold of 0.5, selected via visual inspection of cluster quality. From 9,187 comments, we obtain 14,137 triples and 308 propositional clusters. To identify interconnections between clusters, we represent a proposition with its cluster ID followed by transforming triples of propositions into triples of cluster IDs (TIDs). Each TID, comprised of three cluster IDs, represents a local argument structure mentioned in one or more comments and reveals inferential relations among the corresponding clusters. Overall, we obtain 6,811 unique TIDs, visualized as a hypergraph, where a TID forms a hyperedge and a node is a propositional cluster.

Interpretive analysis of the corpus based on the hypergraph: We find several insights from the obtained hypergraph. Among all the TIDs, 1,862 appear in more than one comment, suggesting that people not only share common beliefs but also use similar argument structures to support their beliefs. Figure 1 shows a fragment of the larger argument hypergraph around the most common argument, (c=P1, r=P2, w=P5), which occurs 373 times; it opposes vaccine approval (c=P1) by saying that children have a low risk from the disease (r=P2). Some comments further elaborate on the backing for P2, by citing low mortality rates from COVID-19 among children (P8), obtained by citing data from government websites. Countering any node in this chain could knock down the entire argument chain. On further exploring the local neighborhood of P1, we find two other frequently mentioned reasons: vaccine side-effects (P7) and lack of long-term testing (P3), consistent with findings from studies of social media discussion on vaccines (Wawrzuta et al., 2021), conferring convergent validity to our approach from a different source.

Explicitly stating warrants also helps reveal the relationship between distinct parts of the hypergraph.[18] Since we cluster all propositions irrespective of their role in a comment, some clusters include both implicit and explicit propositions. For instance, cluster P5 (vaccines are unnecessary for children) includes propositions implied in some comments, while explicit in others. Thus, such clusters bridge distinct parts of the hypergraph.

Overall, we find corpus visualization as a hypergraph promising direction for future work. Graph visualization (among concepts, entities, etc) has been proposed for exploratory corpus analysis (Falke and Gurevych, 2017; Handler and O'Connor, 2018). Complementary to these efforts, our approach can visualize the structure of interrelated arguments, and help analysts research complex questions concerning cluster relations (e.g., 'Why do people think COVID-19 does not affect children?').

7 Conclusion

In this work, we analyze arguments by making explicit their structure and reasoning, employing LMs in a zero-shot setting, and using references to Toulmin's theory as prompts. We validate our approach via robustness across different references and theories.

Here, our case study focuses on applying computational argument modeling to analyze public comments, which is useful for civic decision-making. Argument explication may also have potential in other formal, argument-rich areas, such as law or peer reviews, where authors often include explicit rebuttals and counter-rebuttals. Future work could extend our approach to include these additional components and utilize this explicit argument structure. Overall, the visualization and navigation of recurring, proposition-level argument hypergraphs could aid interpretive work and content analysis in the computational social sciences.

8 Limitations

We list some of the limitations of our study, which we hope will be useful for researchers and practitioners when interpreting our analysis.

(1) Our analyses and experiments only apply to arguments in the English language and the approaches to analyzing non-English argumentative texts using large language models should be explored in future studies.

[17] Obtained via all-mpnet-base-v2 (Reimers and Gurevych, 2019).

[18] A claim-reason pair may be linked by several warrants; for visual clarity, we only display the most frequent one.

(2) Our intrinsic evaluation of the explication triple generation approach does not consider the known political biases of generative language models (Santurkar et al., 2023). While most generated triples are considered acceptable based on our quality evaluation in §5.4, it remains to be studied how these biases could affect the quality of the generated triples, which we will explore as part of future work.

(3) We use a general-purpose sentence encoder (Reimers and Gurevych, 2019) to obtain propositional embeddings in our case study. Fine-tuning the embeddings on the specific domain of interest could enhance both the embeddings and proposition clustering.

(4) Our name reference extraction method relies on a noun-phrase detection algorithm, which can be imperfect, especially when tested out-of-domain. Future research could explore alternative techniques, especially those suited for analyzing informal web text.

(5) An LLM-based approach can be computationally expensive and time-consuming. For instance, for our case study (§6), the proprietary LM (GPT-4) took approximately 1 hour when queried in parallel using OpenAI API. The high-end open-weight LM (Llama-2-70B) would take approximately 5-6 hours for the same case study when queried serially, using public API access (`https://together.ai/`) with Nvidia H100 and A100 GPU support, as stated by the provider. While these estimates are reasonable for our intended use case, larger-scale applications can be both expensive and time-consuming. However, improving the efficiency of LLMs is an active area of research (Wan et al., 2023), with the potential to make LLM-based applications more time and resource-efficient.

(6) Finally, the ARCT and MCT datasets used to evaluate the intrinsic validity of our method (§5.4) may be considered small in size. While the size of these datasets is comparable to datasets in other language model evaluation benchmarks (e.g., BIG-bench has tasks with 100 examples intended to evaluate zero-shot and few-shot capabilities of LMs (Srivastava et al., 2022, Figure 3), similar to our case), this resource constraint highlights the necessity for low-resource or zero-shot techniques for argument analysis. Additionally, we support the evaluation of our method via a case study, whose findings confer with prior manual studies, demonstrating the external validity of our approach (§6).

9 Ethics Statement

Our work is in line with the ACL Ethics Policy. The text and appendix outline all the models, datasets, and evaluation methodologies used in this research. All the datasets used in this research are publicly available and used with the appropriate consent. All comments analyzed in our work are public and were intended for public disclosure in documents or on the Internet, as per regulations.gov.[19] The data collection protocol for human evaluation was approved as exempt from institutional review by the coauthors' institution's human subjects research office. All annotators gave their written consent to disclose their anonymized annotations and comments (Appendix G). Annotators were paid at least $15 per hour, above the local minimum wage, with bonuses for additional time spent on reading guidelines, taking screening tests, and clarifying any questions.

10 Acknowledgements

We are very grateful to the annotators on Upwork for participating in our annotation task and providing annotation and detailed comments. We are grateful to Aditya Jain, Chloe Eggleston, Erica Cai, Maha Alkhairy, Marisa Hudspeth, Marzena Karpinska, Simeng Sun, Swati Gupta, Tessa Massis, and the UMass NLP group for several useful discussions during the course of the project. This material is based upon work supported by National Science Foundation award 1845576 and an IBM Ph.D. Fellowship awarded to AG. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily represent the official views of, nor an endorsement, by any sponsor.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, pages 64–68.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow,

[19]Details under the privacy section at `https://www.regulations.gov/faq?anchor=downloadingdata`

Julien Launay, Quentin Malartic, et al. 2023. Falcon-40B: An open large language model with state-of-the-art performance. *Findings of the Association for Computational Linguistics: ACL*, 2023:10755–10773.

Maria Becker, Ioana Hulpus, Juri Opitz, Debjit Paul, Jonathan Kobbe, Heiner Stuckenschmidt, and Anette Frank. 2020a. Explaining arguments with background knowledge. *Datenbank-Spektrum*, pages 1–11.

Maria Becker, Katharina Korfhage, and Anette Frank. 2020b. Implicit knowledge in argumentative texts: An annotated corpus. *Proceedings of the 12th Conference on Language Resources and Evaluation*.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! Analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354. Association for Computational Linguistics.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit premise generation with discourse-aware commonsense knowledge models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6247–6252. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms.

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

Noam Chomsky. 1957. *Syntactic structures*. Mouton de Gruyter.

Darshan Deshpande, Zhivar Sourati, Filip Ilievski, and Fred Morstatter. 2023. Contextualizing argument quality assessment with relevant knowledge. *ArXiv*, abs/2305.12280.

Or Dinari and Oren Freifeld. 2022. Revisiting dp-means: fast scalable algorithms via parallelism and delayed cluster creation. In *Conference on Uncertainty in Artificial Intelligence*.

Jesse Dodge, Ana Marasović, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Conference on Empirical Methods in Natural Language Processing*.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Lindsay M Ellis. 2015. A critique of the ubiquity of the Toulmin model in argumentative writing instruction in the USA. *Scrutinizing argumentation in practice*, pages 201–213.

Tobias Falke and Iryna Gurevych. 2017. GraphDocExplore: A framework for the experimental comparison of graph-based document exploration techniques. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 19–24, Copenhagen, Denmark. Association for Computational Linguistics.

James B. Freeman. 1991. *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. De Gruyter Mouton, Berlin, Boston.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Abram Handler and Brendan O'Connor. 2018. Relational summarization for corpus analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1760–1769, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Making the implicit explicit: Implicit content as a first class citizen in nlp. *ArXiv*, abs/2305.14583.

Ioana Hulpus, Jonathan Kobbe, Christian Meilicke, Heiner Stuckenschmidt, Maria Becker, Juri Opitz, Vivi Nastase, and Anette Frank. 2019. Towards explaining natural language arguments with background knowledge. In *PROFILES/SEMEX@ISWC*.

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

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *ArXiv*, abs/2304.03245.

Mark Klein. 2012. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. MIT Working Paper.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Brian Kulis and Michael I. Jordan. 2012. Revisiting k-means: New Algorithms via Bayesian Nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning*, page 148. icml.cc / Omnipress.

John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–22.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Inbal Magar and Roy Schwartz. 2022. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. End-to-end Argument Mining with Cross-corpora Multi-task Learning. *Transactions of the Association for Computational Linguistics*, 10:639–658.

S.E. Newman and C.C. Marshall. 1991. Pushing Toulmin too far: Learning from an argument representation scheme.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings* *of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Andreas Peldszus and Manfred Stede. 2015. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–815.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Chris Reed. 2001. Araucaria: Software for puzzles in argument diagramming and XML.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450.

Gil Rocha, Henrique Lopes Cardoso, Jonas Belouadi, and Steffen Eger. 2023. Cross-genre argument mining: Can language models automatically fill in missing discourse markers? *ArXiv*, abs/2306.04314.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *ArXiv*, abs/2303.17548.

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

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163, Brussels, Belgium. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

S. Toulmin, R.D. Rieke, and A. Janik. 1984. *An Introduction to Reasoning*. Macmillan.

Stephen E. Toulmin. 1958. *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, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv e-prints*, page arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno

Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Douglas Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, New York.

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

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2023. Efficient large language models: A survey. *ArXiv*, abs/2312.03863.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Dominik Wawrzuta, Mariusz Jaworski, Joanna Gotlib, and Mariusz Panczyk. 2021. What arguments against covid-19 vaccines run on facebook in poland: content analysis of comments. *Vaccines*, 9(5):481.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn J Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. "according to ..." prompting language models improves quoting from pre-training data. *ArXiv*, abs/2305.13252.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori Hashimoto. 2023. Benchmarking large language models for news summarization. *ArXiv*, abs/2301.13848.

Appendix

A Background on argumentation theories

Toulmin's model of argumentation: Toulmin's model of argumentation consists of six components (Figure 5). The three fundamental components are:

Claim: The claim or conclusion whose merits author is seeking to establish.

Data: Evidence to establish the foundation of the claim, or, as explained by Toulmin, 'the data represent what we have to go on.' The term was later changed to *grounds* in Toulmin et al. (1984).

Warrant: A logical inference from the grounds to the claim. It can be general, hypothetical statements, inference rules, or principles that can act as a bridge between the grounds and the claim. As pointed out by Toulmin (1958), "data are appealed to explicitly, warrants implicitly."

Optional components include *backing* (additional support for warrant), *rebuttal* (a view-point opposing the claim), and *qualifier* (the degree of certainty).

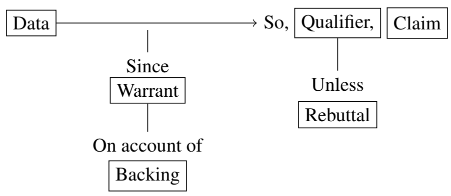

Figure 5: Toulmin's model of argumentation as proposed in Toulmin (1958). Nodes represent argument components, the arrows symbolize the explicit support relation, and the lines indicate the authority conferred by one node to the other.

Argument structure by Freeman: Freeman (1991) proposes some key modifications to address issues observed when applying Toulmin's model to real-life argumentative texts (Newman and Marshall, 1991). In particular, Freeman does not distinguish between *data* and *warrant* and regards any evidence provided to support the *conclusion* (similar to *claim* in Toulmin's terminology) as a *premise*. Other components in Freeman's structure include *rebuttal*, *modality* (how strongly the premises support the conclusion), and *counter-rebuttal* (views opposing rebuttal).

Walton's argumentation schemes: Walton (1996) proposed a set of argumentation schemes or structures of inference. Each scheme represents a form of everyday reasoning and consists of:

Conclusion: The main point of view.

Minor Premise: Provides evidence to support the conclusion.

Major Premise: An inference rule, similar to the warrant in Toulmin's terminology.

The core components of each of the above theories are related to the components of the argument explication task as listed in Table 5.

	Toulmin	**Walton**	**Freeman**
Claim (c)	Claim	Conclusion	Conclusion
Reason (r_i)	Data/Grounds	Minor Premise	Premise
Warrant (w_i)	Warrant	Major Premise	Premise

Table 5: Mapping between the components of the argument explication task and terminology proposed in each of the argumentation theories.

B Experimental Details

B.1 Choice of Temperature

Prior literature has used various decoding strategies when evaluating LLMs for their zero-shot abilities. For instance (Kojima et al., 2022; Wei et al., 2022; Wang et al., 2022) consider greedy decoding. OpenAI also uses greedy decoding as their default setting for conditional text generation (e.g., summarization, translation, grammar correction, etc.) Some work in summarization and machine translation (Zhang et al., 2023; Karpinska and Iyyer, 2023) also considers temperature=0.3. We experimented with three temperatures, 0.0, 0.3, and 0.5 on 50 examples from both ARCT and MCT and found that the generations with different temperatures were semantically very similar to each other, with an average BERTScore (F1) 0.92-0.96 between pair of responses generated by different temperatures. The variations in responses were mostly related to lexical word choice, without altering the overall meaning. Responses were also similar for different samples generated using the same temperature. As a result, for the sake of simplicity, we keep a temperature=0.0 in all our experiments.

C Details of baseline 1

In baseline 1, we experimented with three generic prompts. To count the number of responses that contain any terms relevant to the three core argument components, we searched for the following terms:

1. **Claim:** claim, conclusion, conclude, concludes, assertion, posits, advocating
2. **Reason:** reason, premise, evidence, supports
3. **Warrant:** assumption, warrant, implies, implying, suggests, suggesting, implication

The above terms were curated manually by the author by going through all the LM-generated responses. We included any term that could serve a similar function as the argument component name, including verbs (e.g., 'posits' or 'advocating' for the term 'claim').

D Additional analysis of LM responses generated via the Toulmin prompt

What other terms are present in the LM responses? We also investigate the presence of terms that are not part of Toulmin's theory. Some examples contain the term 'conclusion' (`GPT-4`: 12.16%, `Llama-2-70B`: 48.20%). However, this term is unique, as Toulmin (1958) employs 'conclusion' and 'claim' interchangeably to denote the same concept. Some terms from other argumentation theories are present in a small fraction of `Llama-2-70B`'s responses (premise: 2.48%, modality: 0.23%, counterrebuttal: 4.28%, major premise: 1.13%, minor premise: 0.90%). However, these terms do not appear in GPT-4's responses. Overall, the model's responses most often contain terms from Toulmin's theory and not other theories, suggesting that the LM's responses are predominantly theory-compliant.

E Extraction of theory name references

We extract the most frequent name references to a theory, $N_t = \{n_t^1, n_t^2..n_t^k\}$, from English portion of C4 (C4.EN, Raffel et al., 2020), which is often used for pre-training LMs. For each theory, we retrieve documents[20] containing the theorist's surname[21] and identify sentences mentioning the same surname. For Toulmin, we retrieve 4,805 (4,242 unique) documents; Walton and Freeman yield a large number of matches, we consider the first 10,000 matches, resulting in 9,690 and 9,997 unique documents, respectively.[22] From each sentence containing the theorist's surname, we extract

[20] We use Dodge et al. (2021)'s search engine at https://c4-search.apps.allenai.org/ for retrieval.

[21] Searching via full name filters out relevant documents since informal web discourse may not always use full name references.

[22] Despite more documents for Walton and Freeman, many are false positives as they are more common surnames than Toulmin. According to Forebears (https://forebears.io/), covering 27M surnames of 4B people worldwide, ap-

simple noun phrases containing common terms describing a construct.[23] We use spaCy v3.4.0 (Honnibal and Johnson, 2015) to extract simple noun phrases. After this step, we obtain 888 (127 unique) phrases for Toulmin, 284 (94 unique) for Walton, and 185 (67 unique) for Freeman, all appearing more than once in C4. Notably, noun phrases for Toulmin outnumber those obtained for Walton and Freeman.

After manual filtering for relevance to theorist or argumentation literature (e.g., removal of unrelated references like 'Walton County Local Mitigation Strategy Work Group' and generic/ambiguous phrases like 'argument analysis'), we curate a final list of name references per theory along with their n-gram counts in C4. List of references to Toulmin's theory and Walton's theory are mentioned in Table 4 and Table 6.

Finding references to Freeman's theory is a little challenging. In contrast to Toulmin and Walton, we initially did not find any relevant phrases for Freeman among noun phrases extracted from C4. Among the automatically extracted noun phrases, none refer to James B. Freeman, instead most refer to scientific work by another scientist (e.g., 'Systematic approaches' by Harold S. Freeman, 'Geologic framework' by Philip A. Freeman). This suggests that either Freeman's theory is less frequently referenced on the web or that our noun-phrase extraction misses relevant phrases.

During qualitative analysis, we observed that our noun phrase extraction algorithm (based on spaCy) sometimes fails with colloquial text, which is commonly found on the web. Specifically, it often extracts longer spans than expected, including additional terms such as verbs (e.g., 'the Toulmin model results', 'Toulmin model shows', 'the "Toulmin model" posts', and 'Even the Toulmin model').

We found more success from a different corpus: by extracting phrases from scholarly abstracts, S2ORC (Lo et al., 2020), a dataset of academic literature, also intended for language model pretraining. We use the same noun phrase extraction method to obtain the phrases from S2ORC. However, unlike C4, the text in scholarly abstracts is more formally written, leading to a lower number of errors in the noun phrase extraction step. Table 7 shows the extracted references. These references also have non-zero n-gram counts in C4, indicating that our noun-phrase extraction may overlook some relevant phrases, especially those with low frequency. This suggests a need for refining name reference extraction in future work.

proximately 476 individuals have the surname Toulmin, while 156,730 have the surname Walton, and 331,743 have the surname Freeman.

[23] model(s), method(s), analysis, diagram, scheme(s), schema, framework(s), theory(ies), strategy(ies), approach(es), algorithm(s), structure(s). We curated this list by manually examining noun phrases obtained for all three theories.

Phrase	Frequency
(The) argumentation schemes	907
Walton's theory	32
Walton's approach	15
Walton's critical questions method	13
Douglas Walton('s) logical argumentation theory	3
Walton's schemes	2
Walton Douglas's argumentation schemes	2
Walton's Argumentation Schemes	1

Table 6: References to Walton's theory extracted from C4, with n-gram counts in C4. The most common phrase *'(The) argumentation schemes'* is also the name of the book by Douglas Walton describing various argumentation schemes (Walton et al., 2008).

Phrase	Frequency
Freeman's theory	31
Freeman's model	20
Freeman's method	13
Freeman's Argument Structure Approach	1
Freeman's Argument Structure	1

Table 7: References to Freeman's theory extracted from S2ORC corpus, with non-zero n-gram counts in C4.

F Prevalence of name references to theories across different sources

Table 8 mentions the aggregate frequency of name references to a theory (as mentioned in Tables 4, Tables 6 and Tables 7) across different pre-training corpora and other sources (e.g., Google Scholar citations, Google Books Ngram V3 dataset). For Google Books, we use the service at `https://ngrams.dev/` to extract n-gram counts. We use the n-gram lookup service at `https://wimbd.apps.allenai.org/` for the remaining datasets. We observe that across all the considered sources, name references to Toulmin's theory appear more frequently than the other theories.

Theory	Citations (Google Scholar)	Counts of n-grams			
		Google Books Ngram V3	C4	The Pile	OSCAR
Toulmin	20,703	18640	3401	493	1724
Walton	2218	11522	975	365	328
Freeman	453	2963	66	25	55

Table 8: Aggregate frequency of name references to a theory across different sources/datasets. Name references to Toulmin's theory appear more frequently than the other theories.

Claim	Reason	Gold Warrant	Comment
christians have created a harmful atmosphere for gays.	i find the idea that it is a sin to be born or live a life at all to be preposterous.	being gay is considered a sin	reason irrelevant to the claim
foreign language classes should be mandatory in college.	we should be able to speak other languages rather than expect everyone else to speak english.	students should be taking those classes by force	restatement of claim
With those kinds of amounts you think twice about whether you really want to stay in the flat.	they're very bad however, if the rent suddenly climbs by €100 or €200.	If the rent rises from €100 or €200, many cannot afford to stay in the flat.	incorrect wording, "rent rise from €100 or €200" implies €100 or €200 is the base rent
obamacare is sustainable.	taking a cue from the success of the Swiss and Dutch healthcare models proves Obamacare can work, too.	the Swiss and Dutch government is similar to ours	incorrect wording, similar government does not imply similar healthcare models
Brazil should not postpone Olympics.	the Olympics are a dream for many athletes since they train extremely hard.	the athletes won't get sick going to Brazil	warrant fails to explain the link between claim and reason.
public universities are neglecting in-state students.	they want to take advantage of higher tuitions paid by foreign and out of state students.	universities gain additional funds to make more profit	warrant fails to explain the link between claim and reason.
medicare needs to be reformed.	there needs to be some sort of vetting process for advertisers, some of them attempt to scam the elderly.	the elderly are not the only people that are affected	warrant fails to explain the link between claim and reason.

Table 9: Examples of gold warrants marked unacceptable by our annotators, along with their comments explaining why they marked them as unacceptable.

G Human evaluation of warrants

Informed consent: Before participating in our study, we requested every annotator to provide informed consent. The annotators were informed about the purpose of the research study, any risks associated with it, and the qualifications necessary to participate. The consent form also elaborated on task details describing what they will be asked to do and how long it will take. Annotators were also informed that only satisfactory performance on the screening test would allow them to participate in the annotation task. The annotators were also informed that they could drop out at any time. Annotators were informed that they would be compensated in the standard manner through the Upwork platform, with the amount specified in the initial Upwork contract. As part of this study, we also collected their level of expertise in English composition and rhetoric. We ensured our annotators that this information would remain confidential in the consent form.

Task setup and guidelines: We show 5 claim-reason pairs, each with 3 associated warrants, and asked annotators to mark ALL the warrants that are acceptable for a given pair. In our guidelines, we provided the following constraints to decide the acceptability of a warrant: a) It is relevant to the claim and the reason. b) It explains the underlying assumption or why the claim logically follows from the reason. c) It is NOT a repetition/paraphrase of the claim or the reason. d) It is NOT simply saying: 'If reason then claim'. e) It should hold true for the claim to be inferred from the reason even if it may not align with your personal beliefs. f) Style of the warrant (e.g., better wording, longer length) does not matter, as long as the content of the warrant links the claim-reason pair. We also provided examples explaining each of these constraints in our guidelines. After reading the guidelines, we asked annotators to take a screening test, which asked basic questions related to the guidelines. This test was intended to mainly test their attention. After passing the screening test, they were asked to annotate 5 claim-reason pairs and provide their reasoning as comments for each annotation. We manually reviewed their comments and after ensuring their understanding of the task, they were asked to annotate 150 claim-reason pairs.

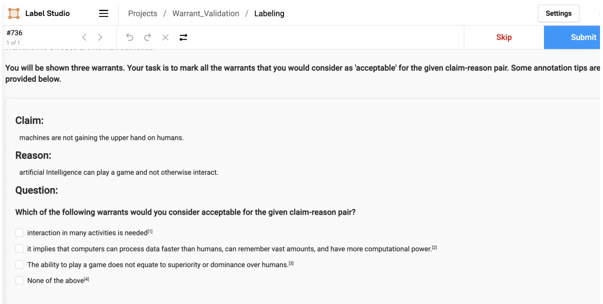

Figure 6: A screenshot of annotation platform for human evaluation of warrants.

Similar to the screening test, during the annotation of the 150 claim-reason pairs, the annotators were shown a claim-reason pair and asked to mark ALL acceptable warrants, with the option to provide comments explaining the reasoning behind their annotations.

Compensation: Each annotator was paid \$0.5 per evaluated claim-reason pair, with an additional \$25 bonus to cover the time spent on reading guidelines, completing screening tests, and clarifying any doubts. Altogether, we paid approximately \$15 per hour, with a total cost of \$200.

Annotation Interface: Figure 6 shows a screenshot of the annotation interface used to collect annotations. The annotators were assigned a unique code to log in to the platform, to maintain their anonymity.

Qualitative annotation analysis: Table 9 provides some examples of gold warrants that were marked as not acceptable by our annotators. Annotators marked a gold warrant unacceptable when it restated the claim, had incorrect wording, was irrelevant to the claim-reason pair or failed to explain the link between the pair.