Beyond Recognising Entailment: Formalising Natural Language Inference from an Argumentative Perspective

Ameer Saadat-Yazdi
School of Informatics
University of Edinburgh
ameer.saadat@ed.ac.uk

Nadin Kökciyan
School of Informatics
University of Edinburgh
nadin.kokciyan@ed.ac.uk

Abstract

In argumentation theory, argument schemes are a characterisation of stereotypical patterns of inference. There has been little work done to develop computational approaches to identify these schemes in natural language. Moreover, advancements in recognizing textual entailment lack a standardized definition of inference, which makes it challenging to compare methods trained on different datasets and rely on the generalisability of their results. In this work, we propose a rigorous approach to align entailment recognition with argumentation theory. Wagemans' Periodic Table of Arguments (PTA), a taxonomy of argument schemes, provides the appropriate framework to unify these two fields. To operationalise the theoretical model, we introduce a tool to assist humans in annotating arguments according to the PTA. Beyond providing insights into non-expert annotator training, we present Kialo-PTA24, the first multi-topic dataset for the PTA. Finally, we benchmark the performance of pre-trained language models on various aspects of argument analysis. Our experiments show that the task of argument canonicalisation poses a significant challenge for state-of-the-art models, suggesting an inability to represent argumentative reasoning and a direction for future investigation.

1 Introduction

When engaging in critical discussion, people often employ stereotypical rules of inference to justify their claims. One of the main areas of study in argumentation theory is categorising these rules as templates to capture commonly employed patterns of reasoning. To this end, various taxonomies have been proposed, the most prevalent being Walton's Argumentation Schemes (Walton et al., 2008). Walton's schemes are conceived of as templates for constructing arguments, with each scheme consisting of a list of (major and minor) premises that support a conclusion. Alongside the premises, argument schemes are associated with several critical questions that serve to test the validity of the argument and identify fallacious reasoning (Walton and Godden, 2005). While Walton conceived of his taxonomy as essential to computational argumentation, little work has been done on the automatic classification of argument schemes given a natural language text. In other words, most of the existing research starts with the assumption that the premises and conclusions and their function as part of a scheme are known a priori.

Similarly, researchers in natural language inference (NLI) have made advances in the task of recognising textual entailment (RTE). Various models appear to achieve remarkable success in capturing textual entailment relationships. Despite these strides, the absence of a standardized definition of entailment hampers the comparability and interpretability of models across different RTE datasets and evaluation metrics (Poliak, 2020). Datasets designed for recognising textual entailment often imply some notion of defeasibility: "*. . . in principle, the hypothesis must be fully entailed by the text. Judgment would be False if the hypothesis includes parts that cannot be inferred from the text. However, cases in which inference is very probable (but not completely certain) are still considered true.*" (Dagan et al., 2006). Yet in the works we have surveyed, this notion remains to be ill-defined. This ambiguity in defining entailment not only impedes progress, but also raises questions about the reliability and generalisability of the models developed for these tasks.

Entailment in natural language involves defeasible inferences that draw on normative and commonsense knowledge. We posit that recognising entailment in natural language text can be more rigorously formulated as the identification of the scheme of inference being employed, and determining whether the hypothesis faithfully applies the scheme to justify the conclusion. In argumentation theoretic terms, recognising entailment involves classifying the argument scheme and then applying the appropriate critical questions to test the argument's validity. This goes beyond simple entailment and provides a measure of the extent to which one statement entails another.

In an attempt to operationalise argumentation theoretic notions of natural language inference, we draw from Wagemans (2016) which describes a taxonomy of argument types named the Periodic Table of Arguments (PTA). Accordingly, we make the following contributions:

1. We conduct an annotation study to rephrase natural language arguments into structured templates and provide insights into how to train non-expert annotators to perform this analysis.

2. We introduce ArgNotator, a tool that assists humans in annotating arguments according to the PTA.

3. We construct Kialo-PTA24 - the first multi-topic dataset of argument types annotated according to the PTA.

4. We compare the performance of state-of-the-art models for two annotation subtasks. For the substance classification task, we benchmark the performance of a number of BERT-based models. For the argument canonicalisation task, we evaluate the performance of two large language models (FLAN-T5, LLAMA2) in both pre-trained and few-shot settings.

The dataset, experimental setup, annotation tool and training materials can all be found on GitLab.[1]

2 Background: Structuring Arguments

The theory of argumentation seeks to provide systematic methods for the analysis, reconstruction, and evaluation of arguments. Throughout history, philosophers have developed many models of argumentation that emphasise different hermeneutical frameworks. The atomic construct studied by all these theories is the *argument*, which is often defined as an inference in which a conclusion is supported by a set of, possibly implicit, premises (Walton et al., 2008).

2.1 Argument Schemes

Building off Aristotle's theory of Topoi (Braet, 2005), Walton's taxonomy of argument schemes (Walton et al., 2008) seeks to identify and codify the structures of inference that exist in various forms of argumentative discourse. This taxonomy describes "stereotypical patterns of reasoning with a corresponding set of critical questions, namely defeasibility conditions." (Walton and Godden, 2005). Walton's schemes are the most prevalent and widely used framework for argument analysis due to their breadth and range of application and have been particularly useful in computational applications of argumentation (Al-Khatib et al., 2020; Kökciyan et al., 2021). Determining the argument scheme used in this taxonomy often requires the argumentation theorist to be familiar with the various possible schemes and be able to distinguish between major and minor premises.

2.2 Periodic Table of Arguments (PTA)

In contrast to the previous taxonomies, the periodic table of arguments (PTA) (Wagemans, 2016), follows a top-down approach, using high-level criteria to reduce the space of possible schemes an argument could belong to. The periodic table asserts that most arguments belong to one of four **'canonical forms'** (alpha, beta, gamma, delta). The premise and conclusion of the argument are then identified to belong to one of three **'substances'**: '*Fact*' (F), '*Value*' (V), or '*Policy*' (P) which alongside the argument form gives the argument type. Each argument type is then associated with a small number of **'concrete levers'** which describe a concrete description of the inference structure employed by the argument. The periodic table not only seeks to provide a theoretically grounded classification of argument schemes but also to develop a classification procedure that can be applied algorithmically with a view towards the automatic classification of argument schemes. A few examples of applying the Argument Type Identification Procedure for the PTA are given in Figure 1.

To classify the argument type, one must first identify the canonical form, rewrite the argument to match the form and then classify the substances of both the premise and the conclusion in the rewritten argument. Recall that the substance of the statement is a classification of the type of statements in an argument. There are three substances posited by

[1] https://git.ecdf.ed.ac.uk/s1707343/kialo-pta24/-/tree/main

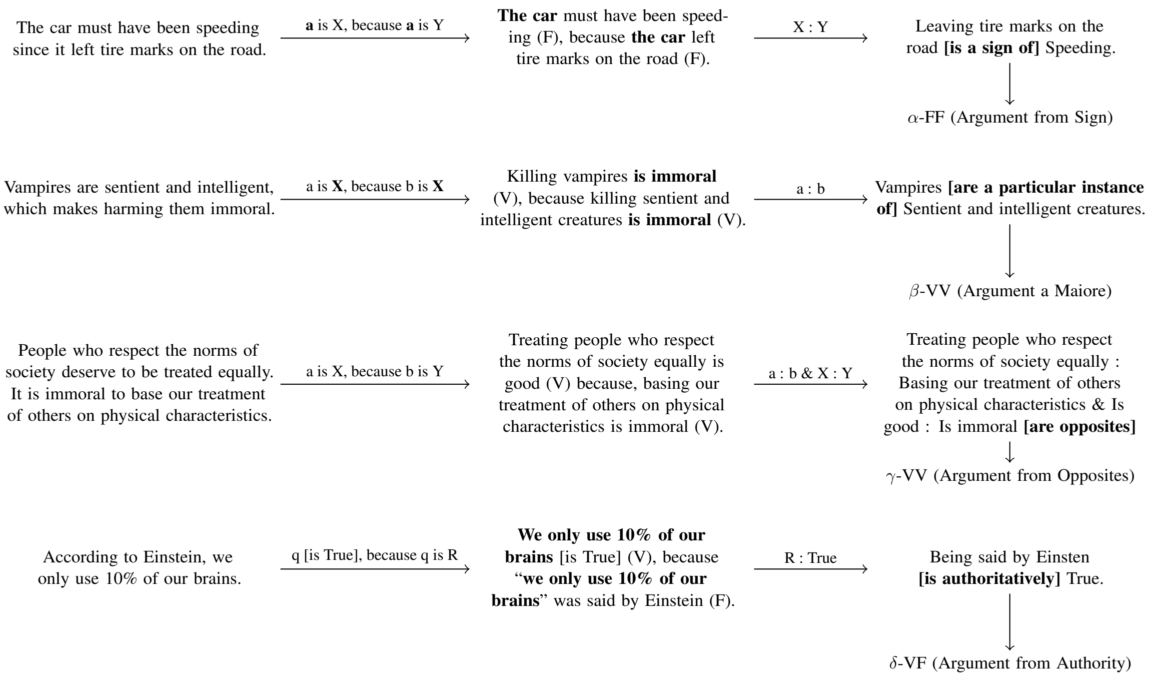

The car must have been speeding since it left tire marks on the road.
a is X, because **a** is Y
The car must have been speeding (F), because **the car** left tire marks on the road (F).
X : Y
Leaving tire marks on the road **[is a sign of]** Speeding.
α-FF (Argument from Sign)
Vampires are sentient and intelligent, which makes harming them immoral.
a is **X**, because b is **X**
Killing vampires **is immoral** (V), because killing sentient and intelligent creatures **is immoral** (V).
a : b
Vampires **[are a particular instance of]** Sentient and intelligent creatures.
β-VV (Argument a Maiore)
People who respect the norms of society deserve to be treated equally. It is immoral to base our treatment of others on physical characteristics.
a is X, because b is Y
Treating people who respect the norms of society equally is good (V) because, basing our treatment of others on physical characteristics is immoral (V).
a : b & X : Y
Treating people who respect the norms of society equally : Basing our treatment of others on physical characteristics & Is good : Is immoral **[are opposites]**
γ-VV (Argument from Opposites)
According to Einstein, we only use 10% of our brains.
q [is True], because q is R
We only use 10% of our brains [is True] (V), because "**we only use 10% of our brains**" was said by Einstein (F).
R : True
Being said by Einsten **[is authoritatively]** True.
δ-VF (Argument from Authority)

Figure 1: Example of arguments analysed for each form. The first step is to rewrite the arguments in canonical form, followed by the identification of the lever (the inference that connects the non-identical aspects of the canonicalised argument). By identifying the lever we obtain the name of the argument from the periodic table.

the PTA[2]:

- Fact: *"a description of a particular state of affairs that is or can be empirically observed in reality or can be imagined to exist in a particular universe of discourse or level of abstraction."*
- Value: *"an evaluative judgment about something based on a definition or assessment criteria"*
- Policy: *"a directive or hortative statement that expresses advice to do something"*

Wagemans defines argument canonicalisation as the process of rewriting an argument in one of four standard forms (Wagemans, 2021). Each canonical form can be represented as 'Conclusion because Premise' as shown in Table 1. Therefore, we experiment with automated argument canonicalisation by framing the problem as a text generation task similar to paraphrasing. In our specific task, we require that the paraphrased argument adheres to the structure of the canonical form.

In Table 1, a and b refer to the argument's subject(s) while X and Y refer to the predicate being applied to the subject. Examples of canonical forms in natural language are given in Figure 1. The delta quadrant is unique in that it conveys arguments that seek to justify the truth of a statement based on some quality of the statement itself e.g. 'q [is True], because everyone says so' or '$\neg q$ [is True], because the person who said it is known to be a liar'.

Alpha: **a** is X, because **a** is Y
Beta: a is **X**, because b is **X**
Gamma: a is X, because b is Y
Delta: **q** [is True] because **q** is R

Table 1: Representation for the four canonical forms

3 Related Work

Annotating argument schemes Experimental work on classifying argument schemes is extremely limited and highly understudied. Walton and Macagno (2015) and Eemeren and Kruiger (2011) both describe classification procedures for identifying schemes; however, few works systematically apply these procedures to real-world data. The periodic table's type identification procedure (ATIP) (Wagemans, 2021) has gone through several iterations of refinement, with an earlier version being compared alongside Walton schemes to measure the ability of annotators to agree on the classification of arguments according to each taxonomy respectively. Visser et al. (2021) annotate

[2] Definitions are taken directly from Wagemans (2021)

a set of US presidential debates from 2016 with argument schemes, both from Walton's taxonomy as well as the periodic table. However, they fail to provide the intermediary steps that are required by the PTA and only give the final classification. Feng and Hirst (2011) uses a small dataset of annotated arguments to train a decision tree to classify arguments in one of five Walton schemes with moderate results. The dataset they used is no longer accessible. Ruiz-Dolz et al. (2024) is the only existing, multi-topic dataset of argument schemes that is publically available, this dataset consists of arguments generated to match Walton's argument schemes.

Defeasible Textual Entailment Defeasibility in non-monotonic logic describes the notion of an inference that is valid on the basis of currently available evidence. Unlike in classical logic, defeasible inference in non-monotonic logic allows the inference to be retracted based on further evidence or in the case of an exception to a rule. Rudinger et al. (2020) recognise the fact that the majority of natural language inference is defeasible and construct a defeasible NLI dataset of defeasible inferences in English across a range of everyday topics. However, their focus is on generating defeasible hypotheses based on a premise rather than determining the nature of the defeasible inference as we focus on here.

Identifying Enthymemes Arguments in which the inference relies on implicit premises are known as *enthymemes*. The identification and disambiguation of enthymemes is quite closely related to that of classifying argument schemes since the argument scheme determines the auxiliary premises that are required for the argument to work. Habernal et al. (2018) make progress towards the automatic reconstruction of implicit warrants by training a model to choose the correct warrant from a list of confounding options. Beyond classification, Chakrabarty et al. (2021) generate implicit premises that support a given conclusion, while Saadat-Yazdi et al. (2022, 2023) generate sequences of commonsense reasoning that connect the premises to the conclusion. We believe that the automatic classification of arguments will help researchers to achieve better results in this task, since identifying the structure of an argument automatically can assist models in discovering implicit premises.

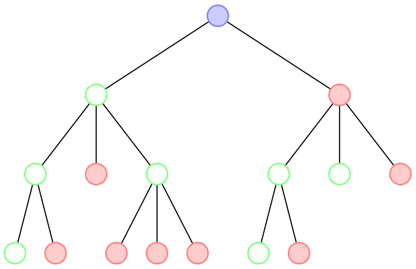


Figure 2: Example of a partially expanded Kialo debate tree. The blue node is the topic of the debate, with pros and cons shown in red and green, respectively.

Argument Substance The detection of argument substance is an existing and well-known task in argument mining (Niculae et al., 2017; Bao et al., 2021). The function of the substance in determining the argument type and its role in downstream tasks is, however, unique to the periodic table.

4 Data

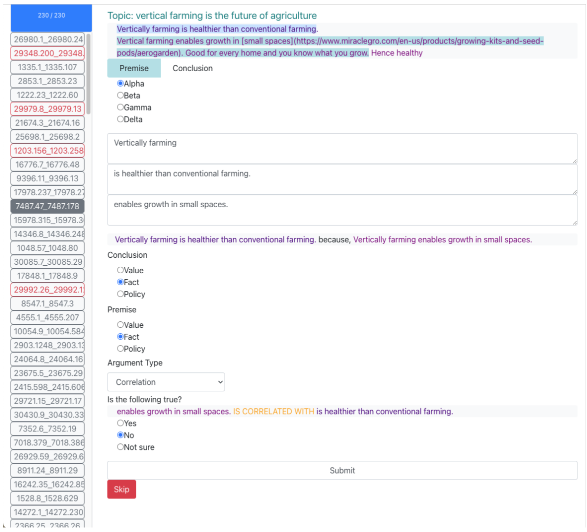
There are no publicly available datasets that researchers could benefit from to work on the automatic classification of argument schemes (Section 3) according to the Periodic Table of Arguments. To construct a new dataset, we began with the annotated data from Jo et al. (2021). This dataset consists of a scrape from Kialo[3]. Kialo is a structured debating website that allows users to provide pro and con claims for various topics. Each topic has a main claim for which pros and cons can be provided. Each pro and con is then viewed as a new claim which can have its own sub-pros and sub-cons. This creates a tree-like debate structure for each topic as in Figure 2.

We chose this dataset due to its structured nature, allowing us to consider claims as conclusions with pros as premises. Additionally, each claim in Kialo is relatively self-contained and can usually be understood without taking into consideration the rest of the discussion. This allows us to focus our study on inferential structures that are explicitly present in the text rather than being implied by context. A datasheet according to Gebru et al. (2018) is provided in Appendix B.

4.1 Dataset Creation

To construct our dataset we began with a sample of 760 topics from the dataset presented in Jo et al. (2021). From each topic, we sampled one pair of supporting claims (i.e., a claim and a pro). Each

[3] https://kialo.com



Topic: vertical farming is the future of agriculture

Vertically farming is healthier than conventional farming.
Vertical farming enables growth in [small spaces](https://www.miraclegro.com/en-us/products/growing-kits-and-seed-pods/aerogarden). Good for every home and you know what you grow. Hence healthy

Premise Conclusion

● Alpha
○ Beta
○ Gamma
○ Delta

Vertically farming

is healthier than conventional farming.

enables growth in small spaces.

Vertically farming is healthier than conventional farming. because, Vertically farming enables growth in small spaces.

Conclusion

○ Value
● Fact
○ Policy

Premise

○ Value
● Fact
○ Policy

Argument Type

Correlation

Is the following true?

enables growth in small spaces. IS CORRELATED WITH is healthier than conventional farming.

○ Yes
● No
○ Not sure

Submit

Skip

Figure 3: Screenshot of the ArgNotator Interface

pair of claims was then annotated for the following features: (i) The canonical form of the argument, (ii) The premise and conclusion rewritten to match the structure of the canonical form, (iii) The substances of rewritten premise and conclusion, (iv) The lever used by the argument in canonical form, and (v) Whether the annotator considers the lever to be valid.

Experts in computational argumentation are rare, particularly those who are knowledgeable about the periodic table of arguments. This rarity raises the cost of paying annotators making the construction of such corpora difficult. To make the construction of argument type corpora more accessible there is a need for annotation tools that simplify argument type identification and training materials that elaborate on the existing Argument Type Identification Procedure (Wagemans, 2021). To address this, we designed a specially tailored web-based annotation tool called ArgNotator and have made it publicly available for use on other domains. In Figure 3, we show an example annotation using ArgNotator. The tool streamlines the process of applying the periodic table and breaks down the process of determining argument types into several subtasks to simplify the annotation procedure for non-experts.

4.2 Annotator Training

Four annotators were involved in this study; the lead author acted as the expert annotator, and three non-expert annotators were recruited from within our institution for this task. Our selection criteria included the following: (i) being fluent in English, (ii) having some familiarity with Western culture and current affairs, since the dataset included a range of topics.

The expert annotator provided training in argumentation theory and the application of the PTA for argument analysis to the non-expert annotators. The training of annotators involved pre-reading a set of prepared and pre-existing material as well as a two-hour in-person seminar. The seminar was designed such that for each of the subtasks that the annotators were required to perform, they were provided with several worked examples that demonstrated how they should deal with ambiguous cases and a problem sheet. After filling in each problem sheet annotators were then asked to discuss their solutions, and seek to agree on an answer before moving on to ensure good agreement when finally performing the annotation.

After the training session, each annotator was given the same 10 examples to work through independently. After annotating these 10, Cohen's Kappa (κ) (McHugh, 2012) was checked between each non-expert annotator and the expert annotator. If the agreement between the non-experts and the expert fell below $\kappa = 0.75$ the annotators were invited to discuss their answers, and resolve disagreement and given another 10 examples to work through. This was repeated until the required agreement was achieved. In the case of our study, this took two rounds. After training and preliminary annotations, each annotator was given 250 examples to work from, 100 of these being shared across all annotators and 150 being unique.

4.3 Dataset Distribution

The analysis of Table 2 reveals valuable insights into the distribution and characteristics of argument forms in Kialo. The alpha quadrant emerges as the most prevalent, with a total of 331 occurrences, emphasizing its significance in the dataset. Within the alpha quadrant, alpha-vf stands out as the most frequent subcategory, demonstrating the diversity of lever types employed in constructing arguments. In contrast, the beta quadrant has a lower total occurrence (42), indicating a lesser prevalence of beta-type arguments, with beta-ff being the dominant subcategory. The gamma quadrant showcases a substantial number of occurrences (105), reflecting the varied nature of gamma-type arguments, with gamma-pf being the most prevalent subcategory. The delta quadrant, with a total of 8 occurrences, suggests that delta-type arguments are less common in the dataset, and delta-ff is the primary

Form	Type	Lever	#
alpha	alpha-ff	sign	60
		cause	57
		correlation	24
		effect	22
		N/A	11
	alpha-fp	N/A	1
	alpha-fv	N/A	3
	alpha-pf	pragmatic	63
		N/A	3
	alpha-pp	evaluation	1
		N/A	1
	alpha-pv	evaluation	7
		deontic	4
	alpha-vf	criterion	67
		N/A	3
	alpha-vp	N/A	1
	alpha-vv	axiological	2
		N/A	1
	total		**331**
beta	beta-ff	example	25
		genus	5
		similarity	4
		N/A	1
	beta-pp	comparison	3
	beta-vv	a maiore	2
		parallel	1
		a minore	1
	total		**42**

Form	Type	Lever	#
gamma	gamma-ff	N/A	15
		petitio principii	11
		opposites	5
		disjunctives	2
	gamma-fv	petitio principii	2
		N/A	1
		disjunctives	1
	gamma-pf	consistency	19
		N/A	9
	gamma-pp	N/A	1
	gamma-pv	N/A	5
	gamma-vf	N/A	13
		tradition	9
	gamma-vv	N/A	4
		opposites	4
		disjunctives	2
		petitio principii	2
	total		**105**
delta	delta-ff	N/A	5
	delta-vf	authority	2
		ad populum	1
	total		**8**

Table 2: Distribution of argument types within the dataset.

	Substance			Form			
Split	F	V	P	α	β	γ	δ
Train	215	96	102	275	37	95	6
Test	40	19	14	56	5	10	2

Table 3: Distribution of substance and form classes for train (413 instances) and test sets (73 instances). Argument forms belong to one of four classes: α, β, γ, and δ. The three substances are either (F)act, (V)alue or (P)olicy (see Section 2.2).

subcategory within this quadrant. The presence of "N/A" entries in the lever column indicates instances where the argument did not fit into one of the existing types of the periodic table, there are a total of 74 arguments for which a lever was not found.

We construct a gold standard test set from the 73 arguments annotated by all annotators. We use the expert annotator's labels to define the labels of this set. The distribution of the data across the training and testing split is shown in Table 3. We note that the training and test sets exhibit similar distributions for all classes, suggesting that the test set accurately represents the data.

4.4 Agreement statistics

Even after training and ensuring high agreement, our annotators still experienced an agreement drift as they proceeded with the annotation. Due to the high disagreement with other annotators, we chose to omit the annotations of one annotator from our dataset. In other words, we used the annotations from one expert and two non-expert annotators. Of the remaining 550 arguments, 54 could not be labelled because they did not contain obvious arguments. What remains is a set of 486 arguments, 73 of which were labelled by all annotators.

Due to the imbalance of classes, we must be careful to choose an agreement statistic that allows for multiple annotators and does not overly penalize mistakes, a result of Cohen's Kappa Paradox (Zec et al., 2017). For this reason, we use Gwet's AC1 (Gwet, 2014) which is more stable than alternative measures to class imbalances. The agreement results are shown in Table 4, we see that for most tasks, we achieve fair to moderate agreement.[4]

It is interesting to note that even the classification of argument form appears to be very challenging, with annotators achieving only moderate agreement. From our analysis of the data, we observe that discrepancies in argument form mainly appear between annotators disagreeing between the *gamma* form and *alpha/beta*. This is often due to annotators disagreeing on whether anaphora between subjects or predicates (a, b or X, Y in Figure 1) can be resolved. This suggests the need to refine the definition of anaphora in this context.

Feature	Agreement	Valuation
Form	58.28	Moderate
Conc. Subs.	74.34	Substantial
Prem. Subs.	75.10	Substantial
Type	24.71	Fair
Validity	50.1	Moderate
Conc. Rewrite*	79.93	-
Prem. Rewrite*	70.36	-

Table 4: Gwet's AC1 agreement statistic for classification tasks. We also provide Landis and Koch (1977)'s interpretations of these values. (*) For rewriting tasks, we use the average Rouge-1 score to measure agreement between annotators.

5 Experiments

In this section, we will evaluate various approaches to learning the steps in the annotation process. In doing so, we wish to obtain a set of benchmarks that characterise the difficulty of the various tasks.

5.1 Detecting the substance

We benchmark the performance of pre-trained language models (PLMs) on detecting argument substance by combining the rewritten premises and conclusions along with their respective substance annotations into a single dataset. We randomly sampled 10% of the training data to use for model validation and trained each model for 10 epochs.

Model	P	R	F1	Acc.
Human*	**81.9**	76.0	78.6	84.0
BERT-Large	76.7	78.6	75.1	84.9
Roberta-Large	78.7	**86.9**	**81.1**	**88.5**
DeBERTa-V3-Large	48.7	65.7	56.4	88.4

Table 5: Macro averaged F1, Precision, Recall and Accuracy of PLMs on the substance classification subtask. (*) Human performance is given by measuring the best performance of the non-expert annotators against the gold standard expert annotations.

Table 5 summarizes supervised classification results for substance classification in an argumentation dataset. Human annotators achieve F1 score 78.6% and Accuracy 84.0%. Among models, RoBERTa-Large performs best with an F1 score of 81.1%, and Accuracy 88.5%. BERT also shows strong performance, while DeBERTa-V3-Large lags behind in Precision (48.7%), Recall (65.7%), F1 score (56.4%), with comparable Accuracy (88.4%). This highlights RoBERTa's effectiveness in capturing argument substance, surpassing both non-expert annotators and other pretrained language models in this classification task. These results reflect the high annotator agreement suggesting that the argument substance is easy to classify.

5.2 Canonicalising arguments

We now investigate the performance of PLMs on argument canonicalisation with further experiments. Argument canonicalisation is the task of rewriting an argument in one of the four standard argument forms: alpha, beta, gamma or delta (Wagemans, 2021). Due to the lack of examples in the beta, gamma, and delta forms, we focus our study in this paper on the alpha quadrant (a is X, because a is Y) to demonstrate the feasibility of this task. Consider the following argument from our dataset:

Claim: *All humans should be vegan.*

Pro: *Veganism reduces both human and animal suffering.*

[4] While we used Cohen's Kappa to measure agreement during training, we were only measuring agreement between two annotators, the non-experts and the expert. Additionally, we were not aware of the class imbalance as the data had not been annotated yet, hence the change of agreement statistic.

	Rouge-1			Rouge-2			RougeL		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Human*	81.3	78.2	78.6	68.9	66.5	66.8	75.8	73.3	73.6
FlanT5-large (fine-tuned)	74.4	73.3	**72.1**	60.1	60.0	**59.5**	58.6	68.4	**67.6**
Llama2-7b (fine-tuned)	42.15	46.41	41.93	22.72	23.98	22.15	31.88	34.31	31.37
FlanT5-large (5-shot)	77.23	50.10	56.80	57.89	36.19	41.64	66.65	41.30	47.77
Llama2-13b (5-shot)	56.90	72.98	56.87	36.96	51.11	39.12	44.52	57.64	44.97
FlanT5-large (10-shot)	77.08	50.09	56.76	57.80	36.19	41.50	66.87	41.39	48.05
Llama2-13b (10-shot)	57.09	72.90	57.06	37.15	51.44	39.32	44.44	57.66	44.85
FlanT5-large (15-shot)	76.98	50.14	56.80	57.68	36.13	41.37	66.77	41.31	47.97
Llama2-13b (15-shot)	56.98	72.52	56.72	36.85	50.80	39.07	44.57	57.58	44.98

Table 6: Performance comparison of models on alpha argument canonicalization showing performance for both supervised fine-tuning and few shot approaches. (*) Human performance is given by measuring the best performance of the non-expert annotators against the gold standard expert annotations.

This argument is canonicalised as: *Veganism should be upheld by all humans, because veganism reduces both human and animal suffering.* This task poses several interesting challenges, one needs to first identify the main subject of the argument (a), identify all the terms that refer to the subject ("*veganism*", "*vegans*"), identify the argument predicates (X and Y), and finally modify the voice of the text to match the structure of the canonical form (alpha).

In order to ensure that our models generate arguments that match the form, we found that requiring the model's output to conform to a list of assignments works relatively well. In other words, instead of generating the canonical argument in free-form, we provide examples that represent the canonical form as a JSON object such as "`{'a': Veganism, 'X': should be practised by all humans, 'Y': reduces both human and animal suffering}`". In our experiments, we compare two open-source large language models (FLAN-T5 & LLAMA2) (Raffel et al., 2023; Touvron et al., 2023) in both pre-trained and few-shot scenarios. For the few-shot setting, we randomly sample 5, 10 & 15 training examples and embed them in an appropriate prompt (see Appendix A). For fine-tuning both models were trained for eight epochs. The model with the best validation Rouge-1 was chosen for evaluation on the test set. Due to the computational costs of training and inference of LLMs, we only report results for a single run.

The results in Table 6 suggest that pre-trained LLMs are not able to model argument canonicalisation well. Fine-tuning appears to be necessary to achieve meaningful performance. Even with fine-tuning Llama2's outputs performing the worst of all the models, an analysis of generated outputs shows that the fine-tuned Llama2 is prone to generating nonsensical phrases while conforming to the required output structure. The number of examples given to the few-shot prompted models also seems to have little impact on the result. The results of fine-tuned Flan-T5, however, suggest that an encoder-decoder framework may be the most suitable for this task and provides a promising direction for future research. Generation outputs of the models are shown in the Appendix C.

	Rouge-1	Rouge-2	Rouge-L
a	64.4	42.5	64.5
X	**75.6**	**67.9**	**75.30**
Y	59.6	50.8	58.0

Table 7: Breakdown of performance of fine-tuned Flan-T5 on test set. After the manual cleanup of 5 incorrectly generated samples.

Table 7 shows that the model performs best at identifying conclusion predicates but struggles with premise predicates. This is corroborated by the low Rouge-1 between annotators for the argument premise in Table 4.

6 Conclusion

We constructed the first multi-topic dataset of argument types by using the theory of the periodic table of arguments. This dataset was created as a result of an annotation study, in which the annotators used our developed tool (ArgNotator). The focus of our experiments was on the canonical representation and classification of argument forms and substances. We showed that substance classification can be done effectively using BERT-based models, while fine-tuned LLMs provide a promising approach to argument canonicalisation.

The next steps involve detecting and classifying the argument lever, considering both its form and substance, to contribute to a more nuanced machine understanding of different argument types. The research also aims to include a critical evaluation of argument quality by posing relevant critical questions that assess the robustness and coherence of identified arguments. Having the means to measure the quality of an argument systematically opens the doors to automatically constructing Weighted Argumentation Frameworks (Amgoud and Ben-Naim, 2018) that allow computational methods to be used to evaluate the strength of arguments within the context of a discourse/debate. This work provides some initial steps towards bridging the gap between natural language inference and symbolic approaches to computational argumentation.

Limitations

The dataset we have constructed is only based on Kialo which encourages users to be more thoughtful about their arguments. This means that the distribution of the dataset is not representative of online argumentation in general where we would expect to see more delta arguments, the majority of which are stereotypically viewed as fallacious. Additionally, while we would like to encourage accessibility of verifying our results, we found that we were not able to train effective models for argument canonicalisation using smaller language models. This meant that we had to focus our canonicalisation experiments on large language models. Due to the distribution of argument forms in our dataset, we are also unable to present a full table of results for the canonicalisation of beta, gamma and delta forms. We have also omitted discussions on lever detection, as we believe that this requires substantial additional work and goes beyond the scope of this current study.

Ethics

Our dataset is built from the publicly available corpus provided by Jo et al. (2021). For the annotation work, ethics approval was obtained from our institution, and annotators were paid a standard rate of £16/hr for their time. This time includes both annotation time and training.

Acknowledgements

The work was supported by the Edinburgh-Huawei Joint Lab and Huawei's grant CIENG4721/LSC. We would like to thank the annotators for their time and diligence as well as Edinburgh University's CHAI Lab for offering feedback in the initial stages of the development of this work.

References

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-End Argumentation Knowledge Graph Construction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7367–7374.

Leila Amgoud and Jonathan Ben-Naim. 2018. Weighted Bipolar Argumentation Graphs: Axioms and Semantics. In *IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5194–5198. Association for Computing Machinery.

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6354–6364, Online. Association for Computational Linguistics.

A. C. Braet. 2005. The Common Topic in Aristotle's Rhetoric: Precursor of the Argumentation Scheme. *Argumentation*, 19(1):65–83.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Frans H van Eemeren and Tjark Kruiger. 2011. 8 Identifying Argumentation Schemes. In *8 Identifying Argumentation Schemes*, pages 70–81. De Gruyter Mouton.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *CoRR*, abs/1803.09010.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Nadin Kökciyan, Isabel Sassoon, Elizabeth Sklar, Sanjay Modgil, and Simon Parsons. 2021. Applying Metalevel Argumentation Frameworks to Support Medical Decision Making. *IEEE Intelligent Systems*, 36(2):64–71. Conference Name: IEEE Intelligent Systems.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Ramon Ruiz-Dolz, Joaquin Taverner, John Lawrence, and Chris Reed. 2024. Nlas-multi: A multilingual corpus of automatically generated natural language argumentation schemes.

Ameer Saadat-Yazdi, Xue Li, Sandrine Chausson, Vaishak Belle, Björn Ross, Jeff Z. Pan, and Nadin Kökciyan. 2022. KEViN: A knowledge enhanced validity and novelty classifier for arguments. In *Proceedings of the 9th Workshop on Argument Mining*, pages 104–110, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kokciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2484–2495, Dubrovnik, Croatia. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng 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.

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

Jean H M Wagemans. 2021. Argument Type Identification Procedure (ATIP) – Version 4.

Jean H.M. Wagemans. 2016. Constructing a Periodic Table of Arguments. *SSRN Electronic Journal*.

Douglas Walton and David Godden. 2005. The Nature and Status of Critical Questions in Argumentation Schemes. *OSSA Conference Archive*.

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. Google-Books-ID: qc3LCgAAQBAJ.

Slavica Zec, Nicola Soriani, Rosanna Comoretto, and Ileana Baldi. 2017. High agreement and high prevalence: The paradox of cohen's kappa. *Open Nurs J*, 11:211–218.

A Prompt Template for Few-Shot Training

For reproducibility, we provide the prompt template used for our few-shot experiments. Text in **bold** indicates values to be filled in during preprocessing.

Rewrite the following argument in canonical form a is X because a is Y, give your answer in JSON format {\'a\': a, \'y\': y, \'x'\': x }

Here are a few examples:
Conclusion: **EXAMPLE1_CLAIM**
Premise: **EXAMPLE1_PRO**
Answer: **EXAMPLE1_GOLD**
Conclusion: **EXAMPLE2_CLAIM**
Premise: **EXAMPLE2_PRO**
Answer: **EXAMPLE2_GOLD**
. . .
Conclusion: **CLAIM**
Premise: **PRO**
Answer:

B Datasheet for Kialo-PTA24

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
There is currently a lack of multi-topic datasets that assign argument schemes to arguments found online. Of the datasets that exist, very few apply Wagemann's Periodic Table of Arguments as the taxonomy of choice and, to our knowledge, none provide intermediate annotations of canonical forms that aid the identification of the argument scheme.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
Anonymous

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
Anonymous

Any other comments?
N/A

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The dataset is comprised of pairs of claims from Kialo.com.

How many instances are there in total (of each type, if appropriate)?
The dataset consists of 497 pairs of claims with corresponding annotations.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset consists of a sample of the claims present on Kialo. The data was sampled randomly from a scrape of Kialo performed by (Jo et al., 2021). The data is representative in the cases where arguments were found since we sampled from the total set of arguments, however, a sizeable portion of the data was not possible to annotate due to the lack of a clear argument. These instances were omitted as they were not relevant to our study.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Claims are given in raw markdown as given by Jo et al. (2021).

Is there a label or target associated with each instance? If so, please provide a description.
Each instance contains the following annotations: (i) Canonical form [one of four classes] (ii) Original text rephrased into the canonical form (iii) Premise and conclusion after canonicalisation (iv) Substance of premise after canonicalization [one of three classes] (v) Substance of conclusion after canonicalization [one of three classes] (vi) Type of argument based on the Periodic Table of Arguments [one of 36 possible classes] (v) Validity of the argument lever [either true, false or n/a]

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Individual instances are taken from separate discussions and so are unrelated.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
We provide a recommended training and testing split. The testing split is the gold standard defined by an expert annotator and has been validating by measuring agreement with the non-expert annotators.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
Two of three annotators were exposed to the annotation scheme for the first time, there is a possibility for misinterpretation of certain class definitions and mistakes in canonicalization. These were mitigated by performing a prior training but may naturally be present regardless.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
Yes, the arguments were obtained from a debating website where many controversial topics are discussed. These are representative of the kinds of discussion that appear on the web and are therefore essential to the study of online argumentation.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes.

Does the dataset identify any subpopulations

(e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
It is possible to identify the original user who posted a claim by cross-referencing the dataset with Kialo.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
The dataset reveals beliefs and personal opinions that people have made public on Kialo, however, these can only be linked to a user's ID and not necessarily their actual name.

Any other comments?
N/A

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The data was directly observed.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.
The original data was crawled in 2020 by Jo et al. (2021). The annotation was performed between November and December of 2023.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
N/A.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*(Strubell et al., 2019) for approaches in this area.)
Approx. $1200 for annotator compensation.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
We first sampled a pair of claims from each discussion with uniform probability. This resulted in a set of roughly 1700 arguments, this was then downsized to a set of 650 arguments by random uniform sampling.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
Three PhD students in computer science were recruited from within our institution, in addition to the main author who acted as the expert annotator. The three students were paid £16/hr for 20 hours of work.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Yes, we completed an ethics request form within our institution which was approved by a panel of experts. The link to the form will be made available.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this

section.
Yes.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Data was obtained via third parties, namely Jo et al. (2021) who scraped the data from the Kialo website.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
The individual posters did not consent to their data being used. However, the annotators for our dataset were given consent forms and a participant information sheet which they were asked to read and sign.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Any other comments?
N/A

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
The data was labelled for argumentative features by annotators as detailed in previous sections. Instances were removed based on there being no relevant argument to analyse.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Yes. Link pending.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Our tool ArgNotator was used for labelling. Link pending.

Any other comments?
N/A

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Those listed in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
Argument identification, generation and evaluation.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that

could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.

Are there tasks for which the dataset should not be used? If so, please provide a description.
This dataset should not be used for de-anonymisation tasks that seek to reveal the identify of Kialo users.

Any other comments?
N/A

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
Yes, the dataset will be maintained on an internally hosted GitLab instance.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
Yes, pending paper acceptance.

When will the dataset be distributed?
Upon the approval of the paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
We aim to publish under an Attribution CC-BY licence.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

Any other comments?
N/A

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
Main author of the paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Main author email will be made available.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
We may provide extensions to the dataset and publicise improvements in peer-reviewed venues.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Yes, all versions of the dataset will be maintained

in parallel.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Others may extend the dataset or continue work on it and publish their results independently.

Any other comments?
N/A

C Examples

The examples in Table 8 demonstrate the behaviour of the various models on argument canonicalisation. FlanT5 appears to consistently produce text in the correct format, and Llama2 with prompting similarly. FlanT5 with prompting however, consistently fails to produce text in the correct structure. However, we see that with finetuning there is a drastic shift, the model consistently generates the output in the correct format while also providing meaningful canonicalisation.

Source	**Text**
Claim **Pro**	Student fees are already too expensive and saddle people with a lifetime of debt. The average student in the UK will owe 50,000 GBP when graduating.
FlanT5-large (finetuned)	’a’: ’Student fees’, ’y’: ’are already too expensive and saddle people with a lifetime of debt’, ’x’: ’are already too expensive and saddle people with a lifetime of debt’
Llama2-7b-hf (finetuned)	’a’: ’students fees’, ’b’: None, ’x’: ’are already too expensive and snake people with a life of debate.’, ’y’: ’will have 10,0% of alternate student in the UK when you are able to be able%%%%%%%% of student unions.
FlanT5-large (5-shot)	’The average student in the UK will owe 50,000 GBP when graduating’
Llama2-7b-hf (5-shot)	’a’: ’Student fees are already too expensive and saddle people with a lifetime of debt.’, ’y’: ’The average student in the UK will owe 50,000 GBP when graduating.’ , ’x’: ’Student fees are already too expensive’
Claim **Pro**	Russia is directly interfering with the governance of other countries through assassination, coup d’etats and intimidation. In 2016, Russian intelligence operatives plotted to overthrow the democratically elected government of Montenegro and murder its prime minister.
FlanT5-large (finetuned)	’a’: ’Russia’, ’y’: ’plotted to overthrow the democratically elected government of Montenegro and murder its prime minister’, ’x’: ’is directly interfering with the governance of other countries through assassination, coup d’etats and intimidation’
Llama2-7b-hf (finetuned)	2016% of Russian intelligence operation and anxietyiety’, ’b’: None, ’x’: ’is directly interfering with the governance of other countries through the Taliban party and anxiety
FlanT5-large (5-shot)	"Russia is directly interfering with the governance of other countries through assassination, coup d’etats and intimidation."
Llama2-7b-hf (5-shot)	’a’: ’Russia is directly interfering with the governance of other countries through assassination, coup d’etats and intimidation.’, ’y’: ’In 2016’
Claim **Pro**	Firebase has good security. If the website itself gets breached, they wont be able to access the sensitive information.
FlanT5-large (finetuned)	’a’: ’Firebase’, ’y’: ’has good security’, ’x’: ’is not vulnerable to breaches’
Llama2-7b-hf (finetuned)	’a’: 19% of alternate reality games’, ’b’: None, ’x’: ’has good security’, ’y’: ’has good security
FlanT5-large (5-shot)	’Firebase has good security. If the website itself gets breached, they wont be able to access the sensitive information.’
Llama2-13b-hf (5-shot)	’a’: ’Firebase has good security’, ’y’: ’the website itself gets breached’, ’x’: ’they wont be able to access the sensitive information’

Table 8: Random sample of generations from the various models in our experiments.