Natural Language Reasoning in Large Language Models: Analysis and Evaluation

Debela Gemechu, Ramon Ruiz-Dolz, Henrike Beyer and **Chris Reed**
Centre for Argument Technology (ARG-tech)
University of Dundee
Dundee DD1 4HN, United Kingdom
{d.t.z.gemechu, rruizdolz001, 2579207, c.a.reed}@dundee.ac.uk

Abstract

While Large Language Models (LLMs) have demonstrated promising results on a range of reasoning benchmarks—particularly in formal logic, mathematical tasks, and Chain-of-Thought prompting—less is known about their capabilities in unconstrained natural language reasoning. Argumentative reasoning, a form of reasoning naturally expressed in language and central to everyday discourse, presents unique challenges for LLMs due to its reliance on context, implicit assumptions, and value judgments. This paper addresses a gap in the study of reasoning in LLMs by presenting the first large-scale evaluation of their unconstrained natural language reasoning capabilities based on natural language argumentation. The paper offers three contributions: (i) the formalisation of a new strategy designed to evaluate argumentative reasoning in LLMs: argument-component selection; (ii) the creation of the Argument Reasoning Tasks (ART) dataset, a new benchmark for argument-component selection based on argument structures for natural language reasoning; and (iii) an extensive experimental analysis involving four different models, demonstrating the limitations of LLMs on natural language reasoning tasks.

1 Introduction

The question of whether Large Language Models (LLMs) can perform reasoning is a thorny one. Not only have there been a wide range of studies exploring the issue (and coming to wildly different conclusions), but techniques such as Chain of Thought prompting (CoT) (Wei et al., 2022) and multi-hop Question-Answering (Yang et al., 2018; Zhu et al., 2024), that purport to place reasoning at the forefront of LLM interaction, have generated remarkable performance enhancements and demanding challenge tasks (Chu et al., 2024). Coupled with high profile marketing touting LLM reasoning capabilities[1] and anecdotal evidence of both spectacular success and spectacular failure, it is no wonder there has been such an explosion of work in trying to fairly assess reasoning competence in LLMs (Miao et al., 2020; Cobbe et al., 2021; Patel et al., 2021; Talmor et al., 2021; Geva et al., 2021; Mirzadeh et al., 2024; Han et al., 2024; Valmeekam et al., 2024; Mehrafarin et al., 2024; Paruchuri et al., 2024; Tyagi et al., 2024a; Samadarshi et al., 2024; Shiri et al., 2024; Chen et al., 2021; Luo et al., 2023).

To date, however, all of this work has largely centred on structured problem-solving tasks such as arithmetic, logic puzzles, and theorem proving, or has focused on single-domain and basic inferential reasoning adapted into natural language. These approaches do not fully capture the complexities of human natural language reasoning as it naturally occurs in argumentation. Though a focus on such problems offers an opportunity to carefully control variability under laboratory conditions, it also risks seriously misrepresenting LLM performance with respect to realistic human reasoning. Even (Guan et al., 2023), who demonstrate deep weaknesses in current LLM capacity, rest their argument on classical planning, a very narrow and tightly constrained type of reasoning. What is required is a vocabulary, a model, a dataset and a set of tasks that also cover natural, in-situ human reasoning, as it is expressed in language. This is the domain of argumentation theory (van Eemeren et al., 2014), and our goal is to leverage recent results in the area to equip us with the tools to assess LLM performance in realistic settings. While argumentative reasoning does not serve as the sole or definitive test of reasoning ability, it provides a valuable lens for examining LLMs' capacity to handle complex, naturalistic reasoning as expressed in everyday language—complementing existing approaches by ad-

[1] openai.com/index/learning-to-reason-with-llms/

dressing a dimension that remains comparatively underexplored.

We address this challenge by presenting the first large-scale evaluation of the natural language reasoning capabilities of LLMs through argumentation, a setting where reasoning inherently unfolds in natural language, extending beyond simple inferential processes. This paper has, therefore, the three following main contributions: (i) we formalise a new task that we define as argument-component selection which is designed to evaluate argumentative reasoning in LLMs; (ii) we create and release publicly the Argument Reasoning Tasks (ART) dataset, a new benchmark for argumentation reasoning consisting of 112,212 multiple-choice questions covering a total of sixteen different tasks addressing structural aspects of argumentation; and (iii) we present a complete set of experiments involving four open- and closed-weight models as well as a thorough analysis of the observed results.

2 Related Work

As pointed out in recent work, CoT reasoning can also be achieved without any prompt engineering, by just modifying the greedy decoding strategy to explore alternative top-k decoding paths (Wang and Zhou, 2024). This finding, though suggesting that the models reason intrinsically, can also be interpreted as the models not reasoning at all. Instead, they select from alternative sequence paths learned during training, and tuning the input prompt or adjusting decoding allows for selecting a different path than the one leading to the direct answer. Following this important finding, rather than focusing on how the output of the model is decoded, the focus should be on studying the model's ability to generalise and keep this behaviour labelled as "reasoning" when addressing problems of different nature and involving more complex and realistic reasoning than the ones that are commonly studied in the literature[2].

This aspect has been discussed in recent work (Valmeekam et al., 2024; Wu et al., 2024), where the reported results show that with minimal variations of the *standard* versions of the tasks included in the most popular benchmarks used for reasoning, the performance of LLMs drops significantly. These findings challenge the claims that LLMs can do reasoning and that it allows them to improve their performance in a broad range of tasks.

In addition to CoT and (multi-hop) QA-related tasks, other datasets and tasks to evaluate reasoning in LLMs have been proposed. ProofWriter (Tafjord et al., 2021) presents a dataset to evaluate deductive logical reasoning through formal logic problems. COPA (Roemmele et al., 2011) and its multilingual version XCOPA (Ponti et al., 2020) are two datasets created to evaluate causal reasoning by providing situations and asking to select the most likely outcome according to a cause-effect relationship. In this same direction, SWAG (Zellers et al., 2018) and HellaSWAG (Zellers et al., 2019) introduce two datasets to evaluate commonsense reasoning inference featuring adversarially generated scenarios in which models need to determine the most plausible option. The aNLI (Bhagavatula et al., 2019) dataset is proposed to investigate abductive reasoning. Again, the models are challenged to identify plausible outcomes for incomplete information scenarios. FOLIO (Han et al., 2024) consists of a collection of first order logic statements to evaluate the reasoning capabilities of LLMs. The models are asked to determine the truth values of a set of conclusions given some premises which are presented in both, natural language and first-order logic statements. From the reported results, it is possible to observe how LLMs struggle to solve this task. Finally, it is also worth mentioning other recent approaches, which have proposed the assessment of the reasoning capacities of LLMs based on games such as Minesweeper, grid puzzles, Sudoku or crosswords among others (Li et al., 2024; Tyagi et al., 2024b; Shah et al., 2024; Saha et al., 2024).

We can observe, however, that despite being focused on reasoning-related tasks, none of them address the problem of non-constrained reasoning in natural language (e.g., in argumentation), which is a fundamental aspect for evaluating and understanding the actual natural language reasoning capabilities of LLMs.

3 Theoretical Background

Arguments combine premises and conclusions to create complex reasoning structures. Argument theory distinguishes different structural combinations of premises and conclusions: serial (premises and/or conclusions are supported by premises themselves) (Beardsley, 1950), linked (multiple premises support a conclusion together in a combined inferential step) (Thomas, 1973), convergent

[2]And that, therefore, will most likely be also included in the training data of the latest versions of the popular LLMs.

(multiple premises independently support the conclusion), and divergent (same premise supports more than one conclusion). With these four types of argument structure, it is possible to analyse and understand argumentation in similar ways as multi-hop and CoT reasoning are commonly studied. These structures are also well-documented in foundational texts on argumentation theory (Walton, 2005; Groarke, 2004).

An important challenge in the evaluation of LLMs on argumentation skills is to ensure that the reasoning capacities are assessed instead of the dialogue generation abilities. Their training enables LLMs to create credible textual output based on probable token combinations. In an argument continuation task without sufficient limitations, the model will produce a probable continuation based on the input text. In this case, it is difficult to evaluate the appropriateness of the created continuation.

Our approach solves the evaluation challenge by introducing a multiple-choice task, a setup similar to the ones LSAT tests already used to measure the reasoning abilities of LLMs in logic games (Malik, 2024). By asking for one or more elements from a complex argumentative graph structure, the model needs to identify the correct continuation among a choice of options from the same argumentative context. This requires the ability to follow and reconstruct an implicit reasoning path. Tasks targeting larger chunks of argumentative elements require a model to choose an appropriate sub-structure as continuation, which demands deep understanding of necessary intermediate reasoning steps, similar to complex multi-hop Q&A tasks. Illustrative examples can be found in Table 23, and a visual representation is provided in Table 24.

4 Method

Aimed at providing, for the first time, a method to consistently evaluate the natural language reasoning capabilities of LLMs in argumentation, we formulate argument-component selection, consisting of a series of sixteen different argumentative reasoning tasks grouped into four different types of argumentative structures. This way, the proposed method allows us to evaluate the natural language reasoning capabilities of LLMs by asking them to build and reconstruct natural language arguments.

4.1 Task Formulation

An argument is represented as a structure consisting of a sequence of argument components $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ and the relations of inference and conflict between them $\mathcal{R} = \{\vdash, \multimap\}, \mathcal{R} : A \times A$. Our proposed tasks leverage the argumentative context, $\mathcal{C}$ to predict or generate a required argument component. To facilitate automatic evaluation and ensure consistency, we restrict the proposed tasks to selecting missing components from a predefined set of options. This constraint is crucial as open-ended generation poses challenges for evaluation, given that multiple valid components could fulfil the argument structure. By limiting the options, we allow the model to focus on identifying the most appropriate components while enabling reliable evaluation against gold-standard answers. We therefore define a series of Argumentative Reasoning Tasks as argument-component selection problems, where the model must identify the correct argument component(s) from a set of candidates to meet some set of structural argumentative criteria. The task requires filling the missing components of specific substructures while considering the entire argument as context. Accordingly, the model is provided with an argument as a context $\mathcal{C}$, a partially specified argument substructure (with missing argument components), and a candidate set $\mathcal{U} = \{u_1, u_2, \ldots, u_k\}$, which includes the correct answer $\hat{u}$. The objective is to select the correct missing component $\hat{u}$ by evaluating the candidates for their relevance and alignment with the given context $\mathcal{C}$. This process is formalized as:

$$\hat{u} = \arg\max_{u \in \mathcal{U}} \text{score}(u \mid \mathcal{C}),$$

where $\text{score}(u \mid \mathcal{C})$ measures the semantic and structural fit of the candidate u within the argument substructure. The next section outlines the instantiation of the argument-component selection formulation into the argumentative reasoning tasks included in our proposed evaluation method.

4.2 Argumentive Reasoning Tasks (ART)

We design a series of sixteen tasks based on four different types of argument structures: serial, linked, convergent, and divergent argument. Aimed at easing its understanding, a visual representation of the designed tasks can be found in Appendix A.

4.2.1 Serial Reasoning

In serial argument, an argument relation of inference ($\vdash$) is applied sequentially. The model is

tasked with identifying a conclusion, premise, or intermediate step based on the argument component(s) and the entire argument as a context. It includes the following six tasks:

One-hop Conclusion. With an argument relation of inference ($\vdash$), between a premise, α and a conclusion $\hat{\beta}$, a set is created of alternative potential conclusions, $\{\beta_1, \ldots, \beta_n\}$ (which when taken together with $\hat{\beta}$ is referred to together as the set B), from which the model must select. Treating the model as a function, f, the inputs are a set of argument components, plus a set of context, $\mathcal{C}$. The argument components in this case are the premise α, and the fact that the role to be played by the model's selection is as the conclusion of a $\vdash$ relation that has α as its premise. This role is expressed in the input by a metavariable X. The model's result is a binding of X to $\hat{\beta}$, one of the elements of B. Formally, given $B = \{\beta_0, \beta_1, \ldots, \beta_n\}$, $f(\{\alpha, \alpha \vdash X\}, \mathcal{C}) = \{X : \hat{\beta}\}$, where $\hat{\beta} \in B$.

One-hop Premise. The task in this case is to identify a premise $\hat{\alpha}$ given a conclusion β, where $\hat{\alpha}$ supports β ($\hat{\alpha} \vdash \beta$), from a set of alternative potential premises $A = \{\alpha_0, \alpha_1, \ldots, \alpha_n\}$, which includes the target premise $\hat{\alpha}$. Given the inputs β, and the fact that the model's role is to select a premise for a ($\vdash$) relation with β as its conclusion (denoted as Y), along with a set of context $\mathcal{C}$, the model f outputs the selected premise. The model's result is a binding of Y to $\hat{\alpha}$, one of the elements of A. Formally, given $A = \{\alpha_0, \alpha_1, \ldots, \alpha_n\}$, $f(\{\beta, Y \vdash \beta\}, \mathcal{C}) = \{Y : \hat{\alpha}\}$, where $\hat{\alpha} \in A$.

Two-hop Conclusion. In a two-hop argument, there are two sequential inference relations, ($\vdash$). The first is between a premise α and an intermediate conclusion β, and the second is between β and a final conclusion $\hat{\gamma}$. A set of alternative potential conclusions, $\{\gamma_0, \gamma_1, \ldots, \gamma_n\}$ (referred to together with $\hat{\gamma}$ as the set D), is created from which the model must select in the given context $\mathcal{C}$. Formally, given $D = \{\gamma_0, \gamma_1, \ldots, \gamma_n\}$: $f(\{\alpha, \beta, \alpha \vdash \beta, \beta \vdash X\}, \mathcal{C}) = \{X : \hat{\gamma}\}$, where $\hat{\gamma} \in D$.

Two-hop Premise. Following a similar formalisation, with two sequential argument relations of inference, ($\vdash$), the first between a premise, $\hat{\alpha}$, and an intermediate conclusion, β, and the second between β and a final conclusion, γ, a set is created of alternative potential premises, $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ (referred to together with $\hat{\alpha}$ as the set A), from which the model must select in the given context $\mathcal{C}$. Formally, given $A = \{\alpha_0, \alpha_1, \ldots, \alpha_n\}$,$f(\{X \vdash \beta, \beta, \beta \vdash \gamma, \gamma\}, \mathcal{C}) = \{X : \hat{\alpha}\}$, where $\hat{\beta} \in A$.

One-Intermediate Conclusion. Similarly, intermediate conclusion involves two sequential argument relations of inference, ($\vdash$). The first is between a premise α and an intermediate conclusion $\hat{\beta}$, and the second is between $\hat{\beta}$ and a final conclusion γ. A set of alternative potential intermediate conclusions, $\{\beta_0, \beta_1, \ldots, \beta_n\}$ (referred to together with $\hat{\beta}$ as the set B), is created from which the model must select in the given context $\mathcal{C}$. Formally, given $B = \{\beta_0, \beta_1, \ldots, \beta_n\}$, $f(\{\alpha, X \vdash \gamma, \gamma\}, \mathcal{C}) = \{X : \hat{\beta}\}$, where $\hat{\beta} \in B$.

Two-Intermediate Conclusions. Two intermediate conclusions involve three sequential argument relations of inference, ($\vdash$). The first is between a premise α and the first intermediate conclusion $\hat{\beta}$, the second is between $\hat{\beta}$ and the second intermediate conclusion $\hat{\gamma}$, and the third is between $\hat{\gamma}$ and the final conclusion ω. Given the context $\mathcal{C}$, the model selects $\hat{\beta}$ and $\hat{\gamma}$ from a set of alternative potential intermediate conclusions, $B = \{\beta_0, \beta_1, \ldots, \beta_n\}$ and $U = \{\gamma_0, \gamma_1, \ldots, \gamma_n\}$, respectively. Formally, given B, U, $f(\{\alpha, \omega, \alpha \vdash X, X \vdash Y, Y \vdash \omega\}, \mathcal{C}) = \{X : \hat{\beta}\}, \{Y : \hat{\gamma}\}$, where $\hat{\beta} \in B$ and $\hat{\gamma} \in U$.

4.2.2 Linked Reasoning

In a linked argument, there exists a support relation where a conclusion β is supported by a premise α in combination with another premise θ. It involves the following variants.

One Linked Premise. Given the context $\mathcal{C}$, the aim of this task is to identify the premise $\hat{\theta}$, such that $\alpha \wedge \hat{\theta} \vdash \beta$ holds, from a set of alternative potential linked premises, $Z = \{\theta_0, \theta_1, \ldots, \theta_n\}$. Formally, the relation is expressed as, $f(\{\alpha, \beta, \alpha \wedge X \vdash \beta\}, \mathcal{C}) = \{X : \hat{\theta}\}$, where $\hat{\theta} \in Z$.

Two Linked Premises. In two linked premise, given the context $\mathcal{C}$, the task is to identify both premises $\hat{\alpha}$ and $\hat{\theta}$, such that $\hat{\alpha} \wedge \hat{\theta} \vdash \beta$, from alternative potential linked premises, $A = \{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ and $Z = \{\theta_0, \theta_1, \ldots, \theta_m\}$. The relation is expressed as, $f(\{\beta, X \wedge Y \vdash \beta\}, \mathcal{C}) = \{(X, Y) : (\hat{\alpha}, \hat{\theta})\}$, where $\hat{\alpha} \in A$, $\hat{\theta} \in Z$.

Linked Reasoning Conclusion. This task aims to identify the conclusion $\hat{\beta}$, such that $\alpha \wedge \theta \vdash \hat{\beta}$ holds, from a set of alternative potential conclusions, $B = \{\beta_0, \beta_1, \ldots, \beta_n\}$, in the given context $\mathcal{C}$. The relation is expressed as, $f(\{\alpha, \theta, \alpha \wedge \theta \vdash X\}, \mathcal{C}) = \{X : \hat{\beta}\}$, where $\hat{\beta} \in B$.

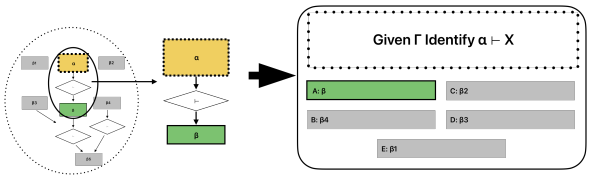


Figure 1: Illustration of the data processing for ART. In the argument graph (left), substructures of the target task are identified (the middle). Based on these, multiple-choice questions on the right are created, where the question contains the context $\Gamma = \{\alpha, \beta, \beta_1, \ldots, \beta_5\}$ and asks for a component X so that $\alpha \vdash X$. The correct answer β (green full outlines) is presented alongside with incorrect options $(\beta_1, \ldots, \beta_5)$ sampled from all other nodes in the graph. Examples of serial, linked, convergent, and divergent argumentation structures are provided in Table 23, alongside a visual representation in Table 24, which explains how traditional box-and-arrow diagrams from argumentation theory correspond to formal graphs, and how these graphs translate into the syntax of premises and conclusions connected through each argument subtype. Complete ART samples are also available in the supplementary material.

4.2.3 Convergent Reasoning

In a convergent argument, multiple premises (α, θ) independently support a conclusion β. It includes the following variants.

One Convergent Premise. The task is to identify a premise $\hat{\alpha}$ that independently supports β, given the conclusion β and the other premise θ that also independently supports β in the context $\mathcal{C}$. The model selects $\hat{\alpha}$ from a set of alternative potential premises, $A = \{\alpha_0, \alpha_1, \ldots, \alpha_n\}$. Formally, the relation is expressed as, $f(\{\theta, \beta, X \vdash \beta\}, \mathcal{C}) = \{X : \hat{\alpha}\}, \quad \text{where } \hat{\alpha} \in A.$

Two Convergent Premises. This task identifies both $\hat{\alpha}$ and $\hat{\theta}$, such that each independently supports β in the given context $\mathcal{C}$. The premises $\hat{\alpha}$ and $\hat{\theta}$ are selected from the sets of alternative potential premises $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ and $T = \{\theta_0, \theta_1, \ldots, \theta_m\}$, respectively. Formally, the relation is expressed as, $f(\{\beta, X \vdash \beta, Y \vdash \beta\}, \mathcal{C}) = \{(X, Y) : \hat{\alpha}, \hat{\theta}\}, \quad \text{where } \hat{\alpha} \in A \text{ and } \hat{\theta} \in T.$

Convergent Reasoning Conclusion. Finaly, this task identifies the conclusion $\hat{\beta}$, which is independently supported by the two premises α and θ. Given the premises α, θ and the context $\mathcal{C}$, the model must select a conclusion $\hat{\beta}$ from a set of potential conclusions B = $\{\beta_0, \beta_1, \ldots, \beta_m\}$. Formally, the relation is expressed as, $f(\{\alpha, \theta, \alpha \vdash X, \theta \vdash X\}, \mathcal{C}) = \{X : \hat{\beta}\}, \quad \text{where } \hat{\beta} \in B.$

Alternative Hop. Given a premise α, an intermediate conclusion β, and a final conclusion ω, each with their respective inference relation, ($\vdash$), the aim is to find an alternative reasoning chain that leads to ω. This chain involves an alternative premise $\hat{\theta}$ that supports an intermediate conclusion $\hat{\gamma}$, which in turn leads to the final conclusion ω. The model identifies an alternative $\hat{\theta}$ such that $\hat{\theta} \vdash \hat{\gamma}$ and $\hat{\gamma} \vdash \omega$ in the given context $\mathcal{C}$. Formally, let $Z = \{\theta_0, \theta_1, \ldots, \theta_n\}$ be the set of potential alternative premises and $U = \{\gamma_0, \gamma_1, \ldots, \gamma_n\}$ the set of potential intermediate conclusions. The model's task is then to find $\hat{\theta} \in Z$ and $\hat{\gamma} \in U$ that satisfy the relation. This is expressed as, $f(\{\alpha, \beta, \omega, \alpha \vdash \beta, \beta \vdash \omega, X \vdash Y, Y \vdash \omega\}, \mathcal{C}) = \{X : \hat{\theta}\}, \{Y : \hat{\gamma}\}, \quad \text{where } \hat{\theta} \in Z \quad \text{and} \quad \hat{\gamma} \in U.$ $\hat{\theta}$ is the selected alternative premise from the set Z and $\hat{\gamma}$ is the selected intermediate conclusion from the set U, such that $\theta \vdash \gamma$ and $\gamma \vdash \omega$ holds true.

4.2.4 Divergent Reasoning

In divergent argument, one premise supports multiple conclusions. It involves the following variants.

One Divergent Reasoning Conclusion. This task identifies one of the conclusions $\hat{\beta}$, $\hat{\gamma}$, which is supported by the premise α. Given the premise α, and one of the conclusions γ and the context $\mathcal{C}$, the model selects $\hat{\beta}$ from a set of potential conclusions $\{B = \beta_0, \beta_1, \ldots, \beta_m\}$. Formally, the relation is expressed as, $f(\{\alpha, \alpha \vdash X, \alpha \vdash \gamma\}, \mathcal{C}) = \{(X : \hat{\gamma}\}, \text{where } \hat{\beta} \in B.$

Two Divergent Reasoning Conclusions. This task identifies two conclusions $\hat{\beta}$ and $\hat{\gamma}$, both of which are supported by the premise α. Given the premise α and context $\mathcal{C}$, the model selects $\hat{\beta}$ and $\hat{\gamma}$ from a set of potential conclusions $\{B = \beta_0, \beta_1, \ldots, \beta_m\}$ and $Z = \{\gamma_0, \gamma_1, \ldots, \gamma_n\}$. Formally, the relation is expressed as, $f(\{\alpha, \alpha \vdash X, \alpha \vdash Y\}, \mathcal{C}) = \{(X, Y) : \{\hat{\beta}, \hat{\gamma}\}, \text{where } \hat{\beta} \in B \text{ and } \hat{\gamma} \in Z.$

Divergent Reasoning Premise. Given the conclusions β and γ within a context $\mathcal{C}$, the model selects $\hat{\alpha}$ from a set of potential premises $A = \{\alpha_1, \ldots, \alpha_n\}$, such that $\hat{\alpha}$ supports both β and γ. Formally, this relation is defined as, $f(\{X \vdash \beta, X \vdash \gamma\}, \mathcal{C}) = \{X : \hat{\alpha}\}, \text{where } \hat{\alpha} \in A.$

4.3 Data

To create a robust and comprehensive evaluation, we incorporate seven corpora spanning diverse domains and argumentative contexts, covering both monologue and dialogue structures. The corpora include MTC (Peldszus and Stede, 2015), AAEC

Dataset	Domain	Inferences	Conflicts	Neutral	Total
MTC	Structured Argumentation	272	108	713	1,093
AAEC	Essay	4,841	497	10,676	16,014
CDCP	Financial	694	82	1,552	2,328
ACSP	Scientific	8,069	697	17,532	26,298
AbstRCT	Medical	2,290	344	4,581	7,215
US2016	Political	Dialogue	3,083	650	
QT30	Question Answering	Dialogue	7,501	737	
Total	-	23,756	3,139	54,763	81,658

Table 1: Summary of the datasets included in the ART.

Tasks		MTC	AAEC	CDCP	ACSP	AbstRCT	US2016	QT30
Type	**Variants**							
Serial	**1H-C**	290	4841	1033	5789	2288	3379	6488
	1H-P	290	4841	1033	5789	2288	3379	6488
	2H-C	57	3279	348	759	327	1009	1118
	2H-P	57	3279	348	759	327	1009	1118
	Int-C	57	3279	348	759	327	1009	1118
	2-Int-C	3	569	89	80	8	249	787
Linked	**1L-P**	17	-	64	-	-	180	511
	2L-P	17	-	64	-	-	180	511
	LR-C	17	-	64	-	-	180	511
Convergent	**1C-P**	96	4735	763	2024	1899	1129	397
	2C-P	96	4735	763	2024	1899	1129	397
	CR-C	96	4735	763	2024	1899	1129	397
	AH	57	3279	348	759	327	1009	1118
Divergent	**1DR-C**	-	-	11	184	48	106	386
	2DR-C	-	-	11	184	48	106	386
	DR-P	-	-	11	184	48	106	386

Table 2: Statistics of task types for each dataset. The task variants are defined as follows: 1H-C (One-hop Conclusion), Int-C (Intermediate Conclusion), 2H-P (Two-hop Premise), 2-Int-C (Two-Intermediate Conclusions); 1L-P (One Linked Premise), 2L-P (Two Linked Premises), LR-C (Linked Reasoning Conclusion); 1C-P (One Convergent Premise), 2C-P (Two Convergent Premises), CR-C (Convergent Reasoning Conclusion), AH (Alternative Hop); 1DR-C (One Divergent Reasoning Conclusion), 2DR-C (Two Divergent Reasoning Conclusions) and DR-P (Divergent Reasoning Premise).

(Stab and Gurevych, 2017), CDCP (Park and Cardie, 2018), ACSP (Lauscher et al., 2018), AbstRCT (Mayer et al., 2020), US2016 (Visser et al., 2020), and QT30 (Hautli-Janisz et al., 2022).

MTC consists of short argumentative texts originally in German and translated into English, annotated according to Freeman's macro-structural theory of argumentation, with argument relations categorized as supports and attacks. **AAEC** consists of persuasive student essays annotated with argumentative relations, including supports and attacks. **CDCP** is a corpus of user comments on the Consumer Debt Collection Practices (CDCP) rule. It includes two types of support relations, categorised as Reason and Evidence which are consolidated into a single support relation. **ACSP** is a corpus of scientific publications in the field of computer graphics, annotated for argumentative relations, including supports, contradictions, and semantic equivalence. **AbstRCT** is a corpus of abstracts from randomized controlled trials in various medical domains, annotated with argument relations such as support, attack, and partial attack. **US2016** comprises transcripts of debates from the 2016 US presidential election and related Reddit discussions, annotated using Inference Anchoring Theory (IAT) with argument relations categorized as supports, attacks, and rephrases. Finally, **QT30** contains transcripts from the UK's *Question Time*, a political talk show, also annotated with IAT to identify supports, attacks, and rephrases. A summary of the dataset is presented in Table 1.

4.4 Data Processing

For each of the sixteen tasks, we systematically navigate through the argument structures available in the seven corpora, extracting all substructures that conform to the task specifications presented above. The resulting multiple-choice questions are organized into an input set and a corresponding target answer. The input set comprises the involved types of argumentative relations and their corresponding components (excluding the target correct answer), alongside the concatenation of all the sentences surrounding the argument component as the context (C). Four alternative incorrect answer options are randomly selected from other arguments outside of the identified argument substructure. For tasks instantiating the argument-component selection formulation, if multiple correct answers are present in an argument, only one correct answer is included in the list of options, while other correct answers are excluded from the pool of incorrect options. For serial reasoning task types, any reasoning chain involving linked arguments is excluded. This exclusion ensures that the substructure adequately captures the logic of the chain, as partial chains that involve only one argument component do not fully represent the structure of linked reasoning. Figure 1 summarises the data processing steps, in which complex and large argument graphs are converted into five-option multiple-choice questions. Refer to Appendix F for more on processing argument diagrams with boxes and labelled arrows.

As a result of this process, we present the Argumentative Reasoning Tasks (ART) dataset[3]. The ART dataset consists of a total of 112,212 multiple-choice questions following the sixteen task definitions, which can also be easily implemented as prompts as exemplified in Appendix C. Table 2 depicts the number of questions divided by task and

[3]The dataset will be publicly released after the acceptance of this paper under a CC BY-NC-SA 4.0 license.

Dataset	Model	Size	Argument-Component Selection			
			Serial	Linked	Convergent	Divergent
AAEC	**Qwen 2.5**	7B	23.78 ± 13.52	-	10.85 ± 11.50	-
		72B	35.59 ± 13.49	-	18.95 ± 19.37	-
	Llama 3.1	8B	12.23 ± 9.87	-	4.15 ± 3.62	-
		70B	38.77 ± 8.12	-	16.08 ± 20.25	-
	Mistral	7B	29.82 ± 14.12	-	10.4 ± 13.46	-
	DeepSeek-R1	70B	46.75±16.65	-	33.91±18.23	-
	GPT	GPT-4o	49.83 ± 17.37	-	35.78 ± 21.50	-
MTC	**Qwen 2.5**	7B	0.2 ± 0.21	-	1.75 ± 2.04	-
		72B	19.51 ± 16.29	-	2.6 ± 3.40	
	Llama 3.1	8B	0.16 ± 0.16	-	1.05 ± 1.50	-
		70B	8.53 ± 11.71	-	5.46 ± 4.56	-
	Mistral	7B	0.16 ± 0.26	-	0.9 ± 1.53	-
	DeepSeek-R1	70B	45.34±10.45	-	15.87±12.34	-
	GPT	GPT-4o	49.73 ± 24.36	-	11.36 ± 11.54	-
CDCP	**Qwen 2.5**	7B	29.97 ± 14.84	35.38 ± 25.32	17.45 ± 20.45	0.86 ± 0.80
		72B	50.28 ± 21.52	51.28 ± 16.59	24.68 ± 28.54	1.2 ± 0.61
	Llama 3.1	8B	10.33 ± 7.95	9.23 ± 12.21	5.85 ± 6.66	0.4 ± 0.4
		70B	40.71 ± 17.94	49.74 ± 21.88	21.47 ± 28.40	0.93 ± 0.53
	Mistral	7B	22.97 ± 12.18	12.82 ± 14.94	8.85 ± 12.64	0.26 ± 0.46
	DeepSeek-R1	70B	61.65±9.78	63.43±12.22	41.56±24.23	7.12±3.44
	GPT	GPT-4o	65.06 ± 13.41	68.87 ± 14.93	44.94 ± 30.31	7.33 ± 2.52
AbstRCT	**Qwen 2.5**	7B	11.46 ± 6.28	-	14.4 ± 18.73	0.933 ± 0.90
		72B	33.96 ± 19.27	-	29.40 ± 33.71	1.46 ± .070
	Llama 3.1	8B	4.7 ± 3.30	-	8.9 ± 7.01	0.4 ± 0.4
		70B	19.05±19	-	11.12.86	1.33 ± 0.80
	Mistral	7B	10.0 ± 5.77	-	6.35 ± 9.19	0.33 ± 0.41
	DeepSeek-R1	70B	46.34±23.45	-	36.56±25.67	10.45±5.56
	GPT	GPT-4o	48.61 ± 28.90	-	34.48 ± 29.19	11.4 ± 3.13
ACSP	**Qwen 2.5**	7B	37.13 ± 19.03	-	16.05 ± 15.38	9.13 ± 6.77
		72B	47.31 ± 23.55	-	25.07 ± 15.23	12.8 ± 6.43
	Llama 3.1	8B	12.3 ± 8.25	-	4.5 ± 4.94	2.4 ± 2.42
		70B	39.64 ± 13.76	-	12.433 ± 18.07	8.86 ± 6.10
	Mistral	7B	26.66 ± 13.47	-	12.4 ± 14.18	5.86 ± 5.08
	DeepSeek-R1	70B	56.78±10.43	-	51.45±9.56	24.78±5.23
	GPT	GPT-4o	90.47 ± 7.34	-	86.38 ± 3.16	41.45 ± 14.34
US2016	**Qwen 2.5**	7B	34.12 ± 19.53	30.55 ± 19.37	20.45 ± 21.46	7.6 ± 5.4
		72B	49.53 ± 27.61	48.33 ± 18.86	30.34 ± 25.69	10.53 ± 6.26
	Llama 3.1	8B	14.41 ± 6.34	11.66 ± 8.67	9.9 ± 12.58	2.86 ± 2.71
		70B	45.51 ± 25.65	45.18 ± 21.37	26.39 ± 26.64	8.06 ± 5.98
	Mistral	7B	37.95 ± 20.51	20.18 ± 17.84	12.8 ± 15.21	4.53 ± 3.70
	DeepSeek-R1	70B	60.34±13.45	47.65±15.34	41.95±13.72	36±14.63
	GPT	GPT-4o	58.47 ± 12.94	53.03 ± 9.32	45.85 ± 15.12	37.78 ± 17.21
QT30	**Qwen 2.5**	7B	31.40±16.96	20.76 ± 18.63	11.4 ± 11.10	20 ± 18.11
		72B	42.45 ± 20.84	45.50 ± 16.24	20.33 ± 17.02	29.0 ± 15.77
	Llama 3.1	8B	9.99 ± 5.15	11.50 ± 10.26	5.8 ± 4.48	12.33 ± 13.52
		70B	36.21 ± 15.94	43.38 ± 20.84	18.10 ± 16.59	23.16 ± 17.40
	Mistral	7B	33.98 ± 17.96	20.76 ± 18.63	6.2 ± 8.22	12.4 ± 11.78
	DeepSeek-R1	70B	55.78±20.35	46.56±14.47	40.92±18.43	38.21±11.45
	GPT	GPT-4o	53.62 ± 23.80	53.04 ± 18.66	46.69 ± 21.44	41.65 ± 15.34

Table 3: Macro averaged F1-scores and standard deviations for the argument-component selection tasks.

corpora that make up our dataset.

5 Experiments

5.1 Experimental Setup

We evaluate the performance of state-of-the-art models, including Qwen 2.5 (Yang et al., 2024), Llama 3.1 (Touvron et al., 2023), Mistral (Jiang et al., 2023), GPT-4 (Achiam et al., 2023), o1[4], and DeepSeek-R1 (Guo et al., 2025), across the reasoning tasks in a prompt-based setting. The specific prompt templates and model hyperparameters, including temperature, top-p sampling, and inference steps, are detailed in the Appendix B for reproducibility and transparency. For evaluating the models on the ART multiple-choice reasoning tasks, we evaluate model performance using macro averaged F1-score. The code and dataset are available at https://github.com/arg-tech/art.

[4] https://openai.com/index/learning-to-reason-with-llms/

5.2 Results and Discussion

Table 3 reports the macro averaged F1-scores and their standard deviations for each model and type of argument structure. The fine-grained results considering each of the ART tasks independently have been included in Appendix D. Having the random chance baseline (i.e., 20%, one correct answer out of five options) as a reference, we can observe how language models could not consistently provide the correct answers for the ART tasks. This implies that while LLMs may exhibit reasoning abilities in areas such as formal logic and mathematical reasoning, they may face challenges when dealing with argumentative reasoning. Even when their outputs appear to reflect reasoning, this can often be attributed to surface-level pattern matching — shaped by token biases (Jiang et al., 2024), shallow heuristics (Gendron et al., 2024), or memorisation and replication of training data patterns (Mirzadeh et al., 2024) — rather than genuine inferential processes. These patterns, often mistaken for reasoning ability, stem from the model's capacity to generate fluent text rather than a true ability to perform any type of (argumentative) reasoning.

Across all tasks, GPT and DeepSeek stand out, with GPT achieving the highest average performance. On average, GPT-4o achieves 54.38 ± 25.30, 49.52 ± 22.96, 52.60 ± 26.53 and 27 ± 10.52 F1-score on serial, linked, convergent and divergent task types respectively, closely followed by DeepSeek. Qwen, Mistral, and Llama models' poor performance was consistent across the board. Despite outperforming others, GPT-4o results show higher standard deviations (increasing with performance), indicating a big performance gap between simple and complex versions of the same task types (e.g., **1H-C**, **1H-P** versus **2H-C**, **2H-C**).

This observation can be generalised to the rest of the models, which also show significant performance variations across task types and corpora. The models struggle with task types involving argument substructures as the right answers (i.e., **2-Int-C** and **AH**), achieving performance lower than the random baseline. Notably, with the exception of GPT-4o, all other models, regardless of their size, performed near zero F1-score when tasked with selecting alternative reasoning hops (AH) and two intermediate conclusions (2-Int-C). For instance, Qwen 2.5:70B achieves 4.25 ± 4.92, 3.47 ± 4.72 in **AH** and **2-Int-C**, respectively. This highlights a significant limitation in handling complex reasoning

structures, even for larger model architectures.

The results for GPT-4o on ACSP constitute significant outliers with a macro-average F1-score of 90.47, 86.38, and 41.45 for serial, linked and divergent types of argument structure respectively. The same model achieves 53.62, 53.04, 41.65 on QT30 for serial, linked and divergent types of argument structure respectively. These results would, in principle, mean that GPT-4o is capable of effectively parse and understand natural language reasoning structures in scientific publications. After a deeper analysis on the data we observed, however, that on average ACSP has 324 argument components per argumentative context C, while US2016, QT30, AAEC, MTC, ABstRACT, and CDCP involve 17, 15, 15, 5, 7 and 26, respectively. Since the incorrect answers are randomly selected from C, ACSP provides larger space of candidate answers involving more semantically diverse and distant sets of answers. This makes the task easier, allowing to distinguish the correct answer by focusing on semantic features of the text.

An additional insight that emerged from our experiments concerns the evaluation of enthymemes—arguments in which one or more components (typically premises or conclusions) are left implicit. Many of the argumentative reasoning tasks explored in this work are effectively instances of enthymeme formulation, where participants or models must infer missing argumentative elements. Reconstruction of enthymemes has been tackled in NLP (Rajendran et al., 2016; Lawrence and Reed, 2020), but always views the enthymematic reconstruction as additions to an argument: here the enthymematic forms result from subtractions from natural text.In contrast, the tasks presented here start from naturally occurring, complete arguments and remove components to create inference challenges. This subtraction-based setup yields a natural gold standard for evaluation, without the need for manual scoring or assess the artificial reconstructions.

5.3 Sensitivity Study

In addition to discussing about the results achieved by LLMs on ART directly, we have also analysed the models' sensitivity to variations in settings including an open ended reasoning with human evaluation setup, model size and prompt template.

Open-Ended Reasoning with Human Evaluation. To investigate on the effect of the multiple-choice setup versus an open-ended one where LLMs can freely reason and be creative, we conduct an evaluation in a language generation setup where human experts assess the generated components. In this setup, the model generates missing argument components conditioned on the argument context and a partially specified substructure with the missing components. Unlike the multiple-choice setup, this allows the model to freely produce content while maintaining logical consistency and coherence with the context. Two expert annotators independently evaluated the correctness and contextual relevance of the generated components. To reduce subjectivity, we defined annotation guidelines for correctness and relevance (see Appendix E.1). One random example per task from each dataset was selected, resulting in a total of 112 samples (7 datasets × 16 tasks). From these samples, to make the human evaluation feasible, we excluded 17 having an excessively large context, making a total of 95 samples. GPT-4o, the best performing model in the multiple-choice setup, was used for generating the argument components. An IAA of $\kappa = 0.44$ was achieved during human evaluation.

We observed an F1 of 25.8 in the open ended reasoning task. This score is notably lower than in the multiple-choice setup despite a more generous evaluation, where human evaluators accepted contextually relevant argument components that didn't exactly match the gold standard. According to the human evaluation, the main reason behind this poor performance is that in most of the cases, the models copy-pasted existing argument components or concatenated them instead of generating original ones aligned with the required structure and context. This further supports our previous claim that LLMs rely on superficial probabilistic language patterns rather than genuine reasoning.

Human Annotations as an Upper Bound. To better understand model performance in the multiple-choice setup, it is useful to consider the nature of the gold-standard annotations used for evaluation. These annotations stem from a prior human study in which participants were not guided by a fixed structure, faced hundreds of potential distractors instead of just four, and had to distinguish between closely related alternatives drawn from the same local argument context. Compared to the constrained nature of multiple-choice formats, this setup presents significantly greater cognitive and inferential challenges. As such, these annotations provide a strong benchmark for assessing model capabilities and the observed inter-annotator

Llama 3.1		GPT	
70B	**405B**	**gpt-4o**	**o1-preview**
9.98	18.73	32.18	41.96

Table 4: Sensitivity to model size across different architectures and variants (**2C-P**).

Model	**Prompt-1**	**Prompt-2**
Llama 3.1:70B	16.01	15.40
Mistral	7.25	7.09
Qwen 2.5:72B	16.29	14.61
GPT-4o	34.32	35.78

Table 5: Prompt Sensitivity: Models performance on Prompt-1 and Prompt-2 (**2H-C**,**2L-P**, **1CP**, **2DR-C**).

agreement of $\kappa = 0.6$ serves as a meaningful upper bound on expected model performance in this task.

Model Size. The assessment of model sizes compares the 70B and 405B parameter versions of Llama 3.1, as well as GPT-4o vs o1-Preview. The parameter sizes for GPT-4o and o1-Preview are undisclosed, but according to OpenAI's release notes, o1-Preview is designed to handle more complex reasoning tasks compared to GPT-4o. Table 4 reports the results of this comparative study on the **2C-P** task, which, as highlighted in the previous results, is among the most challenging. This task requires the correct answer to include two argument components. The results of the model size sensitivity study show that the performance improves with the model size[5]. These findings indicate that the improvement of the task scales with the size of the model. This improvement, however, is still far from claiming a successful performance on the task. Scaling, therefore, seems not to be a solution to problems involving complex reasoning in natural language, having the 405B version of the Llama 3.1 model performing worse than a random baseline. Even o1-preview, a model that has been described as reasoning model, cannot effectively identify the two correct premises in a convergent argument.

Prompt Template. Finally, we also investigate the influence of the prompt phrasing on the model performance by testing another independently developed prompt. The two prompts were created by two different authors of this paper without being able to see each other's prompt, having only available the formal definition of the selected tasks (i.e., **2H-C**, **2L-P**, **1C-P**, and **DR-C**) presented in Section 4. Table 5 reports the results from this study, showing a very similar performance on both prompts, meaning that the phrasing of the prompt used in our experiment does neither harm nor boost the model performance for the multiple-choice argument-component selection task.

[5]The assumption is that o1-preview is the largest model.

6 Conclusion

This paper pushes forward the boundaries of knowledge on the reasoning capabilities of LLMs, a controversial and widely debated topic in the last years. We do so by asking a simple yet relevant question, can LLMs parse and understand argumentative reasoning structures? Given that argumentation is the natural way of reasoning in natural language, if LLMs can reason, they should be able to parse, understand, and build natural language arguments.

Our results show that not only LLMs are not capable of understanding argumentative reasoning structures (let's not forget that this means reasoning in natural language), but also cases where a slightly more challenging argumentative structure is used, they perform worse than a random baseline. This highlights the need to develop challenging tasks to evaluate natural language reasoning, and also question the reasoning capabilities of LLMs, as it has been recently suggested in the literature. While we do not position argumentative reasoning as the sole measure of reasoning ability, it complements existing approaches by offering a valuable lens through which to assess LLMs in more realistic forms of natural language reasoning.

Acknowledgments

This work is funded in part by the 'AI for Citizen Intelligence Coaching against Disinformation (TITAN)' project, funded by the EU Horizon 2020 research and innovation programme under grant agreement 101070658, and by UK Research and innovation under the UK governments Horizon funding guarantee grant numbers 10040483 and 10055990; in part by Volkswagen Stiftung Foundation under grant 98 543, "Deliberation Laboratory"; and in part by the Swiss National Science Foundation under grant 10001FM_200857, "Mining argumentative patterns in context".

Limitations

Due to limitations in the compute budget, this work assesses very large / expensive models like the

405B parameter version of Llama and o1 only on a limited subset of ART multiple-choice questions. Nevertheless, the reported results indicate important trends, revealing that despite showing a slight increase in performance, they are still not capable of addressing tasks involving complex reasoning.

References

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

Monroe Curtis Beardsley. 1950. *Practical Logic*. Englewood Cliffs, NJ, Prentice-Hall.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Leo Groarke. 2004. Good reasoning matters!: a constructive approach to critical thinking.

Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. In *Advances in Neural Information Processing Systems*, volume 36, pages 79081–79094. Curran Associates, Inc.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031. Association for Computational Linguistics.

Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. 2022. Qt30: A corpus of argument and conflict in broadcast debate. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3291–3300. European Language Resources Association (ELRA).

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

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. 2024. A peek into token bias: Large language models are not yet genuine reasoners.

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

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

Yinghao Li, Haorui Wang, and Chao Zhang. 2024. Assessing logical puzzle solving in large language models: Insights from a minesweeper case study. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 59–81. Association for Computational Linguistics.

Man Luo, Shrinidhi Kumbhar, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, Chitta Baral, et al. 2023. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *arXiv preprint arXiv:2310.00836*.

Saumya Malik. 2024. Lost in the logic: An evaluation of large language models' reasoning capabilities on lsat logic games. *arXiv preprint*.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.

Houman Mehrafarin, Arash Eshghi, and Ioannis Konstas. 2024. Reasoning or a semblance of it? a diagnostic study of transitive reasoning in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11647–11662.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. *arXiv preprint arXiv:2406.12830*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

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.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2016. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Soumadeep Saha, Sutanoya Chakraborty, Saptarshi Saha, and Utpal Garain. 2024. Language models are crossword solvers. *arXiv preprint*.

Prisha Samadarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of llms using the new york times connections word game. *arXiv preprint arXiv:2406.11012*.

Kulin Shah, Nishanth Dikkala, Xin Wang, and Rina Panigrahy. 2024. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. *arXiv preprint*.

Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455.

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

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Stephen Naylor Thomas. 1973. *Practical reasoning in natural language*. Englewood Cliffs, Prentice-Hall.

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

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin Rrv, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, and Chitta Baral. 2024a. Step-by-step reasoning

to solve grid puzzles: Where do llms falter? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19898–19915.

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin Rrv, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, and Chitta Baral. 2024b. Step-by-step reasoning to solve grid puzzles: Where do llms falter? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19898–19915. Association for Computational Linguistics.

Karthik Valmeekam, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. 2024. Planning in strawberry fields: Evaluating and improving the planning and scheduling capabilities of lrm o1. *arXiv preprint arXiv:2410.02162*.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Douglas Walton. 2005. *Fundamentals of critical argumentation*. Cambridge University Press.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.

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

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37.

A Task Visualisation

To simplify the understanding of the task formalisations in Section 4, Figure 2 depicts a sub-set of tasks from the ART dataset including **1H-C**, **1Int-C**, **2H-P**, **2H-C**, **1DR-C**, **2DR-C**, **AH**, and **2Int-C**.

B Hyper-Parameters

We utilize the LLaMA 3.1 model in its 8B, 70B, and 405B configurations (Touvron et al., 2023), accessed through the `Ollama` library[6]. Additionally, the 7B configuration of the Mistral model (Jiang et al., 2023) is employed, also via the `Ollama` library[7]. Furthermore, we use the 7B and 72B versions of the Qwen 2.5 model[8], accessed through the `Ollama` library[9]. For GPT variants, we rely on the API provided by OpenAI for interacting with the GPT-4o and o1-Preview models. Across the models we use default parameters including the temperature and top_k predictions. We do not perform any finetuning and only apply prompting to off-the-shelf models.

C Prompt Templates

Aimed at improving the transparency and reproducibility of the results reported in this paper, Table 22 contains the templates of the prompts that we used for the different tasks included in ART.

D Complete Results

This appendix section contains the fine-grained results of the LLMs on the sixteen tasks included in ART.

D.1 Serial

- One-hop conclusion (**1H-C**): Table 6.
- One-hop premise (**1H-P**): Table 7.
- Two-hop conclusion (**2H-C**): Table 8.
- Two-hop premise (**2H-P**): Table 9.
- Intermediate conclusion (**Int-C**): Table 10.
- Two intermediate conclusions (**2-Int-C**): Table 11.

D.2 Linked

- One linked premise (**1L-P**): Table 12.
- Two linked premises (**2L-P**): Table 13.
- Linked reasoning conclusion (**LR-C**): Table 14.

D.3 Convergent

- One convergent premise (**1C-P**): Table 15.
- Two convergent premises (**2C-P**): Table 16.
- Convergent reasoning conclusion (**CR-C**): Table 17.
- Alternative Hop (**AH**): Table 18.

D.4 Divergent

- One divergent reasoning conclusion (**1DR-C**): Table 19.
- Two divergent reasoning conclusions (**2DR-C**): Table 20.
- Divergent reasoning premise (**DR-P**): Table 21.

E Open Eneded Reasoning with Human Evaluation

E.1 Annotation Guidelines

The annotation process has the aim to identify whether the argument-component marked with "COMPONENT" are similar to the gold-standard and if not whether the component is an appropriate continuation of the map. The gold-standard can be found in the annotation .csv file.

The annotation follows a 3-step process:

E.1.1 Step 1

Is the target component similar to the gold-standard component? If yes (annotate Step 1 with 1), stop the annotation here.

If no (annotate Step 1 with 0), continue with Step 2 and 3.

E.1.2 Step 2

Is the target component appropriate in the given position in the given map. (1 for yes and 0 for no). This includes but is not limited to:

C1 Fit: Is the generated component fulfilling the targeted argumentative function? Does it respect the existing relations without creating new ones?

[-] This is not fulfilled if: The component should be a conclusion but rather rephrases or explains one of its premises / The component is rather a premise for another premise than a premise for the conclusion.

[6] https://ollama.com/library/llama3.1
[7] https://ollama.com/library/mistral
[8] https://github.com/QwenLM/Qwen
[9] https://ollama.com/library/qwen2.5

C2 Relevance: Is the component relevant in the context of the given argument map?

C3 Self-contained: Is the component self-contained (a proposition and not a whole argument)?

The components may follow shallow heuristics, which are important to spot and make the generated component often invalid. We list here a few frequent heuristics.

H1 Surrounding components in the graph are concatenated with or without a discourse marker.

H2 Another component from the same graph is copied or slightly rephrased.

H3 Existing surrounding components are extended.

These heuristics can also be combined and are not limited to the ones listed here. If you identified the component as not appropriate, please justify your decision in Step 3.

E.1.3 Step 3

Justify why the component is not an appropriate continuation. Please refer to the criteria and heuristics mentioned in Step 2. You may also add free text in addition. If you spot another heuristics, please describe it in the first instance where you spotted it and name it consistently (H4, H5 . . .).

We employed two PhD students experienced in annotating argument and compensated them at the standard hourly rate equivalent to their experience and qualifications.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	68.90	74.34	98.46	74.98	62.21	69.12	69.66
DeepSeek-R1	65.28	69.25	68.767	66.84	54.43	71.87	72.28
llama3.1:70b	34.80	24.60	37.80	43.60	0.40	34.80	-
llama3.1:8b	19.00	5.60	17.40	15.80	0.20	15.83	14.81
mistral	42.40	17.80	29.80	33.10	0.60	50.27	54.33
qwen2.5:72b	43.14	59.90	54.41	67.83	33.33	55.21	69.19
qwen2.5:7b	32.40	17.40	40.20	40.80	0.40	50.27	54.33

Table 6: 1H-C

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	68.12	71.34	93.41	71.87	58.12	68.90	68.23
DeepSeek-R1	67.36	72.36	57.79	69.18	56.16	70.65	72.55
llama3.1:70b	42.4	15.2	51.6	48.4	0.4	49.01	67.6
llama3.1:8b	25	8.2	17.8	22.4	0.4	10.733	10.218
mistral	27.6	10	36.4	28.6	0	40.25	47.2
qwen2.5:72b	34.79	35.64	52.96	62.02	33.33	52.06	63.60
qwen2.5:7b	37.2	14.2	48	42.4	0.4	42.6	47.65

Table 7: 1H-P

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	25.46	37.62	87.74	53.62	61.66	48.83	54.83
DeepSeek-R1	20.77	31.56	53.65	50.62	59.34	51.36	52.16
llama3.1:70b	51.00	36.00	44.40	32.40	0.40	35.60	35.00
llama3.1:8b	18.00	8.00	22.00	9.80	0.20	14.20	15.40
mistral	20.50	11.40	29.20	26.40	0.40	32.80	32.20
qwen2.5:72b	45.00	33.00	54.60	37.60	0.40	38.20	32.00
qwen2.5:7b	33.00	12.00	44.60	32.60	0.40	30.20	22.20

Table 8: 2H-C.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	41.00	39.10	79.31	70.43	33.33	54.43	56.08
DeepSeek-R1	38.56	36.98	44.76	66.87	24.50	58.65	59.45
llama3.1:70b	35.59	38.11	40.97	58.84	33.33	42.44	59.62
llama3.1:8b	8.00	4.60	10.20	9.40	0.00	9.5707	20.00
mistral	24.80	9.00	29.00	22.60	0.00	32.46	35.76
qwen2.5:72b	32.20	38.11	50.98	57.10	33.33	49.64	62.81
qwen2.5:7b	17.60	9.40	33.60	25.80	0.00	32.46	35.76

Table 9: 2H-P.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	55.88	69.24	97.50	76.16	33.33	71.67	66.66
DeepSeek-R1	52.76	65.15	56.86	72.78	32.27	75.67	76.27
llama3.1:70b	30.07	49.01	49.54	52.17	16.67	47.57	60.10
llama3.1:8b	3.40	1.80	6.40	4.60	0.00	11.90	10.84
mistral	33.80	11.80	35.60	24.80	0.00	47.60	57.00
qwen2.5:72b	38.46	37.13	69.70	64.93	16.67	57.46	68.79
qwen2.5:7b	21.80	15.80	55.20	36.00	0.00	32.40	43.59

Table 10: Int-C.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	39.65	0	86.42	43.33	0	8.78	35.40
DeepSeek-R1	36.76	2.76	60.87	43.65	0	6.43	29.27
Llama 3.1 70B	0.53	0	13.58	8.89	0	2.16	5.24
Llama 3.1 8B	0	0	0	0	0	0.13	0
Mistral	0.70	0	0	2.22	0	0.51	1.21
Qwen 2.5 72B	7.89	0	1.23	12.22	0	2.16	0.81
Qwen 2.5 7B	0.70	0	1.23	2.22	0	0.51	1.21

Table 11: 2-Int-C.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	–	–	–	79.87	–	65.23	73.81
DeepSeek-R1	–	–	–	73.36	–	58.16	73.18
llama3.1:70b	–	–	–	64.62	–	57.54	58.89
llama3.1:8b	–	–	–	4.62	–	13.49	17.22
mistral	–	–	–	9.23	–	25.60	26.67
qwen2.5:7b	–	–	–	49.23	–	39.68	43.89
qwen2.5:72b	–	–	–	58.46	–	53.77	57.22

Table 12: 1L-P

	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	–	–	–	51.87	–	31.56	58.12
DeepSeek-R1	–	–	–	48.39	–	27.84	57.68
llama3.1:70b	–	–	–	24.62	–	19.44	20.56
llama3.1:8b	–	–	–	0	–	0.40	1.67
mistral	–	–	–	0	–	0.20	0
qwen2.5:7b	–	–	–	6.15	–	9.33	8.33

Table 13: 2L-P

Model	AAEC	ABstRACT	ACSP	CDCP	Microtext	QT30	US2016
GPT-4o	–	–	–	74.87	–	62.34	57.23
DeepSeek-R1	–	–	–	68.53	–	53.65	60.65
llama3.1:70b	–	–	–	60.00	–	53.17	56.11
llama3.1:8b	–	–	–	23.08	–	20.63	16.11
mistral	–	–	–	29.23	–	36.51	33.89
qwen2.5:7b	–	–	–	50.77	–	39.29	39.44
qwen2.5:72b	–	–	–	63.08	–	55.95	61.11

Table 14: LR-C.

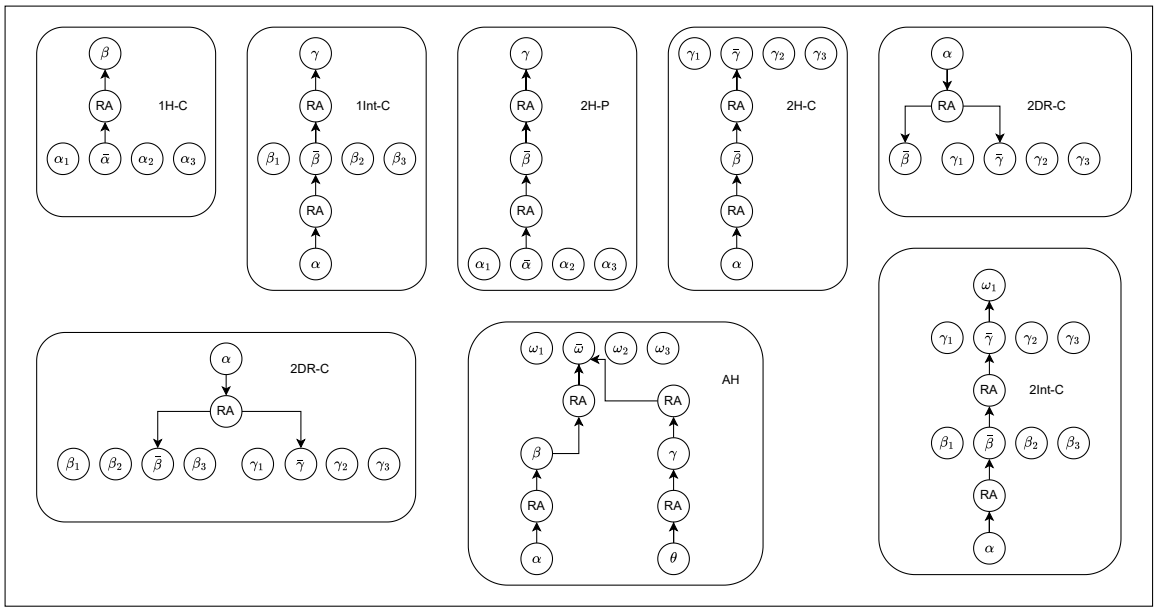


Figure 2: Illustration of selected task types highlighting serial, linked, convergent, and divergent argument structures. The figure includes task types involving single argument components, two argument components, and substructures such as alternative reasoning and two intermediate conclusions.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	44.68	23.44	89.45	46.00	3.46	54.98	51.35
DeepSeek-R1	43.76	25.58	50.87	44.18	6.19	52.12	51.65
llama3.1:70b	15.80	18.80	22.40	17.80	1.40	18.60	29.20
llama3.1:8b	4.40	14.00	6.40	6.40	1.00	7.00	9.40
mistral	9.00	2.40	22.60	7.20	0.20	5.60	18.00
qwen2.5:72b	22.20	21.20	34.60	24.80	1.60	27.00	37.00
qwen2.5:7b	16.60	12.40	26.80	20.80	1.80	14.60	27.20

Table 15: 1C-P

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	20.88	16.90	85.63	18.85	17.65	28.39	36.99
DeepSeek-R1	17.87	18.57	51.44	11.74	22.76	19.34	23.68
llama3.1:70b	8.34	14.58	3.90	5.50	9.80	13.82	13.98
llama3.1:8b	3.40	6.60	1.00	2.00	0.00	5.40	2.40
mistral	2.80	3.00	0.40	1.00	0.20	2.80	1.20
qwen2.5:72b	9.00	15.40	10.60	5.40	1.20	11.20	12.80
qwen2.5:7b	2.60	3.80	6.00	4.00	0.60	5.60	7.00

Table 16: 2C-P

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	61.898	78.09	88.34	87.13	24.34	73.13	64.45
DeepSeek-R1	58.76	80.65	55.25	78.59	18.63	62.52	63.89
llama3.1:70b	40.2	78.60	33.1	62.6	6.6	40	62.2
llama3.1:8b	8.8	15	10.6	15	3.1	10.8	27.8
mistral	29.8	20	26.6	27.2	3.1	18.6	32
qwen2.5:72b	44.6	78.6	41.4	65	7.6	40.8	63.8
qwen2.5:7b	24.2	41.4	31.4	45	4.6	25.4	47.6

Table 17: CR-C.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	15.68	19.51	82.12	27.79	0	30.19	30.63
DeepSeek-R1	15.16	21.35	48.25	31.74	0	29.65	28.57
llama3.1:70b	0	0	0.2	0	0	0	0.2
llama3.1:8b	0	0	0	0	0	0	0
mistral	0	0	0	0	0	0	0
qwen2.5:72b	0	2.44	13.69	3.52	0	2.36	7.79
qwen2.5:7b	0	0	0	0	0	0	0

Table 18: AH.

	AAEC	ABstRACT	ACSP	CDCP	Microtext	QT30	US2016
GPT-4o	-	15.4	52.34	9.34	-	51.34	48.34
Deep-Seek-R1	-	13.83	22.75	7.28	-	50.38	46.45
LLAMA3.1:70B	-	1.6	9.2	1	-	23	9.2
LLAMA3.1:8B	-	0.4	2	0.4	-	10.2	3.2
Mistral	-	0.2	8.8	0	-	10	4.6
Qwen2.5	-	1	12	1	-	16.8	7.6
Qwen2.5:72B	-	2.2	18	1.4	-	35.6	11.6

Table 19: DR-C.

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	-	7.8	26.45	4.56	-	24.56	17.23
DeepSeek-R1	-	7.73	23.36	5.35	-	17.65	21.73
llama3.1:70b	-	0.2	2.6	0.4	-	6	1.6
llama3.1:8b	-	0	0.2	0	-	0	0
mistral	-	0	0	0	-	2	0.8
qwen2.5	-	0	1.4	0	-	4.8	2.2
qwen2.5:72b	-	0.8	5.6	0.6	-	11	3.8

Table 20: 2DR-C

Model	AAEC	ABstRACT	ACSP	CDCP	MTC	QT30	US2016
GPT-4o	-	10.8	45.6	8.32	-	48.5	46.23
DeepSeek-R1	-	9.78	28.19	8.73	-	46.65	40.65
LLAMA3.1:70B	-	1.6	14.8	1.4	-	40.8	13.4
LLAMA3.1:8B	-	0.8	5.0	0.8	-	26.8	5.4
Mistral	-	0.8	8.8	0.8	-	25.2	8.2
Qwen2.5	-	1.8	14.0	1.6	-	40.4	13.0
Qwen2.5:72B	-	1.4	14.8	1.8	-	40.4	16.2

Table 21: DR-P.

Task Type	Prompt
1H-C	A one-hop argument involves a single inference step where a **Premise** directly supports **Conclusion**. Consider the following argument: '{argument}'. Given the **Premise: '{premise}'**, your task is to identify which of the following options represents the **Conclusion** that is directly supported by the **Premise**.
1H-P	A one-hop argument consists of a single inference step where a **Premise** directly supports **Conclusion**. Consider the following argument: '{argument}'. Given the **Conclusion: '{conclusion}'**, your task is to identify which of the following options can serve as the **Premise** that supports this **Conclusion**.
1Int-C	A two-hop serial argument involves two inference steps: a **Premise** supports **Conclusion 1** (the intermediate conclusion), and **Conclusion 1** further supports a final **Conclusion 2** in a chain. Consider the following argument: '{argument}'. Given the **Premise: '{premise}'**, your task is to identify which of the following options can serve as **Conclusion 1** that connects the **Premise** to **Conclusion 2: '{conclusion_2}'**.
2H-C	A two-hop serial argument involves two inference steps: a **Premise** supports **Conclusion 1** (the intermediate conclusion), and **Conclusion 1** further supports a final **Conclusion 2** in a chain. Consider the following argument: '{argument}'. Given the **Premise: '{premise}'** which supports **Conclusion 1: '{conclusion_1}'**, your task is to identify which of the following options can serve as the final **Conclusion 2** that is further supported by **Conclusion 1**.
2H-P	A two-hop serial argument involves two inference steps: a **Premise** supports **Conclusion 1** (the intermediate conclusion), and **Conclusion 1** further supports a final **Conclusion 2** in a chain. Consider the following argument: '{argument}'. Given **Conclusion 2: '{conclusion_2}'** which is supported by **Conclusion 1: '{conclusion_1}'**, your task is to identify which of the following options can serve as the **Premise** that supports **Conclusion 1**.
2Int-C	A three-hop serial argument involves three inference steps: a **Premise** supports **Conclusion 1**, **Conclusion 1** supports **Conclusion 2**, and **Conclusion 2** further supports **Conclusion 3** in a chain. Consider the following argument: '{argument}'. Given the **Premise: '{premise}'**, your task is to identify which one of the following options represents **Conclusion 1** that is logically supported by the **Premise**, and which one represents **Conclusion 2** that is supported by **Conclusion 1**, such that **Conclusion 2** further supports **Conclusion 3: '{conclusion_3}'** in the chain. The missing argument components must logically align with the provided context, ensuring that **Conclusion 1** is supported by the **Premise**, **Conclusion 2** is supported by **Conclusion 1**, and **Conclusion 3** is supported by **Conclusion 2**.
1L-P	In a linked argument, a **conclusion** is supported jointly by multiple premises (**Premise 1**, **Premise 2**). Consider the following argument: '{argument}'. Given the **Premise 1: '{premise_1}'**, your task is to identify which of the following options represents the **Premise 2** that, when used jointly with **Premise 1**, directly supports the **conclusion**: '{conclusion}'.
2L-P	In two linked premises, a **conclusion** is supported jointly by **Premise 1** and **Premise 2**. Consider the following argument: '{argument}'. Identify which one of the following represents **Premise 1** and **Premise 2**, from the given set of alternatives, jointly supporting the **conclusion**: '{conclusion}'.
LR-C	In a linked reasoning argument, a conclusion is supported jointly by **Premise 1** and **Premise 2**. Consider the following argument: '{argument}'. Given the **Premise 1: '{premise_1}'** and **Premise 2: '{premise_2}'**, your task is to identify which one of the following options represents the **Conclusion** that is jointly supported by **Premise 1** and **Premise 2**.
1C-P	In a Convergent argument, a **conclusion** is independently supported by multiple premises (**Premise 1**, **Premise 2**). Consider the following argument: '{argument}'. Given the **Premise 1: '{premise_1}'**, your task is to identify which of the following options represents the **Premise 2** that also independently supports the **Conclusion**: '{conclusion}'.
2C-P	In a Convergent argument, a **conclusion** is independently supported by **Premise 1** and **Premise 2**. Consider the following argument: '{argument}'. Identify which one of the following represents **Premise 1** and **Premise 2**, from the given set of alternatives, independently supporting the **Conclusion**: '{conclusion}'.
CR-C	In a Convergent reasoning argument, a **Conclusion** is independently supported by **Premise 1** and **Premise 2**. Consider the following argument: '{argument}'. Given the **Premise 1: '{premise_1}'** and **Premise 2: '{premise_2}'**, your task is to identify which one of the following options represents the **Conclusion** that is independently supported by **Premise 1** and **Premise 2**.
1DR-C	In divergent reasoning, a **Premise** supports multiple **Conclusions** (**Conclusion 1** and **Conclusion 2**). Consider the following argument: '{argument}'. Given the **Premise: '{premise}'**, and **Conclusion 1: '{conclusion_1}'**, your task is to identify which one of the following options represents the **Conclusion 2** that is also supported by the **Premise**.
2DR-C	In divergent reasoning, a **Premise** supports multiple **Conclusions** (**Conclusion 1** and **Conclusion 2**). Consider the following argument: '{argument}'. Given the **Premise: '{premise}'**, your task is to identify which one of the following represents **Conclusion 1** and **Conclusion 2**, from the given set of alternatives, that are supported by the **Premise** in the provided argument.
DR-P	In divergent reasoning, a **Premise** supports multiple **Conclusions** (**Conclusion 1** and **Conclusion 2**). Consider the following argument: '{argument}'. Given the **Conclusion 1: '{conclusion_1}'** and **Conclusion 2: '{conclusion_2}'**, your task is to identify the **Premise** that supports both **Conclusion 1** and **Conclusion 2** in the provided argument.

Table 22: Task Types and Corresponding Prompts.

Example	Task Type
Please answer the following multiple-choice question: Question: A one-hop argument consists of a single inference step where a premise directly supports a conclusion. Consider the following argument: 'The notification isn't so much a problem at most airports. The electronic boards are usually updated minute by minute. The problem is that the airlines will say "Flight 100, delayed till 7:00pm." then "Flight 100, delayed till 7:05pm". And so on and so forth. They're notifying everyone...with completely useless information. Forcing them to do so more frequently isn't going to fix a thing. Forcing them to come up with an accurate estimate is what is needed. Here is an incentive: if a customer is dissatisfied with flight notifications, they ought to take their business elsewhere.'. Given the premise: 'The electronic boards are usually updated minute by minute.', your task is to identify which one of the following options represents the conclusion that is directly supported by the premise. Options: 1. The problem is that the airlines will say "Flight 100, delayed till 7:00pm." then "Flight 100, delayed till 7:05pm". 2. They're notifying everyone...with completely useless information. 3. Forcing them to do so more frequently isn't going to fix a thing. 4. if a customer is dissatisfied with flight notifications, they ought to take their business elsewhere. 5. The notification isn't so much a problem at most airports. Select the number representing the correct choice. Do not provide any explanations.	1H-C
Please answer the following multiple-choice question: Question: A one-hop argument consists of a single inference step where a premise directly supports a conclusion. Consider the following argument: 'advertising is the major reason for high sales of a product high sales are obviously a reflection of the powerful advertisements The most effective way to convince consumers to purchase a product is through advertising it is not sufficient in itself The product also should satisfy the needs of the consumers advertisers push the limits of creativity to dispose the consumers to purchase the product Advertisement is the most effective way to create a well-known product They make the product preferable We are mainly introduced to products through advertisements When the consumers are impressed by the way a product is advertised, they can be convinced to consider that the product is a need in some cases Recently, there is a very creative advertisements of a soft drink product on TV The story delivers a desired call to drink that soft drink that people tend to drink when the weather is too hot the number of that product being sold will increases the more an advertisement of a product takes place in mass media, the more popular the product becomes Consumers tend to purchase the most known product when it comes to picking one out of two different brands of the same product When a product is commonly used, it becomes trustworthy for the society, no matter what quality it is it also has to be affordable for the consumer advertisements have undeniable affects on the society about the product being advertised'. Given the conclusion: 'the number of that product being sold will increases', your task is to identify which one of the following options can serve as the premise that supports this conclusion. Options: 1. high sales are obviously a reflection of the powerful advertisements 2. the more an advertisement of a product takes place in mass media, the more popular the product becomes 3. it also has to be affordable for the consumer 4. advertising is the major reason for high sales of a product 5. Recently, there is a very creative advertisements of a soft drink product on TV Select the number representing the correct choice. Do not provide any explanations.	1H-P
Please answer the following multiple-choice question: Question: A two-hop serial argument involves two inference steps: a premise supports conclusion 1 (the intermediate conclusion), and conclusion 1 further supports a final conclusion 2 in a chain.Consider the following argument: 'there's a thing happening at the moment we do have to protect our representatives our representatives are our representatives we need to make sure that our representatives are safe, first and foremost Sir David Amess is an extraordinary man Brian Cox didn't know Sir David Amess what comes across about Sir David Amess's life was this was a man who was very ecumenical in his beliefs Sir David Amess had great relationships with our Islamic brethren, for example clearly Sir David Amess was targeted Sir David Amess was targeted for being across the divide we have no idea why Sir David Amess was killed Brian Cox thought Sir David Amess having been targeted for being across the divide was shocking when a group of the Islamic brethren came along and they and Brian Cox were talking about Sir David Amess and how much they respected this man and how wonderful this man was Brian Cox thought, now Sir David Amess is a good guy, this is one of the good guys they're killing one of the good guys them killing one of the good guys is what scared Brian Cox'. Given the premise: 'Sir David Amess had great relationships with our Islamic brethren, for example', your task is to identify which of the following options can serve as conclusion 1, that connects the premise to conclusion 2: 'Sir David Amess was targeted for being across the divide'. Options: 1. Brian Cox thought, now Sir David Amess is a good guy, this is one of the good guys 2. them killing one of the good guys is what scared Brian Cox 3. our representatives are our representatives 4. they're killing one of the good guys 5. what comes across about Sir David Amess's life was this was a man who was very ecumenical in his beliefs Select the number representing the correct choice. Do not provide any explanations.	1Int-C

Example	Task Type
Please answer the following multiple-choice question: Question: A two-hop serial argument involves two inference steps: a Premise supports Conclusion 1 (the intermediate conclusion), and Conclusion 1 further supports a final Conclusion 2 in a chain. Consider the following argument: "HRQOL was better in Japanese postmenopausal women treated with tamoxifen than those treated with exemestane or anastrozole. Given the results of the TEAM trial, upfront use of tamoxifen followed by an aromatase inhibitor (AI) may be an important option for adjuvant endocrine therapy in Japanese postmenopausal women. HRQOL and AEs were similar with exemestane and anastrozole. Given the results of the TEAM trial, upfront use of tamoxifen followed by an aromatase inhibitor (AI) may be an important option for adjuvant endocrine therapy in Japanese postmenopausal women. Arthralgia and fatigue were less frequent, but vaginal discharge was more frequent in the tamoxifen group than in the exemestane group or anastrozole group. HRQOL was better in Japanese postmenopausal women treated with tamoxifen than those treated with exemestane or anastrozole. ES scores and CES-D scores were similar in all treatment groups. HRQOL and AEs were similar with exemestane and anastrozole. FACT-B scores were similar in the exemestane group and anastrozole group. HRQOL and AEs were similar with exemestane and anastrozole. FACT-B scores increased after treatment began and remained significantly higher in the tamoxifen group than in the exemestane group or anastrozole group during the first year (P¬†=¬†0.045). HRQOL was better in Japanese postmenopausal women treated with tamoxifen than those treated with exemestane or anastrozole'. Given the Premise: 'ES scores and CES-D scores were similar in all treatment groups' which supports Conclusion 1: 'HRQOL and AEs were similar with exemestane and anastrozole', your task is to identify which of the following options can serve as the final Conclusion 2 that is further supported by Conclusion 1. Options: 1. Arthralgia and fatigue were less frequent, but vaginal discharge was more frequent in the tamoxifen group than in the exemestane group or anastrozole group. 2. HRQOL was better in Japanese postmenopausal women treated with tamoxifen than those treated with exemestane or anastrozole. 3. HRQOL was better in Japanese postmenopausal women treated with tamoxifen than those treated with exemestane or anastrozole. 4. FACT-B scores were similar in the exemestane group and anastrozole group. 5. Given the results of the TEAM trial, upfront use of tamoxifen followed by an aromatase inhibitor (AI) may be an important option for adjuvant endocrine therapy in Japanese postmenopausal women. Select the number representing the correct choice. Do not provide any explanations.	2H-C
Please answer the following multiple-choice question: Question: A two-hop serial argument involves two inference steps: a Premise supports Conclusion 1 (the intermediate conclusion), and Conclusion 1 further supports a final Conclusion 2 in a chain. Consider the following argument: 'Dog owners should pay higher fines for dog dirt left on pavements, although there aren't enough bins and bag-dispensers for dog dirt. One reason for this is that they have thus far hardly had to fear the consequences despite the obligation to clean up. A higher fine, so dog owners would have to dig deeper into their pockets, is supposed to be a deterrent after all. The city, especially the green spaces, should be kept tidy after all, for they are there for our recreation. Besides you're not allowed to leave other rubbish without punishment'. Given Conclusion 2: 'Dog owners should pay higher fines for dog dirt left on pavements,' which is supported by Conclusion 1: 'The city, especially the green spaces, should be kept tidy after all,', your task is to identify which of the following options can serve as the Premise that supports Conclusion 1 Options: 1. Besides you're not allowed to leave other rubbish without punishment. 2. One reason for this is that they have thus far hardly had to fear the consequences despite the obligation to clean up. 3. A higher fine, so dog owners would have to dig deeper into their pockets, is supposed to be a deterrent after all. 4. although there aren't enough bins and bag-dispensers for dog dirt. 5. for they are there for our recreation. Select the number representing the correct choice. Do not provide any explanations.	2H-P

Example	Task Type
Please answer the following multiple-choice question: Question: A three-hop serial argument involves three inference steps: a Premise supports Conclusion 1, Conclusion 1 supports Conclusion 2, and Conclusion 2 further supports Conclusion 3 in a chain. Consider the following argument: 'given that Russia, India and China didn't turn up, xxx is how successful are the breakthrough pledges of COP-26, a glimmer of hope or blah blah blah the only country with tropical forests, has seen a rapid decline in deforestation Indonesia are continuing with a rapid decline in deforestation Indonesia put a price of carbon in place Indonesia have aggressive government policies to continue in the direction of decreasing deforestation Indonesia know the world doesn't want unsustainable palm oil anymore Indonesia know they are destroying livelihoods of indigenous people if they continue like this so Indonesia is actively involved in preserving our natural capital Indonesia is probably best placed in the future in this world to monetise the enormous natural capital that they have to confer their economy to this greener more sustainable economy the leaders didn't have support Tim Stanley is feeling positive about COP26, actually what COP26 has done is come up with deadlines and goals goals and deadlines push the world in a certain direction once, most catalytically, we begin to invest private capital in the technology, that will develop momentum'. Given the Premise: 'Indonesia know the world doesn't want unsustainable palm oil anymore', your task is to identify which one of the following options represents Conclusion 1 that is logically supported by the Premise, and which one represents Conclusion 2 that is supported by Conclusion 1, such that Conclusion 2 further supports Conclusion 3: 'Indonesia is probably best placed in the future in this world to monetise the enormous natural capital that they have to confer their economy to this greener more sustainable economy.' in the chain. The missing argument components must logically align with the provided context, ensuring that Conclusion 1 is supported by the Premise, Conclusion 2 is supported by Conclusion 1, and Conclusion 3 is supported by Conclusion 2. Options for Conclusion 1: 1. Indonesia put a price of carbon in place 2. Indonesia know they are destroying livelihoods of indigenous people if they continue like this 3. what COP26 has done is come up with deadlines and goals 4. given that Russia, India and China didn't turn up, xxx is how successful are the breakthrough pledges of COP-26, a glimmer of hope or blah blah blah 5. Indonesia have aggressive government policies to continue in the direction of decreasing deforestation Options for Conclusion 2: 1. Indonesia are continuing with a rapid decline in deforestation 2. goals and deadlines push the world in a certain direction 3. Indonesia put a price of carbon in place 4. Tim Stanley is feeling positive about COP26, actually 5. so Indonesia is actively involved in preserving our natural capital Select the correct options for each answer as a list of numbers (the first number represents Argument B, and the second represents Argument C). Do not provide any explanations. Just return the correct numbers as a list.	2Int-C
Please answer the following multiple-choice question: Question: In a linked argument, a conclusion is supported jointly by multiple premises (Premise 1, Premise 2. Consider the following argument: 'the national living wage has already increased by four thousand pounds the amount people are getting the facts are that we need to pay for whatever it is that we do provide we have to work on the facts here the universal credit system is working vastly better than the system that it replaced the universal credit system actually handled the coronavirus increase in the number of people who required assistance very well Grant Shapps hears what Munira Wilson and others say about the trap that people find themselves in people would end up having to pay one per cent extra on their national insurance perhaps Grant Shapps wasn't clear six billion a year is the equivalent of not just one penny on income tax it's a penny on income tax plus three pence on fuel duty you wanted to do that all through income tax people would have even more money to pay in income tax the 20 pounds a week is one part the gentleman said it all sounded very complicated tax and the way that the system works just is very complicated some of which didn't exist before doubling doubling the tax threshold the amount people could earn from six and a half thousand to 12 and a half thousand doubling the tax threshold has allowed people a lot more tax free earnings you do have to look at the overall amount'. Given the Premise 1: 'you wanted to do that all through income tax', your task is to identify which of the following options represents the Premise 2 that, when used jointly with Premise 1, directly supports the conclusion: 'people would have even more money to pay in income tax' Options: 1. the 20 pounds a week is one part 2. tax and the way that the system works just is very complicated 3. the national living wage has already increased by four thousand pounds the amount people are getting 4. perhaps Grant Shapps wasn't clear 5. it's a penny on income tax plus three pence on fuel duty Select the number representing the correct choice. Do not provide any explanations.	1L-P

Example	Task Type
Please answer the following multiple-choice question: Question: In two linked premises, a conclusion is supported jointly by Premise 1 and Premise 2. Consider the following argument: 'While notifying customers of delays as soon as is feasibly possible is an admirable goal, I wonder if delays of 30 minutes would actually affect passenger behavior. In my experience it usually takes about 30 minutes to get to major airports, in which case, killing 30 minutes at home or at the airport makes little difference especially when security lines make you leave way earlier than your expected flight. Maybe it would be a better use of the airlines resources to focus efforts on notifying passengers of delays that are 2 hours or more as soon as possible. As a very frequent traveler, the biggest issue is not having updated information about flight status. beyond that i believe that adding all of these unnecessary burdens to airlines often of which they have no control such as weather and unexpected mechanical problems is both costly and ridiculous. you would think that with the economic turndown, our society would abate their temper tantrams and demanding attitudes. let's be civil to one another and not set unrealistic expectations it costs money to accommodate your pampering and often results in unintended consequences such as the recent 3 hour tarmac limit. this will result in more flight cancellations. i wonder how many of you who complained about that and now have gotten what you asked for, will see yourselves as the cause of this ludicrous rule. Last winter I had to make two roundtrips to an airport in a blizzard to take family members for a flight posted as "on time". They had nonrefundable tickets and saw no option but to go or lose their money. Seven hours later their flight was canceled. I feel that, in this case, the airline acted in their own interest giving no thought to the impact on the passengers. Surely they had better information. That said, I don't have a suggestion as I don't know enough about airline information systems'. Identify which one of the following represents Premise 1 and Premise 2, from the given set of alternatives, jointly supporting the conclusion: ' I wonder if delays of 30 minutes would actually affect passenger behavior' Options for Premise 1: 1. let's be civil to one another and not set unrealistic expectations 2. Seven hours later their flight was canceled. 3. I feel that, in this case, the airline acted in their own interest giving no thought to the impact on the passengers. 4. i wonder how many of you who complained about that and now have gotten what you asked for, will see yourselves as the cause of this ludicrous rule. 5. In my experience it usually takes about 30 minutes to get to major airports Options for Premise 2: 1. beyond that i believe that adding all of these unnecessary burdens to airlines often of which they have no control such as weather and unexpected mechanical problems is both costly and ridiculous. 2. this will result in more flight cancellations. 3. As a very frequent traveler, the biggest issue is not having updated information about flight status. 4. I feel that, in this case, the airline acted in their own interest giving no thought to the impact on the passengers. 5. in which case, killing 30 minutes at home or at the airport makes little difference especially when security lines make you leave way earlier than your expected flight. Select the correct options for each answer as a list of numbers (the first number represents Premise 1, and the second represents Premise 2). Do not provide any explanations.	2L-P

Example	Task Type
In a linked reasoning argument, a conclusion is supported jointly by Premise 1 and Premise 2. Consider the following argument: 'Arlene Foster is critical about the protocol Arlene Foster did describe the renegotiated protocol as a serious and sensible way forward no, Arlene Foster did not describe the renegotiated protocol as a serious and sensible way forward Arlene Foster does not think she described the renegotiated protocol as a serious and sensible way forward of course Boris Johnson decided not to those matters and it was just around SPS and animal checks at that time xxx is teething problems or deeper issues with the protocol xxx is a combination of things xxx is partly teething problems some of the problems we are having will be overcome it is a question of traders learning how to use the forms, fill them in, but there are other factors that come into play at the moment we have a series of grace periods for food imports into Northern Ireland the grace periods for food imports into Northern Ireland ends on 1 April it will get harder again there will be more delay at the borders the real test of this agreement is going to come when we come out the other side of the pandemic, trade and travel come back to closer to their normal level, at which point it is only then we won't be able to judge at the moment we are dealing with a period of much reduced trade and travel COVID restrictions have reduced travel we shouldn't expect things to be the same as before 31 December what Brexit means is there are checks between Great Britain and Northern Ireland checks between Great Britain and Northern Ireland is going to mean in cases it might mean shortages what we have at the moment is a whole range of protocol actions that causes damage to Northern Ireland and which is not what we were talking about in October 2019 when we were saying in 2019 that it was important that the Northern Ireland assembly had a say in those matter'. Given the Premise 1: 'the grace periods for food imports into Northern Ireland ends on 1 April' and Premise 2: 'at the moment we have a series of grace periods for food imports into Northern Ireland ', your task is to identify which one of the following options represents the Conclusion that is jointly supported by Premise 1 and Premise 2. Options: 1. no, Arlene Foster did not describe the renegotiated protocol as a serious and sensible way forward 2. what we have at the moment is a whole range of protocol actions that causes damage to Northern Ireland and which is not what we were talking about in October 2019 3. checks between Great Britain and Northern Ireland is going to mean in cases it might mean shortages 4. Arlene Foster is critical about the protocol 5. it will get harder again Select the number representing the correct choice. Do not provide any explanations.	LR-C
Please answer the following multiple-choice question: Question: In a Convergent argument, a conclusion is independently supported by Premise 1 and Premise 2. Consider the following argument: 'settling is is not admitting wrongdoing It can cost more to fight a lawsuit to prove your innocence than to just settle with no admission of wrong-doing Not saying that's what happened in TRUMP's case though Thank you for clarifying for me technically no that's part of the settlement agreement Pay hush money and technically people have to shut up about it and you don't take blame Depends on the agreement TRUMP's specified that they weren't'. Given the Premise 1:'that's part of the settlement agreement', your task is to identify which of the following options represents the Premise 2 that also independently supports the Conclusion: 'technically no'? Options: 1. TRUMP's specified that they weren't 2. settling is / is not admitting wrongdoing 3. Thank you for clarifying for me 4. Depends on the agreement 5. Pay hush money and technically people have to shut up about it and you don't take blame Select the number representing the correct choice. Do not provide any explanations.	1C-P

Example	Task Type
Please answer the following multiple-choice question: Question: In a Convergent argument, a conclusion is independently supported by Premise 1 and Premise 2. Consider the following argument: 'Absolutely we've seen this week actually that the police have all the powers they need to deal with Insulate Britain it's important to say, Munira Wilson doesn't agree with the means at all the cause Munira Wilson does agree with we're a leader in climate change policies these people are campaigning to insulate Britain when the Liberal Democrats were in government between 2010 and 2015, we had a zero carbon home standard the zero carbon homes standard was got rid of once the Tories were on their own not a single zero carbon home has been built since the Tories were on their own the Liberal Democrats had an obligation on energy companies to pay for and support insulation of homes, that was scrapped energy bills are going up apart from the fact that global prices have gone up our homes are not insulated very well at all we want to cut fuel poverty we want to cut bills we want to cut emissions we need to have an emergency insulation programme Munira Wilson absolutely agrees with Insulate Britain's aims Munira Wilson doesn't agree with the means no, Grant Shapps is saying the police don't have all the powers they need to deal with Insulate Britain'. Identify which one of the following represents Premise 1 and Premise 2, from the given set of alternatives, independently supporting the Conclusion: 'we need to have an emergency insulation programme' Options for Premise 1: 1. not a single zero carbon home has been built since the Tories were on their own 2. it's important to say, Munira Wilson doesn't agree with the means at all 3. the cause Munira Wilson does agree with 4. Munira Wilson doesn't agree with the means 5. we want to cut emissions Options for Premise 2: 1. we want to cut emissions 2. we're a leader in climate change policies 3. the zero carbon homes standard was got rid of once the Tories were on their own 4. it's important to say, Munira Wilson doesn't agree with the means at all 5. we want to cut fuel poverty Select the correct options for each answer as a list of numbers (the first number represents Premise 1, and the second represents Premise 2). Do not provide any explanations.	2C-P
Please answer the following multiple-choice question: Question: In a Convergent argument, a conclusion is independently supported by Premise 1 and Premise 2. Consider the following argument: 'Absolutely we've seen this week actually that the police have all the powers they need to deal with Insulate Britain it's important to say, Munira Wilson doesn't agree with the means at all the cause Munira Wilson does agree with we're a leader in climate change policies these people are campaigning to insulate Britain when the Liberal Democrats were in government between 2010 and 2015, we had a zero carbon home standard the zero carbon homes standard was got rid of once the Tories were on their own not a single zero carbon home has been built since the Tories were on their own the Liberal Democrats had an obligation on energy companies to pay for and support insulation of homes, that was scrapped energy bills are going up apart from the fact that global prices have gone up our homes are not insulated very well at all we want to cut fuel poverty we want to cut bills we want to cut emissions we need to have an emergency insulation programme Munira Wilson absolutely agrees with Insulate Britain's aims Munira Wilson doesn't agree with the means no, Grant Shapps is saying the police don't have all the powers they need to deal with Insulate Britain'. Given the Premise 1: 'we want to cut emissions' and and Premise 2: 'we want to cut fuel poverty', your task is to identify which one of the following options represents the Conclusion that is independently supported by Premise 1 and Premise 2. Options: 1. not a single zero carbon home has been built since the Tories were on their own 2. it's important to say, Munira Wilson doesn't agree with the means at all 3. the cause Munira Wilson does agree with 4. Munira Wilson doesn't agree with the means 5. we need to have an emergency insulation programme Select the number representing the correct choice. Do not provide any explanations.	CR-C

Example	Task Type
Please answer the following multiple-choice question: Question: In divergent reasoning, a Premise supports multiple Conclusions (Conclusion 1 and Conclusion 2). Consider the following argument: 'Headache, fatigue, and drowsiness were similar in the 2 groups. Brimonidine is safe and effective in lowering IOP in glaucomatous eyes. Headache, fatigue, and drowsiness were similar in the 2 groups. Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response. Allergy was seen in 9% of subjects treated with brimonidine. Brimonidine is safe and effective in lowering IOP in glaucomatous eyes. Allergy was seen in 9% of subjects treated with brimonidine. Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response. No evidence of tachyphylaxis was seen in either group. Brimonidine is safe and effective in lowering IOP in glaucomatous eyes. No evidence of tachyphylaxis was seen in either group. Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response. Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P < .03) Brimonidine is safe and effective in lowering IOP in glaucomatous eyes. Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P < .03) Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response. Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg. Brimonidine is safe and effective in lowering IOP in glaucomatous eyes. Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg. Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response'. Given the Premise: 'Argument unit 1 Allergy was seen in 9% of subjects treated with brimonidine', and Conclusion 1: 'Brimonidine is safe and effective in lowering IOP in glaucomatous eyes', your task is to identify which one of the following options represents the Conclusion 2 that is also supported by the Premise. Options: 1. No evidence of tachyphylaxis was seen in either group. 2. Headache, fatigue, and drowsiness were similar in the 2 groups. 3. Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 (P < .03) 4. Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg. 5. Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response. Select the number representing the correct choice. Do not provide any explanations.	1DR-C
Please answer the following multiple-choice question: Question: In divergent reasoning, a Premise supports multiple Conclusions (Conclusion 1 and Conclusion 2). Consider the following argument: 'maybe slightly unpopular, we should extend the deadline we have a solution and it is the vaccine we should not put more at risk for the sake of potentially a few more weeks we should not put more people at risk 130,000-odd dead already Anthony Costello wants to have his cake and eat it on the one hand Anthony Costello is saying we should maintain a lockdown, then he says in the Far East they manage to continue to boom without a lockdown AudienceMember 20210610QT04 is as confused as Jenni as on the one hand Anthony is saying we should maintain a lockdown, then he says in the Far East they manage to continue to boom without a lockdown the vaccination programme was going really well we were on top of the vaccination programme and getting better by the day we are told that the circumstance is going to go on forever at some point we've really got to regain our lives we should have learned from what happened in Vietnam we should have been wearing face masks much earlier early in the lockdowns the scientists were poo-pooing masks saying this disease is transmitted by touch the advice just keeps changing AudienceMember 20210610QT04 wishes all these people would get together, work out, learn from what others have been doing we haven't done well but we haven't done as badly as some other countries let's move forward, get the population vaccinated, get out of it and move forward with sensible ideas that will allow us to regain our lives.' Given the Premise: 'early in the lockdowns the scientists were poo-pooing masks saying this disease is transmitted by touch', your task is to identify which one of the following represents Conclusion 1 and Conclusion 2 from the given set of alternatives, that are supported by the Premise in the provided argument. Options for Conclusion 1: 1. maybe slightly unpopular, we should extend the deadline 2. on the one hand Anthony Costello is saying we should maintain a lockdown, then he says in the Far East they manage to continue to boom without a lockdown 3. we haven't done well but we haven't done as badly as some other countries 4. 130,000-odd dead already 5. the advice just keeps changing Options for Conclusion 1: 1. we haven't done well but we haven't done as badly as some other countries 2. we were on top of the vaccination programme and getting better by the day 3. AudienceMember 20210610QT04 wishes all these people would get together, work out, learn from what others have been doing 4. we should not put more at risk for the sake of potentially a few more weeks 5. we should have been wearing face masks much earlier Select the correct options for each answer as a list of numbers (the first number represents Conclusion 1, and the second represents Conclusion 1). Do not provide any explanations.	2DR-C

Example	Task Type
Please answer the following multiple-choice question: Question: In divergent reasoning, a Premise supports multiple Conclusions (Conclusion 1 and Conclusion 2). Consider the following argument: 'successful people can learn important and very valuable points from trying new things great success requires taking great risks success and trying new things and taking risks go hand in hand and to achieve success you should consider them if you want to know how to act effectively when communicating with people, you can try something new like salesmanship in various stores and places By selling goods or other commodities you will face different people with different behaviors and personalities you will learn how to behave people in different situations and jobs in the future These experiences can only be obtained by trying new things and can't be always found in books There are always interesting and useful experiences that successful people can learn them only by trying new things If you want to gain very high profits in investments, you should use great and very high amounts of money in very risky financial decisions and dealings in which you may lose much amount of money If you want to achieve a remarkable success in an important exam, you should risk studying all the time and sacrificing your free time and your favorite hobbies The more you take risks, the greater your successes will be it is important to take risks If you don't take any risks, you will have an ordinary life with average successes'.Given the Conclusion 1: 'success and trying new things and taking risks go hand in hand and to achieve success you should consider them', and Conclusion 2: 'it is important to take risks.', your task is to identify the Premise that supports both Conclusion 1 and Conclusion 2 in the provided argument. Options: 1. There are always interesting and useful experiences that successful people can learn them only by trying new things 2. successful people can learn important and very valuable points from trying new things 3. if you want to know how to act effectively when communicating with people, you can try something new like salesmanship in various stores and places 4. The more you take risks, the greater your successes will be 5. great success requires taking great risks Select the number representing the correct choice. Do not provide any explanations.	DR-P

Table 23: Examples of selected ART types.

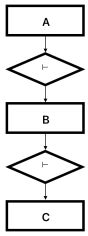
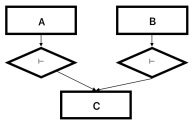
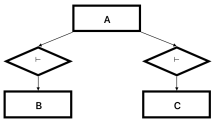
Reasoning structure	Argument diagram	Directed Graph	Premise-Conclusion
Serial Reasoning		$G = \{V, E\}$ $V = \{A, B, C\}$ $E = \{(A, B)_{\vdash}, (B, C)_{\vdash}\}$	$A \vdash B \vdash C$
Linked Reasoning		$G = \{V, E\}$ $V = \{A, B, C\}$ $E = \{(A, C)_{\vdash}, (B, C)_{\vdash}\}$	$(A \wedge B) \vdash C$
Divergent Reasoning		$G = \{V, E\}$ $V = \{A, B, C\}$ $E = \{(A, B)_{\vdash}, (A, C)_{\vdash}\}$	$(A \vdash B) \wedge (A \vdash B) \vdash C$

Table 24: Illustration of transformation from argument diagrams (left) over formal directed graphs (middle) to premise-conclusion representations (right) of the reasoning structures in ART.

F From Argument Graphs to Premise-Conclusion Structures

The complexity of argumentative reasoning is frequently represented in so called argument diagrams, which consists of boxes representing argumentative propositions and (labelled) arrows linking boxes to one another and representing argumentative relations between the propositions. Table 24 contains examples of such argument diagram in the leftmost column, representing all types of reasoning analysed in this paper. These diagram can be formalised as a directed graph as shown in the middle column. To allow for labelled edges differentiating different argumentative relations, the edge tuples receive a subscript detailing the relation. Finally, the edges can be rewritten into a linear representation of premises and conclusions as shown in the rightmost column.