

ARGQA: EVALUATION OF REASONING OVER ELEMENTARY LOGICAL STRUCTURES IN ARGUMENTS

Anonymous authors

Paper under double-blind review

ABSTRACT

As large language models advance in their reasoning capabilities, their adequate evaluation is becoming increasingly important. Existing logical reasoning benchmarks are typically constructed by automatically converting symbolic logic into natural language or curating questions from standardized exams, such as LSAT. However, both synthetic and exam-style questions contain unnatural language, thereby limiting their applicability to real-world contexts. Also, the systematic assessment of reasoning over diverse logical structures remains underexplored. Therefore, we present ARGQA, a novel dataset of 3,807 multiple-choice questions based on authentic arguments from four distinct domains—product reviews, argumentative essays, e-rulemaking comments, and medical research abstracts. Each question is designed to assess the ability to recognize and reconstruct one of three elementary logical structures—linear, convergent, and divergent—whose understanding is a prerequisite to both simple and complex reasoning. Experiments show that even the strongest LLMs still have considerable room for improvement with the overall 9-shot accuracy ranging from 29.2% (Qwen-2) to 61.8% (GPT-o3).

1 INTRODUCTION

As large language models (LLMs) demonstrate increasingly sophisticated reasoning, adequately assessing their logical reasoning abilities has become crucial for continuing the progress (OpenAI, 2025a;b; Yang et al., 2025). In line with the need, a range of benchmarks have been developed in recent years, typically by automatically converting propositions in symbolic logic into natural language (Saparov & He, 2023; Parmar et al., 2024) or by adopting exams designed for people like LSAT and GMAT (Yu et al., 2020; Zhong et al., 2022). Note, these so-called *logical* reasoning benchmarks focus purely on assessing logical reasoning capabilities. This is distinct from other reasoning benchmarks, which require general or domain-specific knowledge and the ability to reason based on such knowledge. For instance, popular benchmarks HellaSwag (Zellers et al., 2019) and WinoGrande (Sakaguchi et al., 2021) test commonsense reasoning, and GSM8k (Cobbe et al., 2021) and AIME (Patel et al., 2024) assess mathematical reasoning.

For the purpose of evaluating LLMs’ reasoning in real-world contexts, however, existing logical reasoning benchmarks are insufficient. For one, they consist of text that bears little resemblance to text in the wild, thereby limiting their relevance to real-world contexts. As one may expect, the disparity is more pronounced for synthetically generated benchmarks, which are comprised of sentences like “Sawyer is a poet.” and “Sawyer is either a musician or a poet, but not both.” (Qi et al., 2025). However, even those made of exam questions can contain language rarely used in real-life: “Seven directors—A, B, C, D, E, F, and G—serves on the X committee or the Y committee.” (Zhong et al., 2022). ReClor (Yu et al., 2020) does have many questions based on realistic arguments, since it adopts LSAT’s logical reasoning questions, which are about “arguments as they occur in ordinary language” (LSAC, 2025). However, not all questions follow this style, and the overall linguistic and topical diversity is limited. Also, the systematic assessment of reasoning over diverse logical structures remains underexplored, as questions do not target specific logical structures.

To address these issues, we present ArgumentationQA (ARGQA), a novel dataset of 3,807 multiple-choice questions (MCQs) to assess the ability to recognize and reconstruct elementary logical structures in realistic arguments. As shown in Figure 1, it was constructed based on authentic arguments from four domains—product reviews, argumentative essays, e-rulemaking comments,

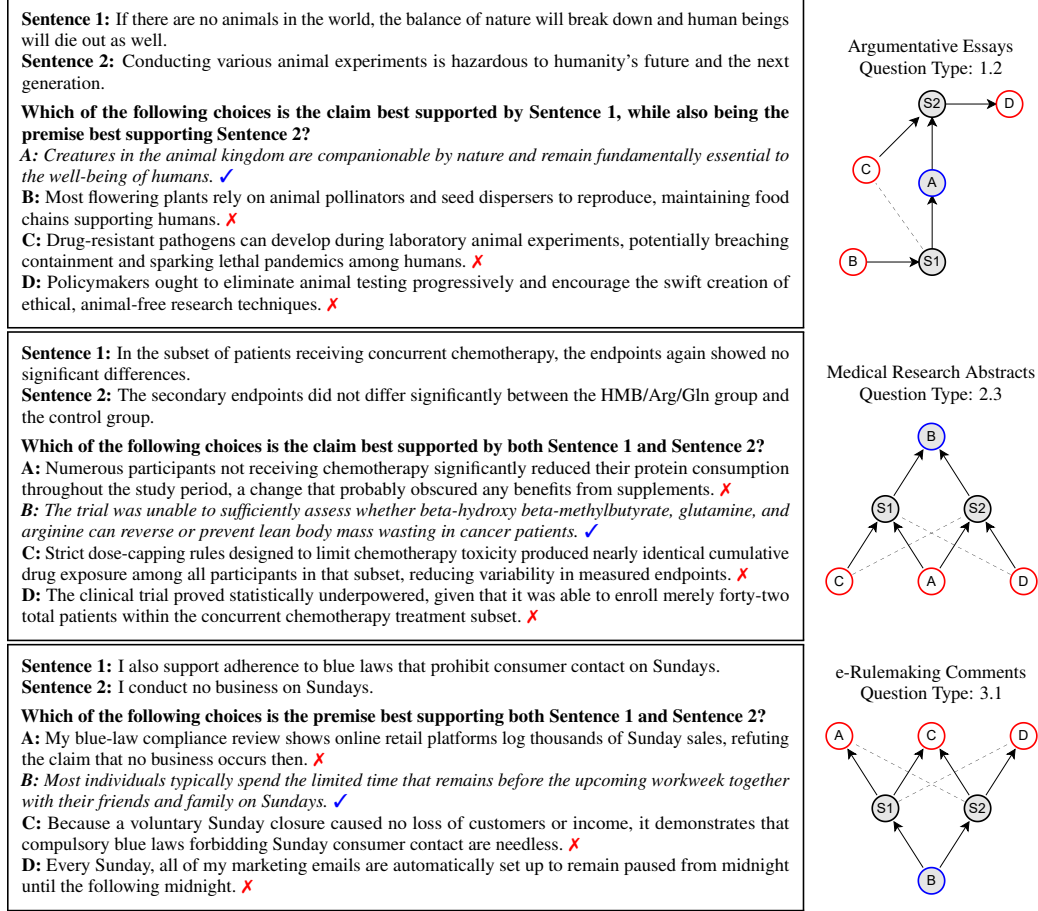


Figure 1: Examples from ARGQA. Each question assesses the ability to understand a particular elementary logical structure. Blue represents the correct option, and red, the wrong options. Gray nodes in the graphs represent the propositions comprising a real argument from the respective domain, which were rephrased by GPT-o3 to make them self-contained, e.g. by replacing pronouns with proper nouns. White nodes represent propositions newly generated with GPT-o3 as the wrong options.

and medical research abstracts—to ensure broad coverage of topics and a spectrum of authorship. The questions were designed around three elementary logical structures—linear, convergent, and divergent, as illustrated in Figure 2—which serve as the basic building blocks of arguments in real-life (Groarke et al., 1997). More specifically, there are nine question types—three types for each of the three elementary logical structures—each of which specifies how the correct and incorrect options are logically linked to the provided context, as shown in Figure 3. Because each incorrect option captures a particular misunderstanding of the logic, additional insights can be gained by analyzing incorrect responses. Lastly, the questions are in multiple-choice format to facilitate an easy integration of ARGQA into the suite of popular MCQ benchmarks that new LLMs are tested on at the time of their release, such as MMLU (Hendrycks et al., 2021) and GSM8k (Cobbe et al., 2021).

With ARGQA, we evaluate the latest LLMs’ abilities to understand elementary logical structures across various domains. In particular, we experiment with popular open-source LLMs—Mistral-7B-it-v0.3 (Jiang et al., 2023), Llama-3.1-8B-it (Grattafiori et al., 2024), Qwen-2-7B-it (Yang et al., 2024), and Gemma-7B-it (Team et al., 2024)—as well as strong proprietary models—GPT-o4-mini and GPT-o3 (OpenAI, 2025a). Experimental results show that even the latest models have considerable room for improvement, with the average accuracy ranging from 29.24% (Qwen-2) to 61.81% (GPT-o3). As expected, GPT-o3, the model with one of the strongest reasoning abilities outperforms the rest by a large margin. Among the 7B to 8B parameter models, Gemma achieves the highest overall accuracy of 44.62%, followed by GPT-o4-mini scoring 43.27%. Models exhibit similar performance

on question types sharing the same logical structure, though the propositions comprising them are drastically different. For question types with the same logical structure but with edge direction flipped, models are substantially better at identifying a premise supporting the context, rather than a claim supported by the context. Lastly, medical research abstract is the easiest domain with more clear logical relations, as opposed to argumentative essays with more obscure connections.

2 RELATED WORKS

Table 1: Comparison of ARGQA with existing logical reasoning datasets. [†]: Product reviews, argumentative essays, medical research abstracts, and e-Rulemaking comments. [‡]: The context and correct option are a real argument from the source domain paraphrased to be self-contained. Each incorrect option was generated to be logically linked to the context, but to form a logical structure different from the one specified in the question text. *: Entailment/Contradiction/Neutral/Paradox.

Dataset	Size	Source Text	Construction Method	Text Structure	Task
CLUTRR	6k	N/A	Machine-Generated	Synthetic Story	T/F
RuleTaker	500k	N/A	Machine-Generated	Set of Propositions	T/F
ProofWriter	500k	N/A	Machine-Generated	Set of Propositions	T/F with Proof
LogicNLI	20k	N/A	Machine-Generated	Set of Propositions	E/C/N/P*
SimpleLogic	560k	N/A	Machine-Generated	Set of Propositions	T/F
PrOntoQA	7.9k	N/A	Machine-Generated	Set of Propositions	T/F with Proof
LogicBench	2k	N/A	Machine-Generated	Synthetic Story	Y/N and MC
ProverQA	1.5k	N/A	Machine-Generated	Set of Propositions	T/F with Proof
FOLIO	1.4k	N/A	Human-Written	Set of Propositions	T/F/Unknown
LogiQA	8.6k	NCSE	Curated (Human-Written)	Mixed	MC
ReClor	6.1k	GMAT, LSAT	Curated (Human-Written)	Mixed	MC
AR-LSAT	2k	LSAT	Curated (Human-Written)	Mixed	MC
ARGQA (ours)	3.8k	Various[†]	Paraphrased or Generated[‡]	Argument	MC

Logical Reasoning Datasets. Given the growing need for LLMs to perform sophisticated logical reasoning, numerous datasets focusing specifically on assessing their logical reasoning abilities have been developed in recent years, as shown in Table 1. One group of logical reasoning datasets adopt questions from exams designed for people: LogiQA (Liu et al., 2021) and LogiQA 2.0 (Liu et al., 2023) consist of questions from National Civil Servants Exam (NCSE) of China, whereas ReClor (Yu et al., 2020) and AR-LSAT (Zhong et al., 2022) incorporate US-based standardized tests: LSAT and GMAT for the former, and LSAT for the latter. These datasets cover various types of questions, including drawing a conclusion from a set of premises and inferring an event from a synthetic scenario like “Seven directors—A, B, C, D, E, F, and G—serves on the X committee or the Y committee.”

The other group comprises synthetic sentences generated with rules and templates or using LLMs. These datasets were typically generated based on propositions in symbolic logic, where logical structures among the propositions are easier to control: With such high degree of control, questions were carefully designed to assess the ability to identify logical paths linking provided propositions and propositions to be verified. In RuleTaker (Clark et al., 2021), ProofWriter (Tafjord et al., 2021), LogicNLI (Tian et al., 2021), and SimpleLogic (Zhang et al., 2023), the questions are formulated as confirming new facts using provided facts and logical rules. PrOntoQA (Saparov & He, 2023) and PrOntoQA-OOD (Saparov et al., 2023) only focus on *modus ponens* deduction rule. ProverQA (Qi et al., 2025) makes the problem more difficult with so-called distractors, which are propositions unnecessary for the proof. Unlike other datasets in this group, FOLIO (Han et al., 2024) was manually written based on real world knowledge available in Wikipedia. Lastly, CLUTRR (Sinha et al., 2019) and LogicBench (Parmar et al., 2024) assess logical reasoning in the context of synthetic scenarios, rather than sets of propositions.

ARGQA complements these datasets with questions based on real arguments from various domains, carefully designed to assess the ability to understand elementary logical structures. This preserves transferability to real-world contexts, while allowing more fine-grained analyses of models with respect to different domains and logical structures.

Argument Mining Datasets. As will be discussed in Section 3.1, an argument in natural language is a set of premises supporting a claim, where the premises and the claim are propositions that are either true or false. Argument mining is a task of identifying and extracting arguments in text, which can assist deeper comprehension and critical evaluation, as well as the generation of relevant text, such as a counter-argument. To support research in this direction, many argument mining datasets have been developed over the years. Because the structure and style of writing vary from one domain to another, a dataset typically targets a single domain, such as news articles (e.g. Eckle-Kohler et al. 2015; Al-Khatib et al. 2016; Ein-Dor et al. 2020), legal documents (e.g. Poudyal et al. 2020; Grundler et al. 2022), political debates (e.g. Haddadan et al. 2019; Visser et al. 2019; Hautli-Janisz et al. 2022), student essays (e.g. Stab & Gurevych 2017; Alhindi & Ghosh 2021; Schaller et al. 2024), and user-generated content online (e.g. Boltužić & Šnajder 2014; Habernal & Gurevych 2017; Bhatti et al. 2021). In addition to the differences in domain, the authors may be driven by distinct goals, leading to different annotated components (e.g. different types of premises) and relations (e.g. different types of support relations). Also, the task itself can be posed as a structured prediction problem (e.g. input: an entire document, output: a directed graph representing the logical structure) or a binary classification problem (e.g. input: an ordered pair of propositions, output: a binary verdict on whether the first proposition supports the second), among others. Due to the limited standardization, it can be difficult to use these datasets for assessing the logical reasoning abilities of models. Also, analyzing models’ mistakes to gain additional insights is cumbersome. To the best of our knowledge, ARGQA is the first dataset in a standardized format that has been designed to evaluate the capacity to recognize elementary logical structures in real arguments and analyze mistakes in a convenient manner.

3 THE ARGUMENTATIONQA (ARGQA) DATASET

3.1 THEORETICAL BACKGROUND

Elementary Logical Structures. *Argumentation* refers to the process of constructing a natural language argument, a set of propositions (*premises*), and a target proposition (*claim*) logically supported by the premises (Stede & Schneider, 2018). With natural language arguments, unlike in formal-logic, the same sentence can be interpreted in many different ways, and their logical relationships are often not as straight-forward as, say, deductive inference. Also, some of the premises are typically left implicit—resulting in arguments known as *enthymemes* (Blair & Johnson, 1987; Walton, 2009)—because it is prohibitive to list all premises needed to support a given claim. Such fluidity of language and practical constraints limit the applicability of formal logic, whose rules are clean yet rigid. Thus, argumentation theory aims to bridge the gap between formal logic and everyday reasoning with additional flexibility from relatively under-specified logical relations like “support.” Through the study of reasoning in natural language, argumentation theorists have identified a few elementary logical structures that serve as the building blocks for all arguments (Groarke et al., 1997; Rahwan, 2008; Lawrence & Reed, 2019). Here are three common elementary logical structures in practical argumentation (Examples can be found in Figure 2):

1. **Linear Structure.** A proposition supports another proposition, which in turn, supports yet another proposition.¹
2. **Convergent Structure.** Multiple propositions independently support a proposition.
3. **Divergent Structure.** A proposition independently supports multiple propositions.

There is a fourth type of elementary logical structure, namely the *linked* structure. This structure is similar to the convergent structure in that multiple propositions support a proposition. In this case, however, the supporting propositions work collectively, thus all supporting propositions are needed for the argument to work. We exclude this structure from our work, due to its infrequent use in practice, which limits its relative significance and poses challenges for data collection.

Question Types. Recognizing or constructing elementary logical structures—which serve as the building blocks for both simple and complex logical structures—is a prerequisite for strong logical reasoning. One way to assess this ability in a standardized format is by designing MCQs that require selecting a proposition which, when combined with two provided propositions, forms a target

¹Here, the second proposition is a *claim* with respect to the first, and *premise* in relation to the third.

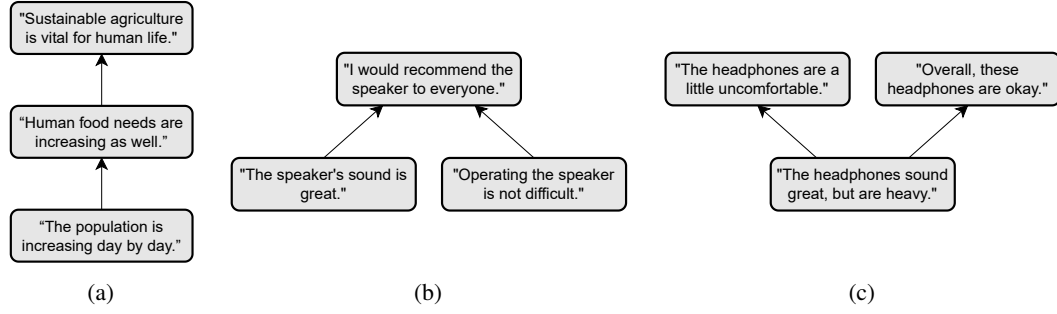


Figure 2: Examples of three elementary logical structures found in practical argumentation: (a) linear, (b) convergent, and (c) divergent structure. Directed edges represent support relations.

elementary logical structure. Since each structure consists of three propositions, we define three question types for each structure for a total of nine question types, as presented in Figure 3.

The logical distractors are designed so that they are logically related to one or both of the context propositions. This results in their forming various logical structures with the context, requiring the capability to correctly recognize the specific logical structure described in the question. In addition, this naturally prevents the correct option from being the only option that is logically related to the context—or, even worse, the only option that is topically relevant. Both cases can be easily exploited, potentially hindering a proper evaluation of the logical reasoning abilities.

Note, we assume the transitivity of support relations. That is, if proposition a supports proposition b , which in turn supports proposition c , then proposition a also supports proposition c . This is typically, though not necessarily, the case in real arguments. As a side-effect, this limits the ways in which distractors are related to the context. For instance, in question type 3.2, a distractor cannot be supported by $S1$, because it would form the correct structure with the context. More over, it cannot be supported by $S2$, because that would mean it is also supported by $S1$ —by transitivity—which forms the correct structure. Thus, the only possible logical relations for distractors are supporting $S1$ or $S2$.

3.2 DATASET CONSTRUCTION

We aim for questions based on real arguments to assess the capacity to understand elementary logical structures. Following the aforementioned question design, we construct the ARGQA dataset through a multi-stage pipeline, with each step optimized through pilot experiments:

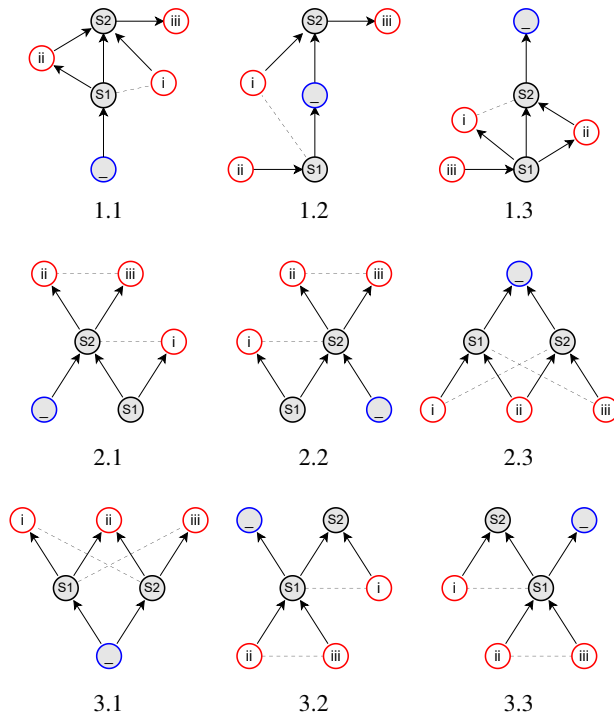


Figure 3: Question types and their logical structures. Directed edges represent support relations, and dashed edges represent the explicit lack of support relations in either direction. Grey nodes constitute the target elementary logical structure formed by the context $S1$, $S2$, and the correct option $_$ in blue. The three nodes in red are incorrect options, which are logical distractors related to the context in a structure different from what is described in the question.

Step 1: Source Selection. We select argument mining datasets to acquire raw text from diverse domains, along with their corresponding logical relation annotations. We leverage publicly available argument mining datasets, which annotate documents with logical structures. However, we exclude those with synthetic text (e.g. Peldszus & Stede 2015), with an annotation scheme disallowing one or more of the three elementary logical structures (e.g. Poudyal et al. 2020), and without enough context to make propositions interpretable (e.g. Hautli-Janisz et al. 2022). From over 20 datasets we considered, only the following meet the criteria: (1) AAEC2 (Stab & Gurevych, 2014) consists of argumentative essays by students on contentious issues, often using implicit reasoning allowing for multiple interpretations; (2) CDCP (Park & Cardie, 2018) is composed of user comments on e-Rulemaking, presenting opinions and related experiences regarding consumer debt collection practices; (3) AbstrCT (Mayer et al., 2020) features abstracts of Randomized Controlled Trials (RCT) from the MEDLINE database, often displaying clear and concise logic; and (4) AM² (Chen et al., 2022) is a collection of product reviews from Amazon, characterized by grammatical flexibility.

Step 2: Triplet Extraction. From each dataset, we extract triplets of propositions that form elementary logical structures of our interest: *linear*, *convergent*, and *divergent*. For this, we first construct directed graphs of logical relations from the annotations, then extract triplets of propositions tagged with the elementary logical structure they form. The triplets are then filtered to prevent semantic overlaps while balancing the number of triplets across the logical structures: All triplets in the divergent structure—the most rare one—are selected, then the same number of triplets for each of linear and convergent structure are selected, with each proposition appearing in one triplet only.

Step 3: Proposition Paraphrasing. We paraphrase each proposition to make them self-contained. Argument mining datasets typically keep the original text, thus some propositions cannot be interpreted without the context. However, because our questions contain triplets without the context, they need to be self-contained. Thus, we use GPT-o3 to paraphrase each proposition to a self-contained and grammatically complete sentence, through co-reference resolution and sentence completion using the original context. For example, “JUNK!” is paraphrased as “These headphones are junk!” based on the context of the product review in which it appears. (See Table 7 for the prompt.)

Step 4: Triplet Deduplication. We filter triplets comprised of semantically similar propositions. We do so by measuring the cosine similarity between the SBERT embeddings (Reimers & Gurevych, 2019) of all pairs of propositions within each triplet. When the similarity is above a threshold, the triplet is discarded. This is to exclude triplets where semantically identical or similar propositions have been wrongly annotated as in support relations.

Step 5: Distractor Generation. For each triplet, we construct three *context-options* pairs by splitting the triplet into context—two propositions—and one correct option in three ways. Each pair is tagged with the corresponding question type. Then, for each pair, we generate three logical distractors with GPT-o3 by feeding in the context and a description of the logical relation between the context and the proposition to be generated, specified by the question type. These distractors constitute the incorrect options and are added to the respective *context-options* pairs. (See Table 8 for the prompt.)

Step 6: Option Rephrasing. For each *context-options* pair, we rephrase both correct and incorrect options to similar lengths. This is to prevent potential differences in lengths revealing the correct option. For this, we compute the median word count across the four options and paraphrase each option to have its length fall within two words of the median. (See Table 9 for the prompt.)

Step 7: Question Construction. Finally, we construct a question for each *context-options* pair by combining it with a pre-written question describing the logical relationship between the context and the correct option, such as “Which of the following choices is the premise best supporting Sentence 1?” for question type 1.1. (See Table 4 for the full list of question text.) Each question is composed of two context propositions, one question text, and four options—of which one is correct—in a randomized order. Note, each question is tagged with domain, as well as the question type.

3.3 RESULTING DATASET

An summarized in Table 2, ARGQA consists of 3,807 MCQs, evenly split across the nine question types, and in turn the three elementary logical structures. For each domain, nine questions are reserved as the development set used as examples during few-shot experiments. The remainder are randomly split into 10% validation and 90% test set. Domains are distributed as follows: 1,620

questions on product reviews (42.6%), 1,350 on argumentative essays (35.5%), 567 on e-Rulemaking comments (14.9%), and 270 on medical research abstracts (7.1%). Interestingly, their ranking by length is the opposite: medical research abstracts has the longest context and option propositions, followed by e-Rulemaking comments, argumentative essays, and product reviews. The same goes for ordering by linguistic diversity as measured by vocabulary size per instance.

To confirm the quality, we manually analyzed 108 validation instances randomly selected through stratified sampling, i.e., 27 per domain. 93.8% of the options are in the correct relation to the context as specified by the question type. The common error patterns differ for the correct and the incorrect options. For correct options, the lack of explicit context often obscures the logic. For instance, the relation between two propositions each discussing a company’s and a student’s perspective is not as clear without knowing that the student is an intern. For incorrect options, it is common to have a reason embedded in the option, e.g. “B, because A.” Then, it is not supported by a context proposition stating A—though it was supposed to—since A is already stated in the option.

Table 2: Overview of ARGQA. Source arguments were collected from argument mining datasets across four domains for a wide range of topics and writing styles. Each domain-split contains the same number of questions for each of the nine question types. For instance, the validation set for e-Rulemaking comments comprises six questions per question type, totaling 54. *: number of words. ♣: Chen et al. (2022). ♦: Stab & Gurevych (2014). ♥: Park & Cardie (2018). ♠: Mayer et al. (2020).

Domain	Source Dataset	# of Instances				Vocab*	Avg. Sentence Len.	
		Dev	Val	Test	Total		Context*	Option*
Product Reviews	AM ² ♣	9	162	1,449	1,620	8,616	11.6	15.7
Argumentative Essays	AAEC2 ♦	9	135	1,206	1,350	11,465	16.8	20.3
e-Rulemaking Comments	CDCP ♥	9	54	504	567	6,211	18.0	21.1
Medical Research Abstracts	AbstRCT ♠	9	27	234	270	4,640	26.9	24.4
ARGQA		36	378	3,393	3,807	18,452	15.5	18.8

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We use ARGQA to assess the logical reasoning abilities of popular open-source LLMs—Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2-7B-Instruct (Yang et al., 2024), and Gemma-7B-Instruct (Team et al., 2024)—as well as strong proprietary models—GPT-o4-mini and GPT-o3 (OpenAI, 2025a). We focus on 7B to 8B parameter models for a fair comparison. However, GPT-o3 is also included as a representative of the LLMs with the strongest reasoning capabilities; among the available GPT models we pilot-tested, GPT-o3 performed the best. Experiments were conducted on the held-out test set in zero-shot and nine-shot settings, using the development set as examples of the nine question types. The performance was measured in accuracy, i.e., the percentage of correct responses. We adapted a multiple-choice prompting protocol from MMLU-Pro (Wang et al., 2024), instructing models to return answers in the format “the answer is (X)”, and automatically extracting the answer using regular expressions. The exact wording (See Figure 10 for the prompt), as well as maximum output length (32 tokens) and repetition penalty (1.0) were finalized based on the validation performance. Lastly, open-source models were evaluated with greedy decoding, with temperature set to 0, for maximally deterministic responses.

4.2 RESULTS & ANALYSIS

As presented in Table 3², the overall performance of the LLMs measured in accuracy on the ARGQA test set for all domains ranges from 29.24% (Qwen-2) to 61.81% (GPT-o3), with GPT-o3 considerably outperforming the rest of the models. This is in line with expectations, since GPT-o3 is a substantially larger model specializing in logical reasoning. Among the other models—which all have 7B to 8B parameters—Gemma is generally the best performer, followed by GPT-o4-mini. Qwen-2 and Mistral rank the lowest with only a modest 4% improvement from the random baseline of 25%,

²See Table 6 for 0-shot results. As expected, the performance generally degrades without examples.

Table 3: Nine-shot performance of LLMs on the ARGQA test set for all domains. Each cell reports accuracy (%), the percentage of questions that were correctly answered. Open-source models are evaluated with greedy decoding, temperature set to 0, for maximally deterministic responses. LLMs show considerable room for improvement, with GPT-o3 substantially outperforming the rest.

Logical Structure	Mistral	LLaMA-3.1	Qwen-2	Gemma	GPT-o4-mini	GPT-o3	Avg.
Linear	30.77	31.92	24.84	36.69	41.47	55.71	36.90
Q-type: 1.1	37.67	43.50	29.97	46.95	54.38	59.95	45.40
Q-type: 1.2	29.97	33.16	27.85	41.38	43.24	65.00	40.10
Q-type: 1.3	24.67	19.10	16.71	21.75	26.79	42.18	25.20
Convergent	30.15	41.03	29.97	52.96	50.22	64.99	44.89
Q-type: 2.1	34.75	51.46	36.87	63.13	60.74	68.97	52.65
Q-type: 2.2	32.89	44.56	35.28	61.01	58.09	64.72	49.43
Q-type: 2.3	22.81	27.06	17.77	34.75	31.83	61.27	32.58
Divergent	26.97	31.48	32.89	44.21	38.11	64.72	39.73
Q-type: 3.1	48.01	58.89	46.68	74.54	75.07	80.64	63.97
Q-type: 3.2	16.98	18.04	27.85	29.18	19.10	57.03	28.03
Q-type: 3.3	15.92	17.51	24.14	28.91	20.16	56.50	27.19
All (<i>Q-type Avg.</i>)	29.30	34.81	29.24	44.62	43.27	61.81	40.51

which corresponds to randomly selecting the correct answer from four options. Overall, these results indicate that LLMs currently have insufficient abilities to recognize elementary logical structures, even those comprised of only three propositions. In other words, understanding logical structures, no matter how simple, remains a challenging problem to modern LLMs.

Analyses of the results with respect to the logical structures provide additional insights. For instance, consider Q-types 2.1 and 2.2, which share the same logical structure, but consist of different propositions: Of the six propositions making up a question, one is the same (a context proposition), two are paraphrases and play different roles (one context proposition and the correct option, and vice versa for the other Q-type), and three are entirely distinct. (Figure 6 shows how different the questions are.) Even with the considerable differences in the propositions forming the question, the performance on these question types is quite similar across the models. On the other hand, the performance is drastically different on Q-type 2.3, which shares the same type of similarities in propositions, but has a different logical structure. This suggests that the logical structure has more impact on the model response than the particular propositions forming the structure, which aligns with the goal of assessing logical reasoning with an emphasis on the structure. The same trend can be observed for Q-types 3.2 and 3.3, which also share the same logical structure but different propositions.

Also, consider Q-types 2.3 and 3.1. The logical structures for these Q-types are mirror-images of each other, i.e., they are identical, except the edges point in the opposite direction. Unlike the cases where the logical structures are the same, the performance varies drastically between the pair, with the models averaging 32.58% for the former, and 63.97% for the latter. A similar pattern can be observed between another pair with the same logical structure but flipped edges: Q-types 1.1 and 1.3. One plausible explanation is that the models are better at identifying a premise supporting a given claim(s), as opposed to recognizing a claim supported by a given premise(s). This observation is also consistent with the fact that the Q-types with the highest performance reported—Q-types 3.1, 2.1, 2.2, and 1.1—all require identifying a premise supporting claim(s) provided as context.

Among the domains, argumentative essays are the most difficult for the models, with accuracies averaging 33.78%. Arguments on contentious topics typically involves so-called *defeasible* reasoning in argumentation theory, which can be challenged in several ways, e.g. *rebutting* the claim with a different reason or *undercutting* the logical connection. For instance, in Example 1 in Figure 1, one may undercut the link between sentence 2 and option A by saying that animals used for experiments constitute only a negligibly small portion of all animals. In this way, logical connections are less concrete in this domain. On the other hand, the models perform best on medical research abstracts, with an average accuracy of 50.0%. This is partly because abstracts summarizing research papers tend to have clearer logical relations. For instance, the supported proposition can simply be generalizing the propositions that support it with more specific cases, as shown in Example 2 in Figure 1.

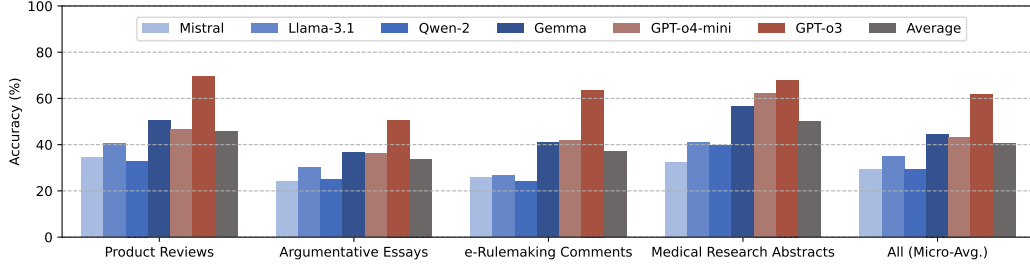


Figure 4: Nine-shot performance of LLMs on the ARGQA test set for each domain. On average, as shown in gray bars, LLMs perform best on medical research abstracts, followed by product reviews, e-Rulemaking comments and argumentative essays.

The selection rates for the logical distractors are presented in Figure 5. Note that distractors i, ii, and iii can be any of the options A, B, C, or D in a given question, since the order is randomized. Thus, patterns shown in this figure do not result from locational bias. One trend is the models’ strong preference toward the distractor linked to both context propositions. Distractor ii in Q-types 1.1, 1.3, 2.3, and 3.1 is one such distractor, and we can observe a marked inclination toward choosing it. It seems that when models fail to identify a proposition logically related to the context in a way described in the question text, they favor the proposition related to more context propositions. This does not reflect a unconditional predisposition to select the proposition associated with both context propositions, because if that were the case, models would have performed much better on Q-type 1.2, where the only proposition related to both context propositions is the correct option. Regarding Q-types with shared structures—Q-types 2.1 and 2.2, as well as 3.2 and 3.3—we not only observe similar performance, as previously discussed, but also similar distractor selection patterns; the stacked bars are near identical for these pairs across the models. This again confirms that the logical structure has a stronger impact on the choice than the particular propositions comprising the structure, which is desirable for evaluating reasoning over various logical structures.

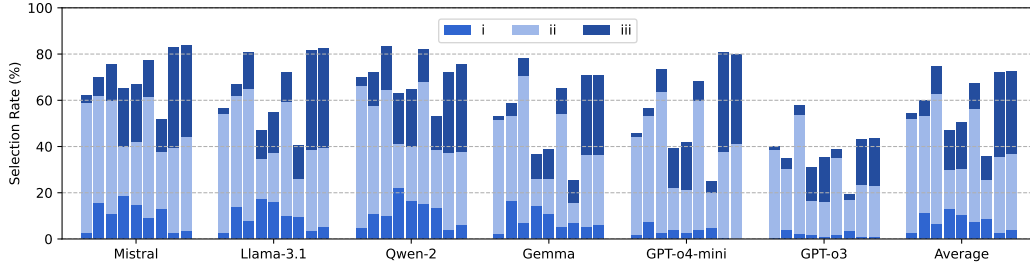


Figure 5: Selection rates for the distractors. Each group of nine stacked bars represent the percentages of selecting the three distractors for the nine question types. The stacks do not add up to 100%, because the percentage of selecting the correct option is not shown. Note, generalized claims about a particular distractor number, e.g. distractor i, cannot be made based on this figure, because the distractors’ logical relationship to the context is different for each question type, as shown in Figure 3.

5 CONCLUSIONS

In this work, we introduce ARGQA, a novel dataset designed to evaluate logical reasoning abilities over elementary logical structures: linear, convergent, and divergent. By moving beyond synthetic and exam-style questions to realistic arguments from various domains, ARGQA addresses the critical issue of limited applicability suffered by existing logical reasoning benchmarks. Also, the systematic design of question types enables convenient analyses of error patterns with respect to each elementary logical structure. Experiments on the latest LLMs reveal that even the strongest reasoning models like GPT-o3 have considerable room for improvement. With its standardized format, we hope ARGQA becomes a valuable resource for assessing LLMs’ reasoning capability.

6 ETHICS STATEMENT

ARGQA is constructed by leveraging publicly available argument-mining datasets (AM², CDCP, AAEC2 and AbstRCT) to paraphrase and generate new propositions. Our project did not involve any human participants, and no private or personally identifying information was collected or published.

7 REPRODUCIBILITY STATEMENT

Included with our submission are all the code for dataset construction and experiments, along with the data. This is to ensure the transparency in data construction, verification of the experimental process, and reproducibility of the results. Upon acceptance, we plan to formally release the code and the dataset via an open source repository.

REFERENCES

- Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In Yuji Matsumoto and Rashmi Prasad (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3433–3443, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL <https://aclanthology.org/C16-1324/>.
- Tariq Alhindi and Debanjan Ghosh. “Sharks are not the threat humans are”: Argument component segmentation in school student essays. In Jill Burstein, Andrea Horbach, Ekaterina Kochmar, Ronja Laarmann-Quante, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch (eds.), *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 210–222, Online, April 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.bea-1.22/>.
- Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. Argument mining on Twitter: A case study on the planned parenthood debate. In *Proceedings of the 8th Workshop on Argument Mining*, pp. 1–11, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.argmining-1.1>.
- J. Anthony Blair and Ralph H. Johnson. The current state of informal logic. *Informal Logic*, 9(2), 1987. doi: 10.22329/il.v9i2.2671.
- Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker (eds.), *Proceedings of the First Workshop on Argumentation Mining*, pp. 49–58, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2107. URL <https://aclanthology.org/W14-2107>.
- Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. Argument mining for review helpfulness prediction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8914–8922, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.609. URL <https://aclanthology.org/2022.emnlp-main.609>.
- Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*, 2021. ISBN 9780999241165.
- Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL <https://api.semanticscholar.org/CorpusID:239998651>.
- Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In Lluís Màrquez, Chris Callison-Burch,

and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2236–2242, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1267. URL <https://aclanthology.org/D15-1267/>.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7683–7691. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6270. URL <https://doi.org/10.1609/aaai.v34i05.6270>.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparth, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpiere Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido,

Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Sweeney, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojuan Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

Leo Groarke, Christopher Tindale, and Linda Fisher. *Good reasoning matters! : a constructive approach to critical thinking*. Oxford University Press, Toronto, 1997.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. Detecting arguments in CJEU decisions on fiscal state aid. In Gabriella Lapesa, Jodi Schneider, Yohan Jo, and Sougata Saha (eds.), *Proceedings of the 9th Workshop on Argument Mining*, pp. 143–157, Online and in Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics. URL <https://aclanthology.org/2022.argmining-1.14/>.

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017. doi: 10.1162/COLI_a_00276. URL <https://aclanthology.org/J17-1004>.

- Shohreh Haddadan, Elena Cabrio, and Serena Villata. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4684–4690, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1463. URL <https://aclanthology.org/P19-1463/>.
- 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. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL <https://aclanthology.org/2024.emnlp-main.1229/>.
- Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris Reed. QT30: A corpus of argument and conflict in broadcast debate. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3291–3300, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.352>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4): 765–818, December 2019. doi: 10.1162/coli_a_00364. URL <https://aclanthology.org/J19-4006/>.
- Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2947–2962, July 2023. ISSN 2329-9290. doi: 10.1109/TASLP.2023.3293046. URL <https://doi.org/10.1109/TASLP.2023.3293046>.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*, 2021. ISBN 9780999241165.
- Law School Admission Council LSAC, 2025. URL <https://www.lsac.org/lsat/taking-lsat/test-format/logical-reasoning>. Accessed: 2025-09-10.
- Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based Argument Mining for Healthcare Applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, Santiago de Compostela / Online, Spain, August 2020. URL <https://hal.science/hal-02879293>.
- OpenAI. OpenAI o3 and o4-mini system card. System card, OpenAI, 2025a. URL <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.
- OpenAI. GPT-5 system card. System card, OpenAI, 2025b. URL <https://cdn.openai.com/gpt-5-system-card.pdf>.

- Joonsuk Park and Claire Cardie. A corpus of eRulemaking user comments for measuring evaluability of arguments. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1257/>.
- Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.739. URL <https://aclanthology.org/2024.acl-long.739/>.
- Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. Aime: Ai system optimization via multiple llm evaluators, 2024. URL <https://arxiv.org/abs/2410.03131>.
- Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pp. 801–815, 2015.
- Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. ECHR: Legal corpus for argument mining. In Elena Cabrio and Serena Villata (eds.), *Proceedings of the 7th Workshop on Argument Mining*, pp. 67–75, Online, December 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.argmining-1.8>.
- Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. Large language models meet symbolic provers for logical reasoning evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=C25SgeXWjE>.
- Iyad Rahwan. Mass argumentation and the semantic web. *Web Semant.*, 6(1):29–37, February 2008. ISSN 1570-8268. doi: 10.1016/j.websem.2007.11.007. URL <https://doi.org/10.1016/j.websem.2007.11.007>.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL <https://arxiv.org/abs/1908.10084>.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL <https://doi.org/10.1145/3474381>.
- Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations*, 2023.
- Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. DARIUS: A comprehensive learner corpus for argument mining in German-language essays. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4356–4367, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.389/>.

- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL <https://aclanthology.org/D19-1458/>.
- Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In Junichi Tsujii and Jan Hajic (eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1501–1510, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL <https://aclanthology.org/C14-1142/>.
- Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, September 2017. doi: 10.1162/COLI_a_00295. URL <https://aclanthology.org/J17-3005>.
- Manfred Stede and Jodi Schneider. *Argumentation Mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2018.
- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL <https://aclanthology.org/2021.findings-acl.317/>.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshv, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi,

- and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.
- Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3738–3747, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.303. URL <https://aclanthology.org/2021.emnlp-main.303/>.
- Jacky Visser, John Lawrence, Jean Wagemans, and Chris Reed. An annotated corpus of argument schemes in us election debates. In *Proceedings of the 9th Conference of the International Society for the Study of Argumentation (ISSA), 3-6 July 2018*, pp. 1101–1111, 2019.
- Douglas Walton. Argumentation theory: A very short introduction. In Iyad Rahwan and Guillermo R Simari (eds.), *Argumentation in Artificial Intelligence*, pp. 1–22. Springer, 2009. doi: 10.1007/978-0-387-98197-0_1.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL <https://arxiv.org/abs/2406.01574>.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL <https://arxiv.org/abs/2407.10671>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HJgJtT4tvB>.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472/>.
- Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 3365–3373. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/375. URL <https://doi.org/10.24963/ijcai.2023/375>. Main Track.

WanJun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Yining Chen, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. Analytical reasoning of text. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2306–2319, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.177. URL <https://aclanthology.org/2022.findings-naacl.177/>.

A APPENDIX

A.1 ADDITIONAL DETAILS

Table 4: Question text for each question type.

Type	Question Text
1.1	Which of the following choices is the premise best supporting Sentence 1?
1.2	Which of the following choices is the claim best supported by Sentence 1, while also being the premise best supporting Sentence 2?
1.3	Which of the following choices is the claim best supported by Sentence 2?
2.1	Which of the following choices is the premise best supporting Sentence 2?
2.2	Which of the following choices is the premise best supporting Sentence 2?
2.3	Which of the following choices is the claim best supported by both Sentence 1 and Sentence 2?
3.1	Which of the following choices is the premise best supporting both Sentence 1 and Sentence 2?
3.2	Which of the following choices is the claim best supported by Sentence 1?
3.3	Which of the following choices is the claim best supported by Sentence 1?

Table 5: Content of each ARGQA instance. The order of the four answer choice options is pre-randomized in the released file, ensuring that solvers cannot benefit from a positional bias.

Field	Type	Description
docID	integer	Index that links the item to its source document in the original argument mining corpus.
instanceID	string	Globally unique identifier formed from the split label (dev, val, test) and a counter.
structure	string	The logical structure label, one of lin, conv, div for linear, convergent, or divergent arguments.
q_type	string	Question type code, for example "1.1" for proposition prediction.
context	array of 2 strings	Two sentences that make up the argument fragment shown to the solver.
choices	array of 4 objects	Each object has text and type.
text	string	Answer text shown to the solver.
type	string	Categorical label where _ marks the single gold answer whose reasoning chain matches the target structure, and i, ii, iii mark other distractor subtypes (for example simple backward, complex forward, complex linear).

Table 6: Zero-shot performance of LLMs on the ARGQA test set for all domains. Each cell reports accuracy (%), the percentage of questions that were correctly answered. Greedy decoding, with temperature set to 0, was used to minimize variance. All LLMs tested show considerable room for improvement, with GPT-o3 substantially outperforming the rest.

Logical Structure	Mistral	LLaMA-3.1	Qwen-2	Gemma	GPT-o4-mini	GPT-o3	Avg.
Linear	26.79	27.23	24.14	27.76	39.79	46.42	32.02
Q-Type: 1.1	27.59	36.60	30.50	38.73	50.66	51.99	39.35
Q-Type: 1.2	32.10	22.55	24.40	23.34	42.71	57.29	33.73
Q-Type: 1.3	20.69	22.55	17.51	21.22	25.99	29.97	22.99
Convergent	31.04	38.02	34.30	39.88	49.43	49.87	40.42
Q-Type: 2.1	32.10	40.58	43.50	45.89	58.36	56.23	46.11
Q-Type: 2.2	30.77	41.91	40.58	44.56	53.58	53.85	44.21
Q-Type: 2.3	30.24	31.56	18.83	29.18	36.34	39.52	30.95
Divergent	30.86	31.12	33.42	33.24	38.37	42.62	34.94
Q-Type: 3.1	42.18	48.54	53.58	57.82	68.70	68.97	56.63
Q-Type: 3.2	23.08	22.81	24.14	20.69	22.55	29.97	23.87
Q-Type: 3.3	27.32	22.02	22.55	21.22	23.87	28.91	24.32
All (Micro-Avg.)	29.56	32.12	30.62	33.63	42.53	46.30	35.79

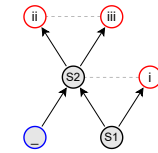
A.2 EXAMPLE QUESTIONS

Sentence 1: Wiggling the headphone jack caused both speakers to become intermittent.
Sentence 2: The QC-25 headset proved to be a disappointment.

Which of the following choices is the premise best supporting Sentence 2?

- A: After fewer than three minutes, the left speaker completely ceased to operate. ✓
 B: I sent the QC-25 headset back to the store and obtained a complete refund. ✗
 C: A new cable instantly restored perfect sound, proving the QC-25 itself was blameless. ✗
 D: QC-25's poor performance drove our audio-engineering club to schedule a seminar on noise-cancellation design flaws. ✗

Product Reviews
 Question Type: 2.1



Sentence 1: The left speaker stopped working after less than three minutes.

Sentence 2: The QC-25 headset proved to be a disappointment.

Which of the following choices is the premise best supporting Sentence 2?

- A: I asked the retailer to provide me with a complete refund. ✗
 B: I quickly returned defective headphones, bought a QC-25 that has long performed flawlessly. ✗
 C: The engineering group launched an internal probe into the product's acoustic flaws. ✗
 D: Jiggling the headphone jack caused both of the speakers to cut out intermittently. ✓

Product Reviews
 Question Type: 2.2

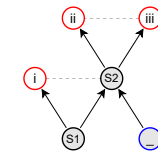


Figure 6: Examples with Equal Structure but Different Propositions

A.3 EXAMPLE PROMPTS

Table 7: Prompt for Proposition Paraphrasing for AM²**Developer Prompt:**

Identity

You are an expert editor that rewrites the sentence to become grammatically complete if necessary.

Instructions

Your task is to rewrite the original sentence, such that:

1. The generated sentence becomes a grammatically complete sentence that can stand on its own.
2. The generated sentence preserves the original meaning without adding new details or elaboration.
3. All pronouns (e.g., "he", "she", "it", "they") are replaced with appropriate proper nouns or clear references from the given Context.
4. Remove connectors (e.g., "because", "but", "so", "in order that") if necessary.
5. The original point of view (e.g., first-person) must be preserved.
6. Only output a single rewritten sentence. Do not include explanations, formatting, or additional commentary.

Examples

<product_review id="example-1">

Context: The speakers aren't even oriented around your ears, they're cockeyed.

Original Sentence: they're cockeyed.

</product_review>

<assistant_response id="example-1">

The speakers are cockeyed.

</assistant_response>

<product_review id="example-2">

Context: I hate this headset. Connection is terrible.

Original Sentence: I hate this headset.

</product_review>

<assistant_response id="example-2">

I hate this headset.

</assistant_response>

<product_review id="example-3">

Context: Spend a bit more money and get something better. I will, I have to now.

Original Sentence: I have to now.

</product_review>

<assistant_response id="example-3">

I have to purchase better quality headphones now.

</assistant_response>

<product_review id="example-4">

Context: Had to return it because the sound quality was not good.

Original Sentence: because the sound quality was not good.

</product_review>

<assistant_response id="example-4">

The sound quality was not good.

</assistant_response>

User Prompt:

Context: {full_context}

Original Sentence: {original_sentence}

Completed Sentence:

Table 8: Prompt for Generating Distractors for AM²**Developer Prompt:****# Identity**

You are an expert natural language logician. Your task is to generate a sentence that serves as the logical bridge between two given sentences. Given two sentences, A and B, your task is to generate a new sentence C, such that A "is a reason" for C, and C "is a reason" for B (A \rightarrow C \rightarrow B). You must make sure neither reverse relation holds. B must NOT be a reason for C, and C must NOT be a reason for A.

Instructions

* Sentence A must be a reason for sentence C in natural language, and sentence C must simultaneously be a reason for sentence B in natural language. Imagine the full sequence as "[Sentence A]. Because of this reason, [Sentence C]. Because of this reason, [Sentence B]."

* The relation must only go one way. Sentence B must NOT be a reason for sentence C, and sentence C must NOT be a reason for sentence A. If sentence B is a reason for sentence C, or sentence C is a reason for sentence A, your output is invalid.

* Sentence C must be distinct in meaning from sentence A or B. It must contain a new proposition without repetition from sentence A or B.

* Sentence C must be a sentence that can stand on its own. It must not have any unresolved references like pronouns that rely on sentence A or B (e.g., "it", "they", "them").

* Your response must be the single generated sentence C, with no additional formatting or explanation.

Examples

<sentence id="good-example-1">

Sentence A: The Bluetooth signal comes solely from the right earphone.

Sentence B: The issue with this Bluetooth headset can be easily overcome.

</sentence>

<assistant_response id="good-example-1">

The Bluetooth headset only has skip issues when the cell phone is in the left pocket.

</assistant_response>

<sentence id="good-example-2">

Sentence A: The instruction booklet was easy to read and understand.

Sentence B: I am glad I bought these headphones.

</sentence>

<assistant_response id="good-example-2">

I had no trouble getting the headphones out and figuring out how to use them.

</assistant_response>

<sentence id="good-example-3">

Sentence A: I am very disappointed.

Sentence B: Spend the extra money.

</sentence>

<assistant_response id="good-example-3">

Go with the more expensive alternative.

</assistant_response>

<sentence id="bad-example-3">

Sentence A: One speaker went out in less than one month.

Sentence B: The second speaker failed soon after.

</sentence>

<assistant_response id="bad-example-3">

Don't waste your money on them.

</assistant_response>

This is a bad example because both A and B support C. A, B are both reasons for C, which is unacceptable. Additionally, this example uses an unresolved reference ("them"), which is also unacceptable.

User Prompt:

Sentence A: {first_sentence}

Sentence B: {second_sentence}

Table 9: Prompt for Option Rephrasing for AAEC2

Developer Prompt:**# Identity**

You are an expert editor who rewrites sentences to precisely match a target word length.

Instructions

Your task is to rewrite the Original Sentence such that:

- * The rewritten sentence conveys exactly the same meaning. No information must be added, removed, or altered.
- * The rewritten sentence must contain about the same number of words as the target, within a two-word tolerance. Verify the word count of your rewritten sentence and revise it until it matches that target. However, do not add or remove any content in order to meet this word count — preserving the original meaning is more important than exact length.
- * The original point of view and tense must remain unchanged. You must NOT add any new unresolved references like pronouns.
- * Your response must be exactly one grammatically complete and independent sentence. Do not split it into multiple sentences.
- * Output only the rewritten sentence without any explanations, formatting, or additional commentary.

Examples

<sentence id="good-example-1">

Sentence: Budget constraints make hiring temporary replacements financially impossible for management.

Number of words targeted: 30

</sentence>

<assistant_response id="good-example-1">

Because the already strained budget leaves no available funds, management finds that hiring short-term replacement workers is entirely out of reach financially, rendering any temporary staffing plan impossible.

</assistant_response>

<sentence id="good-example-2">

Sentence: Analyses of thousands of successful career trajectories show that bold self-direction paired with meticulous preparation—not mere optimism that any gamble will succeed—is the chief driver of achievement, thereby challenging the claim that simply taking chances and believing they will work out is essential.

Number of words targeted: 15

</sentence>

<assistant_response id="good-example-2">

Analyses of thousands of careers show accomplishment arises from bold, prepared self-direction, not optimistic gamble-taking.

</assistant_response>

<sentence id="good-example-3">

Sentence: I now feel confident living independently.

Number of words targeted: 20

</sentence>

<assistant_response id="good-example-3">

At this point in my life, I genuinely feel very self-assured about managing everything and living entirely on my own.

</assistant_response>

<sentence id="bad-example-3">

Sentence: I now feel confident living independently.

Number of words targeted: 20

</sentence>

<assistant_response id="bad-example-3">

I now feel confident in my ability to manage everything financially and live independently in my dream city.

</assistant_response>

This is a bad example because the original sentence does not talk about financial management and "dream city", while the response does. It adds NEW information, which is unacceptable.

<sentence id="bad-example-4">

Sentence: My uncanny capacity to recall every television channel while forgetting textbook passages prompted my teachers to suspect a specific reading disorder, and subsequent evaluations showed that my academic results were solid rather than disastrous, contradicting the claim that insufficient studying caused my performance.

Number of words targeted: 20

<assistant_response id="bad-example-4">

My recall of TV channels while forgetting textbooks made them suspect a reading disorder until evaluations showed solid grades, debunking insufficient-study claims.

</assistant_response>

This is a bad example because the original noun "teachers" was removed after the rewritten work. The pronoun "them" is unresolved in the response, which is unacceptable.

<sentence id="bad-example-5">

Sentence: My cousin borrowed the headphones after hearing my consistently positive comments.

Number of words targeted: 25

<assistant_response id="bad-example-5">

A considerate roommate is extremely important for the quality of our life at school.

</assistant_response>

This is a bad example because the rewritten sentence contains 14 words, which is not within the permitted range of 23-27. It is thus unacceptable.

User Prompt:

Sentence: {original_sentence}

Number of words targeted: {target_length}

Table 10: Nine-shot Multiple-choice Prompt for CDCP

User Prompt:

The following are multiple choice questions. Answer with "the answer is (X)" where X is the correct letter choice.

Question

Sentence 1: Consumers should not have to pay additional costs for cell phone calls, texting, or voice messaging related to collections.

Sentence 2: Debt collectors and creditors must cease all calls and texts to any device the debtor identifies as not providing free communications.

Which of the following choices is the premise best supporting Sentence 1?

(A) The greater the costs imposed on a debtor, the less money that debtor retains for paying creditors.

(B) Persistent debt collection phone calls may mute or postpone critical emergency notifications on smartphones, thus jeopardizing public safety.

(C) Each unsolicited call or text reaching a device with paid communications should be kept as evidence of noncompliance.

(D) Collection agencies are required to restrict their collection communications to channels that place no monetary cost on the consumer.

Answer: (A)

Question

Sentence 1: The more costs assessed to a debtor, the less money the debtor has to pay creditors.

Sentence 2: Debt collectors and creditors must cease all calls and texts to any device the debtor identifies as not providing free communications.

Which of the following choices is the claim best supported by Sentence 1, while also being the premise best supporting Sentence 2?

(A) Consumers ought not be charged extra fees for collection-related cell phone calls, text messages, or voice mails.

(B) The debtor shall promptly provide a written notice that enumerates every phone number on which charges are incurred.

(C) Frequent unwanted calls to pay-per-use mobile phones may obstruct crucial emergency messages that need to reach the debtor.

(D) Before distributing any payments to creditors, court fees and collection charges are subtracted from the debtor's already limited funds.

Answer: (A)

Question

Sentence 1: The more costs assessed to a debtor, the less money the debtor has to pay creditors.

Sentence 2: Consumers should not have to pay additional costs for cell phone calls, texting, or voice messaging related to collections.

Which of the following choices is the claim best supported by Sentence 2?(A) Reducing supplementary fees that are imposed on debtors ultimately maximizes the financial resources they can allocate toward repaying their original obligations.(B) Fees levied by the court and statutory penalties usually receive payment priority over creditor claims, instantly shrinking funds remaining for unpaid debts.(C) Creditors and debt collectors must stop all calls and texts to any device the debtor designates as lacking free communication.

(D) Because each extra fee cuts creditors' recovery, they already aim to reduce collection call and message costs, so banning charge passing is unjustified.

Answer: (C)

Question

Sentence 1: The Fair Debt Collection Practices Act needs to be updated for modern times.

Sentence 2: Electronic communication is the preferred method of communication for consumers without a doubt.

Which of the following choices is the premise best supporting both Sentence 1 and Sentence 2?

(A) Congress should propose a law that expressly permits debt collectors to interact with consumers through email, text messages, and additional approved electronic communication methods nationwide.

(B) In general, the majority of consumers would unquestionably prefer receiving an email or text message instead of being contacted through a traditional phone call.

(C) Widespread consumer preference for electronic communication demonstrates that debt collectors already interact effectively within the current statutory framework, making revision of the Fair Debt Collection Practices Act unwarranted.

(D) Recognizing that the Fair Debt Collection Practices Act remains rooted in a pre-internet era demonstrates that communication habits are too varied to assert consumers prefer electronic messages.

Answer: (B)

Question

Sentence 1: Most consumers would certainly prefer to receive an email or text rather than a phone call.

Sentence 2: Electronic communication is the preferred method of communication for consumers without a doubt.

Which of the following choices is the claim best supported by Sentence 1?

(A) Ongoing staffing deficits and logistical bottlenecks have rendered traditional postal delivery progressively more unreliable and slow.

(B) The Fair Debt Collection Practices Act now requires updating to keep pace with contemporary societal realities.

(C) Digital communications furnish a written record consumers can readily store and consult should any misunderstandings emerge.

(D) Texts and emails let recipients read and respond when convenient, avoiding interruption of their ongoing activities.

Answer: (B)

Question

Sentence 1: Most consumers would certainly prefer to receive an email or text rather than a phone call.

Sentence 2: The Fair Debt Collection Practices Act needs to be updated for modern times.

Which of the following choices is the claim best supported by Sentence 1?

(A) Electronic written records, including emails and texts, create an easily searchable trail that consumers might require later as reference during disputes.

(B) Swift progress in artificial intelligence now lets debt collectors deploy automated chatbots and algorithmic dialers that lawmakers in 1977 could never have envisioned.

(C) Undeniably, consumers overwhelmingly favor electronic channels as their primary means of staying in touch and relaying information over alternative approaches.

(D) Emails and text messages let recipients examine information whenever convenient and spare them the sudden intrusion of a ringing phone.

Answer: (C)

Question

Sentence 1: If a problem arises with the representative, the customer can review the recording to reveal the truth.

1188 Sentence 2: Automated dialing systems include many built-in controls that protect consumers.
 1189 Which of the following choices is the premise best supporting Sentence 2?
 1190 (A) Having direct access to verifiable evidence weakens the notion consumers chiefly depend on automated dialer safeguards for protection.
 1191 (B) Automated dialers may use filters limiting calls to customers by location, time of day, or number of prior attempts.
 1192 (C) Regulatory authorities ought to allow compliant companies to keep using their automated dialing systems without interruption.
 1193 (D) Community college consumer rights courses should contain a module that explains the functioning of these built-in controls.
 1194 ## Answer: (B)

1194 ## Question
 1195 Sentence 1: Automated dialers can incorporate filters that restrict calls to customers by location, time of day, or the number of prior attempts.
 1196 Sentence 2: Automated dialing systems include many built-in controls that protect consumers.
 1197 Which of the following choices is the premise best supporting Sentence 2?
 1198 (A) A university ethics board approved a comprehensive study testing whether elderly patients feel comfortable receiving medication reminders
 1199 via automated dialers with built-in consumer protections.
 1200 (B) Because automated dialers' flexibility permits adding location, time-of-day, and attempt filters, aggressive collectors can likewise disable
 1201 them, leaving consumers largely unprotected.
 1202 (C) Regulatory authorities ought to allow businesses to depend on automated dialing technologies when delivering timely notifications and
 1203 important alerts to their customers.
 1204 (D) Should any issue ever arise regarding the representative's conduct, the customer is entitled to examine the recording in order to uncover the
 1205 actual facts.
 1206 ## Answer: (D)

1204 ## Question
 1205 Sentence 1: Automated dialers can incorporate filters that restrict calls to customers by location, time of day, or the number of prior attempts.
 1206 Sentence 2: If a problem arises with the representative, the customer can review the recording to reveal the truth.
 1207 Which of the following choices is the claim best supported by both Sentence 1 and Sentence 2?
 1208 (A) Federal consumer-protection rules impose hefty penalties on firms calling outside permitted hours or surpassing contact quotas.
 1209 (B) Automated dialing systems incorporate numerous internal safeguards that are specifically designed to protect consumers from harm.
 1210 (C) Recent consumer-protection regulations impose limits on call frequency and require that verification of conversations be recorded.
 1211 (D) Every customer service call is automatically recorded and stored in a secure database for ninety days.
 1212 ## Answer: (B)

1211 ## Question
 1212 Sentence 1: {first_context_sentence}
 1213 Sentence 2: {second_context_sentence}
 1214 {question}
 1215 (A) {choice_A}
 1216 (B) {choice_B}
 1217 (C) {choice_C}
 1218 (D) {choice_D}
 1219 ## Answer

1217
 1218
 1219
 1220
 1221
 1222
 1223
 1224
 1225
 1226
 1227
 1228
 1229
 1230
 1231
 1232
 1233
 1234
 1235
 1236
 1237
 1238
 1239
 1240
 1241