

000 001 002 003 004 005 ARGQA: EVALUATION OF REASONING OVER ELEMEN- 006 TARY LOGICAL STRUCTURES IN ARGUMENTS 007 008 009

010 **Anonymous authors**
011 Paper under double-blind review
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026

ABSTRACT

027 As large language models advance in their reasoning capabilities, their adequate
028 evaluation is becoming increasingly important. Existing logical reasoning benchmarks
029 are typically constructed by automatically converting symbolic logic into
030 natural language or curating questions from standardized exams, such as LSAT.
031 However, both synthetic and exam-style questions contain unnatural language,
032 thereby limiting their applicability to real-world contexts. Also, the systematic
033 assessment of reasoning over diverse logical structures remains underexplored.
034 Therefore, we present ARGQA, a novel dataset of 3,807 multiple-choice ques-
035 tions based on authentic arguments from four distinct domains—product reviews,
036 argumentative essays, e-rulemaking comments, and medical research abstracts.
037 Each question is designed to assess the ability to recognize and reconstruct one
038 of three elementary logical structures—linear, convergent, and divergent—whose
039 understanding is a prerequisite to both simple and complex reasoning. Experiments
040 show that even the strongest LLMs still have considerable room for improvement
041 with the overall 9-shot accuracy ranging from 29.2% (Qwen-2) to 61.8% (GPT-o3).
042
043

1 INTRODUCTION

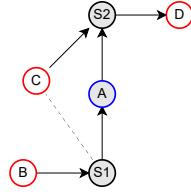
044 As large language models (LLMs) demonstrate increasingly sophisticated reasoning, adequately
045 assessing their logical reasoning abilities has become crucial for continuing the progress (OpenAI,
046 2025a;b; Yang et al., 2025). In line with the need, a range of benchmarks have been developed
047 in recent years, typically by automatically converting propositions in symbolic logic into natural
048 language (Saparov & He, 2023; Parmar et al., 2024) or by adopting exams designed for people like
049 LSAT and GMAT (Yu et al., 2020; Zhong et al., 2022). Note, these so-called *logical* reasoning
050 benchmarks focus purely on assessing logical reasoning capabilities. This is distinct from other
051 reasoning benchmarks, which require general or domain-specific knowledge and the ability to reason
052 based on such knowledge. For instance, popular benchmarks HellaSwag (Zellers et al., 2019) and
053 WinoGrande (Sakaguchi et al., 2021) test commonsense reasoning, and GSM8k (Cobbe et al., 2021)
054 and AIME (Patel et al., 2024) assess mathematical reasoning.

055 For the purpose of evaluating LLMs’ reasoning in real-world contexts, however, existing logical
056 reasoning benchmarks are insufficient. For one, they consist of text that bears little resemblance
057 to text in the wild, thereby limiting their relevance to real-world contexts. As one may expect,
058 the disparity is more pronounced for synthetically generated benchmarks, which are comprised of
059 sentences like “Sawyer is a poet.” and “Sawyer is either a musician or a poet, but not both.” (Qi et al.,
060 2025). However, even those made of exam questions can contain language rarely used in real-life:
061 “Seven directors—A, B, C, D, E, F, and G—serves on the X committee or the Y committee.” (Zhong
062 et al., 2022). ReClor (Yu et al., 2020) does have many questions based on realistic arguments, since it
063 adopts LSAT’s logical reasoning questions, which are about “arguments as they occur in ordinary
064 language” (LSAC, 2025). However, not all questions follow this style, and the overall linguistic
065 and topical diversity is limited. Also, the systematic assessment of reasoning over diverse logical
066 structures remains underexplored, as questions do not target specific logical structures.

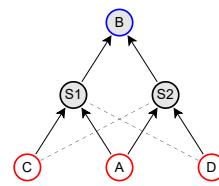
067 To address these issues, we present ArgumentationQA (ARGQA), a novel dataset of 3,807 multiple-
068 choice questions (MCQs) to assess the ability to recognize and reconstruct elementary logical
069 structures in realistic arguments. As shown in Figure 1, it was constructed based on authentic
070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

054	Sentence 1: If there are no animals in the world, the balance of nature will break down and human beings will die out as well.
055	Sentence 2: Conducting various animal experiments is hazardous to humanity's future and the next generation.
056	Which of the following choices is the claim best supported by Sentence 1, while also being the premise best supporting Sentence 2?
057	A: <i>Creatures in the animal kingdom are companionable by nature and remain fundamentally essential to the well-being of humans.</i> ✓
058	B: Most flowering plants rely on animal pollinators and seed dispersers to reproduce, maintaining food chains supporting humans. ✗
059	C: Drug-resistant pathogens can develop during laboratory animal experiments, potentially breaching containment and sparking lethal pandemics among humans. ✗
060	D: Policymakers ought to eliminate animal testing progressively and encourage the swift creation of ethical, animal-free research techniques. ✗
061	
062	
063	
064	
065	Sentence 1: In the subset of patients receiving concurrent chemotherapy, the endpoints again showed no significant differences.
066	Sentence 2: The secondary endpoints did not differ significantly between the HMB/Arg/Gln group and the control group.
067	Which of the following choices is the claim best supported by both Sentence 1 and Sentence 2?
068	A: Numerous participants not receiving chemotherapy significantly reduced their protein consumption throughout the study period, a change that probably obscured any benefits from supplements. ✗
069	B: <i>The trial was unable to sufficiently assess whether beta-hydroxy beta-methylbutyrate, glutamine, and arginine can reverse or prevent lean body mass wasting in cancer patients.</i> ✓
070	C: Strict dose-capping rules designed to limit chemotherapy toxicity produced nearly identical cumulative drug exposure among all participants in that subset, reducing variability in measured endpoints. ✗
071	D: The clinical trial proved statistically underpowered, given that it was able to enroll merely forty-two total patients within the concurrent chemotherapy treatment subset. ✗
072	
073	
074	
075	Sentence 1: I also support adherence to blue laws that prohibit consumer contact on Sundays.
076	Sentence 2: I conduct no business on Sundays.
077	Which of the following choices is the premise best supporting both Sentence 1 and Sentence 2?
078	A: My blue-law compliance review shows online retail platforms log thousands of Sunday sales, refuting the claim that no business occurs then. ✗
079	B: <i>Most individuals typically spend the limited time that remains before the upcoming workweek together with their friends and family on Sundays.</i> ✓
080	C: Because a voluntary Sunday closure caused no loss of customers or income, it demonstrates that compulsory blue laws forbidding Sunday consumer contact are needless. ✗
081	D: Every Sunday, all of my marketing emails are automatically set up to remain paused from midnight until the following midnight. ✗
082	
083	

Argumentative Essays
Question Type: 1.2



Medical Research Abstracts
Question Type: 2.3



e-Rulemaking Comments
Question Type: 3.1

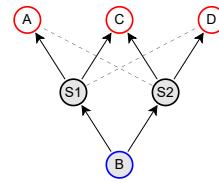


Figure 1: Examples from ARGQA. Each question assesses the ability to understand a particular elementary logical structure. **Blue** represents the correct option, and **red**, the wrong options. Gray nodes in the graphs represent the propositions comprising a real argument from the respective domain, 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

108 on question types sharing the same logical structure, though the propositions comprising them are
 109 drastically different. For question types with the same logical structure but with edge direction flipped,
 110 models are substantially better at identifying a premise supporting the context, rather than a claim
 111 supported by the context. Lastly, medical research abstract is the easiest domain with more clear
 112 logical relations, as opposed to argumentative essays with more obscure connections.
 113

114 2 RELATED WORKS

115
 116 Table 1: Comparison of ARGQA with existing logical reasoning datasets. [†]: Product reviews,
 117 argumentative essays, medical research abstracts, and e-Rulemaking comments. [‡]: The context and
 118 correct option are a real argument from the source domain paraphrased to be self-contained. Each
 119 incorrect option was generated to be logically linked to the context, but to form a logical structure
 120 different from the one specified in the question text. ^{*}: Entailment/Contradiction/Neutral/Paradox.
 121

122 Dataset	123 Size	124 Source Text	125 Construction Method	126 Text Structure	127 Task
128 CLUTRR	129 6k	130 N/A	131 Machine-Generated	132 Synthetic Story	133 T/F
134 RuleTaker	135 500k	136 N/A	137 Machine-Generated	138 Set of Propositions	139 T/F
140 ProofWriter	141 500k	142 N/A	143 Machine-Generated	144 Set of Propositions	145 T/F with Proof
146 LogicNLI	147 20k	148 N/A	149 Machine-Generated	150 Set of Propositions	151 E/C/N/P*
152 SimpleLogic	153 560k	154 N/A	155 Machine-Generated	156 Set of Propositions	157 T/F
158 PrOntoQA	159 7.9k	160 N/A	161 Machine-Generated	162 Set of Propositions	163 T/F with Proof
164 LogicBench	165 2k	166 N/A	167 Machine-Generated	168 Synthetic Story	169 Y/N and MC
170 ProverQA	171 1.5k	172 N/A	173 Machine-Generated	174 Set of Propositions	175 T/F with Proof
176 FOLIO	177 1.4k	178 N/A	179 Human-Written	180 Set of Propositions	181 T/F/Unknown
182 LogiQA	183 8.6k	184 NCSE	185 Curated (Human-Written)	186 Mixed	187 MC
188 ReClor	189 6.1k	190 GMAT, LSAT	191 Curated (Human-Written)	192 Mixed	193 MC
194 AR-LSAT	195 2k	196 LSAT	197 Curated (Human-Written)	198 Mixed	199 MC
ARGQA (ours)		3.8k	Various[†]	Paraphrased or Generated[‡]	Argument
					MC

138 **Logical Reasoning Datasets.** Given the growing need for LLMs to perform sophisticated logical
 139 reasoning, numerous datasets focusing specifically on assessing their logical reasoning abilities have
 140 been developed in recent years, as shown in Table 1. One group of logical reasoning datasets adopt
 141 questions from exams designed for people: LogiQA (Liu et al., 2021) and LogiQA 2.0 (Liu et al.,
 142 2023) consist of questions from National Civil Servants Exam (NCSE) of China, whereas ReClor (Yu
 143 et al., 2020) and AR-LSAT (Zhong et al., 2022) incorporate US-based standardized tests: LSAT
 144 and GMAT for the former, and LSAT for the latter. These datasets cover various types of questions,
 145 including drawing a conclusion from a set of premises and inferring an event from a synthetic scenario
 146 like “Seven directors—A, B, C, D, E, F, and G—serves on the X committee or the Y committee.”

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

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

162 **Argument Mining Datasets.** As will be discussed in Section 3.1, an argument in natural language
 163 is a set of premises supporting a claim, where the premises and the claim are propositions that are
 164 either true or false. Argument mining is a task of identifying and extracting arguments in text, which
 165 can assist deeper comprehension and critical evaluation, as well as the generation of relevant text,
 166 such as a counter-argument. To support research in this direction, many argument mining datasets
 167 have been developed over the years. Because the structure and style of writing vary from one domain
 168 to another, a dataset typically targets a single domain, such as news articles (e.g. Eckle-Kohler et al.
 169 2015; Al-Khatib et al. 2016; Ein-Dor et al. 2020), legal documents (e.g. Poudyal et al. 2020; Grundler
 170 et al. 2022), political debates (e.g. Haddadan et al. 2019; Visser et al. 2019; Hautli-Janisz et al.
 171 2022), student essays (e.g. Stab & Gurevych 2017; Alhindi & Ghosh 2021; Schaller et al. 2024), and
 172 user-generated content online (e.g. Boltužić & Snajder 2014; Habernal & Gurevych 2017; Bhatti et al.
 173 2021). In addition to the differences in domain, the authors may be driven by distinct goals, leading to
 174 different annotated components (e.g. different types of premises) and relations (e.g. different types of
 175 support relations). Also, the task itself can be posed as a structured prediction problem (e.g. input: an
 176 entire document, output: a directed graph representing the logical structure) or a binary classification
 177 problem (e.g. input: an ordered pair of propositions, output: a binary verdict on whether the first
 178 proposition supports the second), among others. Due to the limited standardization, it can be difficult
 179 to use these datasets for assessing the logical reasoning abilities of models. Also, analyzing models’
 180 mistakes to gain additional insights is cumbersome. To the best of our knowledge, ARGQA is the
 181 first dataset in a standardized format that has been designed to evaluate the capacity to recognize
 182 elementary logical structures in real arguments and analyze mistakes in a convenient manner.

3 THE ARGUMENTATIONQA (ARGQA) DATASET

3.1 THEORETICAL BACKGROUND

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

- 201 **Linear Structure.** A proposition supports another proposition, which in turn, supports yet
 202 another proposition.¹
- 203 **Convergent Structure.** Multiple propositions independently support a proposition.
- 204 **Divergent Structure.** A proposition independently supports multiple propositions.

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

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

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

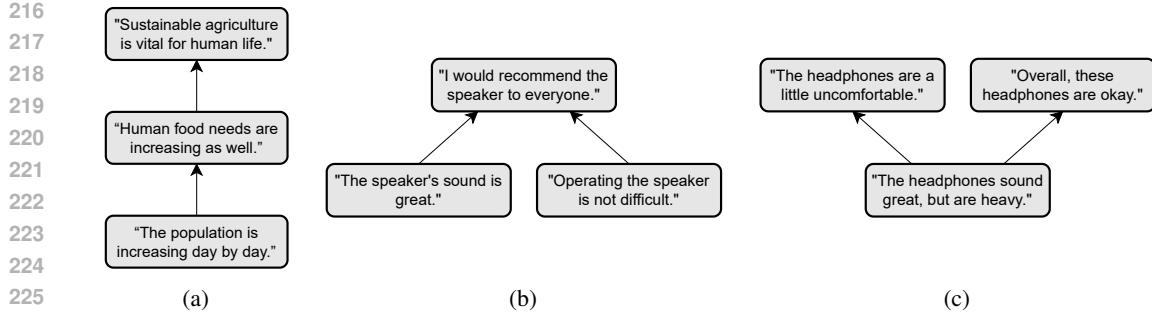


Figure 2: Examples of three elementary logical structures found in practical argumentation: (a) linear, (b) convergent, and (c) divergent structure. Directed edges represent support relations.

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:

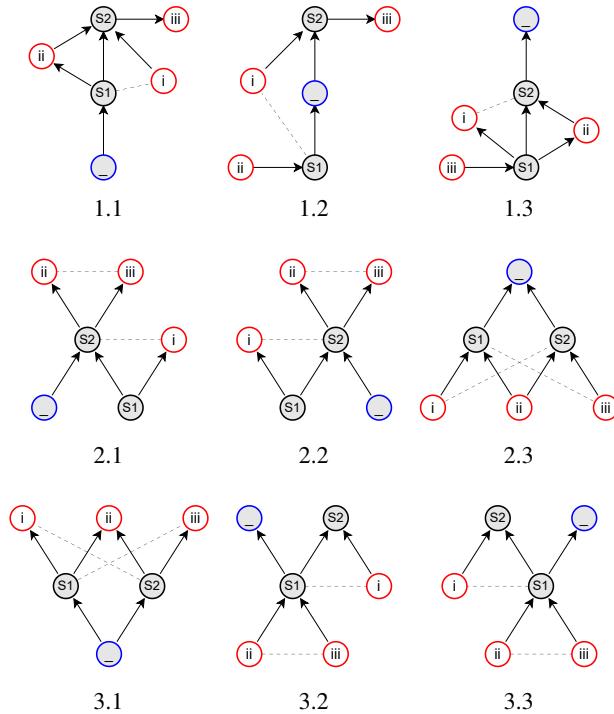


Figure 3: Question types and their logical structures. Directed edges represent support relations, and dashed edges represent the explicit lack of support relations in either direction. Grey nodes constitute the target elementary logical structure formed by the context $S1$, $S2$, and the correct option $_$ 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.

270 **Step 1: Source Selection.** We select argument mining datasets to acquire raw text from diverse
 271 domains, along with their corresponding logical relation annotations. We leverage publicly available
 272 argument mining datasets, which annotate documents with logical structures. However, we exclude
 273 those with synthetic text (e.g. Peldszus & Stede 2015), with an annotation scheme disallowing one
 274 or more of the three elementary logical structures (e.g. Poudyal et al. 2020), and without enough
 275 context to make propositions interpretable (e.g. Hautli-Janisz et al. 2022). From over 20 datasets
 276 we considered, only the following meet the criteria: (1) AAEC2 (Stab & Gurevych, 2014) consists
 277 of argumentative essays by students on contentious issues, often using implicit reasoning allowing
 278 for multiple interpretations; (2) CDCP (Park & Cardie, 2018) is composed of user comments on
 279 e-Rulemaking, presenting opinions and related experiences regarding consumer debt collection
 280 practices; (3) AbstRCT (Mayer et al., 2020) features abstracts of Randomized Controlled Trials
 281 (RCT) from the MEDLINE database, often displaying clear and concise logic; and (4) AM² (Chen
 282 et al., 2022) is a collection of product reviews from Amazon, characterized by grammatical flexibility.
 283

284 **Step 2: Triplet Extraction.** From each dataset, we extract triplets of propositions that form
 285 elementary logical structures of our interest: *linear*, *convergent*, and *divergent*. For this, we first
 286 construct directed graphs of logical relations from the annotations, then extract triplets of propositions
 287 tagged with the elementary logical structure they form. The triplets are then filtered to prevent
 288 semantic overlaps while balancing the number of triplets across the logical structures: All triplets in
 289 the divergent structure—the most rare one—are selected, then the same number of triplets for each of
 290 linear and convergent structure are selected, with each proposition appearing in one triplet only.
 291

292 **Step 3: Proposition Paraphrasing.** We paraphrase each proposition to make them self-contained.
 293 Argument mining datasets typically keep the original text, thus some propositions cannot be inter-
 294 preted without the context. However, because our questions contain triplets without the context, they
 295 need to be self-contained. Thus, we use GPT-o3 to paraphrase each proposition to a self-contained
 296 and grammatically complete sentence, through co-reference resolution and sentence completion using
 297 the original context. For example, “*JUNK!*” is paraphrased as “*These headphones are junk!*” based
 298 on the context of the product review in which it appears. (See Table 7 for the prompt.)
 299

300 **Step 4: Triplet Deduplication.** We filter triplets comprised of semantically similar propositions. We
 301 do so by measuring the cosine similarity between the SBERT embeddings (Reimers & Gurevych,
 302 2019) of all pairs of propositions within each triplet. When the similarity is above a threshold, the
 303 triplet is discarded. This is to exclude triplets where semantically identical or similar propositions
 304 have been wrongly annotated as in support relations.
 305

306 **Step 5: Distractor Generation.** For each triplet, we construct three *context-options* pairs by splitting
 307 the triplet into context—two propositions—and one correct option in three ways. Each pair is tagged
 308 with the corresponding question type. Then, for each pair, we generate three logical distractors with
 309 GPT-o3 by feeding in the context and a description of the logical relation between the context and the
 310 proposition to be generated, specified by the question type. These distractors constitute the incorrect
 311 options and are added to the respective *context-options* pairs. (See Table 8 for the prompt.)
 312

313 **Step 6: Option Rephrasing.** For each *context-options* pair, we rephrase both correct and incorrect
 314 options to similar lengths. This is to prevent potential differences in lengths revealing the correct
 315 option. For this, we compute the median word count across the four options and paraphrase each
 316 option to have its length fall within two words of the median. (See Table 9 for the prompt.)
 317

318 **Step 7: Question Construction.** Finally, we construct a question for each *context-options* pair by
 319 combining it with a pre-written question describing the logical relationship between the context and
 320 the correct option, such as “Which of the following choices is the premise best supporting Sentence
 321 1?” for question type 1.1. (See Table 4 for the full list of question text.) Each question is composed
 322 of two context propositions, one question text, and four options—of which one is correct—in a
 323 randomized order. Note, each question is tagged with domain, as well as the question type.
 324

325 3.3 RESULTING DATASET

326 An summarized in Table 2, ARGQA consists of 3,807 MCQs, evenly split across the nine question
 327 types, and in turn the three elementary logical structures. For each domain, nine questions are
 328 reserved as the development set used as examples during few-shot experiments. The remainder are
 329 randomly split into 10% validation and 90% test set. Domains are distributed as follows: 1,620
 330

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

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

337 Table 2: Overview of ARGQA. Source arguments were collected from argument mining datasets
 338 across four domains for a wide range of topics and writing styles. Each domain-split contains the
 339 same number of questions for each of the nine question types. For instance, the validation set for
 340 e-Rulemaking comments comprises six questions per question type, totaling 54. *: number of words.
 341 ♣: 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

356 We use ARGQA to assess the logical reasoning abilities of popular open-source LLMs—Mistral-
 357 7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2-7B-
 358 Instruct (Yang et al., 2024), and Gemma-7B-Instruct (Team et al., 2024)—as well as strong proprietary
 359 models—GPT-o4-mini and GPT-o3 (OpenAI, 2025a). We focus on 7B to 8B parameter models for a
 360 fair comparison. However, GPT-o3 is also included as a representative of the LLMs with the strongest
 361 reasoning capabilities; among the available GPT models we pilot-tested, GPT-o3 performed the best.
 362 Experiments were conducted on the held-out test set in zero-shot and nine-shot settings, using the
 363 development set as examples of the nine question types. The performance was measured in accuracy,
 364 i.e., the percentage of correct responses. We adapted a multiple-choice prompting protocol from
 365 MMLU-Pro (Wang et al., 2024), instructing models to return answers in the format “the answer is
 366 (X)”, and automatically extracting the answer using regular expressions. The exact wording (See
 367 Figure 10 for the prompt), as well as maximum output length (32 tokens) and repetition penalty (1.0)
 368 were finalized based on the validation performance. Lastly, open-source models were evaluated with
 369 greedy decoding, with temperature set to 0, for maximally deterministic responses.

4.2 RESULTS & ANALYSIS

371 As presented in Table 3², the overall performance of the LLMs measured in accuracy on the ARGQA
 372 test set for all domains ranges from 29.24% (Qwen-2) to 61.81% (GPT-o3), with GPT-o3 considerably
 373 outperforming the rest of the models. This is in line with expectations, since GPT-o3 is a substantially
 374 larger model specializing in logical reasoning. Among the other models—which all have 7B to
 375 8B parameters—Gemma is generally the best performer, followed by GPT-o4-mini. Qwen-2 and
 376 Mistral rank the lowest with only a modest 4% improvement from the random baseline of 25%,
 377

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

378 Table 3: Nine-shot performance of LLMs on the ARGQA test set for all domains. Each cell reports
 379 accuracy (%), the percentage of questions that were correctly answered. Open-source models are
 380 evaluated with greedy decoding, temperature set to 0, for maximally deterministic responses. LLMs
 381 show considerable room for improvement, with GPT-o3 substantially outperforming the rest.
 382

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 (Q-type Avg.)	29.30	34.81	29.24	44.62	43.27	61.81	40.51

397
 398
 399 which corresponds to randomly selecting the correct answer from four options. Overall, these results
 400 indicate that LLMs currently have insufficient abilities to recognize elementary logical structures,
 401 even those comprised of only three propositions. In other words, understanding logical structures, no
 402 matter how simple, remains a challenging problem to modern LLMs.

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

415 Also, consider Q-types 2.3 and 3.1. The logical structures for these Q-types are mirror-images of
 416 each other, i.e., they are identical, except the edges point in the opposite direction. Unlike the cases
 417 where the logical structures are the same, the performance varies drastically between the pair, with
 418 the models averaging 32.58% for the former, and 63.97% for the latter. A similar pattern can be
 419 observed between another pair with the same logical structure but flipped edges: Q-types 1.1 and 1.3.
 420 One plausible explanation is that the models are better at identifying a premise supporting a given
 421 claim(s), as opposed to recognizing a claim supported by a given premise(s). This observation is also
 422 consistent with the fact that the Q-types with the highest performance reported—Q-types 3.1, 2.1,
 423 2.2, and 1.1—all require identifying a premise supporting claim(s) provided as context.

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

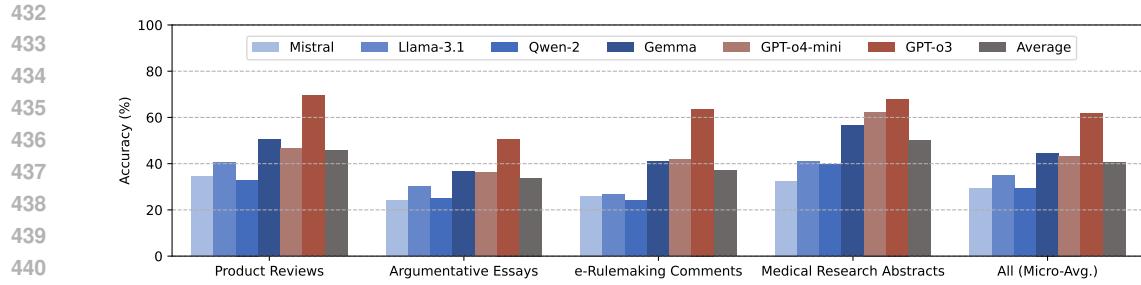


Figure 4: Nine-shot performance of LLMs on the ARGQA test set for each domain. On average, as shown in gray bars, LLMs perform best on medical research abstracts, followed by product reviews, e-Rulemaking comments and argumentative essays.

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 an 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.

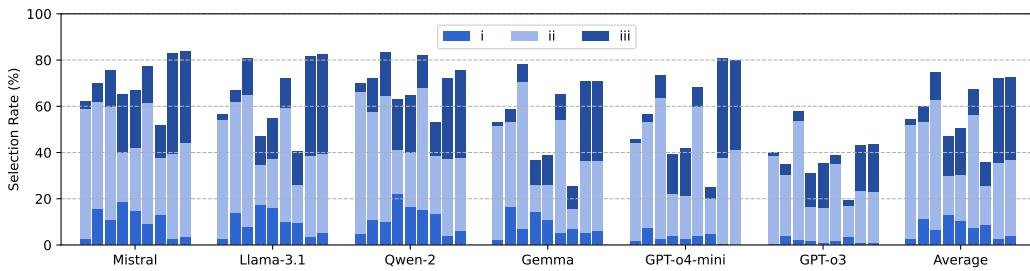


Figure 5: Selection rates for the distractors. Each group of nine stacked bars represent the percentages of selecting the three distractors for the nine question types. The stacks do not add up to 100%, because the percentage of selecting the correct option is not shown. Note, generalized claims about 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.

486 **6 ETHICS STATEMENT**
487488 ARGQA is constructed by leveraging publicly available argument-mining datasets (AM², CDCP,
489 AAEC2 and AbstRCT) to paraphrase and generate new propositions. Our project did not involve any
490 human participants, and no private or personally identifying information was collected or published.
491492 **7 REPRODUCIBILITY STATEMENT**
493494 Included with our submission are all the code for dataset construction and experiments, along with
495 the data. This is to ensure the transparency in data construction, verification of the experimental
496 process, and reproducibility of the results. Upon acceptance, we plan to formally release the code
497 and the dataset via an open source repository.
498499 **REFERENCES**
500

501 Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news
502 editorial corpus for mining argumentation strategies. In Yuji Matsumoto and Rashmi Prasad (eds.),
503 *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/>.

506 Tariq Alhindi and Debanjan Ghosh. “Sharks are not the threat humans are”: Argument component
507 segmentation in school student essays. In Jill Burstein, Andrea Horbach, Ekaterina Kochmar,
508 Ronja Laarmann-Quante, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis,
509 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
510 Linguistics. URL <https://aclanthology.org/2021.bea-1.22/>.

512 Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. Argument mining on
513 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
514 Computational Linguistics. URL <https://aclanthology.org/2021.argmining-1.1>.

517 J. Anthony Blair and Ralph H. Johnson. The current state of informal logic. *Informal Logic*, 9(2),
518 1987. doi: 10.22329/il.v9i2.2671.

524 Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In
525 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
526 for Computational Linguistics. doi: 10.3115/v1/W14-2107. URL <https://aclanthology.org/W14-2107>.

532 Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. Argument
533 mining for review helpfulness prediction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang
534 (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*,
535 pp. 8914–8922, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational
536 Linguistics. doi: 10.18653/v1/2022.emnlp-main.609. URL <https://aclanthology.org/2022.emnlp-main.609>.

538 Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language.
539 In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI’20*, 2021. ISBN 9780999241165.

546 Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias
547 Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman.
548 Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL <https://api.semanticscholar.org/CorpusID:239998651>.

556 Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. On the role of discourse markers for discriminating
557 claims and premises in argumentative discourse. In Lluís Márquez, Chris Callison-Burch,
558

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

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

550 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
 551 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,
 552 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev,
 553 Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru,
 554 Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak,
 555 Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu,
 556 Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle
 557 Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego
 558 Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,
 559 Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel
 560 Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon,
 561 Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan
 562 Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet,
 563 Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde,
 564 Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie
 565 Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua
 566 Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak,
 567 Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley
 568 Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence
 569 Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas
 570 Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri,
 571 Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie
 572 Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes
 573 Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne,
 574 Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal
 575 Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,
 576 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,
 577 Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie
 578 Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana
 579 Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie,
 580 Sharan Narang, Sharath Raparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon
 581 Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan,
 582 Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas
 583 Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami,
 584 Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti,
 585 Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenjin Fu, Whitney Meers, Xavier
 586 Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao
 587 Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song,
 588 Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe
 589 Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya
 590 Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenber, Alexei
 591 Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu,
 592 Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit
 593 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,

594 Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu
 595 Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer,
 596 Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu,
 597 Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc
 598 Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily
 599 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers,
 600 Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank
 601 Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee,
 602 Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan,
 603 Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph,
 604 Helen Suk, Henry Aspereen, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog,
 605 Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James
 606 Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny
 607 Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings,
 608 Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai
 609 Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik
 610 Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle
 611 Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng
 612 Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish
 613 Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim
 614 Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle
 615 Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,
 616 Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam,
 617 Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier,
 618 Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia
 619 Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro
 620 Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani,
 621 Pritish Yuvaraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy,
 622 Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin
 623 Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu,
 624 Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh
 625 Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay,
 626 Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang,
 627 Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie
 628 Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta,
 629 Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman,
 630 Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun
 631 Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria
 632 Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,
 633 Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz,
 634 Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv
 635 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
 636 llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

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

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

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

648 Shohreh Haddadan, Elena Cabrio, and Serena Villata. Yes, we can! mining arguments in 50 years
 649 of US presidential campaign debates. In Anna Korhonen, David Traum, and Lluís Màrquez
 650 (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,
 651 pp. 4684–4690, Florence, Italy, July 2019. Association for Computational Linguistics. doi:
 652 10.18653/v1/P19-1463. URL <https://aclanthology.org/P19-1463/>.

653 Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James
 654 Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah
 655 Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor,
 656 Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej
 657 Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex
 658 Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order
 659 logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024*
 660 *Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami,
 661 Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/
 662 2024.emnlp-main.1229. URL <https://aclanthology.org/2024.emnlp-main.1229/>.

663 Annette Hautli-Janisz, Zlata Kikteva, Wassiliki Siskou, Kamila Gorska, Ray Becker, and Chris
 664 Reed. QT30: A corpus of argument and conflict in broadcast debate. In Nicoletta Calzolari,
 665 Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara
 666 Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios
 667 Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*,
 668 pp. 3291–3300, Marseille, France, June 2022. European Language Resources Association. URL
 669 <https://aclanthology.org/2022.lrec-1.352>.

670 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
 671 Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.

672 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
 673 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,
 674 Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas
 675 Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.

676 John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):
 677 765–818, December 2019. doi: 10.1162/coli_a_00364. URL <https://aclanthology.org/J19-4006/>.

678 Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa
 679 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM*
 680 *Trans. Audio, Speech and Lang. Proc.*, 31:2947–2962, July 2023. ISSN 2329-9290. doi: 10.1109/
 681 TASLP.2023.3293046. URL <https://doi.org/10.1109/TASLP.2023.3293046>.

682 Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: a
 683 challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of*
 684 *the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI’20, 2021. ISBN
 685 9780999241165.

686 Law School Admission Council LSAC, 2025. URL <https://www.lsac.org/lsat/taking-lsat/test-format/logical-reasoning>. Accessed: 2025-09-10.

687 Tobias Mayer, Elena Cabrio, and Serena Villata. Transformer-based Argument Mining for Healthcare
 688 Applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, Santiago de
 689 Compostela / Online, Spain, August 2020. URL <https://hal.science/hal-02879293>.

690 OpenAI. OpenAI o3 and o4-mini system card. System card, OpenAI, 2025a.
 691 URL <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.

692 OpenAI. GPT-5 system card. System card, OpenAI, 2025b. URL <https://cdn.openai.com/gpt-5-system-card.pdf>.

702 Joonsuk Park and Claire Cardie. A corpus of eRulemaking user comments for measuring evaluability
 703 of arguments. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck,
 704 Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo,
 705 Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of
 706 the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*,
 707 Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://aclanthology.org/L18-1257/>.

709 Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty,
 710 Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical
 711 reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek
 712 Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Compu-
 713 tational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, Bangkok, Thailand, August
 714 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.739. URL
 715 <https://aclanthology.org/2024.acl-long.739/>.

716 Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh
 717 Manocha. Aime: Ai system optimization via multiple llm evaluators, 2024. URL <https://arxiv.org/abs/2410.03131>.

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

721 Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo
 722 Quaresma. ECHR: Legal corpus for argument mining. In Elena Cabrio and Serena Villata (eds.),
 723 *Proceedings of the 7th Workshop on Argument Mining*, pp. 67–75, Online, December 2020. Associa-
 724 tion for Computational Linguistics. URL <https://aclanthology.org/2020.argmining-1.8>.

725 Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He.
 726 Large language models meet symbolic provers for logical reasoning evaluation. In *The Thirteenth
 727 International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=C25SgeXWjE>.

728 Iyad Rahwan. Mass argumentation and the semantic web. *Web Semant.*, 6(1):29–37, February
 729 2008. ISSN 1570-8268. doi: 10.1016/j.websem.2007.11.007. URL <https://doi.org/10.1016/j.websem.2007.11.007>.

730 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks,
 731 2019. URL <https://arxiv.org/abs/1908.10084>.

732 Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adver-
 733 sarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN
 734 0001-0782. doi: 10.1145/3474381. URL <https://doi.org/10.1145/3474381>.

735 Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis
 736 of chain-of-thought. In *International Conference on Learning Representations*, 2023.

737 Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi,
 738 Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language
 739 models using ood examples. In *Proceedings of the 37th International Conference on Neural
 740 Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

741 Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer
 742 Meyer, and Thorben Jansen. DARIUS: A comprehensive learner corpus for argument mining in
 743 German-language essays. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro
 744 Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International
 745 Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING
 746 2024)*, pp. 4356–4367, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.389/>.

756 Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR:
 757 A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent
 758 Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in*
 759 *Natural Language Processing and the 9th International Joint Conference on Natural Language*
 760 *Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association
 761 for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL <https://aclanthology.org/D19-1458/>.

763 Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive
 764 essays. In Junichi Tsujii and Jan Hajic (eds.), *Proceedings of COLING 2014, the 25th International*
 765 *Conference on Computational Linguistics: Technical Papers*, pp. 1501–1510, Dublin, Ireland,
 766 August 2014. Dublin City University and Association for Computational Linguistics. URL
 767 <https://aclanthology.org/C14-1142/>.

768 Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Com-
 769 putational Linguistics*, 43(3):619–659, September 2017. doi: 10.1162/COLI_a_00295. URL
 770 <https://aclanthology.org/J17-3005>.

772 Manfred Stede and Jodi Schneider. *Argumentation Mining*. Number 40 in Synthesis Lectures on
 773 Human Language Technologies. Morgan & Claypool, 2018.

774 Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and
 775 abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto
 776 Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp.
 777 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/
 778 2021.findings-acl.317. URL <https://aclanthology.org/2021.findings-acl.317/>.

780 Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya
 781 Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan
 782 Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar,
 783 Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin,
 784 Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur,
 785 Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison,
 786 Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia
 787 Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris
 788 Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger,
 789 Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric
 790 Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary
 791 Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra,
 792 Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha
 793 Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost
 794 van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed,
 795 Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia,
 796 Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago,
 797 Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel
 798 Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow,
 799 Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan,
 800 Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad
 801 Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda,
 802 Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep
 803 Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh
 804 Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien
 805 M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan
 806 Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kociský, Tulsee Doshi,
 807 Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei,
 808 Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei,
 809 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,

810 and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL
 811 <https://arxiv.org/abs/2408.00118>.

812

813 Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the
 814 first-order logical reasoning ability through LogicNLI. In Marie-Francine Moens, Xuanjing Huang,
 815 Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical
 816 Methods in Natural Language Processing*, pp. 3738–3747, Online and Punta Cana, Dominican
 817 Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.
 818 emnlp-main.303. URL <https://aclanthology.org/2021.emnlp-main.303/>.

819

820 Jacky Visser, John Lawrence, Jean Wagemans, and Chris Reed. An annotated corpus of argument
 821 schemes in us election debates. In *Proceedings of the 9th Conference of the International Society
 822 for the Study of Argumentation (ISSA)*, 3-6 July 2018, pp. 1101–1111, 2019.

823

824 Douglas Walton. Argumentation theory: A very short introduction. In Iyad Rahwan and Guillermo R
 825 Simari (eds.), *Argumentation in Artificial Intelligence*, pp. 1–22. Springer, 2009. doi: 10.1007/
 978-0-387-98197-0_1.

826

827 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming
 828 Ren, Aaran Arulraj, Xuan He, Ziyian Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi
 829 Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language
 830 understanding benchmark, 2024. URL <https://arxiv.org/abs/2406.01574>.

831

832 An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li,
 833 Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong
 834 Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu,
 835 Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin
 836 Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao,
 837 Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin
 838 Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng
 839 Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu,
 840 Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL
<https://arxiv.org/abs/2407.10671>.

841

842 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
 843 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,
 844 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin
 845 Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,
 846 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui
 847 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang
 848 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger
 849 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan
 850 Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

851

852 Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset
 853 requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL
<https://openreview.net/forum?id=HJgJtT4tvB>.

854

855 Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a
 856 machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez
 857 (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,
 858 pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi:
 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472/>.

859

860 Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the
 861 paradox of learning to reason from data. In Edith Elkind (ed.), *Proceedings of the Thirty-Second
 862 International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 3365–3373. International
 863 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.

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

880 A APPENDIX

885 A.1 ADDITIONAL DETAILS

886
 887
 888 Table 4: Question text for each question type.
 889

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?

900
 901
 902 Table 5: Content of each ARGQA instance. The order of the four answer choice options is pre-
 903 randomized in the released file, ensuring that solvers cannot benefit from a positional bias.
 904

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).

918
 919 Table 6: Zero-shot performance of LLMs on the ARGQA test set for all domains. Each cell reports
 920 accuracy (%), the percentage of questions that were correctly answered. Greedy decoding, with
 921 temperature set to 0, was used to minimize variance. All LLMs tested show considerable room for
 922 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

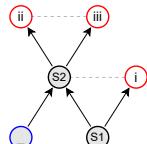
A.2 EXAMPLE QUESTIONS

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

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

957 A: After fewer than three minutes, the left speaker completely ceased to operate. ✓
 958 B: I sent the QC-25 headset back to the store and obtained a complete refund. ✗
 959 C: A new cable instantly restored perfect sound, proving the QC-25 itself was blameless. ✗
 960 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



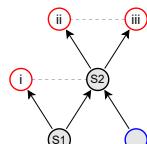
962 Sentence 1: The left speaker stopped working after less than three minutes.

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

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

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

Product Reviews
 Question Type: 2.2



971 Figure 6: Examples with Equal Structure but Different Propositions

972 A.3 EXAMPLE PROMPTS
973974 Table 7: Prompt for Proposition Paraphrasing for AM²
975

976 **Developer Prompt:**
977 # Identity
978 You are an expert editor that rewrites the sentence to become grammatically complete if necessary.
979
980 # Instructions
981 Your task is to rewrite the original sentence, such that:
982 1. The generated sentence becomes a grammatically complete sentence that can stand on its own.
983 2. The generated sentence preserves the original meaning without adding new details or elaboration.
984 3. All pronouns (e.g., "he", "she", "it", "they") are replaced with appropriate proper nouns or clear references from the given Context.
985 4. Remove connectors (e.g., "because", "but", "so", "in order that") if necessary.
986 5. The original point of view (e.g., first-person) must be preserved.
987 6. Only output a single rewritten sentence. Do not include explanations, formatting, or additional commentary.
988
989 # Examples
990 <product_review id="example-1">
991 Context: The speakers aren't even oriented around your ears, they're cockeyed.
992 Original Sentence: they're cockeyed.
993 </product_review>
994 <assistant_response id="example-1">
995 The speakers are cockeyed.
996 </assistant_response>
997
998 <product_review id="example-2">
999 Context: I hate this headset. Connection is terrible.
999 Original Sentence: I hate this headset.
1000 </product_review>
1001 <assistant_response id="example-2">
1002 I hate this headset.
1003 </assistant_response>
1004
1005 <product_review id="example-3">
1006 Context: Spend a bit more money and get something better. I will, I have to now.
1007 Original Sentence: I have to now.
1008 </product_review>
1009 <assistant_response id="example-3">
1010 I have to purchase better quality headphones now.
1011 </assistant_response>
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

1008 **User Prompt:**
1009 Context: {full_context}
1010 Original Sentence: {original_sentence}
1011 Completed Sentence:

1026
1027
1028
1029
1030
1031
1032
1033
1034

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}

1073
1074
1075
1076
1077
1078
1079

1080

1081

1082

Table 9: Prompt for Option Rephrasing for AAEC2

1083

Developer Prompt:

1084

Identity

1085

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

1086

Instructions

1087

Your task is to rewrite the Original Sentence such that:

1088

* The rewritten sentence conveys exactly the same meaning. No information must be added, removed, or altered.

1089

* 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.

1090

* The original point of view and tense must remain unchanged. You must NOT add any new unresolved references like pronouns.

1091

* Your response must be exactly one grammatically complete and independent sentence. Do not split it into multiple sentences.

1092

* Output only the rewritten sentence without any explanations, formatting, or additional commentary.

1093

Examples

1094

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

1095

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

1096

</sentence>

1097

<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.

1098

</assistant_response>

1099

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

1100

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.

1101

Number of words targeted: 15

1102

</sentence>

1103

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

1104

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

1105

</assistant_response>

1106

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

1107

Sentence: I now feel confident living independently.

1108

Number of words targeted: 20

1109

</sentence>

1110

<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.

1111

</assistant_response>

1112

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

1113

Sentence: I now feel confident living independently.

1114

Number of words targeted: 20

1115

</sentence>

1116

<assistant_response id="bad-example-3">
I now feel confident in my ability to manage everything financially and live independently in my dream city.

1117

</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.

1118

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

1119

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.

1120

Number of words targeted: 20

1121

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

1122

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

1123

</assistant_response>

1124

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.

1125

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

1126

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

1127

Number of words targeted: 25

1128

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

1129

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

1130

</assistant_response>

1131

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.

1132

User Prompt:

1133

Sentence: {original_sentence}

Number of words targeted: {target_length}

1134

1135

1136

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

1137

User Prompt:

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

1139

Question

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

1140

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?

1142

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

1143

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

1144

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

1145

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

Answer: (A)

1146

Question

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

1148

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?

1149

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

1150

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

1151

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

1152

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

Answer: (A)

1153

Question

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

1154

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

1155

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.

1156

(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)

1157

Question

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

1158

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

1159

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

1160

(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.

1161

(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.

1162

(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.

1163

(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)

1164

Question

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

1165

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

1166

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

1167

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

1168

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

1169

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

1170

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

Answer: (B)

1171

Question

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

1172

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

1173

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

1174

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

1175

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

1176

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

1177

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

Answer: (C)

1178

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.
1193 ## 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.
1196 Which of the following choices is the premise best supporting Sentence 2?
1197 (A) A university ethics board approved a comprehensive study testing whether elderly patients feel comfortable receiving medication reminders via automated dialers with built-in consumer protections.
1198 (B) Because automated dialers' flexibility permits adding location, time-of-day, and attempt filters, aggressive collectors can likewise disable them, leaving consumers largely unprotected.
1199 (C) Regulatory authorities ought to allow businesses to depend on automated dialing technologies when delivering timely notifications and important alerts to their customers.
1200 (D) Should any issue ever arise regarding the representative's conduct, the customer is entitled to examine the recording in order to uncover the actual facts.
1201 ## Answer: (D)

1202 ## Question
1203 Sentence 1: Automated dialers can incorporate filters that restrict calls to customers by location, time of day, or the number of prior attempts.
1204 Sentence 2: If a problem arises with the representative, the customer can review the recording to reveal the truth.
1205 Which of the following choices is the claim best supported by both Sentence 1 and Sentence 2?
1206 (A) Federal consumer-protection rules impose hefty penalties on firms calling outside permitted hours or surpassing contact quotas.
1207 (B) Automated dialing systems incorporate numerous internal safeguards that are specifically designed to protect consumers from harm.
1208 (C) Recent consumer-protection regulations impose limits on call frequency and require that verification of conversations be recorded.
1209 (D) Every customer service call is automatically recorded and stored in a secure database for ninety days.
1209 ## Answer: (B)

1210 ## Question
1211 Sentence 1: {first_context_sentence}
1212 Sentence 2: {second_context_sentence}
1213 {question}
1214 (A) {choice_A}
1215 (B) {choice_B}
1215 (C) {choice_C}
1216 (D) {choice_D}
1216 ## 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