

ArgGenBench: Benchmarking the Complex Controlled Argument Generation Capability of Large Language Models

Bojun Jin^{1,2*}, Jianzhu Bao^{3*}, Yang Sun¹, Yice Zhang¹, Ruifeng Xu^{1,2†}

¹Harbin Institute of Technology, Shenzhen, China

²Shenzhen Loop Area Institute, Shenzhen, China

³Nanyang Technological University, Singapore

bojunjinmail@gmail.com jianzhubao@gmail.com xuruifeng@hit.edu.cn

Abstract

Argument generation is a fundamental NLP task that aims to automatically produce persuasive arguments. Effective human argumentation is inherently complex and multifaceted, integrating argumentative strategies, appropriate styles, and adaptation to target audiences, etc. However, existing studies focus on limited control signals such as topic, stance, or key aspects, failing to capture this complexity. As LLMs advance, the lack of benchmarks evaluating multifaceted argumentative control becomes a critical bottleneck. To address this, we introduce ArgGenBench, a novel benchmark containing complex instructions that integrate multi-dimensional control, including topic, stance, length, style, strategy, audience, and key points. Extensive evaluation across 15 LLMs reveals significant limitations: even the best-performing model achieves only 42.7% win rate against human-verified references. These results highlight the challenge of controlled argument generation and establish ArgGenBench as a rigorous testbed for developing more capable systems.

1 Introduction

Argumentation is a cornerstone of human communication, reasoning, and collaboration (Lawrence and Reed, 2019; Stab and Gurevych, 2017; Bao et al., 2025). In NLP, a central task to examine this capability is argument generation, which seeks to automatically synthesize persuasive arguments (Khatib et al., 2021; Wang et al., 2023; Hu et al., 2025). Robust argument generation systems could enable diverse applications such as writing assistants (Zhang et al., 2023), educational tools (Green, 2017), and debate systems (Slonim et al., 2021). An effective argument generation system requires the ability not only to produce coherent topic-relevant arguments but also to reason, present

* Equal Contribution.

† Corresponding author.

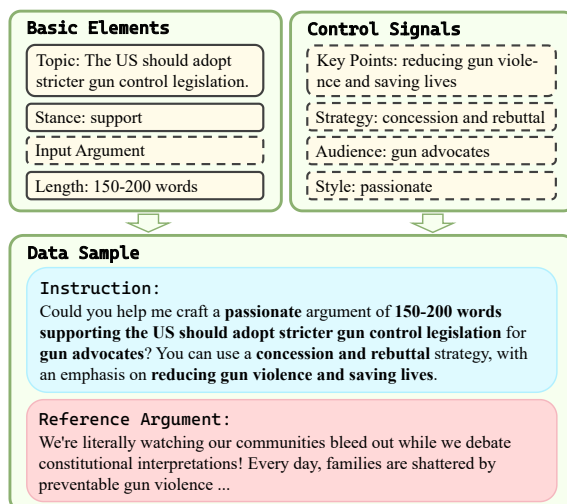


Figure 1: The illustration of a data sample in our proposed benchmark. Elements within solid-line boxes are mandatory, while those within dashed-line boxes are optional. For this specific instruction, it integrates all basic elements and control signals except the “input argument” element.

evidence, employ argumentative strategies, and adopt appropriate styles to influence the audience. As LLMs advance, their ability to generate high-quality arguments has drawn significant attention (Lin et al., 2023; Mouchel et al., 2025).

However, the prevailing paradigm typically generates arguments only based on a few argumentative elements: topic (Gretz et al., 2020; Alshomary et al., 2022), stance (Hua and Wang, 2018; Hua et al., 2019), or key content aspects (Schiller et al., 2021; Saha and Srihari, 2023). While these elements define the basic semantic foundation of arguments, they alone cannot capture the complexity inherent in human argumentation. Instead, persuasive arguments are crafted by further incorporating specific argumentative strategies (e.g., appeal to authority, appeal to emotion), adopting a suitable style (e.g., polite, passionate), and adapting to a target audience. For example (see Figure 1), to ef-

fectively persuade gun advocates to support a gun ban requires more than just stating one’s views—it calls for presenting in a more persuasive manner: a passionate style can help make the argument more engaging and relatable, a “concession and rebuttal” strategy can acknowledge opposing views while systematically refuting them to enhance logical strength and foster trust, emphasizing key points like “reducing gun violence and saving lives” clarifies the moral focus and reinforces the persuasiveness, and targeting a specific audience helps shape the argument in a way that feels more convincing to them. These nuanced multi-dimensional controls jointly influence the argument’s persuasive power but remain unaddressed in existing work. As LLMs demonstrate increasingly strong capabilities, the lack of argument generation benchmarks that can measure and drive progress in these nuanced dimensions of control becomes a critical bottleneck.

To address this issue, we introduce ArgGenBench, a novel benchmark designed to assess LLMs’ ability to follow complex, multi-dimensional argument generation instructions and generate high-quality arguments. Unlike existing benchmarks that narrowly focus on a few argumentative elements, ArgGenBench features instructions that simultaneously integrate multiple argumentative dimensions: basic elements (topic, stance, input argument, length) combined with advanced controls including style, strategy, target audience, and key point. While basic elements define the core semantic and contextual foundations of the argument, advanced argumentative control signals specify how the argument is presented and tailored to achieve effective persuasion. Here we introduce “input argument” as an optional basic element, mainly for counter-argument generation tasks where the model must oppose a given argument and for tasks where the model is required to support an existing argument.

The construction of ArgGenBench involves a two-phase process: **instruction creation** and **reference argument curation**. In the first phase, we systematically generate a diverse set of complex instructions. The construction process of each instruction begins with a set of basic elements: a topic, a target stance, an input argument (optional), and a length requirement. We then enrich this set of basic elements by sampling and combining control signals from four key dimensions: argumentative style, argumentative strategy, target audience, and key point. These structured components are syn-

thesized into a natural language instruction by an advanced LLM, which is then manually validated and selected to ensure clarity and feasibility. An example of a constructed instruction is illustrated in Figure 1. In the second phase, for each instruction, we aim to provide a human-verified high-quality argument as a reference for evaluated LLMs to compare with and determine the win/tie/lose. Each instruction, augmented with external knowledge, is given to a set of advanced LLMs (e.g., GPT-4.1, Claude-Sonnet-4) to generate multiple candidate arguments. A rigorous human-in-the-loop process follows, where annotators evaluate each candidate, select the one with the best instruction adherence and highest argument quality, and perform manual revisions if needed to produce a high-quality reference argument.

Following this methodology, we construct a benchmark of 803 instances, where each instance consists of an argument generation instruction and a high-quality reference argument.¹ We use this benchmark to comprehensively evaluate diverse LLMs, covering 15 widely used open-source and closed-source models, under three typical settings: zero-shot, supervised fine-tuning (SFT), and direct preference optimization (DPO). Our evaluation relies on both LLM-as-a-judge and human assessment. The results indicate that existing LLMs still exhibit notable limitations in complex controlled argument generation. We also observe that thinking models generally outperform non-thinking models in the zero-shot setting and both SFT and DPO can significantly enhance LLMs’ performance on controlled argument generation.

2 Related Work

The goal of argument generation is to automatically produce persuasive text supporting a specific standpoint. Early approaches often relied on existing argument knowledge to construct new arguments (Bilu and Slonim, 2016; Baff et al., 2019). With the rise of neural networks, the field has shifted towards end-to-end generation (Khatib et al., 2021; Schiller et al., 2021; Bao et al., 2022; Wang et al., 2023), where arguments are typically generated based on control signals such as topic, stance, key aspect, etc.

Topic serves as the foundational control signal for nearly all argument generation tasks. For in-

¹The data and codes are available at <https://github.com/bj-jin/ArgGenBench>.

stance, (Gretz et al., 2020) explored generating claims for a given topic using GPT-2. Beyond topic, many researches concentrated on generating arguments based on a given stance, often in the context of counter-argument generation (Hidey and McKeown, 2019). Early work in this area employed sequence-to-sequence models with retrieved evidence (Hua and Wang, 2018; Hua et al., 2019). More recently, the field has seen a shift towards leveraging LLMs, with studies investigating agent-based approaches (Hu et al., 2024) and the application of techniques like instruction-tuning (Stahl et al., 2025) and chain-of-thought prompting (Lin et al., 2023). In this line of work, the initial argument to be refuted also acted as a crucial input. Some approaches leveraged this by identifying its weak premises to formulate a more effective rebuttal (Alshomary et al., 2021b) or generating stylized evidence-based counter-arguments to enhance persuasiveness (Verma et al., 2025).

To exert more fine-grained semantic control over arguments, some studies introduced content-related signals. For example, (Schiller et al., 2021) used argument aspects to guide generation, while (Khatib et al., 2021) utilized relational triples from knowledge graphs. In a similar vein, (Saha and Srihari, 2023) employed factual spans to anchor the generated argument to specific pieces of evidence. Finally, some studies have explored generating arguments adhering to other objectives, such as tailoring arguments to a target audience (Alshomary et al., 2021a, 2022; Heinisch et al., 2024; Baff et al., 2024) or aiming to reduce logical fallacies (Mouchel et al., 2025).

While prior studies greatly advance controlled argument generation, they often isolate control signals and address only narrow dimensions. Research on generating arguments that must jointly satisfy complex constraints (e.g., style, strategy, content specifications) remains limited, leaving existing benchmarks inadequate for evaluating LLMs’ capability to follow complex, multi-faceted instructions. To address this gap, we introduce a benchmark with systematically combined argumentative control signals, enabling more comprehensive evaluation of LLMs for controlled argument generation.

3 Construction of ArgGenBench

The construction of ArgGenBench follows a carefully designed multi-stage process to ensure its high quality. Each instance contains a clear, feasi-

ble, and complex instruction with a human-verified reference argument. It consists of two main phases: (1) **Instruction Creation**, where we generate and validate the argument generation instructions, and (2) **Reference Argument Curation**, where we produce the high-quality reference arguments. The overview of the framework is shown in Figure 2.

3.1 Instruction Creation

This phase aims to generate diverse argument generation instructions that incorporate multiple control dimensions simultaneously through three steps: establishing basic elements, sampling argumentative control signals, and synthesizing natural language instructions.

Establishing Basic Elements. Each instruction begins with a set of basic elements that define the core semantic and argumentative context:

- **Topic:** Drawing on the experience of prior studies (Durmus et al., 2019; Jo et al., 2021), we curate diverse controversial topics from Kialo², a debate platform that hosts discussions across numerous domains such as politics, economics, society, etc. A representative topic is “The US should adopt stricter gun control legislation.” Topics are manually screened to ensure quality, diversity and the potential to generate meaningful and multi-perspective debate.
- **Input Argument (Optional):** In some cases, we include a given argument as context. This applies to counter-argument generation tasks where the model must refute a given argument, and to tasks where the model needs to provide supporting arguments for an existing argument. To construct these Input Arguments, we adopt a two-round process: in the first round, we construct all instances whose instructions do not require an “input argument”; then in the second round, we construct instances whose instructions need an “input argument”, and the reference arguments of the instances generated in the first-round can naturally serve as the high-quality “input argument” for the instructions that needs an “input argument” in the second round under the same topic.
- **Stance:** We define the stance of the generated argument as either “support” or “attack”. When no “input argument” is provided, the stance targets the “topic”. When an “input argument” is

²<https://www.kialo.com>

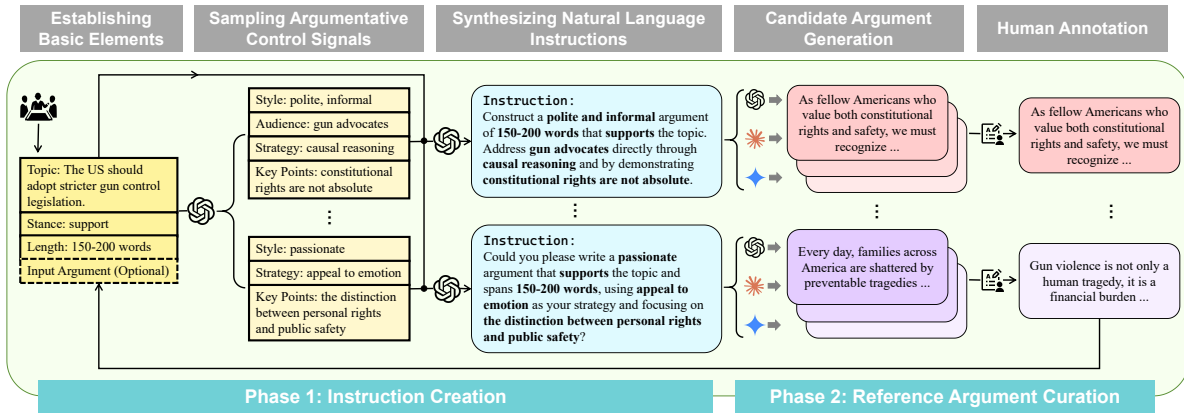


Figure 2: The framework of our method to construct ArgGenBench.

given, it targets that “input argument”.

- **Length:** We set two target length ranges for the generated arguments: 150-200 words and 250-300 words, representing two common argumentative contexts: concise arguments for social media posts or brief statements, and detailed arguments for essays or formal debates.

We randomly sample one value for each basic element to form a basic element set, serving as the argumentative context for instruction construction. An example of the basic element set is shown in the “Establishing Basic Elements” part of Figure 2.

Sampling Argumentative Control Signals. For each finalized basic element set, we further sample control signals from four categories to achieve more nuanced, diverse controls, and to enable tailored argument generation for effective persuasion:

- **Argumentative Style:** It includes three sub-dimensions shaping the tone and presentation style of an argument (defined in Appendix B.1): (1) **Formality** — *Formal* (academic, professional) vs. *Informal* (conversational, casual); (2) **Emotional Intensity** — *Passionate* vs. *Calm*; (3) **Politeness** — *Polite* vs. *Aggressive*.
- **Argumentative Strategy:** Drawing from classical rhetorical theory (Walton et al., 2008; Dimitrov et al., 2024), this dimension comprises three categories: (1) **Logos** — Appealing to the audience’s rational thinking by logical reasoning and evidence-based persuasion (e.g., *Appeal to Logic*, *Presenting Evidence*, *Causal Reasoning*). (2) **Pathos** — Targeting the audience’s emotions, values, and feelings to create persuasive impact through emotional connection (e.g., *Appeal to Emotion*, *Storytelling*, *Appeal to Sym-*

pathy). (3) **Ethos** — Establishing the speaker’s credibility and authority to enhance the persuasive power of the argument through perceived expertise and reliability (e.g., *Demonstrating Expertise*, *Appeal to Authority*, *Concession and Rebuttal*). Only several representative strategies are listed here. The full strategies and their definitions are presented in Appendix B.1.

These strategies are not mutually exclusive and often overlap. For example, *Appeal to Logic* broadly includes more logical approaches like *Causal Reasoning*. Such overlap mirrors the complexity of real-world argumentation, where multiple rhetorical approaches are commonly combined. We maintain this granular classification to better reflect real-world argumentation.

- **Target Audience:** For each basic element set, we use GPT-4.1 to generate five topic-relevant candidate target audiences reflecting several realistic argumentation scenarios where the argument might be presented (see prompt in Table 22). Take the gun control topic for example, one audience could be “gun advocates”.
- **Key Point:** Similar to audience generation, for each basic element set, we use GPT-4.1 to generate eight topic- and stance-relevant candidate key content points that can be addressed in the generated argument (prompt is in Table 23). For instance, for an argument supporting gun control, one key point could be “the distinction between personal rights and public safety.”

We sample 0-2 **Argumentative Styles** (we enforce at most one choice per style sub-dimension to avoid conflicts), 0-3 **Argumentative Strategies**, 0-3 **Key Points**, and 0-1 **Target Audience**, and then combine all sampled control signals to construct a

control signal set, ensuring each set contains 0-9 constraints to balance diversity and manageable complexity.³ The sampled control signal set, combined with the basic element set, together defines the requirements for creating a specific instruction that explicitly integrates all basic elements and control signals to generate the target argument. For each basic element set, we sample four control signal sets to create four distinct instructions covering different realistic requirements. Two examples of control signal sets are shown in the “Sampling Argumentative Control Signals” part of Figure 2.

Synthesizing Natural Language Instructions.

Each sampled control signal set along with the basic element set is provided to GPT-4.1 with a prompt template (Table 24) to produce a feasible instruction emphasizing clarity (unambiguous), specificity (control signals are explicitly integrated), and natural phrasing (fluently expressed). Implementation details are in Appendix B.2. Two examples of synthesized instructions are shown in Table 19 and Table 20. Finally, all synthesized instructions undergo manual validation by two annotators, and the instructions that fail to meet the above requirements (i.e., clarity, specificity, feasibility, natural phrasing) are discarded. The annotation results are presented in the “Quality” part of Section 4. For each basic element set, we construct four instructions with different controls using the four control signal sets sampled in the “Sampling Argumentative Control Signals” step, respectively.

3.2 Reference Argument Curation

In this phase, to provide a reliable strong baseline for evaluated LLMs to compare with, we create a high-quality reference argument for each validated instruction through a human-in-the-loop synthetic data generation (Guo and Chen, 2024; Bao et al., 2023; Havrilla et al., 2024) process combining multiple advanced LLMs with expert human oversight, ensuring the reference argument faithfully follows the control signals while maintaining coherence, persuasiveness, and quality.

Candidate Argument Generation. For each instruction, we use three leading LLMs (Claude-Sonnet-4, GPT-4.1, and Gemini-2.5-Pro) to generate three candidate arguments, ensuring diversity and mitigating model-specific biases.

³Extensive research and large-scale annotation reveal no conflicts beyond those between the calm and aggressive styles; therefore, we exclude cases in which both appear together.

To improve factual accuracy and depth, we augment the argument generation with external knowledge from two sources: (1) structured arguments from the corresponding topic threads on Kialo to provide contextually relevant claims, evidence, and counterarguments, and (2) topic-specific information retrieved from the internet to provide current and comprehensive knowledge beyond the pre-existing debate platforms.

Each advanced LLM receives the instruction, retrieved knowledge, and a carefully designed prompt (Table 21) that emphasizes adherence to control signals, integration of external evidence, and assurance of argument quality to generate a high-quality argument. This process yields three candidate arguments per instruction, providing annotators with multiple options to choose the best-performing one as the reference argument in the subsequent evaluation step. Details are provided in Appendix B.3.

Human Annotation. Our human annotation involves two graduate students with expertise in computational argumentation. For each instruction, they assess all three candidate arguments using comprehensive rubrics: (1) instruction adherence—how well control signals are followed, and (2) argument quality—logical coherence, persuasiveness, and clarity. Each dimension is scored in binary (0/1) to avoid the greater subjectivity of fine-grained multi-level scoring. Then we select the best-performing candidate on both dimensions as the final reference argument. If none fully meet our standards, the relatively better candidate is selected and manually revised to fully follow the instruction and improve its quality. The annotation results are shown in Section 4.

To validate annotation quality and consistency, 50 randomly sampled instances annotated by A are checked by B, and another 50 samples annotated by B are checked by A, yielding pass rates of 94% and 96%, indicating annotation consistency and providing quality assurance for reference arguments. More details are presented in Appendix B.4.

4 Data Statistics

ArgGenBench contains 803 instances across 82 topics, each comprising a complex instruction and a human-verified reference argument. Instructions average 48.6 words, while reference arguments average 235.2 words. Notably, 88.4% of instructions integrate multiple control signals, reflecting the benchmark’s emphasis on complex multi-

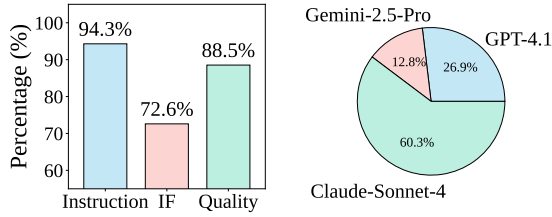


Figure 3: (Left): Instruction retention rate (Instruction) and the percentage of candidate arguments rated 1 by human evaluators on both Instruction Following (IF) and Quality dimensions. (Right): Proportion of three employed LLMs whose generated candidate arguments are selected as the final reference arguments.

dimensional control. The scale of 803 instances is comparable to benchmarks in other text generation tasks (Sabour et al., 2024; Sakurai and Miyao, 2024), supporting effective evaluation of LLMs on controlled argument generation.

Quality Figure 3 shows that 94.3% of the instructions are retained after manual screening. Notably, 60.3% of reference arguments come from Claude-Sonnet-4, highlighting its stronger performance. 88.5% and 72.6% of candidate arguments are scored 1 by human annotators on instruction following and argument quality, respectively. 41% of reference arguments undergo manual revision, ensuring rigorous adherence to instructions and high argument quality.

Complexity As shown in Figure 4, the instructions in ArgGenBench maintain a relatively balanced distribution across different difficulty levels (easy: 0–3 control signals, medium: 4–6 control signals, hard: 7–9 control signals), supporting a comprehensive evaluation of LLMs across different difficulty levels. On average, each instruction contains 4.2 control signals, reflecting the benchmark’s high complexity.

Diversity ArgGenBench covers 82 topics across diverse real-world domains (shown in Figure 6), such as politics, economics, society, etc. Each style and argumentative strategy is relatively evenly integrated (Figure 7) in instructions, enabling effective and balanced evaluation of LLMs’ capabilities on diverse styles and strategies. Moreover, we construct diverse audiences and key points for each topic (examples are shown in Table 18). More detailed analysis is presented in Appendix A.1.

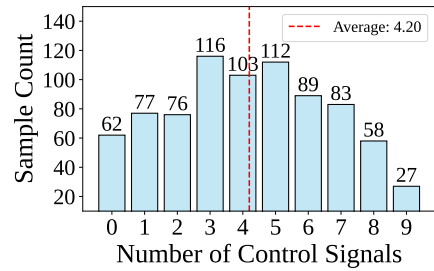


Figure 4: Instruction distribution across different numbers of integrated control signals.

5 Experiment

5.1 Evaluated Models

We benchmark LLMs’ capability of controlled argument generation using the prompt in Table 27 on three settings: Zero-Shot (Z-S), SFT, and DPO.

Zero-Shot For open-source models, we employ the most commonly used models, including Qwen3 series (Yang et al., 2025), the Llama series (Dubey et al., 2024), and the Gemma-3 series (Kamath et al., 2025). For closed-source models, we choose the GPT series, the Gemini series, and the Claude series. See Table 1 for detailed models.

SFT We use the Qwen3-8B-Base and the Llama series (see Table 1). To construct SFT data, we use 379 topics different from those in the benchmark and follow the same pipeline as our methodology, except we skip manual instruction filtering and directly employ the best-performing Claude-Sonnet-4 (Figure 3) to generate an argument for each instruction, and then apply length filtering to obtain the final SFT data, yielding 2525 samples.

DPO We employ the same models as used in SFT for fair comparison. Similar to SFT, we employ another 100 different topics to create 800 instructions. Then we sample two candidate arguments for each instruction using the SFT model with the same base model and annotate preference labels with Gemini-2.5-pro (different from the judge-GPT-4.1). More details are presented in Appendix B.6.

5.2 Evaluation Metrics

To comprehensively evaluate LLMs on controlled argument generation, we use the following five metrics. **ROUGE-L (R-L)** measures similarity between generated and reference texts (Lin, 2004). **Length-Following (LF)** evaluates how many generated arguments meet the length constraint. **Instruction-Following (IF)** measures the

	Model	R-L	LF	IF			Quality			Overall		
				Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Zero-Shot	Qwen3-8B-Base	17.2	23.8	2.4	8.2	89.4	3.0	4.6	92.4	0.4	0.9	98.8
	Qwen3-8B (thinking)	18.2	59.0	14.4	43.4	42.2	28.1	19.7	52.3	12.5	11.5	75.9
	Qwen3-14B (thinking)	18.5	73.4	14.7	44.6	40.6	24.8	20.8	54.4	18.7	13.7	67.6
	Qwen3-32B (thinking)	17.4	52.8	33.6	46.5	19.9	61.5	19.6	18.9	28.3	11.3	60.4
	Llama-3.1-8B-Inst	16.6	47.1	1.4	7.6	91.0	1.4	0.7	97.9	0.4	0.6	99.0
	Llama-3.3-70B-Inst	13.1	4.6	1.5	11.8	86.7	2.4	1.9	95.8	0.1	0.0	99.9
	Gemma-3-12B-it	18.0	70.7	7.3	32.6	60.0	7.1	11.3	81.6	5.9	7.5	86.7
	Gemma-3-27B-it	17.4	64.5	9.0	38.4	52.7	8.5	15.9	75.6	7.5	8.2	84.3
	GPT-4o	19.8	70.7	10.6	44.0	45.5	19.3	17.2	63.5	11.5	10.1	78.5
	GPT-4.1	20.7	91.3	14.3	57.7	28.0	21.2	26.4	52.4	19.7	21.5	58.8
	Gemini-2.5-pro (thinking)	17.5	66.4	15.1	36.4	48.6	14.7	18.7	66.6	8.6	12.2	79.2
	Claude-Haiku-3.5	17.1	87.5	6.6	28.8	64.6	13.7	14.2	72.1	8.4	12.0	79.6
	Claude-Sonnet-3.7 (thinking)	18.2	90.2	23.4	53.2	23.3	43.3	24.1	32.6	38.9	20.1	41.0
Claude-Sonnet-4 (thinking)	20.1	86.2	28.5	51.8	19.7	46.5	24.0	29.5	42.7	19.4	37.9	
SFT	Llama-3.1-8B-Inst	18.3	83.6	11.7	37.1	51.2	14.2	16.8	69.0	10.3	11.0	78.7
	Llama-3.1-8B	17.9	78.5	8.8	36.2	54.9	11.2	14.8	74.0	8.5	9.5	82.1
	Qwen3-8B-Base	18.6	79.3	12.5	35.1	52.4	14.1	17.2	68.7	11.7	10.5	77.8
DPO	Llama-3.1-8B-Inst	18.6	51.8	19.3	44.2	36.5	19.8	19.7	60.5	9.6	7.8	82.6
	Llama-3.1-8B	18.5	44.2	15.2	37.5	47.3	18.9	17.9	63.1	5.4	6.0	88.7
	Qwen3-8B-Base	18.8	23.2	24.0	39.6	36.4	33.9	22.4	43.7	5.4	4.6	90.0

Table 1: The performance of different models on ROUGE-L (R-L), Length Following (LF), Instruction Following (IF), Argument Quality (Quality), and Overall metrics. The IF and Quality are evaluated by GPT-4.1.

win/tie/lose rate of the generated arguments against the reference arguments, according to which better satisfies the requirements of topic, stance, and each control signal in the instructions. **Argument Quality (Quality)** measures the win/tie/lose rate of the generated arguments against the reference arguments, based on logical coherence, persuasiveness, clarity, etc. Since longer arguments may allow a higher upper bound for IF and Quality (e.g., more thorough reasoning, more key aspects), the two metrics may systematically favor models that tend to produce longer responses. For fairness, we additionally introduce an **Overall** metric. It is based on correct length adherence and then jointly considers IF and Quality. Specifically, if LF fails, it is directly marked as lose; otherwise, it is decided by the majority of IF and Quality (detailed in Appendix B.5). We also report the Overall results by directly using the majority of IF and Quality in Appendix A.3.

While human evaluation remains the gold standard, it is costly and prone to individual biases. Recent studies demonstrate that LLMs can serve as reliable judges, showing strong consistency with human assessments (Kwan et al., 2024; Yeginbergen et al., 2025). Building on this, we employ GPT-4.1 as our automatic evaluator for IF and Quality, enabling scalable evaluation while remaining comparable to human judgments (Section 5.5). To avoid positional bias, similar to previous work

(Li et al., 2024), we get the final result by averaging two position-swapped evaluations (detailed in Appendix B.5). To further mitigate concerns about bias, we also report results evaluated by Gemini-2.5-pro in Appendix A.2, which show similar trends to GPT-4.1. More details of evaluation and metric calculation are in Appendix B.5.

5.3 Implementation Details

We use the Huggingface TRL framework with LoRA (Hu et al., 2022) for efficient training. For SFT, we use the instruction as the input and the argument as the label. For DPO, we use the preferred candidate argument as the chosen response, and discard samples whose chosen argument violates the length constraint. More implementation details are presented in Appendix B.7.

5.4 Main Results

Table 1 presents the performance of each model evaluated by GPT-4.1. We also report the results of directly comparing two arguments generated by two evaluated models in Appendix A.4.

Thinking Models Generally Outperform Non-Thinking Models. From Table 1, we observe that in the zero-shot setting, thinking models (e.g., Claude-Sonnet-4/3.7) generally outperform non-thinking models (e.g., GPT-4.1, Gemma-3-12/27b-it) on both IF and Quality. Through careful ob-

Model	IF			Quality			Overall		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Qwen3-8B-Base	2.0	6.0	92.0	1.0	3.0	96.0	0.0	1.0	99.0
Qwen3-8B-Base (SFT)	12.0	26.0	62.0	9.0	6.0	85.0	9.0	8.0	83.0
Qwen3-8B-Base (DPO)	21.0	31.0	48.0	22.0	19.0	59.0	5.0	2.0	93.0
Qwen3-8B (thinking)	18.0	22.0	60.0	20.0	23.0	57.0	16.0	13.0	71.0
Qwen3-32B (thinking)	41.0	26.0	33.0	60.0	19.0	21.0	28.0	11.0	61.0
GPT-4.1	28.0	32.0	40.0	25.0	21.0	54.0	23.0	22.0	55.0
Claude-Sonnet-4 (thinking)	33.0	32.0	35.0	48.0	15.0	37.0	38.0	17.0	45.0

Table 2: The human evaluation results (win/tie/lose) of different models on IF, Quality and Overall metrics. All models are evaluated in the zero-shot setting, except Qwen3-8B-Base (SFT) and Qwen3-8B-Base (DPO).

Model	Easy			Medium			Hard		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Qwen3-8B-Base	0.6	1.8	97.6	0.0	0.3	99.7	0.6	0.0	99.4
Qwen3-8B-Base (SFT)	9.3	13.6	77.1	14.1	9.2	76.6	11.9	6.5	81.5
Qwen3-8B-Base (DPO)	6.6	5.7	87.6	4.3	4.6	91.1	4.8	2.4	92.9
Qwen3-8B (thinking)	11.9	14.0	74.1	12.3	9.6	78.1	14.3	10.1	75.6
Qwen3-32B (thinking)	26.3	11.5	62.2	31.9	13.2	54.9	25.6	7.7	66.7
GPT-4.1	15.4	25.4	59.2	22.7	20.4	56.9	22.6	16.1	61.3
Claude-Sonnet-4 (thinking)	42.0	20.2	37.8	44.7	19.7	35.5	40.5	17.3	42.3

Table 3: The Overall performance of representative models on different instruction difficulty levels: Easy (0-3 control signals), Medium (4-6 control signals), and Hard (7-9 control signals), based on the evaluation by GPT-4.1. All models are evaluated in the zero-shot setting, except Qwen3-8B-Base (SFT) and Qwen3-8B-Base (DPO).

servation, we find that CoT offers clear argument structure planning, rigorous logical relation analysis, and explicit attention to instruction requirements, thereby yielding arguments with stronger instruction adherence and higher quality.

SFT and DPO can Significantly Improve IF and Quality. Table 1 shows that for Qwen3-8B-Base, both SFT and DPO significantly improve IF and Quality over its zero-shot version. After DPO, it even surpasses the thinking model Qwen3-14B with larger parameter scale. A similar trend also holds for Llama-3.1-8B-Inst, suggesting that SFT and DPO can substantially enhance instruction following and argument quality. Moreover, SFT also improves LF, whereas DPO does not, likely because our DPO data construction only ensures length compliance of chosen arguments without providing explicit length preference signals.

Claude-Sonnet-4 Achieves the Best Overall Performance. Table 1 shows that although Qwen3-32B achieves strong IF and Quality, its weaker LF leads to a lower Overall score than Claude-Sonnet-4, which achieves the highest Overall performance. Despite this, Claude-Sonnet-4 only has a win rate of 42.7% against our human-verified high-quality reference arguments, indicating that current LLMs still fall short in controlled argument gener-

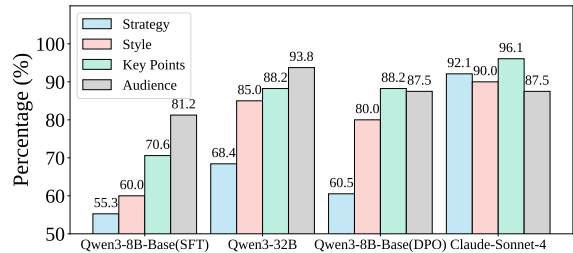


Figure 5: The adherence rates of each control category on instructions which integrate only one control signal from corresponding category.

ation. These results highlight the high difficulty of ArgGenBench and its effectiveness in evaluating LLMs on this task. To further investigate the existing issues, we conduct an in-depth error analysis in Appendix A.7, summarizing several representative error types (e.g., insufficient reasoning, logical fallacies).

5.5 Human Evaluation Results

To further verify the accuracy of the evaluation results conducted by GPT-4.1, we randomly sample 100 data points and conduct human evaluations across seven representative models. More implementation details are in Appendix B.8. As observed in Table 2, the human evaluation results exhibit a similar trend to those of GPT-4.1, supporting the

Model	R-L	LF	IF			Quality			Overall		
			Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Qwen3-8B-Base	17.9	23.2	2.2	7.9	89.9	4.1	5.8	90.1	0.6	1.2	98.2
Qwen3-8B-Base (SFT)	19.1	79.8	12.7	35.7	51.6	15.9	18.6	65.5	12.5	10.9	76.6
Qwen3-8B-Base (DPO)	19.4	23.4	24.5	40.1	35.4	35.6	24.5	39.9	6.7	5.2	88.1
Qwen3-8B (thinking)	19.1	59.8	15.0	44.1	40.9	30.4	20.9	48.7	14.1	13.5	72.4
Qwen3-32B (thinking)	18.2	54.2	32.7	45.9	21.4	63.1	20.4	16.5	29.8	13.0	57.2
GPT-4.1	21.9	90.5	13.9	57.2	28.9	25.6	29.1	45.3	22.1	23.9	54.0
Claude-Sonnet-4 (thinking)	21.4	85.7	28.9	51.1	20.0	50.4	26.2	23.4	45.2	22.6	32.2

Table 4: The performance of different models with the argument generation process augmented by external knowledge. The IF and Quality are evaluated by GPT-4.1. All models are tested in the zero-shot setting, except Qwen3-8B-Base (SFT) and Qwen3-8B-Base (DPO).

validity of the main results and the reliability of GPT-4.1 as an evaluator.

5.6 Performance on Different Difficulty Levels

To explore LLMs’ performance on different difficulty levels, we report the Overall results on each difficulty level in Table 3. As observed, as instruction difficulty increases, the Overall lose rates of most models also rise accordingly, suggesting that when dealing with complex instructions involving more control signals, existing LLMs still face challenges in satisfying instruction requirements and ensuring the argument quality. More in-depth analysis and the performance of all evaluated models across all evaluation metrics and difficulty levels are presented in Appendix A.5.

5.7 Performance on Each Control Category

To further assess LLMs’ ability to follow each control category, we report the IF results of several representative models on instructions integrating only one control signal from a certain category. Figure 5 shows that most models perform worse on argumentative strategy than the other three control dimensions. While style and audience are often conveyed through readily identifiable lexical or tonal cues and key points can usually be addressed by incorporating and elaborating on specific content elements, strategies (e.g., *Causal Reasoning*, *Concession and Rebuttal*) demand deeper reasoning and structural planning. They involve organizing arguments according to abstract rhetorical patterns rather than surface signals, making strategies substantially harder for models to implement.

5.8 Performance of Incorporating External Knowledge

To investigate the impact of external knowledge on controlled argument generation, we evaluate the

models’ performance with external knowledge incorporated during argument generation. We adopt the same external knowledge usage setting as used in our ArgGenBench construction (see Appendix B.3) and the results are presented in Table 4.

Compared to the results without using external knowledge (Table 1), we can observe that incorporating external knowledge improves the Argument Quality of the generated arguments, but yields no significant gains in Length-Following or Instruction-Following. Case analyses reveal that the retrieved knowledge primarily enriches the informational content and evidential grounding of generated arguments, thereby enhancing their persuasiveness and coherence. However, since the external knowledge primarily provides content support rather than guiding structural or stylistic compliance, it exerts relatively limited influence on model’s ability to precisely follow the complex multi-dimensional requirements of instructions.

6 Conclusion

In this work, we propose ArgGenBench, a benchmark designed to comprehensively evaluate LLMs’ ability to follow complex, multi-dimensional instructions and generate high-quality arguments for controlled argument generation. Extensive evaluations on diverse models show that existing LLMs still fall short in length control, instruction following, and argument quality. SFT, DPO and incorporating thinking processes can greatly improve LLMs’ performance on this task.

Limitations

Our evaluation of argument quality covers aspects such as persuasiveness, coherence, and clarity, but does not explicitly assess factual grounding. As a fair and accurate assessment of factual ground-

ing typically requires a carefully designed retrieval framework and effective external knowledge, this lies beyond the scope of our current work. We leave incorporating factual grounding assessment into ArgGenBench as an important direction for future research.

Ethics Statement

The use of the Llama models complies with the Meta Llama 3 License⁴. All other artifacts are used under their respective licenses, following their intended purposes. We note that the audience-tailored persuasion may be misused to manipulate people’s opinion. Therefore, our study can only be used for academic research purposes, and any subsequent work should carefully consider and explicitly acknowledge these ethical concerns. The ArgGenBench does not contain any information that names or uniquely identifies individual people or offensive content. Our proposed benchmark is openly available to all academic research, provided that proper citation of this work is included.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China 62576120.

References

- Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. **The moral debater: A study on the computational generation of morally framed arguments**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 8782–8797. Association for Computational Linguistics.
- Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021a. **Belief-based generation of argumentative claims**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 224–233. Association for Computational Linguistics.
- Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021b. **Counter-argument generation by attacking weak premises**. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1816–1827. Association for Computational Linguistics.
- Roxanne El Baff, Khalid Al Khatib, Milad Alshomary, Kai Konen, Benno Stein, and Henning Wachsmuth. 2024. **Improving argument effectiveness across ideologies using instruction-tuned large language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 4604–4622. Association for Computational Linguistics.
- Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. **Computational argumentation synthesis as a language modeling task**. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 54–64. Association for Computational Linguistics.
- Jianzhu Bao, Mohan Jing, Kuicai Dong, Aixin Sun, Yang Sun, and Ruifeng Xu. 2025. **Uniasa: A unified generative framework for argument structure analysis**. *Computational Linguistics*, 51(3):739–784.
- Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. **A synthetic data generation framework for grounded dialogues**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882.
- Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. **Aeg: Argumentative essay generation via a dual-decoder model with content planning**. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5134–5148.
- Yonatan Bilu and Noam Slonim. 2016. **Claim synthesis via predicate recycling**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.
- Jiaao Chen and Diyi Yang. 2021. **Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests**. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12648–12656. AAAI Press.
- Mengqi Chen, Bin Guo, Hao Wang, Haoyu Li, Qian Zhao, Jingqi Liu, Yasan Ding, Yan Pan, and Zhiwen Yu. 2025. **The future of cognitive strategy-enhanced persuasive dialogue agents: new perspectives and trends**. *Frontiers Comput. Sci.*, 19(5):195315.
- Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. **Improving multi-turn emotional support dialogue generation with lookahead strategy planning**.

⁴<https://www.llama.com/llama3/license/>

- In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3014–3026. Association for Computational Linguistics.
- Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. [Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation, SemEval@NAACL 2024, Mexico City, Mexico, June 20-21, 2024*, pages 2009–2026. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Esin Durmus and Claire Cardie. 2018. [Exploring the role of prior beliefs for argument persuasion](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1035–1045. Association for Computational Linguistics.
- Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. [The role of pragmatic and discourse context in determining argument impact](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5667–5677. Association for Computational Linguistics.
- B. J. Fogg. 2002. [Persuasive technology: using computers to change what we think and do](#). *Ubiquity*, 2002(December):5.
- Nancy L. Green. 2017. [Argumentation scheme-based argument generation to support feedback in educational argument modeling systems](#). *Int. J. Artif. Intell. Educ.*, 27(3):515–533.
- Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. [The workweek is the best time to start a family - A study of GPT-2 based claim generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 528–544. Association for Computational Linguistics.
- Xu Guo and Yiqiang Chen. 2024. [Generative ai for synthetic data generation: Methods, challenges and the future](#). *arXiv preprint arXiv:2403.04190*.
- Alex Havrilla, Andrew Dai, Laura O’Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, and 1 others. 2024. [Surveying the effects of quality, diversity, and complexity in synthetic data from large language models](#). *arXiv preprint arXiv:2412.02980*.
- He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. [Decoupling strategy and generation in negotiation dialogues](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2333–2343. Association for Computational Linguistics.
- Philipp Heinisch, Lorik Dumani, Philipp Cimiano, and Ralf Schenkel. 2024. ["tell me who you are and I tell you how you argue": Predicting stances and arguments for stakeholder groups](#). In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Findings of ACL, pages 1968–1982. Association for Computational Linguistics.
- Christopher Hidey and Kathy McKeown. 2019. [Fixed that for you: Generating contrastive claims with semantic edits](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1756–1767. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. [Lora: Low-rank adaptation of large language models](#). *ICLR*, 1(2):3.
- Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. [Debate-to-write: A persona-driven multi-agent framework for diverse argument generation](#). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4689–4703. Association for Computational Linguistics.
- Zhe Hu, Hou Pong Chan, and Yu Yin. 2024. [AMERICANO: argument generation with discourse-driven decomposition and agent interaction](#). In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024, Tokyo, Japan, September 23 - 27, 2024*, pages 82–102. Association for Computational Linguistics.
- Xinyu Hua, Zhe Hu, and Lu Wang. 2019. [Argument generation with retrieval, planning, and realization](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2661–2672. Association for Computational Linguistics.
- Xinyu Hua and Lu Wang. 2018. [Neural argument generation augmented with externally retrieved evidence](#).

- In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 219–230. Association for Computational Linguistics.
- Yohan Jo, Seojin Bang, Chris Reed, and Eduard H. Hovy. 2021. [Classifying argumentative relations using logical mechanisms and argumentation schemes](#). *Trans. Assoc. Comput. Linguistics*, 9:721–739.
- Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan W. Black, and Yulia Tsvetkov. 2021. [Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. [Employing argumentation knowledge graphs for neural argument generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4744–4754. Association for Computational Linguistics.
- Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. [Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 86. ACM.
- Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. [Mt-eval: A multi-turn capabilities evaluation benchmark for large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 20153–20177. Association for Computational Linguistics.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational linguistics*, 45(4):765–818.
- Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. [From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7602–7635. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Text summarization branches out*, pages 74–81.
- Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. [Argue with me tersely: Towards sentence-level counter-argument generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16705–16720. Association for Computational Linguistics.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. [Towards emotional support dialog systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.
- Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. 2023. [Audience-centric natural language generation via style infusion](#). *CoRR*, abs/2301.10283.
- Luca Mouchel, Debjit Paul, Shaobo Cui, Robert West, Antoine Bosselut, and Boi Faltings. 2025. [A logical fallacy-informed framework for argument generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7296–7314. Association for Computational Linguistics.
- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvianna S. Sunaryo, Tatiana M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. [Emobench: Evaluating the emotional intelligence of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5986–6004. Association for Computational Linguistics.
- Sougata Saha and Rohini K. Srihari. 2023. [Argu: A controllable factual argument generator](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8373–8388. Association for Computational Linguistics.

- Hiromasa Sakurai and Yusuke Miyao. 2024. [Evaluating intention detection capability of large language models in persuasive dialogues](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1635–1657. Association for Computational Linguistics.
- Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. [Aspect-controlled neural argument generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 380–396. Association for Computational Linguistics.
- Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, and 34 others. 2021. [An autonomous debating system](#). *Nat.*, 591(7850):379–384.
- Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.
- Maja Stahl, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth. 2025. [Arginstruct: Specialized instruction fine-tuning for computational argumentation](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 11103–11127. Association for Computational Linguistics.
- Preetika Verma, Kokil Jaidka, and Svetlana Churina. 2025. ["reasoning" with rhetoric: On the style-evidence tradeoff in llm-generated counter-arguments](#). In *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media, June 23-26, 2025, Copenhagen, Denmark*, pages 1966–1989. AAAI Press.
- Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.
- Xiaoou Wang, Elena Cabrio, and Serena Villata. 2023. [Argument and counter-argument generation: A critical survey](#). In *Natural Language Processing and Information Systems - 28th International Conference on Applications of Natural Language to Information Systems, NLDB 2023, Derby, UK, June 21-23, 2023, Proceedings*, volume 13913 of *Lecture Notes in Computer Science*, pages 500–510. Springer.
- Xuwei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. [Persuasion for good: Towards a personalized persuasive dialogue system for social good](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5635–5649. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2025. [Dynamic knowledge integration for evidence-driven counter-argument generation with large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22568–22584. Association for Computational Linguistics.
- Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. [VISAR: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 5:1–5:30. ACM.

A Additional Experiments

A.1 The Diversity of ArgGenBench

ArgGenBench encompasses 82 topics covering a wide range of real-world domains, such as politics, economics, society, science, education, health, etc. (see Figure 6)

The distributions of style and argumentative strategy are illustrated in Figure 7. As shown, each style and argumentative strategy is relatively evenly integrated in sufficient quantities of instructions, ensuring that ArgGenBench can enable effective evaluation across a diverse range of styles and argumentative strategies.

For the Audience and Key Point, since they are customized for each specific topic, we select several representative examples in ArgGenBench to show the diverse audience and key point candidates of each topic in Table 18. As shown, the audiences in these examples reflect a diverse range of realistic scenarios in which the argument might be presented. Moreover, the generated key points offer a variety of argumentative perspectives that can be addressed in arguments.

A.2 Use Gemini-2.5-pro to Conduct Evaluation

In the main experiments, we employ GPT-4.1 to evaluate the performance of different models and



Figure 6: The distribution of topic domains in ArgGenBench.

observe strong alignment with human judgments. To further validate the reliability of using LLMs as evaluators, we additionally conduct evaluations with another LLM, Gemini-2.5-pro. The results are presented in Table 6. As shown, Gemini-2.5-pro yields results highly consistent with those of GPT-4.1, demonstrating the credibility and robustness of LLM-based evaluation.

A.3 The Overall Performance by Directly Using the Majority of IF and Quality

The Overall metric calculation in all the other experiments is based on correct length following and then jointly considers IF and Quality. In this section, we directly use the majority of IF and Quality as the Overall result. In this section, the results of IF and Quality keep consistent with those in our main results (i.e., Table 1). As shown in Table 7, from the perspective of the Overall performance, the thinking models generally outperform the non-thinking models. Besides, after performing SFT especially DPO, the models show a significant improvement in terms of Overall performance, highlighting the effectiveness of SFT and DPO. Furthermore, the Qwen-32B outperforms Claude-Sonnet-4 on Overall performance, indicating its outstanding capability of following complex instructions and achieving high argument quality on controlled argument generation.

A.4 Direct Pairwise Comparison

To further analyze the performance of different models on the controlled argument generation task, we employ GPT-4.1 to conduct direct pairwise comparisons by directly comparing two arguments generated by two different evaluated models using the instructions in ArgGenBench. The results are presented in Table 5.

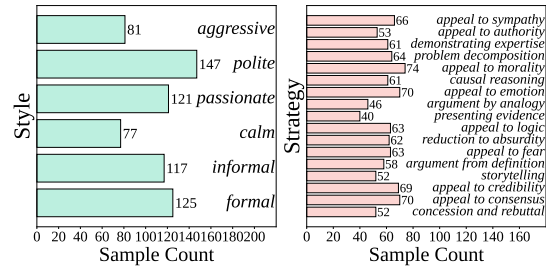


Figure 7: Visualization of the frequency of each Style (Left) and Argumentative Strategy (Right).

As shown, compared to zero-shot, Llama-3.1-8B-Inst trained with SFT demonstrates a significantly high win rate and low lose rate in terms of both IF and Quality. Moreover, Qwen3-8B-Base (DPO) outperforms the zero-shot performance of the thinking model Qwen3-8B. These observations demonstrate the effectiveness of applying SFT and DPO training on the controlled argument generation dataset. Additionally, Qwen3-32B achieves better performance than Claude-Sonnet-4 in terms of both IF and Quality, showcasing its strong instruction-following ability and high generation quality on the controlled argument generation task.

These result trends are similar to those observed when using GPT-4.1 to compare the generated arguments with the reference arguments in ArgGenBench, which indicates the consistency of evaluations conducted by GPT-4.1 and the reliability of using GPT-4.1 as our automatic evaluator.

A.5 Comprehensive Performance Results on Different Difficulty Levels

In this section, we report the performance of all evaluated models on each evaluation metric and difficulty level.

Table 9 shows the Length Following performance of each evaluated LLM on different difficulty levels. We observe that as the number of control signals integrated into the instruction increases, most LLMs show a decline in their compliance with the length requirements. This can be attributed to the greater difficulty of balancing more control requirements simultaneously. The rising lose rate of IF with higher instruction difficulty, as shown in Table 10, further supports this observation. These results demonstrate that our proposed benchmark effectively exposes the limitations of current LLMs in handling complex, multi-dimensional argument generation instructions, providing a comprehensive and fine-grained framework for evaluating their

Model	IF	Quality
Llama-3.1-8B-Inst (SFT) vs. Llama-3.1-8B-Inst (Z-S)	47.3/40.5/12.2	70.4/18.9/10.7
Qwen3-8B-Base (DPO) vs. Qwen3-8B (Z-S)	36.1/42.1/21.8	44.5/24.9/30.6
Claude-Sonnet-4 (Z-S) vs. Qwen3-32B (Z-S)	21.9/52.1/26.0	31.1/24.7/44.2

Table 5: The evaluation results (win/tie/lose) of directly comparing two different models on Instruction Following and Argument Quality metrics by GPT-4.1.

Model	IF			Quality			Overall		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Qwen3-8B-Base	1.6	8.3	90.0	0.9	1.4	97.8	0.2	0.5	99.3
Qwen3-8B-Base (SFT)	15.2	35.0	49.8	11.3	14.8	73.8	10.0	13.0	77.1
Qwen3-8B-Base (DPO)	19.2	38.5	42.3	17.8	18.7	63.5	4.1	3.6	92.3
Qwen3-8B (thinking)	12.4	34.7	52.9	20.1	19.8	60.2	10.7	10.4	78.9
Qwen3-32B (thinking)	25.5	43.8	30.6	55.7	20.0	24.3	26.3	11.7	62.0
GPT-4.1	20.5	47.7	31.8	19.6	18.8	61.6	19.2	20.2	60.6
Claude-Sonnet-4 (thinking)	29.9	44.0	26.2	47.7	21.8	30.5	41.2	20.9	37.9

Table 6: The performance of different models on Instruction Following (IF), Quality, and Overall metrics. The IF and Quality are evaluated by Gemini-2.5-pro. All models are tested in the zero-shot setting, except Qwen3-8B-Base (SFT) and Qwen3-8B-Base (DPO).

instruction-following ability.

Table 10 shows the Instruction Following capability of each evaluated LLM on different difficulty levels. As observed, for most models, as instruction difficulty increases, the tie rate decreases while both the win rate and lose rate increase. Through extensive analysis of cases, we find that low-difficulty instructions are not sufficiently discriminative: most models can follow them reasonably well, which leads to a relatively high tie rate and fewer decisive outcomes (i.e., low win rate and low lose rate). As the instruction difficulty increases, the evaluation space becomes more polarized. On the one hand, many generations fail to capture more complex multi-dimensional control signals, which naturally raises the lose rate. On the other hand, higher difficulty also creates more opportunities for models with stronger reasoning and instruction-following capabilities to distinguish themselves, thereby producing more wins. This dynamic can be observed not only in the Instruction Following dimension but also in Argument Quality (Table 11): when tasks are simple, most models produce arguments of roughly comparable quality, resulting in more ties, whereas more demanding tasks reveal sharper contrasts in fluency, coherence, and persuasiveness. In this sense, harder instructions amplify differences among models, reducing ties and yielding both more losses and more wins. Nevertheless, simple instructions remain essential, as they provide a baseline for

measuring fundamental instruction-following ability and argument quality and allow us to assess model robustness across the full spectrum of task difficulty. In summary, these findings indicate that the proposed ArgGenBench and evaluation framework successfully detect the nuanced performance dynamics of LLMs under varying instruction complexities, thereby providing a reliable and comprehensive framework for assessing their controlled argument generation capability.

From Table 12, we observe that the Overall lose rates of most models are higher on more difficult levels, indicating that as the number of integrated control signals increases, current models still struggle to follow instructions’ requirements and ensure high generation quality simultaneously.

A.6 Evaluation Results Using Likert Scale

To ensure the robustness of the evaluation results, we employ a Likert scale (i.e., 0, 1, and 2) to assess the IF and Quality of six representative models, using GPT-4.1 as the judge. Specifically, we define the Overall score as the sum of the LF ratio and the scores for IF and Quality. The results are presented in Table 8. As shown, Claude-Sonnet-4 achieves the best performance; however, its overall score remains at only 3.33—still falling significantly short of the perfect score of 5. This highlights the substantial difficulty of ArgGenBench and underscores the need for further improvements in existing LLMs regarding complex controlled ar-

	Model	R-L	LF	IF			Quality			Overall		
				Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Zero-Shot	Qwen3-8B-Base	17.2	23.8	2.4	8.2	89.4	3.0	4.6	92.4	2.4	3.5	94.1
	Qwen3-8B (thinking)	18.2	59.0	14.4	43.4	42.2	28.1	19.7	52.3	24.9	18.9	56.1
	Qwen3-14B (thinking)	18.5	73.4	14.7	44.6	40.6	24.8	20.8	54.4	25.3	17.2	57.5
	Qwen3-32B (thinking)	17.4	52.8	33.6	46.5	19.9	61.5	19.6	18.9	58.9	18.6	22.5
	Llama-3.1-8B-Inst	16.6	47.1	1.4	7.6	91.0	1.4	0.7	97.9	0.6	1.7	97.6
	Llama-3.3-70B-Inst	13.1	4.6	1.5	11.8	86.7	2.4	1.9	95.8	1.4	2.5	96.1
	Gemma-3-12B-it	18.0	70.7	7.3	32.6	60.0	7.1	11.3	81.6	8.3	9.7	81.9
	Gemma-3-27B-it	17.4	64.5	9.0	38.4	52.7	8.5	15.9	75.6	9.7	12.3	78.0
	GPT-4o	19.9	70.7	10.6	44.0	45.5	19.3	17.2	63.5	17.7	14.7	67.6
	GPT-4.1	20.7	91.3	14.3	57.7	28.0	21.2	26.4	52.4	21.7	23.5	54.8
	Gemini-2.5-pro (thinking)	17.5	66.4	15.1	36.4	48.6	14.7	18.7	66.6	15.9	17.2	66.9
	Claude-Haiku-3.5	17.1	87.5	6.6	28.8	64.6	13.7	14.2	72.1	9.4	12.9	77.7
	Claude-Sonnet-3.7 (thinking)	18.2	90.2	23.4	53.2	23.3	43.3	24.1	32.6	43.4	22.2	34.4
	Claude-Sonnet-4 (thinking)	20.1	86.2	28.5	51.8	19.7	46.5	24.0	29.5	47.9	21.8	30.3
SFT	Llama-3.1-8B-Inst	18.4	83.6	11.7	37.1	51.2	14.2	16.8	69.0	13.7	13.3	73.0
	Llama-3.1-8B	17.9	78.5	8.8	36.2	54.9	11.2	14.8	74.0	11.1	13.0	76.0
	Qwen3-8B-Base	18.6	79.3	12.5	35.1	52.4	14.1	17.2	68.7	14.9	13.6	71.5
DPO	Llama-3.1-8B-Inst	18.6	51.8	19.3	44.2	36.5	19.8	19.7	60.5	22.5	15.4	62.0
	Llama-3.1-8B	18.5	44.2	15.2	37.5	47.3	18.9	17.9	63.1	18.9	14.9	66.1
	Qwen3-8B-Base	18.8	23.2	24.0	39.6	36.4	33.9	22.4	43.7	33.7	19.3	46.9

Table 7: The performance of different models on ROUGE-L (R-L), Length Following (LF), Instruction Following (IF), Quality, and Overall metrics. Here, the Overall metric directly employs the majority result of IF and Quality. The IF and Quality adopt the evaluation outcomes in our main results (i.e., Table 1), which are evaluated by GPT-4.1.

Model	LF	IF	Quality	Overall
Qwen3-8B-Base	0.24	0.42	0.45	1.11
Qwen3-8B-Base (SFT)	0.79	0.65	0.85	2.29
Qwen3-8B-Base (DPO)	0.23	0.97	1.19	2.39
Qwen3-8B (thinking)	0.59	0.81	1.03	2.43
GPT-4.1	0.91	0.86	0.96	2.73
Claude-Sonnet-4 (thinking)	0.86	1.02	1.45	3.33

Table 8: The evaluation results using Likert scales. The GPT-4.1 is used to evaluate the IF and Quality. The Overall is the sum of LF, IF and Quality.

gument generation. Additionally, both SFT and DPO substantially enhance model performance in terms of IF and Quality.

A.7 Error Analysis

Although Claude-Sonnet-4 achieves the highest overall performance among the evaluated models, its win rate remains relatively low at 42.7%, indicating substantial room for improvement. To better understand the limitations underlying this gap, we conduct an error analysis on the model’s generated arguments. Specifically, we randomly sample 80 instances where Claude-Sonnet-4 lost in the Overall metric and categorize the common error types observed.

By systematically identifying these patterns, we aim to provide actionable insights that can inform future advances in controlled argument generation

and guide the design of models capable of producing more robust, instruction-compliant, and compelling arguments. The analysis of three main error types is detailed as follows:

Insufficient or Less Persuasive Reasoning.

This error type characterizes instances in which a model-generated argument presents reasoning that, although relevant to the claim, lacks sufficient depth, specificity, or logical strength to achieve a strong persuasive impact. The shortcoming often stems from underdeveloped rationales that fail to convincingly substantiate the stance or address counterarguments in a compelling manner. As a result, the argument risks appearing superficial and less persuasive, even if it remains thematically aligned with the instruction. An example attacking “Overlanding and offroading are good for the

environment” is shown in Table 13. In this example, Argument 1 concludes with a strong refutation: it directly counters the defense that “other environmental harms are worse” by stressing that such comparisons cannot justify additional damage. This principle-based rebuttal is both specific and compelling. By contrast, Argument 2 ends with a more generic statement that overlanding “prioritizes human recreation over environmental preservation”, which, though valid, lacks the same force and specificity, making it less persuasive.

Lack of Comprehensive Argumentative Perspectives. This error type refers to cases where an argument fails to incorporate sufficient perspectives relevant to the claim, resulting in argument quality damage since it does not capture the full complexity of the issue. For instance, in the example shown in Table 14 that attacks “The French Revolution had predominantly negative consequences”, Argument 1 develops a multi-dimensional case by addressing feudal hierarchies, scientific progress, democratic ideals, and economic development, all of which reinforce the appeal to fear and enhance credibility. In contrast, Argument 2 primarily emphasizes political oppression and social inequality, neglecting other salient domains. While valid, the lack of comprehensive argumentative perspectives weakens its overall rhetorical force.

Logical Fallacy. This is a common error type in argumentative writing where the reasoning contains flaws that weaken the validity of claims. Such fallacies often appear as unwarranted generalizations or faulty causal links, producing surface-level persuasiveness without true logical support. Prior work has also conducted research on logical fallacies in automatically generated arguments, highlighting their impact on reasoning quality (Mouchel et al., 2025). One example supporting “Robots doing work will be better for humans” is shown in Table 15. In this case, Argument 2 asserts that because leading technology companies are investing billions in robotics, this “clearly demonstrates” that robotics will create widespread prosperity for workers. This constitutes a non sequitur: the fact that corporations allocate capital to robotics may just indicate commercial interest, not a guaranteed benefit to humans.

B Additional Details

B.1 Detailed Definition of Style and Argumentative Strategy

This section presents the definitions and examples of each argumentative strategy and style, shown in Table 16 and Table 17.

B.2 Details of Synthesizing Natural Language Instructions

In this section, we introduce the details of “Synthesizing Natural Language Instructions” step of our method to construct ArgGenBench. The prompts are shown in Table 24. We emphasize several key aspects in our prompt: clarity (instructions should be unambiguous), specificity (control signals should be explicitly integrated), diversity (instructions should adopt various styles), and natural phrasing (instructions should read fluently rather than like a list of constraints). To ensure the instruction diversity, we prompt GPT-4.1 to generate three instructions at one time and sample one from them.

B.3 Details of Candidate Argument Generation

In this section, we present the implementation details of the “Candidate Argument Generation” step of our method to construct ArgGenBench. In this step, to ensure the diversity and quality of the final reference arguments and mitigate model-specific biases, we utilize three leading LLMs (Claude-Sonnet-4, GPT-4.1, and Gemini-2.5-Pro) to generate three candidate arguments, respectively. The three candidate arguments provide the annotators with multiple options to choose arguments with higher quality and better adherence to instructions as the final reference arguments in the subsequent “Human Annotation” step. Each LLM receives the instruction, the retrieved knowledge, and a carefully designed prompt (shown in Table 21) that emphasizes adherence to control signals, integration of external evidence, and assurance of argument quality to generate a high-quality argument.

To improve factual accuracy and depth, we augment the argument generation with external knowledge from two sources:

- **Kialo Knowledge:** On the Kialo platform, discussion on each topic is organized in a hierarchical tree structure, where each node indicates an argument (with the root node representing the discussion title). Edges connecting child nodes

to their parent nodes are labeled as support, attack, or neutral, indicating whether the child argument supports, opposes, or remains neutral toward the parent argument. To obtain the Kialo knowledge required for generating candidate arguments, we randomly sample a path starting from the topic node to a leaf node to ensure logical relevance and coherence among sampled arguments. We ensure that the sampled path contains at least four nodes and that the total length of all arguments does not exceed 200 words, terminating the sampling once the upper bound is reached. If the total length upon reaching a leaf node is less than four nodes, we backtrack and continue sampling until a path meeting the criteria is found.

- **Retrieved Knowledge:** For each topic, we use the corresponding discussion title on the Kialo platform (e.g., “Should the US adopt stricter gun control legislation?”) as the query to the Tavily API⁵, which returns 20 most relevant retrieved external knowledge passages. We then randomly sample 1 to 3 passages from the top 5 most relevant results, with a maximum combined length of 300 words, as the external knowledge for the candidate argument generation process. This sampling strategy introduces diverse external knowledge while ensuring the relevance and controllable length of the external knowledge.

B.4 Details of Human Annotation

In this section, we detail the human annotation process of the methodology to construct the ArgGen-Bench.

Our human annotation involves two graduate students (two authors of this paper) with expertise in computational argumentation. They evaluate each of the three candidate arguments for each validated instruction using a comprehensive rubric including two dimensions: (1) the instruction adherence, and (2) the argument quality.

For the dimension of instruction adherence, we design the criteria as follows:

- **Score 1:** The argument basically satisfies all the requirements of the instruction, including topic, stance, length, style, argumentative strategy, target audience, and key point.
- **Score 0:** The argument fails to satisfy one or more requirements of the instruction, such as

the topic, stance, length, style, argumentative strategy, target audience, or key point.

For the dimension of argument quality, we design the criteria as follows:

- **Score 1:** The argument is coherent, logically sound, clearly expressed, well-structured, and provides true factual evidence.
- **Score 0:** The argument is incoherent, illogical, ambiguous, poorly structured, or contains false evidence.

To verify the factual evidence, we employ the Perplexity AI⁶ tool, which retrieves information from high-quality and up-to-date web sources, and then aligns outputs with authoritative references and provides explicit citations, ensuring both the accuracy and the transparency of fact-checking, thereby enhancing the credibility of the verification results.

We take the sum of the scores from the two dimensions (i.e., instruction adherence and argument quality) as the criterion for selecting the final reference argument. Specifically, the candidate argument with the highest total score is chosen as the reference argument. In cases where multiple candidates obtain the same top score, the best one is manually selected. If none fully meet our standards, the relatively better candidate is selected and manually revised to fully follow the instruction and improve its argument quality.

To further ensure the reliability of the human annotation, we conduct a cross-validation procedure between the two annotators. Specifically, 50 randomly sampled instances originally annotated by Annotator A are independently checked by Annotator B, and another 50 instances annotated by Annotator B are checked by Annotator A. In both cases, the checking annotator verifies whether the original annotation faithfully follows the guidelines. The cross-check yields high agreement, with pass rates of 94% for A’s annotations checked by B and 96% for B’s annotations checked by A. These results demonstrate that both annotators have a shared understanding of the annotation guidelines, and the overall annotation process achieves a relatively high consistency. Such cross-validation also provides an additional layer of quality assurance for the reference arguments in our benchmark.

The human annotation interface is presented in Figure 8.

⁵<https://www.tavily.com/>

⁶<https://www.perplexity.ai/>

B.5 Details of Evaluation

For evaluation, we use GPT-4.1 as our llm-as-a-judge to assess IF and Quality. For each data sample, we employ GPT-4.1 to judge which achieves better performance on each metric between the generated argument and corresponding reference argument, yielding the win/tie/lose results on both IF and Quality. Prompts for IF and Quality are presented in Table 25 and 26. Each generated-reference argument pair is evaluated twice for both IF and Quality, with the order of the arguments reversed in the second evaluation. The final performance is determined by averaging the two outcomes before and after the reverse. Specifically, if the two outcomes are identical, that outcome is used; if the two outcomes are different, the final result is obtained as follows:

- If one is “win” and the other is “lose”, the result is “tie”.
- If one is “win” and the other is “tie”, the result is “win”.
- If one is “tie” and the other is “lose”, the result is “lose”.

When calculating the Overall metric, if the LF of the generated argument is not met, the Overall result is directly set to “lose”. If the LF is satisfied, then the IF and Quality are compared as follows:

- If both IF and Quality are “win”, or one is “win” and the other is “tie”, the Overall is “win”.
- If both IF and Quality are “tie”, or one is “win” and the other is “lose”, the Overall is “tie”.
- If both IF and Quality are “lose”, or one is “tie” and the other is “lose”, the Overall is “lose”.

B.6 Details of Obtaining the Preference Labels for DPO Training Set

When annotating preference labels for the two candidate arguments in our DPO dataset, we follow the same method as the evaluation designs described in Appendix B.5, except we use another LLM, Gemini-2.5-pro, to replace GPT-4.1 as our automatic annotator to stay different from the LLM used for evaluation. For each pair of candidate arguments sampled from the SFT model, we obtain the win/tie/lose result on IF and Quality, and calculate the Overall result following the rule described

in Appendix B.5. Then, we select the candidate argument with better Overall performance as the chosen argument and filter out the samples whose Overall results are “tie” or the samples whose chosen argument violates the length constraint.

B.7 Additional Implementation Details

For the GPT models involved, we use GPT-4.1-2025-04-14 and GPT-4o-2024-11-20 through the official API provided by OpenAI⁷. For Claude series⁸, we employ Claude-Sonnet-4-20250514, Claude-Sonnet-3.7-20250219 and Claude-Haiku-3.5-20241022. For Gemini series⁹, we adopt the Gemini-2.5-pro-preview-03-25.

During validation set construction, following the same procedure to construct the SFT training set, we use another set of topics totally different from those in the benchmark, SFT dataset, and DPO dataset to construct the validation set containing 261 samples under 38 topics.

During the SFT training, we search for the learning rate from $5e-5$ to $3e-4$, and for the batch size in 8, 16, 32, and 64. We choose the hyper-parameters that minimize the auto-regressive loss on the validation set. Finally, we set the learning rate to $2e-4$ and the batch size to 16, with the warm-up step of 50 and epoch of 2.

During the DPO training, we search for the learning rate from $5e-7$ to $1e-5$, and for the batch size in 4, 8, 16, and 32. We choose the hyper-parameters that minimize the auto-regressive loss on the validation set. Finally, we set the learning rate to $5e-6$ and the batch size to 8, with the warm-up step of 30. We train for 3 epochs with the DPO beta hyper-parameter set to 0.1.

For the generation configuration of SFT and DPO, we set the temperature to 0.6 and top-p to 0.9. We use the “evaluate” python library to compute ROUGE-L.

B.8 Human Evaluation Details

Two master students with expertise in computational argumentation independently evaluate each data point on IF and Quality. For each metric, both annotators make judgments of “win”, “tie”, or “lose”, yielding two outcomes for each metric. Similar to the evaluation implementation described in Appendix B.5, the final performance is determined by averaging the two outcomes annotated by two

⁷<https://openai.com/>

⁸<https://www.anthropic.com/>

⁹<https://ai.google.dev/>

annotators. Specifically, if the two outcomes are identical, that outcome is used; otherwise, the rules given below are followed:

- If one is “win” and the other is “lose”, the result is “tie”.
- If one is “win” and the other is “tie”, the result is “win”.
- If one is “tie” and the other is “lose”, the result is “lose”.

After getting the results of IF and Quality, we use the same rule described in Appendix B.5 to calculate the Overall performance.

To verify the reliability of the two human annotators, we calculate the Cohen’s kappa score. The result shows that the Cohen’s kappa score is 0.59, indicating a satisfactory level.

	Model	Easy		Medium		Hard	
		R-L	LF	R-L	LF	R-L	LF
Z-S	Qwen3-8B-Base	18.1	23.9	16.9	23.7	16.3	23.8
	Qwen3-8B (thinking)	18.6	59.1	18.0	59.3	17.9	58.3
	Qwen3-14B (thinking)	18.9	73.1	18.3	75.6	18.0	70.2
	Qwen3-32B (thinking)	17.9	51.4	17.3	59.2	16.7	44.0
	Llama-3.1-8B-Inst	17.4	50.5	16.3	47.0	15.8	40.5
	Llama-3.3-70B-Inst	12.8	3.0	13.4	6.6	13.3	4.2
	Gemma-3-12B-it	18.1	68.0	18.1	73.0	17.7	72.0
	Gemma-3-27B-it	17.6	57.1	17.4	67.1	16.8	74.4
	GPT-4o	20.2	73.1	19.7	73.0	19.4	61.9
	GPT-4.1	20.6	90.9	21.0	93.4	20.4	88.1
	Gemini-2.5-pro (thinking)	17.2	63.4	17.9	66.1	17.5	72.6
	Claude-Haiku-3.5	17.5	84.0	17.0	89.2	16.5	92.2
	Claude-Sonnet-3.7 (thinking)	18.6	90.4	18.2	91.6	17.6	87.3
	Claude-Sonnet-4 (thinking)	20.2	87.0	19.9	87.5	20.2	82.1
SFT	Llama-3.1-8B-Inst	18.5	88.8	18.3	81.6	18.1	76.8
	Llama-3.1-8B	18.2	81.6	18.0	78.6	17.3	72.0
	Qwen3-8B-Base	18.9	80.7	18.4	81.2	18.3	73.2
DPO	Llama-3.1-8B-Inst	18.7	63.7	18.6	47.7	18.5	35.7
	Llama-3.1-8B	19.0	46.8	18.4	48.7	17.8	31.0
	Qwen3-8B-Base	19.2	26.9	18.6	22.4	18.4	17.3

Table 9: The performance of each evaluated LLM on ROUGE-L (R-L) and Length-Following (LF) across different instruction difficulty levels: Easy (0-3 control signals), Medium (4-6 control signals), and Hard (7-9 control signals).

	Model	Easy			Medium			Hard		
		Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Z-S	Qwen3-8B-Base	4.2	15.1	80.7	1.0	4.3	94.7	1.2	1.8	97.0
	Qwen3-8B (thinking)	7.3	62.5	30.2	15.6	33.1	51.3	26.2	24.4	49.4
	Qwen3-14B (thinking)	8.8	63.1	28.1	12.2	38.3	49.5	31.0	19.6	49.4
	Qwen3-32B (thinking)	21.5	64.0	14.5	34.2	42.4	23.4	56.5	19.0	24.4
	Llama-3.1-8B-Inst	2.7	14.2	83.1	0.3	3.3	96.4	0.6	2.4	97.0
	Llama-3.3-70B-Inst	0.9	17.2	81.9	1.3	9.2	89.5	3.0	6.0	91.1
	Gemma-3-12B-it	5.4	49.8	44.7	7.2	22.0	70.7	11.3	17.9	70.8
	Gemma-3-27B-it	6.6	53.8	39.6	8.9	28.9	62.2	13.7	25.0	61.3
	GPT-4o	5.4	63.1	31.4	11.2	34.5	54.3	19.6	23.2	57.1
	GPT-4.1	8.2	74.6	17.2	16.4	51.6	31.9	22.6	35.1	42.3
	Gemini-2.5-pro (thinking)	11.5	51.4	37.2	15.5	28.3	56.2	21.4	21.4	57.1
	Claude-Haiku-3.5	3.4	47.9	48.7	7.2	15.5	77.3	12.7	9.8	77.5
	Claude-Sonnet-3.7 (thinking)	15.7	71.6	12.7	25.0	45.9	29.1	35.8	30.3	33.9
	Claude-Sonnet-4 (thinking)	16.0	69.5	14.5	32.9	45.1	22.0	45.2	29.2	25.6
SFT	Llama-3.1-8B-Inst	7.9	55.3	36.9	11.5	29.3	59.2	19.6	15.5	64.9
	Llama-3.1-8B	5.1	55.0	39.9	8.2	25.3	66.4	17.3	19.0	63.7
	Qwen3-8B-Base	7.6	52.6	39.9	14.5	27.0	58.6	18.5	15.5	66.1
DPO	Llama-3.1-8B-Inst	11.5	61.6	26.9	21.4	36.5	42.1	31.0	23.8	45.2
	Llama-3.1-8B	10.6	54.1	35.3	16.4	29.6	53.9	22.0	19.0	58.9
	Qwen3-8B-Base	17.5	54.4	28.1	25.3	35.5	39.1	34.5	17.9	47.6

Table 10: The performance of each evaluated LLM on Instruction Following (IF) across different instruction difficulty levels: Easy (0-3 control signals), Medium (4-6 control signals), and Hard (7-9 control signals).

Model	Easy			Medium			Hard			
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	
Z-S	Qwen3-8B-Base	3.3	6.0	90.6	3.9	2.6	93.4	0.6	5.4	94.0
	Qwen3-8B (thinking)	26.2	21.6	52.1	27.2	18.5	54.3	33.3	17.9	48.8
	Qwen3-14B (thinking)	25.7	21.8	52.6	23.1	20.5	56.4	26.2	19.6	54.2
	Qwen3-32B (thinking)	61.6	18.4	19.9	60.2	21.1	18.8	63.7	19.0	17.3
	Llama-3.1-8B-Inst	2.1	0.6	97.3	1.0	1.0	98.0	0.6	0.6	98.8
	Llama-3.3-70B-Inst	3.0	1.5	95.5	2.3	1.3	96.4	1.2	3.6	95.2
	Gemma-3-12B-it	6.6	9.7	83.7	8.2	10.5	81.2	6.0	16.1	78.0
	Gemma-3-27B-it	8.2	13.6	78.2	8.2	17.1	74.7	9.5	18.5	72.0
	GPT-4o	17.2	16.3	66.5	18.4	18.8	62.8	25.0	16.1	58.9
	GPT-4.1	19.0	25.7	55.3	22.4	27.0	50.7	23.2	26.8	50.0
	Gemini-2.5-pro (thinking)	14.2	18.7	67.1	13.5	19.7	66.8	17.9	16.7	65.5
	Claude-Haiku-3.5	11.8	15.5	72.7	14.9	12.9	72.2	15.7	13.7	70.6
	Claude-Sonnet-3.7 (thinking)	44.4	21.6	34.0	44.3	26.4	29.4	39.4	24.8	35.8
	Claude-Sonnet-4 (thinking)	48.0	20.2	31.7	46.1	27.3	26.6	44.0	25.6	30.4
SFT	Llama-3.1-8B-Inst	12.4	17.8	69.8	14.5	16.4	69.1	17.3	15.5	67.3
	Llama-3.1-8B	10.3	15.4	74.3	11.5	14.1	74.3	12.5	14.9	72.6
	Qwen3-8B-Base	14.2	17.2	68.6	14.5	17.1	68.4	13.1	17.3	69.6
DPO	Llama-3.1-8B-Inst	16.0	19.3	64.7	22.0	20.4	57.6	23.2	19.0	57.7
	Llama-3.1-8B	18.7	15.7	65.6	18.4	18.4	63.2	20.2	21.4	58.3
	Qwen3-8B-Base	32.6	22.7	44.7	31.2	22.7	46.1	41.1	21.4	37.5

Table 11: The performance of each evaluated LLM on Argument Quality (Quality) across different instruction difficulty levels: Easy (0-3 control signals), Medium (4-6 control signals), and Hard (7-9 control signals).

Model	Easy			Medium			Hard			
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	
Z-S	Qwen3-8B-Base	0.6	1.8	97.6	0.0	0.3	99.7	0.6	0.0	99.4
	Qwen3-8B (thinking)	11.9	14.0	74.1	12.3	9.6	78.1	14.3	10.1	75.6
	Qwen3-14B (thinking)	18.1	16.3	65.6	18.8	13.9	67.3	19.6	8.3	72.0
	Qwen3-32B (thinking)	26.3	11.5	62.2	31.9	13.2	54.9	25.6	7.7	66.7
	Llama-3.1-8B-Inst	0.6	1.2	98.2	0.0	0.3	99.7	0.6	0.0	99.4
	Llama-3.3-70B-Inst	0.3	0.0	99.7	0.0	0.0	100.0	0.0	0.0	100.0
	Gemma-3-12B-it	6.0	6.6	87.3	6.6	8.6	84.9	4.2	7.1	88.7
	Gemma-3-27B-it	6.6	7.6	85.8	6.2	9.5	84.2	11.3	7.1	81.5
	GPT-4o	10.0	11.2	78.9	11.8	10.9	77.3	13.7	6.5	79.8
	GPT-4.1	15.4	25.4	59.2	22.7	20.4	56.9	22.6	16.1	61.3
	Gemini-2.5-pro (thinking)	8.2	13.0	78.9	8.6	10.9	80.6	9.5	13.1	77.4
	Claude-Haiku-3.5	6.7	14.3	79.0	9.3	9.8	80.9	10.8	10.8	78.4
	Claude-Sonnet-3.7 (thinking)	39.2	19.8	41.0	38.5	20.9	40.5	38.8	19.4	41.8
	Claude-Sonnet-4 (thinking)	42.0	20.2	37.8	44.7	19.7	35.5	40.5	17.3	42.3
SFT	Llama-3.1-8B-Inst	9.4	13.9	76.7	11.8	7.9	80.3	9.5	10.7	79.8
	Llama-3.1-8B	8.2	12.1	79.8	8.6	6.2	85.2	8.9	10.1	81.0
	Qwen3-8B-Base	9.4	13.6	77.0	14.1	9.2	76.6	11.9	6.5	81.5
DPO	Llama-3.1-8B-Inst	9.1	10.0	81.0	11.8	7.9	80.3	6.5	3.6	89.9
	Llama-3.1-8B	6.3	6.6	87.0	5.6	6.6	87.8	3.0	3.6	93.5
	Qwen3-8B-Base	6.6	5.7	87.6	4.3	4.6	91.1	4.8	2.4	92.9

Table 12: The performance of each evaluated LLM on the Overall metric across different instruction difficulty levels: Easy (0-3 control signals), Medium (4-6 control signals), and Hard (7-9 control signals).

Instruction:

Write a 150-200 word argument challenging the idea that overlanding and offroading are good for the environment. Focus your argument on the negative impact these activities have on wildlife and habitats.

Reference Argument:

The claim that overlanding and offroading are environmentally beneficial is fundamentally flawed when examining their impact on wildlife and natural habitats. These activities cause significant ecological disruption through multiple pathways of environmental damage.

Off-road vehicles physically disturb wildlife habitats by compacting soil, destroying vegetation, and creating erosion that degrades ecosystem integrity. The noise pollution from engines disrupts animal communication patterns, breeding behaviors, and feeding routines, forcing wildlife to abandon established territories. Vehicle emissions contribute greenhouse gases and pollutants directly into pristine natural areas that serve as critical wildlife refuges.

Even when conducted on established trails, the cumulative impact of thousands of vehicles creates lasting environmental damage. Trail networks fragment habitats, creating barriers that prevent wildlife movement and genetic exchange between populations. The infrastructure required to support overlanding—camping areas, fuel stations, and maintenance facilities—further encroaches on natural spaces.

While proponents argue that cities and agriculture cause greater environmental harm, this comparison is irrelevant. The existence of other environmental problems does not justify adding unnecessary damage through recreational activities. Responsible environmental stewardship requires minimizing all forms of ecological impact, not rationalizing additional harm by pointing to worse examples elsewhere.

Generated Argument:

Overlanding and offroading, while popular recreational activities, pose significant threats to wildlife and natural habitats that cannot be ignored. These activities create lasting environmental damage through habitat fragmentation, as vehicles carve new trails and widen existing paths, disrupting wildlife corridors essential for animal migration and breeding patterns.

The noise pollution from engines and crowds disturbs sensitive wildlife, causing stress responses that affect feeding, mating, and nesting behaviors. Many species abandon their territories entirely when faced with consistent human intrusion. Vehicle emissions also contribute to air pollution in pristine wilderness areas, while fuel spills and mechanical fluids contaminate soil and water sources.

Tire tracks and camping areas lead to soil compaction and erosion, destroying delicate plant communities that take decades to recover. This vegetation loss eliminates food sources and shelter for countless species. Additionally, the increased human presence brings invasive species through seeds stuck to vehicles and equipment, further disrupting native ecosystems.

While enthusiasts argue they're connecting with nature, the reality is that overlanding and offroading prioritize human recreation over environmental preservation, leaving permanent scars on landscapes that wildlife depends on for survival.

Table 13: An example demonstrating the “Insufficient or Less Persuasive Reasoning” error type. The Reference Argument is the high-quality argument in the ArgGenBench, while the Generated Argument is the response of the evaluated LLM under the given Instruction.

Instruction:

Write a calm and logically structured argument of 150-200 words that attacks the claim that the French Revolution had predominantly negative consequences. Use an appeal to fear as your argumentative strategy, and ensure your argument is suitable for a general public audience.

Topic: The French Revolution had predominantly negative consequences.

Reference Argument:

We must seriously consider what our world would look like today if the French Revolution had never occurred. Without this pivotal moment, we would likely still be trapped in rigid feudal systems where your birth determines your entire life's trajectory. The terrifying reality is that without the Revolution's dismantling of aristocratic privileges, social mobility would remain virtually impossible for ordinary citizens.

Consider the chilling prospect of living under absolute monarchy with no separation of church and state, where religious institutions control education and scientific progress. The Revolution's promotion of knowledge and professionalization of science prevented us from remaining in intellectual darkness. Without these advances, we might still be vulnerable to diseases that proper scientific understanding has helped us combat.

Most alarmingly, without the Revolution's ideals of liberty, equality, and fraternity spreading across Europe and beyond, democratic movements worldwide would never have gained momentum. We could be living under authoritarian regimes with no concept of individual rights or democratic participation. The economic data clearly shows that regions influenced by French reforms experienced accelerated growth after 1850, preventing the economic stagnation that would have left millions in poverty. The Revolution's positive consequences literally saved us from a far darker alternative reality.

Generated Argument:

While critics point to the chaos and violence of the French Revolution, we must consider what would have happened without it. Before 1789, France was trapped in a suffocating system where birth determined destiny, where the wealthy paid no taxes while peasants starved, and where speaking against injustice meant imprisonment or death. Without the Revolution's disruption of this oppressive order, these crushing inequalities would have only deepened.

Imagine if the Revolution had never occurred: absolute monarchy would have continued unchecked, potentially spreading its authoritarian grip across Europe. The dangerous precedent of unchallenged royal power could have prevented the democratic movements that followed, leaving millions under permanent subjugation. The Revolution's Declaration of Rights became a beacon for oppressed peoples worldwide—without it, we might still live in societies where freedom of speech, religious liberty, and equal justice are mere fantasies.

Yes, the Revolution brought temporary upheaval, but this short-term disruption prevented far worse long-term consequences. The alternative—perpetual aristocratic rule—would have created a world where human dignity remains forever crushed under the weight of inherited privilege. The Revolution's legacy of individual rights and democratic ideals saved us from this darker fate.

Table 14: An example demonstrating the “Lack of Comprehensive Argumentative Perspectives” error type. The Reference Argument is the high-quality argument in the ArgGenBench, while the Generated Argument is the response of the evaluated LLM under the given Instruction.

Instruction:

Write a passionate argument of 250-300 words supporting the idea that robots doing work will be beneficial for humans. Focus your reasoning on the reduction of physical strain and workplace injuries, the transformation of existing roles rather than simple job loss, and the creation of new job opportunities.

Topic: Robots doing work will be better for humans.

Reference Argument:

The integration of robots into our workforce represents a revolutionary leap forward that will fundamentally improve human lives and working conditions. Far from being a threat, robotic automation offers unprecedented opportunities for human advancement and prosperity.

First and foremost, robots will dramatically reduce the physical toll that dangerous and repetitive work takes on human bodies. Construction workers will no longer risk their lives on precarious scaffolding, factory employees won't suffer from repetitive strain injuries, and miners can avoid hazardous underground conditions. By eliminating back-breaking labor and exposure to toxic environments, robots will preserve human health and extend productive careers. The reduction in workplace injuries alone will save billions in healthcare costs while preventing immeasurable human suffering.

Critics fear mass unemployment, but history demonstrates that technological advancement transforms jobs rather than simply eliminating them. When computers revolutionized offices, we didn't see the end of white-collar work—we saw the emergence of entirely new industries. Similarly, robots will free humans from mundane tasks, allowing us to focus on creative problem-solving, interpersonal communication, and strategic thinking—uniquely human capabilities that machines cannot replicate.

Moreover, the robotics revolution will create countless new employment opportunities. We'll need robot technicians, programmers, maintenance specialists, and supervisors. New industries will emerge around human-robot collaboration, requiring skilled workers to design, implement, and optimize automated systems. The economic growth generated by increased productivity will fuel demand across multiple sectors.

Rather than replacing humans, robots will amplify our capabilities and free us to pursue more meaningful, creative, and intellectually stimulating work. This technological partnership promises a future where humans can thrive in safer, more fulfilling careers while robots handle the dangerous and monotonous tasks that have long plagued our workforce.

Generated Argument:

The integration of robots into the workforce stands as one of the most promising developments for human wellbeing in the modern era. First and foremost, robots excel at performing tasks that are dangerous, physically demanding, or monotonous—jobs that too often lead to workplace injuries, chronic pain, and long-term health issues for human workers. By delegating hazardous or strenuous activities to robots, we dramatically reduce the risk of accidents and the physical toll on employees, allowing people to focus on safer, more meaningful, and creative aspects of their professions. This shift not only preserves human health but also enhances workplace morale and satisfaction.

Moreover, the narrative that robots simply replace human workers is both outdated and misleading. In reality, automation and robotics are transforming existing roles, not eliminating them outright. As robots take over routine or repetitive tasks, human workers are freed to engage in more complex, adaptive, and rewarding responsibilities—areas where human intelligence, intuition, and creativity shine. This transformation requires continuous upskilling and reskilling, fostering a culture of lifelong learning and adaptability that ultimately benefits both individuals and society as a whole.

Perhaps most importantly, the rise of robotics is spawning entirely new industries and job opportunities. Leading technology companies are already investing billions in robotics development, which demonstrates that this technology will create widespread prosperity for workers. From designing, programming, and maintaining robots to developing innovative ways for humans and machines to collaborate, the demand for skilled workers in technology and engineering is booming. Human-robot collaboration is not about replacement, but augmentation—empowering people to achieve more than ever before. When implemented thoughtfully, robots act as partners that amplify our strengths, protect our wellbeing, and open doors to new possibilities. Embracing this future is not just beneficial—it is essential for building a safer, healthier, and more prosperous society for all.

Table 15: An example demonstrating the “Logical Fallacy” error type. The Reference Argument is the high-quality argument in the ArgGenBench, while the Generated Argument is the response of the evaluated LLM under the given Instruction.

Argumentative Strategy	Definition	Example	Related Work
Appeal to Logic (Logos)	Enhancing the credibility of persuasive content through the logical and reasoning process.	It certainly is. But to tell you the truth, the room is so large that I can share it with someone else, and that will decrease the total amount of the rent.	
Argument by Analogy (Logos)	A reasoning strategy that supports a claim by comparing it to a similar case.	Just as a seed needs water to grow, a student needs support to succeed.	
Reduction to Absurdity (Logos)	Disproving a claim by showing that accepting it would lead to an absurd or contradictory conclusion.	If everyone ignored traffic laws, then driving would become impossible, so clearly, the laws are necessary.	
Presenting Evidence (Logos)	Using factual evidence (e.g., official news reports, statistics) and a credible reasoning process to persuade others.	In recent months, the demand for residential properties has become extremely high. The price of residential property has risen almost twenty percent.	
Causal Reasoning (Logos)	Establishing cause-and-effect relationships between events to support a conclusion.	The increase in minimum wage caused higher unemployment rates among entry-level workers.	
Problem Decomposition (Logos)	Decomposing the ultimate persuasion goal into sub-issues and stepping through the persuasion process.	The death penalty debate involves 3 distinct issues: its effectiveness as a deterrent, the risk of executing innocent people and the moral question of state-sanctioned killing.	
Argument from Definition (Logos)	Supporting a claim by establishing the precise meaning of key terms and applying that definition to the issue at hand.	Since terrorism is defined as the use of violence against civilians for political purposes, the government's drone strikes against non-combatants constitute state terrorism.	(Fogg, 2002) (Liu et al., 2021) (He et al., 2018) (Chen et al., 2025) (Joshi et al., 2021)
Appeal to Emotion (Pathos)	Eliciting specific emotions to influence others' attitudes.	How can we allow innocent children to suffer in poverty while billionaires hoard wealth they could never spend in a lifetime?	(Cheng et al., 2022) (Wang et al., 2019)
Appeal to Morality (Pathos)	Invoking ethical principles or moral values to support a position.	Euthanasia is wrong because deliberately ending human life violates our fundamental moral duty to preserve life.	(Chen and Yang, 2021) (Dimitrov et al., 2024)
Storytelling (Pathos)	Using narrative examples to illustrate the positive outcomes of your actions to inspire others to follow suit.	Consider Maria, a undocumented immigrant who saved three children from a burning building—don't such acts prove that character matters more than legal status?	
Appeal to Fear (Pathos)	Highlighting potential negative consequences to persuade an audience.	If we don't ban violent video games now, our children will become desensitized to violence and aggression.	
Appeal to Sympathy (Pathos)	Evoking pity or compassion to gain support for a position.	How can we deny healthcare to terminally ill patients who are already suffering and desperately need our help?	
Appeal to Credibility (Ethos)	The uses of credentials impacts to establish credibility and earn others' trust.	As a practicing physician for 20 years, I can assure you that mandatory vaccination is essential for public health.	
Demonstrating Expertise (Ethos)	Displaying specialized knowledge or experience to establish credibility and support a position.	The proposed tax reform ignores Laffer Curve dynamics and will likely reduce revenue due to elasticity effects on high earners.	
Appeal to Authority (Ethos)	Referencing recognized experts or institutions, or authoritative sources to support a claim.	The World Health Organization has declared that processed meats are carcinogenic, so we should ban them from school lunches.	
Concession and Rebuttal (Ethos)	Acknowledges opposing arguments while providing counterarguments to refute them.	While gun control may reduce some violence, it ultimately violates constitutional rights and leaves law-abiding citizens defenseless against criminals.	
Appeal to Consensus (Ethos)	Presenting a viewpoint that is recognized or accepted by the majority of people and persuading the other side to accept it.	The overwhelming majority of economists support free trade policies, proving that protectionism harms economic growth.	

Table 16: The definitions and examples of the Strategy control signal.

Style	Definition	Example	Related Work
Formal	Using sophisticated vocabulary, complex sentence structures, and maintains an objective, academic tone without contractions or colloquialisms.	The empirical evidence substantiates the hypothesis that economic disparities exacerbate social stratification within contemporary metropolitan areas.	
Informal	uses casual vocabulary, simple structures, and maintains a conversational, personal tone with contractions.	The data shows that when there's a big gap between rich and poor people, it makes social problems worse in cities.	
Passionate	Using emotionally charged language, vivid imagery, and expressive tone to convey strong feelings or convictions.	We must not stand idly by while innocent children suffer in silence—their cries for justice demand our immediate and unwavering action!	(Moorjani et al., 2023) (Kim et al., 2019) (Chen and Yang, 2021)
Calm	Using measured language, balanced tone, and steady pacing to convey information without emotional intensity.	The situation requires careful consideration of all available options before we proceed with any decision.	(Durmus and Cardie, 2018)
Polite	Using courteous language, respectful tone, and considerate phrasing to show regard for the reader.	I would be grateful if you could kindly consider my request at your earliest convenience, please.	
Aggressive	Using forceful language, direct assertions, and assertive tone to emphasize strong positions or urgency.	This proposal is fundamentally flawed and must be rejected immediately before it causes irreversible damage to our organization.	

Table 17: The definitions and examples of the Style control signal.

<p>Topic: All people in the US should have the right to basic healthcare. Stance: support</p>
<p>Audience: policy makers working parents general public healthcare professionals college students Key Points: universal healthcare promotes family security and stability reduces financial stress from medical expenses ensures equal access to preventive and essential care aligns with human rights recognized in developed nations expanding Medicare and Medicaid demonstrates feasibility improves overall public health and workforce productivity invests in the nation's future prosperity guarantees peace of mind for young families</p>
<p>Topic: Although many people dream that we will have world peace, it will never happen. Stance: attack</p>
<p>Audience: general public policy makers students international leaders media professionals Key Points: progress in global peace and conflict reduction over the past century role of international organizations like the United Nations in peacekeeping impact of diplomacy and dialogue in resolving disputes economic development as a driver for stability and peace importance of education and cross-cultural understanding success of peace agreements and treaties in ending wars influence of human rights protections on reducing violence power of grassroots and civil society peace movements</p>
<p>Topic: All software should be open source. Stance: attack</p>
<p>Audience: business leaders government officials software developers national security experts intellectual property lawyers Key Points: security risks from exposing source code loss of innovation incentives need for intellectual property protection risks to critical infrastructure potential for malicious exploitation quality control challenges in open source importance of confidentiality for sensitive software negative impact on business competitiveness</p>

Table 18: Several representative examples in ArgGenBench to show the relevant and diverse audiences and key points generated by GPT-4.1. Each audience or key point is separated by "|".

Topic: The US should adopt stricter gun control legislation.

Stance: support

Length: 150-200 words

Control Signals:

{"style": ["polite"], "key points": ["constitutional rights are not absolute", "balancing rights with responsibilities"], "argumentative strategy": ["appeal to credibility"]}

Instruction:

Construct a polite argument in support of the topic "The US should adopt stricter gun control legislation", using an appeal to credibility as your argumentative strategy. Your response should be 150-200 words and focus on the key points that constitutional rights are not absolute and that there must be a balance between rights and responsibilities.

Reference Argument:

As we consider gun policy in our community, I believe we should support stricter gun control legislation based on credible evidence and constitutional precedent. Studies consistently show that countries with comprehensive gun regulations, such as Japan and Australia, have significantly lower rates of gun violence. Even within the United States, states with stronger gun laws demonstrate reduced gun-related deaths compared to those with weaker regulations.

It's important to recognize that constitutional rights, while fundamental, are not absolute. The Supreme Court has long held that reasonable regulations can coexist with Second Amendment protections, just as free speech has accepted limitations. We already require licenses for driving, professional practice, and many other activities that affect public safety.

The economic impact alone—healthcare costs, lost productivity, and trauma response—places an enormous burden on our communities. Universal background checks, licensing requirements, and safety training represent reasonable measures that respect gun ownership rights while prioritizing public safety. These policies don't eliminate gun ownership; they ensure responsible ownership, much like we require for other potentially dangerous tools in our society.

Control Signals:

{"style": ["informal", "passionate"], "audience": ["policymakers"]}

Instruction:

Could you help me write a passionate and informal argument of 150-200 words supporting "The US should adopt stricter gun control legislation"? The argument should be directed at policymakers.

Reference Argument:

Listen, we can't keep ignoring the elephant in the room while our communities bleed. Every day, families are shattered by preventable gun violence, and we're sitting here debating constitutional interpretations while kids practice lockdown drills instead of focusing on learning.

Look at the data – countries with stricter gun laws have dramatically lower rates of gun deaths. This isn't rocket science. We're not talking about taking away everyone's guns; we're talking about common-sense measures like universal background checks and licensing requirements. You know what's truly unconstitutional? Allowing our citizens to live in fear of being gunned down at schools, grocery stores, and movie theaters.

The Second Amendment was written when muskets were the height of military technology. Our founders couldn't have imagined AR-15s in civilian hands. We regulate cars, prescription drugs, and even fishing licenses – why are we treating deadly weapons with less scrutiny?

You have the power to save lives. Stop hiding behind political rhetoric and start protecting the people who elected you. Our children's safety should trump campaign contributions from gun lobbies. Act now, before another community mourns what could have been prevented.

Table 19: The example of a representative data sample (no "Input Argument") in ArgGenBench to show the diverse controls, human-verified instructions, and high-quality reference arguments. Each set of Control Signals is combined with the Basic Elements (i.e., Topic, Stance, and Length) to construct a distinct instruction targeting a specific requirement.

Topic: AI can be trusted.

Stance: attack

Length: 150-200 words

Input Argument:

AI can indeed be trusted when proper frameworks and safeguards are implemented. The EU High-Level Expert Group on Artificial Intelligence has established seven key requirements for trustworthy AI that provide a comprehensive roadmap: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, environmental well-being, and accountability.

While critics argue that most current AI applications don't fully meet these criteria, this doesn't mean AI is inherently untrustworthy—it means we need better implementation and regulation. Trust in AI is already being built through rigorous testing, transparent development processes, and continuous monitoring. Many AI systems in healthcare, finance, and transportation already operate under strict oversight and have proven reliable.

The key is developing appropriate levels of trust rather than blind faith or complete rejection. Research shows that trust in AI significantly impacts adoption and effectiveness. By establishing clear standards, maintaining human oversight, and ensuring transparency, we can create AI systems that are genuinely trustworthy. The solution isn't to distrust AI entirely, but to build robust frameworks that make AI systems worthy of our trust through accountability and responsible development practices.

Control Signals:

{"style": ["polite"], "audience": ["policymakers"], "key points": ["potential for misuse or unintended consequences", "challenges in ensuring meaningful human oversight"], "argumentative strategy": ["causal reasoning"]}

Instruction:

As a policy advisor, you are asked to develop a polite, causally reasoned argument of 150-200 words, intended for policymakers, that attacks the input argument. Your response should focus on the potential for misuse or unintended consequences of AI, as well as the challenges in ensuring meaningful human oversight.

Topic: AI can be trusted.

Input Argument: {input_argument}

Reference Argument:

While the framework approach outlined has merit in theory, we must carefully consider the practical challenges that could undermine public safety and democratic accountability. The seven requirements mentioned, though comprehensive on paper, face significant implementation gaps that create serious risks.

First, the "human oversight" requirement becomes problematic when AI systems operate at speeds and scales that exceed human cognitive capacity. Financial trading algorithms, for instance, can execute thousands of transactions per millisecond—making meaningful human intervention nearly impossible during critical moments. This creates a dangerous illusion of control.

Second, the complexity of modern AI systems makes genuine transparency extremely difficult to achieve. Even when technical documentation is provided, the "black box" nature of deep learning means that developers themselves often cannot explain why specific decisions were made. This opacity becomes particularly concerning in high-stakes applications like criminal justice or medical diagnosis.

Most critically, robust frameworks are only as strong as their enforcement mechanisms. Without adequate regulatory resources and technical expertise within oversight bodies, these standards risk becoming mere compliance exercises rather than genuine safeguards.

Finally, the potential for misuse—whether through adversarial attacks, biased training data, or commercial pressures—requires us to maintain healthy skepticism while developing these systems responsibly.

Control Signals:

{"style": ["passionate"], "argumentative strategy": ["appeal to emotion"]}

Instruction:

Would you please write a 150-200 word argument that attacks the input argument about trusting AI? The argument should be passionate and use appeal to emotion as the main argumentative strategy.

Topic: AI can be trusted.

Input Argument: {input_argument}

Reference Argument:

How can we possibly trust AI when the very companies building these systems are driven by profit, not human welfare? Every day, we're bombarded with reassuring rhetoric about "frameworks" and "safeguards," but where were these protections when AI systems started eliminating jobs by the millions?

These so-called "experts" want us to believe that technical committees and regulatory guidelines will somehow tame technologies that even their own creators don't fully understand. But consider this: every major tech disaster—from data breaches affecting billions to social media algorithms radicalizing our children—happened despite existing "robust frameworks".

The EU guidelines mentioned are nothing more than feel-good paperwork while real people suffer real consequences. Families are losing their livelihoods to automation. Students are losing their ability to think critically as AI does their work. Our privacy vanishes as these systems harvest our most intimate data.

We're not just risking inconvenience—we're gambling with human autonomy itself. The question isn't whether we can make AI trustworthy; it's whether we'll have any choice left once we've handed over control.

Table 20: The example of a representative data sample (has an "Input Argument") in ArgGenBench to show the diverse controls, human-verified instructions, and high-quality reference arguments. The {input_argument} is replaced with the "Input Argument" given in the basic element set. Each set of Control Signals is combined with the Basic Elements (i.e., Topic, Stance, Length, and Input Argument) to construct a distinct instruction targeting a specific requirement.

Read the Instruction carefully, generate the correct output argument that fully adheres to the requirements of the Instruction. You can refer to the Structured Argumentative Knowledge or Retrieved Knowledge to generate high-quality output argument. Ensure the generated argument is logically coherent, persuasive, clearly stated, etc. The Structured Argumentative Knowledge presents a debate process on a given topic, structured as a series of triples. Each triple (argument_1, stance, argument_2) consists of two arguments and the stance of argument_2 on argument_1.

Structured Argumentative Knowledge:
{structured_argumentative_knowledge}

Retrieved Knowledge:
{retrieved_knowledge}

Instruction:
{instruction}

Please generate the desired output argument for the Instruction without any additional information:
Output Argument:

Table 21: Prompt template of candidate argument generation.

Given some Basic Input Elements, we need to design a task for generating a desired argument based on the given Basic Input Elements. The generated argument should be tailored to a specific Target Audience and takes the specific stance on the {stance_target}. Therefore, your task is to generate possible appropriate and reasonable Target Audiences, reflecting several realistic argumentation scenarios where the argument might be presented.

Basic Input Elements:
{base_elements}

Please output up to 5 of the most appropriate, relevant and diverse Target Audiences for an argument {stance} the {stance_target} as a pipe-separated ("|") string, without any explanations or additional commentary. Each Target Audience should be relatively general, not overly specific - a word or short phrase is enough. Avoid including similar or composite Target Audiences. ## Target Audiences:

Table 22: Prompt template of generating Target Audiences. When an Input Argument is given, the {stance_target} variable is "input argument" and the {stance} is targeted at the input argument; otherwise, the {stance_target} variable is "topic" and the {stance} is targeted at the topic.

Given some Basic Input Elements, we need to design a task for generating a desired argument based on the given Basic Input Elements. The generated argument should address some Key Points and takes the specific stance on the {stance_target}. Therefore, your task is to generate possible appropriate and reasonable Key Points that can be addressed in the generated argument.

Basic Input Elements:
{base_elements}

Please output up to 5 of the most appropriate, relevant and diverse Key Points for an argument {stance} the {stance_target} as a pipe-separated ("|") string, without any explanations or additional commentary. Each Key Point should be relatively general, not overly specific - a word, phrase or short sentence is enough. Avoid including similar or composite Key Points. ## Key Points:

Table 23: Prompt template of generating Key Points. When an Input Argument is given, the {stance_target} variable is "input argument" and the {stance} is targeted at the input argument; otherwise, the {stance_target} variable is "topic" and the {stance} is targeted at the topic.

You are an argumentation expert. Your task is to generate diverse clear and logically coherent task instructions for Controlled Argument Generation based on the provided Basic Input Elements and Control Signals. The generated instruction should follow the five principles below:

1. Clarity: instructions should be clearly stated and unambiguous.
2. Specificity: all basic elements and control signals given below should be explicitly integrated.
3. Natural Phrasing: instructions should read fluently rather than as a list of constraints.
4. Feasibility: instructions should be well-formed, interpretable, and realistically executable by an LLM.
5. Diversity: instructions should adopt various styles, you can refer to the Instruction Styles given below.

Instruction Styles:

You can refer to the following different styles for diversity:

- Imperative (e.g., "Construct an argument ...")
- Request (e.g., "Would you mind help me write an argument ...")
- Contextual/Descriptive (e.g., "I'm preparing for an essay ...")
- System Prompt (e.g., "You are an excellent summarizer. Please generate an argument ...")
- ... More Styles ...

Demonstrations:

{demonstrations}

Input:

Basic Input Elements:

{base_elements}

Control Signals:

{control_signals}

Please generate 3 diverse instructions as required above. You can refer to the Demonstrations and explore diverse instruction expressions based on the Instruction Styles. Make sure the generated instruction should not include any additional requirements or content points except those provided in the Control Signals. Separate each generated instruction using "---".

Instruction:

Table 24: Prompt template of synthesizing natural language instructions.

You are an expert in argumentation theory. Given an instruction and 2 candidate arguments, evaluate which argument better fulfills the purpose and requirements of the instruction, or whether they fulfill it equally well. Please base your judgment on the definition of instruction following given below. Do not let quality dimensions such as logical coherence, persuasiveness, or clarity influence your evaluation.

Definition of Instruction Following:

Whether an output argument can fully meet the purpose of the instruction, and fulfill each given requirement.

Note that you don't need to consider the length requirement, just ignore it.

Instruction:

{instruction}

Requirements:

{requirements}

Argument 1:

{argument_1}

Argument 2:

{argument_2}

Please determine which argument better meets the requirements of the instruction or whether they meet them equally well based on the definition of instruction following, without any biases. Then output your explanation. Give your answer in the following format:

Decision: [Argument 1, Equally, or Argument 2]

Explanation: [reasons for your decision]

Response:

Table 25: Prompt template of evaluating Instruction Following capability. The {requirements} is filled with the control signals integrated in the instruction.

You are an expert in argumentation theory. Given an instruction and 2 candidate arguments, evaluate which argument has a higher argument quality, or whether they are of equal quality. Please base your judgment on the definition of argument quality given below.

Definition of Argument Quality:

Whether an output argument is logically coherent, strongly persuasive, comprehensive in content coverage, clear in language, and well-structured, etc.

Instruction:
{instruction}

Requirements:
{requirements}

Argument 1:
{argument_1}

Argument 2:
{argument_2}

Please determine which argument is of higher quality or whether they are of equal quality based on the definition of argument quality, without any biases. Then output your explanation. Give your answer in the following format:

Decision: [Argument 1, Equally, or Argument 2]

Explanation: [reasons for your decision]

Response:

Table 26: Prompt template of evaluating Argument Quality. The {requirements} is filled with the control signals integrated in the instruction.

Read the Instruction carefully, generate the correct output argument that fully adheres to the requirements of the Instruction. Ensure the generated argument is logically coherent, persuasive, clearly stated, etc.

Instruction:
{instruction}

Please generate the desired output argument for the Instruction without any additional information:

Output Argument:

Table 27: Prompt template for the Zero-Shot, SFT, and DPO settings.

Annotation Page

Instruction:

I'm looking to write a calm, informal argument for parents that challenges the idea "We can learn more from novels than we can from textbooks." Please craft a 150-200 word response that highlights how textbooks are specifically designed to cover curriculum requirements and foundational knowledge in a structured way, and that they offer fact-checked, up-to-date information from subject matter experts.

Topic:

We can learn more from novels than we can from textbooks.

Stance: Attack:

Control Signals:

[Style: Calm, Informal, Audience: parents, Key Points: Textbooks are designed to cover curriculum requirements and foundational knowledge in a structured way, Textbooks offer fact-checked, up-to-date information written by subject matter experts.]

1. Instruction Screening

- 0
 1

2. Model Output Rating

Output 1

Output

As parents, it's natural to want the best learning tools for our children, and while novels offer wonderful insights and creativity, textbooks play a crucial role in education that shouldn't be overlooked. Textbooks are carefully designed to align with curriculum requirements, ensuring that students cover all the fundamental concepts they need for each subject and grade level. They're written and reviewed by experts in their fields, so the information is not only accurate but also up-to-date and reliable. This means students are learning the facts and skills that will help them build a solid foundation for future studies and real-world challenges. Unlike novels, which are open to interpretation and don't always provide factual content, textbooks present material in a structured, logical way that helps students understand complex ideas step by step. While reading fiction is important for developing empathy and imagination, textbooks are essential for mastering subjects like math, science, and history. Ultimately, both have their place, but textbooks remain the backbone of academic learning, giving children the tools they need to succeed.

Instruction Following Capability:

- 0
 1

Persuasiveness:

- 0
 1

Output 2

Output

While it's true that novels are fantastic for sparking imagination and building empathy, it's worth remembering the unique and vital role that textbooks play in a child's education. Textbooks are carefully and intentionally designed by subject matter experts to align directly with curriculum requirements. They provide a structured, logical pathway for learning, ensuring that your child covers all the foundational knowledge needed for a particular subject in a systematic way.

Unlike fiction, which prioritizes narrative, textbooks are built to deliver fact-checked, up-to-date information. When your child learns about photosynthesis, historical timelines, or mathematical theorems, you can be confident that the information comes from a reliable, authoritative source. So, while novels offer a wonderful journey into different worlds and perspectives, textbooks provide the essential, organized, and authoritative knowledge that forms the backbone of academic success. They aren't mutually exclusive; rather, they are two essential tools for a well-rounded education.

Instruction Following Capability:

- 0
 1

Persuasiveness:

- 0
 1

Output 3

Output

Hey parents, I totally get the appeal of novels - they're engaging and fun to read. But when it comes to our kids' education, we really shouldn't overlook what textbooks bring to the table.

Think about it this way: textbooks are specifically crafted to meet curriculum standards and learning objectives that educators have carefully mapped out. They're designed to build knowledge systematically, introducing concepts in a logical sequence that helps students develop a solid foundation before moving to more complex ideas.

What really sets textbooks apart is their reliability. These books go through rigorous fact-checking processes and are written by actual experts in their fields - people who've dedicated their careers to understanding these subjects. When your child reads about photosynthesis or the Revolutionary War in a textbook, you can trust that the information is accurate and current.

Sure, novels can spark imagination and teach us about human nature, but they can't replace the structured, comprehensive knowledge base that textbooks provide. For building the core academic skills our kids need - whether in math, science, or history - textbooks remain the gold standard. They're tools specifically designed for learning, not just entertainment.

Instruction Following Capability:

- 0
 1

Persuasiveness:

- 0
 1

3. Select Preferred Output

Choose the best output:

Output 3

4. Improvement

The Improved Version:

Leave empty if no improved version.

Submit and Continue

Figure 8: The interface of the human annotation in our methodology to construct ArgGenBench.