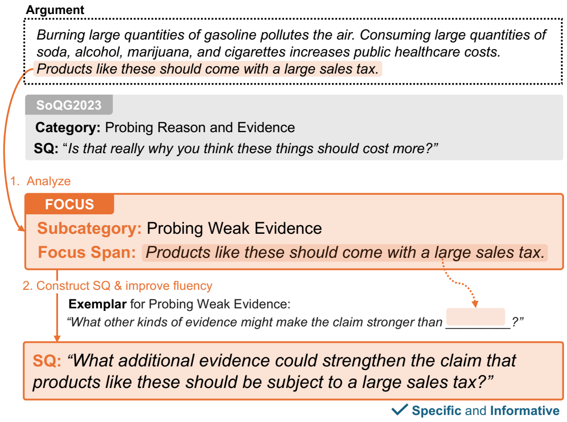
FOCUS: A Benchmark for Targeted Socratic Question Generation via Source-Span Grounding

Surawat Pothong[1] Machi Shimmei[3,4] Naoya Inoue[1,3] Paul Reisert[2] Ana Brassard[3]
Wenzhi Wang[4,3] Shoichi Naito[3,5] Jungmin Choi[3] Kentaro Inui[6,4,3]
[1]JAIST [2]Beyond Reason [3]RIKEN
[4]Tohoku University [5]Ricoh Company, Ltd. [6]MBZUAI
{spothong, naoya-i}@jaist.ac.jp, beyond.reason.sp@gmail.com, machi.shimmei.e6@tohoku.ac.jp
ana.brassard@riken.jp, wang.wenzhi.r7@dc.tohoku.ac.jp, shohichi.naitoh@jp.ricoh.com
jungmin.choi@riken.jp, kentaro.inui@mbzuai.ac.ae

Abstract

We present FOCUS, a benchmark and task setting for Socratic question generation that delivers more informative and targeted feedback to learners. Unlike prior datasets, which rely on broad typologies and lack grounding in the source text, FOCUS introduces a new formulation: each Socratic question is paired with a fine-grained, 11-type typology and an explicit source span from the argument it targets. This design supports clearer, more actionable feedback and facilitates interpretable model evaluation. FOCUS includes 440 annotated instances with moderate partial-match agreement, establishing it as a reliable benchmark. Baseline experiments with representative state-of-the-art models reveal, through detailed error analysis, that even strong models struggle with span selection and context-sensitive categories. An extension study on the LogicClimate dataset further confirms the generalizability of the task and annotation framework. FOCUS sets a new standard for pedagogically grounded and informative Socratic question generation.



Argument
Burning large quantities of gasoline pollutes the air. Consuming large quantities of soda, alcohol, marijuana, and cigarettes increases public healthcare costs. Products like these should come with a large sales tax.
SoQG2023
Category: Probing Reason and Evidence
SQ: "Is that really why you think these things should cost more?"
1. Analyze
FOCUS
Subcategory: Probing Weak Evidence
Focus Span: Products like these should come with a large sales tax.
2. Construct SQ & improve fluency
Exemplar for Probing Weak Evidence:
"What other kinds of evidence might make the claim stronger than ________?"
SQ: "What additional evidence could strengthen the claim that products like these should be subject to a large sales tax?"
✓ Specific and Informative

Figure 1: Overview of the **FOCUS** task and dataset. FOCUS extends SoQG-2023 with a fine-grained typology of **argumentative weaknesses**, with each instance manually annotated with both the weakness type and its corresponding **source span** within the argument. Given an argument, the model is tasked with classifying the weakness, extracting the corresponding span, and inserting it into a predefined exemplar to generate a Socratic question.

1 Introduction

Socratic Questioning (SQ) is a structured method of inquiry that guides students' thinking to reduce cognitive biases and foster critical thinking through self-reflection (Guerraoui et al., 2023; Paul and Binker, 1990). By prompting deeper reflection and clearer explanations, it helps learners identify and close gaps in their reasoning. Widely embraced in education (King, 1994; Azzopardi, 2021), automating SQ offers a promising path for scalable, instructional support in the classroom (Paul and Elder, 2007; Chew et al., 2019). Building on the previous work on SoQG-2023 (Ang et al., 2023), which focuses on monologic settings and the generation and categorization of Socratic questions, our work concentrates on developing automatic SoQG-2023 resources and interpreting the underlying intent behind each question. We believe that generating targeted SQs requires a structured component analysis of arguments. By systematically examining the process of crafting Socratic questions for students, we aim to semantically identify the core issues and weaknesses in arguments, localize them through span annotations, and generate targeted Socratic questions accordingly. An additional benefit of using spans is that educational research has also shown that highlighting the text span that prompts a question greatly improves learners' understanding and performance (Ho et al., 2023; Naito et al., 2022). Therefore, we propose FOCUS, a novel task and benchmark dataset in which each question is linked to a *source span* and an *expanded 11-type typology.* This richer annotation can better guide the students' revision with clear, actionable feedback.

Specifically, the FOCUS dataset builds on SoQG-2023 by introducing (i) a fine-grained yet easy-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2938–2958
December 20-24, 2025 ©2025 Association for Computational Linguistics

to-annotate typology inspired by critical-question and Socratic frameworks (Focus Socratic Question Types (FSQType)), and (ii) source-span annotations that explicitly indicate the section of the argument it addresses. The spans are then placed into predefined exemplar templates and refined by a large language model (LLM) to generate fluent, focused Socratic questions (Calvo Figueras and Agerri, 2024). For example, in Figure 1, the original question from the SoQG-2023 dataset (*"Is that really why you think these things should cost more?"*) is labeled as "Probing Reason and Evidence" without mention of which section the question intends to address. In contrast, FOCUS specifies that it concerns "Probing Weak Evidence" and concretely asks *"What additional evidence could strengthen the claim that products like these should be subject to a large sales tax?"* This question avoids co-reference ambiguity, and both the fine-grained label and source spans narrow the target of revision while leaving room to explore ways to address the gap in reasoning. The dataset, containing 440 annotated instances, was developed by three expert annotators over multiple rounds, resulting in consistent and high-quality annotations. As for the FOCUS task, models must correctly identify the type of weakness (FSQType) and the associated span. This setup makes model evaluation more feasible compared to evaluating questions generated in an open-ended manner, as well as introducing a degree of transparency on why a Socratic question was generated.

We report key findings from our annotation study and baseline experiments. Inter-annotator agreement for the FSQType labels, measured by PABAK (Byrt et al., 1993), ranged from 0.63 to 1.00, while unigram-based partial match scores for the spans ranged from 0.52 to 1.00, both indicating moderate to high consistency. These results suggest that the task and typology are well-defined and suitable for reliable annotation. We also examine the FSQType that could not be instantiated and discuss the challenges encountered during the annotation process. On the modeling side, we conducted baseline experiments using state-of-the-art models, including GPT-4 (OpenAI, 2023), Mistral (Jiang et al., 2024), LLaMA2 (Touvron et al., 2023), and LLaMA3 (Dubey et al., 2024). Results show that span selection remains challenging, with Mistral achieving a BERTScore between 0.34 and 0.68. Additionally, we demonstrate the applicability of our framework on the LogicClimate dataset (Jin et al., 2022), achieving inter-annotator agreement ranging from 0.505 to 1.00. These results suggest that our approach can extend beyond the initial dataset, with strong potential for broader applications in education and argumentation mining.

In summary, we make four key contributions:

1. A **novel fine-grained schema** for SQ grounded in theoretical foundations;

2. A **new dataset** of 440 annotated instances extending SoQG-2023, with significant inter-annotator agreement and partial match similarity;

3. A **novel Socratic Question Generation (SQG) task** that links each question to its source span and FSQ type, supported by baseline experiments; and

4. **Empirical validation of generalizability**, demonstrating that our typology applies to other domains with substantial agreement and consistent task performance.

Overall, our findings highlight significant room for improvement and pose an open challenge for advancing explainable SQG. FOCUS is publicly available on GitHub.[1]

2 Educational Theories in Socratic Feedback

Socratic Questions (SQs) and Critical Questions (CQs) offer two foundational approaches to probing reasoning. CQs, rooted in argumentation schemes (Walton, 2008), enable formal evaluation but are often too rigid for educational settings. For example, the question from (Calvo Figueras and Agerri, 2024), classified as a CQs ("What evidence is there that implicit bias is a problem for everyone, not just police?"), contrasts with SQs (Paul and Elder, 2019), as illustrated in SoQG-2023 (Figure 1), which encourage open-ended, reflective inquiry—making them well-suited for fostering metacognition and critical thinking in learning environments (Al-Hossami et al., 2023; Kim et al., 2023). To align SQ with educational practice, FOCUS draws from learning science theories such as feedback intervention and metacognitive scaffolding (Nicol and Macfarlane-Dick, 2006; Hattie

[1] https://github.com/FOCUSSocratic2025/focus-socratic-question

and Timperley, 2007). By anchoring each question to a specific span in the argument, FOCUS provides interpretable, actionable feedback. Unlike prior SQG datasets like QED (Lamm et al., 2021), which focus on formal logic, FOCUS targets reasoning weaknesses in a pedagogically grounded way. It also contrasts with span-selection tasks in QA and explanation generation, which lack reflective typologies. Recent SQG research has emphasized structured annotations to improve question quality. SoQG-2023 (Ang et al., 2023) introduced five question types, while Kumar and Lan (2024) added invalid examples to fine-tune LLaMA 2 for debugging tasks. Shridhar et al. (2022) explored subquestion generation for multi-hop QA. Though effective, these works overlook argument-focused, feedback-oriented supervision—a gap FOCUS addresses through span-typology alignment (Gu et al., 2021; Yang et al., 2018).

3 Annotation Study

FOCUS extends previous work with a set of 440 instances hand-annotated with fine-grained typology labels and source spans. Here we outline the typology creation, describe the annotation process, and report analysis results.

3.1 Source Data

To build on previous work, we utilize the SoQG-2023 dataset by (Ang et al., 2023), which consists of 110,000 instances. Each instance includes an argument, a Socratic question, and a label indicating the Socratic question type (SQ type). The arguments are collected from Reddit's Change My View subreddit,[2] which features discussions on a wide range of topics. SoQG-2023 categorizes questions into five SQ types, based on the framework proposed by (Paul and Elder, 2019), as illustrated in the upper row of Figure 2. While this existing typology provides useful insights into Socratic question classification, it is overly broad for pedagogical purposes; it makes annotation and evaluation difficult and limits the typology's ability to support the generation of thought-provoking questions. For instance, the line between "Clarification" and "Probing Reason and Evidence" often blurs when vague claims are partially supported but still require elaboration. This ambiguity leads to inconsistencies in annotation and hampers model training, as the input–label relationship is ill-defined.

3.2 FOCUS Typology

From the example in Section 2, we observe that the SoQG-2023 question offers a broad but flexible inquiry, while CQs provide more specific prompts focused on general evidence probing. However, they are less informative in distinguishing whether the argument contains weak evidence or lacks evidence altogether (Figure 1). To improve the specificity of the typology, we introduce a new fine-grained framework called the FOCUS Socratic Question (FSQ) Typology, which consists of eleven distinct types (lower row of Figure 2; FSQType). Ten of these are designed to capture more specific and pedagogically meaningful variations in Socratic questioning, and the eleventh type, "None of the Above," is included to handle edge cases where none other applies (Ziegenbein et al., 2023).

To develop the new typology, we first drew on the strengths of two theoretical foundations: CQs (also known as Walton's scheme) and SQs. We analyzed transcripts and observations from classroom dialogues in (Paul and Binker, 1990; Paul and Elder, 2019), which illustrate how teachers use Socratic questioning to probe students' thoughts and ideas. Based on this analysis, we created an initial typology consisting of 16 question types. We then mapped the tentative question types to major Walton argumentation schemes, including Argument from Consequences, Counter-Argument, and Argument from Ignorance. Through collaborative discussions with annotators, we examined the distinguishing characteristics of each type and refined the framework, ultimately narrowing it down to 10 types. During this process, we developed (1) distinct types, (2) definitions for each type, (3) concrete examples, (4) illustrative exemplars. These four components were then incorporated into a tentative annotation guideline used to support consistent labeling. See Table 12 in the Appendix for a detailed overview of each.

3.3 Data Prepossessing

Upon examining the characteristics of SoQG-2023, we observed that some arguments were contextually insufficient, which often led to questions that were neither meaningful nor thought-provoking. For example, *"I cannot. I don't know what you mean."* lacks the necessary context to support the generation of a substantive question. To improve the quality of our dataset, we aimed to remove these context-poor arguments. For this, we first defined

[2] https://www.reddit.com/r/changemyview/

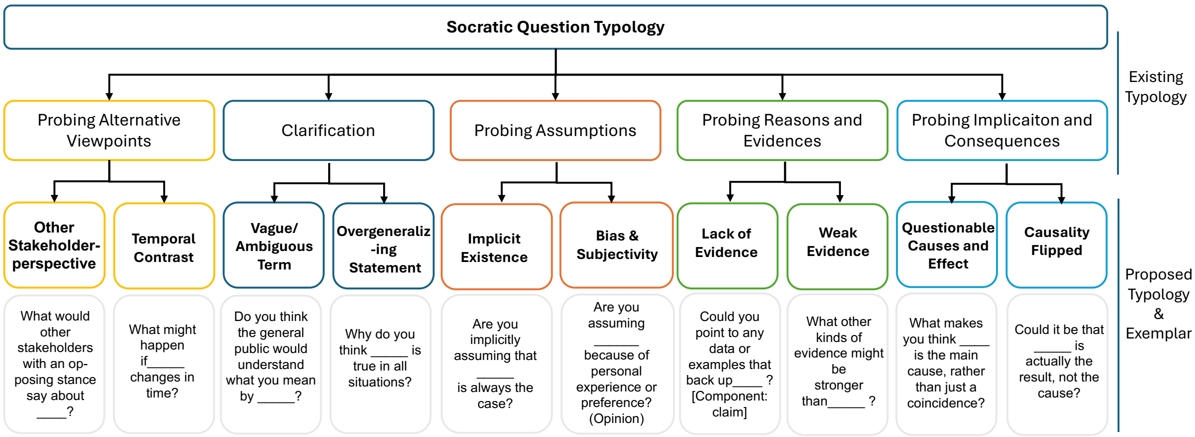

Socratic Question Typology
Existing Typology
Probing Alternative Viewpoints
Clarification
Probing Assumptions
Probing Reasons and Evidences
Probing Implicaiton and Consequences
Proposed Typology & Exemplar
Other Stakeholder-perspective
Temporal Contrast
Vague/ Ambiguous Term
Overgeneraliz-ing Statement
Implicit Existence
Bias & Subjectivity
Lack of Evidence
Weak Evidence
Questionable Causes and Effect
Causality Flipped
What would other stakeholders with an op-posing stance say about ____?
What might happen if_____ changes in time?
Do you think the general public would understand what you mean by _____?
Why do you think _____ is true in all situations?
Are you implicitly assuming that _____ is always the case?
Are you assuming ______ because of personal experience or preference? (Opinion)
Could you point to any data or examples that back up____ ? [Component: claim]
What other kinds of evidence might be stronger than_____ ?
What makes you think ____ is the main cause, rather than just a coincidence?
Could it be that _____ is actually the result, not the cause?

Figure 2: The proposed typology of Socratic questions builds on the framework introduced by Paul and Elder (2007), Paul and Elder (2019), and extended by Azzopardi (2021). It expands existing categories—such as probing alternative viewpoints, clarification, probing assumptions, and probing reasons and evidence—by introducing more fine-grained subcategories.

what qualifies as a contextually sufficient argument. It is an argument that provides enough background for the reader to understand and evaluate the claim without relying on external knowledge. We then randomly sampled 70 instances from the original 110K dataset to manually identify representative examples of insufficient context. From this sample, we selected five examples to build a few-shot prompt for automatic filtering using GPT-4 across the entire dataset. This filtering process resulted in 23,599 candidate instances. We then randomly sampled 1,400 instances from this filtered output for manual validation. After evaluating contextual sufficiency and grouping arguments by topic to streamline the annotation process, we retained 798 contextually sufficient instances. These were then split into a development set of 140 instances and a test set of 300 instances.

Topic	Ctx. Sufficient	Topic Related
T0 Politics & Ideology	103	144
T1 Beliefs & Philosophy	98	111
T2 Social Perception	116	114
T3 Race & Ethnicity	127	109
T4 Gender & Identity	119	131
T5 Economics & Society	117	106
T6 Humanity	118	110
Total	798	825

Table 1: A total of 1,400 instances were annotated by two annotators to filter out contextually insufficient arguments. The table reports the number of instances that passed both contextual sufficiency and topic relevance filtering, grouped by topic category. Each topic initially contained 200 instances prior to annotation.

3.4 Annotation Procedure

Our annotation strategy is inspired by label selection and span selection techniques from prior work (Reisert et al., 2015; Robbani et al., 2024; Naito et al., 2024). The procedure consists of two main steps: (1) label selection, formulated as a binary classification task, and (2) span selection, which involves identifying a relevant span of text within the argument (Reisert et al., 2018). The annotator first determines whether a specific FSQ-Type (e.g., “Lacks Evidence”) is applicable. In this label selection step, annotators are presented with an argument and a specific FSQType and asked to determine whether the argument can be instantiated with that type by selecting “yes” or “no.” If “yes” is selected, they proceed to span selection, where they highlight the portion of the argument that supports the chosen question type.

We recruited three annotators, all of whom are also authors of the paper, with expertise in argumentation mining: a postdoctoral researcher, a faculty member, and a computer science student. One annotator is a native English speaker, while the other two have high proficiency in English. All three contributed to annotating the development set (140 instances), and two of them annotated the test set (300 instances). Before annotating the development set, all annotators received training and participated in calibration exercises. After each round of annotation, disagreements were systematically reviewed and resolved to ensure consistency and a shared understanding of the annotation guidelines.

To establish a standardized annotation guideline, we first split the development set of 140 instances into two subsets: Dev I (70 instances) and Dev II (70 instances). We used Dev I to standardize the guideline, starting from the tentative version introduced in Section 3.2. Using this tentative guideline, three annotators independently annotated the Dev I subset. We then calculated inter-annotator agreement (IAA) and partial match similarity scores. Based on the observed disagreements, the annotators discussed and provided feedback using concrete examples to refine the guideline. The standard guideline consists of the background of FOCUS, the task objective, and step-by-step instructions, including: (1) reading the argument, (2) selecting spans, (3) if multiple possible spans exist, choosing the one that appears first in the text, and (4) identifying the FSQ type with its description, exemplar, and example. The finalized version was then used to annotate the Dev II subset. With observed improvements in IAA and consistency across annotators, we adopted this version as the final annotation guideline.

With the finalized guidelines in place, two annotators conducted the full annotation of the test set (300 samples). In total, we conducted three rounds of annotation to ensure the quality of the dataset, each showing consistent improvements in both Inter-Annotator Agreement (IAA) and partial match (PM) scores.

3.5 Annotation Result Analysis

To ensure a reliable evaluation of our annotations, we measured annotator agreement using multiple metrics: (1) for the binary classification of FSQType, we note the limitations of the Kappa score (Li et al., 2023) under data skewness, which can yield artificially low values. Therefore, we emphasize that our moderate-to-high observed agreement and PABAK should serve as the primary metrics for measuring IAA; and (2) for span selection, we used partial match scores based on Jaccard and BERTScore to assess overlap and semantic similarity between annotators. These metrics collectively offer a comprehensive view of annotation quality, capturing both literal and conceptual alignment (Zhang et al., 2020; Papineni et al., 2002; Lin, 2004). The results for FSQType labeling are presented in Table 2, and the span selection agreement results are shown in Table 3.

We conducted a detailed evaluation of both label and span-level agreement. Table 2 reports agreement scores for both the development (dev) and test sets. Notably, observed agreement improves in the test set, suggesting increased annotator consistency. For example, for the "Other Stakeholder Perspective" FSQType, observed agreement rises from 0.704 (dev) to 0.833 (test). To address class imbalance and provide a more stable measure of reliability, we also report PABAK scores alongside observed agreement. Overall, the improvement in agreement scores across annotation rounds demonstrates the effectiveness of our guidelines and confirms the reliability of the task design. Furthermore, Table 3 presents span similarity metrics—Jaccard, BERTScore, and ROUGE—for both the dev and test sets, categorized by Socratic question types. The results show generally higher similarity scores in the test set across all metrics. For instance, in the Clarification category, the ROUGE score improves from 0.656 (dev set, Overgeneralized Statement) to 0.995 (test set), indicating more consistent span annotations. Taken together, the inter-annotator agreement (IAA) and performance metrics (PM) demonstrate moderate to high scores, reflecting the robustness of the proposed typology in capturing multiple aspects of argumentative weakness. These findings suggest that the typology is well-designed to support sustainable, targeted feedback through span-level annotations and that the task itself is well-defined and objective-oriented.

3.6 Disagreement Analysis

We examine span-level disagreement among three annotators on the development set. Despite following strict guidelines, some arguments were interpreted differently. To account for this, we define the gold span as agreement between at least two annotators, and the disagree span as one selected by only a single annotator. These disagreement spans are retained as supplementary references to capture interpretive diversity and support fairer model evaluation. Table 20 presents both gold and disagree spans, each semantically valid and appropriate as the focus of a Socratic question. The presence of non-overlapping disagree spans underscores that a single argument can support multiple valid interpretations.

4 Baseline Experiment

FOCUS introduces the task of identifying which part of an argument should be the focus when generating a SQ. Each FSQType is intended to reflect a distinct way of probing weaknesses, assumptions, or ambiguities within the argument.

Metric	**Alternative Viewpoint**		**Assumption**		**Clarification**		**Implication and Consequences**		**Reason and Evidence**	
	Other Stakeholder Perspective	*Temporal Contrast*	*Implicit Assumption*	*Bias and Subjectivity*	*Vague or Ambiguity*	*Overgeneralized Statement*	*Questionable Cause-Effect*	*Causality Flipped*	*Lacks Evidence*	*Weak Evidence*
Observed Agreement (Dev)	0.704	0.686	0.686	0.834	0.723	0.815	0.852	0.908	0.815	0.815
Observed Agreement (Test)	0.833	0.833	0.850	0.817	0.883	0.950	1.000	0.917	0.817	0.817
Fleiss' Kappa (Dev)	-0.143	0.072	0.157	0.518	0.258	0.313	0.501	0.325	0.421	0.489
PABAK (Test)	0.667	0.833	0.700	0.633	0.767	0.900	0.967	1.000	0.833	0.833

Table 2: Label selection agreement metrics (Observed Agreement, Fleiss' Kappa, and PABAK) for each FSQ type in development and test sets.

Metric		Alternative Viewpoint		Assumption		Clarification		Implication and Consequences		Reason and Evidence	
		Other Stakeholder Perspective	*Temporal Contrast*	*Implicit Assumption*	*Bias and Subjectivity*	*Vague or Ambiguity*	*Overgeneralized Statement*	*Questionable Cause-Effect*	*Causality Flipped*	*Lacks Evidence*	*Weak Evidence*
Jaccard	Dev	0.605	0.511	0.521	0.501	0.512	0.604	0.674	1.000	0.597	0.539
BERTScore	Dev	0.773	0.692	0.696	0.715	0.669	0.742	0.796	1.000	0.718	0.703
ROUGE	Dev	0.679	0.580	0.572	0.573	0.579	0.656	0.730	1.000	0.650	0.639
Jaccard	Test	0.855	0.791	0.881	0.772	0.899	0.991	0.756	1.000	0.845	0.758
BERTScore	Test	0.864	0.876	0.931	0.842	0.937	0.994	0.897	0.988	0.904	0.863
ROUGE	Test	0.870	0.833	0.907	0.809	0.910	0.995	0.828	1.000	0.877	0.807

Table 3: Span agreement metrics (Jaccard, BERTScore, ROUGE) for dev and test sets, categorized by Socratic question types.

4.1 Task Formulation

Let A be an input argument consisting of a sequence of tokens:

$$A = [w_1, w_2, \ldots, w_n]$$

The model is prompted to jointly predict:

1. A set of **FSQTypes**:

$$\mathcal{Y} = \{y_1, y_2, \ldots, y_k\} \subseteq \mathcal{L}$$

where $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_{11}\}$ is a predefined set of 11 Socratic question types, and $k \leq 2$ in our experimental setup.

2. A corresponding set of **justifying spans**:

$$\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$$

where each $s_i = A_{[i:j]}$ is a contiguous subsequence of tokens from A, i.e.,

$$s_i = [w_i, w_{i+1}, \ldots, w_j]$$

that serves as the rationale for selecting label y_i.

The overall task is modeled as learning a function:

$$f_\theta : A \to (\mathcal{Y}, \mathcal{S})$$

where f_θ is implemented using prompting and fine-tuning techniques, designed to output both the relevant reasoning categories and their corresponding evidence spans. Table 13 demonstrates the prompt format used in the experiment.

4.2 Data Splitting and Baseline Model

To establish a baseline, we split the FOCUS dataset (440 instances) into three subsets: Focus-dev (140 instances) for fine-tuning (10% for validation), Focus-test (300 instances) for evaluation, and few-shot (1-shot, 5-shot) samples from Focus-dev. We evaluate different kind of state-of-the-art generative models: LLaMA-2 (meta-llama/Llama-2-7b-chat-hf), LLaMA-3 (Meta-Llama-3-8B-Instruct), Mistral (Mistral-7B-Instruct-v0.2), OLMo (OLMo-2-1124-7B-Instruct), and Qwen (Qwen2.5-14B-Instruct, Qwen2.5-7B-Instruct), which are run via Hugging Face using similar decoding settings (temperature = 0.3). GPT-4 and GPT3.5 are queried via the OpenAI API using 1-shot and 5-shot setups (temperature = 0.3, max_tokens = 512). Fine-Tuning is implemented using Hugging Face's transformer. PEFT, and TRL with Lora-base tuning.

4.3 Classification Result

Our classification experiments have two goals: (1) evaluate whether models can recognize argument weaknesses and select appropriate probing questions, and (2) establish baseline performance relative to human annotators. Multi-label metrics are computed using scikit-learn's MultiLabelBinarizer.

Table 4 presents per-class F1 scores for focus-type classification (seed 42), revealing complementary strengths across reasoning types. Qwen3-8B excels in Vague and Ambiguous Term and Overgeneralized Statement, Llama2-13B in Other Stakeholder Perspective, and OLMo-2 in Implicit Assumption, Bias and Subjectivity, and Null. Mistral

Focus Typology	Qwen3-8B	Qwen3-14B	Llama2-7B	Llama2-13B	Mistral	CoT-1shot	GPT4-1shot	OLMo-2
Other Stakeholder Perspective	0.255	0.030	0.039	**0.260**	0.255	0.147	**0.222**	0.020
Temporal Contrast	0.194	0.042	0.056	0.314	0.306	0.000	0.000	0.028
Implicit Assumption	0.071	0.200	0.000	0.171	0.095	0.173	0.217	0.357
Bias and Subjectivity	0.104	0.067	0.021	0.192	0.146	0.154	0.120	0.250
Vague and Ambiguous Term	**0.423**	0.103	0.077	0.115	0.077	0.225	0.203	0.040
Overgeneralized Statement	0.364	0.067	0.046	0.227	0.000	0.200	0.182	0.000
Questionable Cause-Effect Rel.	0.000	0.000	0.000	0.000	0.083	**0.125**	0.125	0.067
Causality Flipped	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Lack of Evidence	**0.478**	0.156	**0.739**	0.130	0.565	0.132	0.089	**0.522**
Weak Evidence	0.069	0.051	0.000	0.103	**0.448**	0.191	0.061	0.037
Null	0.597	**0.797**	0.741	0.303	0.282	0.468	0.100	**0.853**

Table 4: Per-class F1 scores on focus-type classification across eight models (seed 42). Bolded values indicate the best-performing model for each reasoning type.

Focus Typology	GPT-4 1-shot		GPT-4 5-shot		GPT-3.5 Turbo 5-shot		Qwen3-8B LoRA		LLaMA2 LoRA		LLaMA3 LoRA		Mistral LoRA	
	J	B	J	B	J	B	J	B	J	B	J	B	J	B
Other Stakeholder Perspective	**0.472**	**0.623**	0.357	0.541	0.473	0.620	0.275	0.432	0.326	0.527	0.399	0.540	0.505	0.667
Temporal Contrast	**0.418**	**0.583**	0.341	0.542	0.367	0.538	0.209	0.281	0.352	0.544	0.387	0.556	0.177	0.549
Implicit Assumption	0.303	0.524	0.372	0.554	0.348	0.525	0.187	0.343	0.274	0.441	0.429	0.607	**0.461**	**0.682**
Bias and Subjectivity	0.263	0.428	0.257	0.415	0.294	0.484	0.094	0.208	0.332	0.485	**0.367**	**0.513**	0.192	0.393
Vague and Ambiguous Term	0.075	0.264	0.116	0.281	0.152	0.335	0.017	0.204	0.041	0.223	0.023	0.213	**0.194**	**0.386**
Overgeneralization	0.357	0.560	0.403	0.591	0.254	0.474	0.124	0.238	0.299	0.543	**0.660**	**0.713**	0.136	0.342
Questionable Cause–Effect Relationship	0.216	0.493	0.168	0.410	0.174	0.400	0.158	0.400	0.273	0.530	**0.336**	**0.558**	0.181	0.455
Causality Flipped	0.000	0.244	0.000	0.000	0.000	0.000	0.000	0.032	0.000	0.109	**1.000**	**0.988**	0.000	0.000
Lack of Evidence	0.462	0.591	**0.531**	**0.679**	0.483	0.639	0.223	0.363	0.431	0.517	0.457	0.567	0.466	0.584
Weak Evidence	0.354	0.485	**0.376**	**0.495**	0.346	0.490	0.051	0.155	0.315	0.490	0.264	0.376	0.259	0.485

Table 5: Span-level similarity (Jaccard "J" and BERTScore "B") between human-annotated and model-predicted spans across ten focus typologies. Bold values mark the highest per typology–metric pair.

Metric	GPT-4	GPT-3.5	GPT-4	GPT-3.5
	Gold	Gold	All	All
$\text{Jaccard}_{\text{avg}}$	0.156	0.189	**0.194**	0.227
$\text{BERTScore}_{\text{avg}}$	0.507	0.432	**0.565**	0.564

Table 6: Span-level evaluation comparing model-generated spans on the dev set with gold-standard spans annotated by two annotators (Gold) versus all annotators (All). Bold indicates the best score in each metric.

performs well on Weak Evidence and Temporal Contrast, while GPT-4 (1-shot) leads in Questionable Cause–Effect Relationship. Overall, no model dominates across all types, highlighting diverse weaknesses among architectures.

Table 23 summarizes model performance averaged across three random seeds. Qwen3-8B shows the highest Macro Precision, Recall, and F1, while Qwen2.5-14B-Instruct achieves the best Micro F1, performing well on frequent classes. The discrepancy between Macro and Micro scores highlights model performance skew across certain classes. Yet, both fall short of human reliability (PABAK 0.63–1.00), with Macro F1 reaching only 0.172 and 0.533, respectively, indicating limited generalization despite partial label-level understanding.

4.4 Span Generation Result

Our span-generation baseline evaluates how well models can extract argument-weakness spans that align with human annotations. Table 5 reports Jaccard similarity (J) and BERTScore (B) between model-predicted and human-annotated spans across eleven FSQTypes, comparing seven setups: GPT-4 (1-shot/5-shot), GPT-3.5 Turbo (5-shot), and four LoRA-fine-tuned models (Qwen3-8B, LLaMA2, LLaMA3, Mistral,). Boldface indicates the best performance for each metric–typology pair.

GPT-4 (1-shot) performs best on span alignment for categories like "Other Stakeholder Perspective" and "Temporal Contrast," while its 5-shot version does well on "Lack of Evidence" and "Weak Evidence." LLaMA3-LoRA stands out across several categories, especially "Implicit Assumption," "Overgeneralization," and "Questionable Cause–Effect Relationship." It is also the only model to achieve nearly perfect alignment on

Error Name	Description	Argument	Error Steps
Overgeneralization Bias	The model overuses "generalization" to explain reasoning flaws, labeling arguments as overgeneralized even when it is not the main issue.	Yes, but in prison, you can, in theory, learn from your actions and become a better person. In hell, you will never redeem yourself, and are tortured forever.	The model misinterprets the word "never" as the main error signal, while the true flaw is the argument's failure to consider alternative perspectives and temporal contrast.
Missed Clarification Probing	The model either over-questions clear concepts or overlooks vague ones, showing weak judgment about what needs clarification.	Theft is immoral by definition, though. You can't establish theft as moral.	The model incorrectly treats the concept of theft as debatable instead of focusing on the more context-dependent notion of morality.
Incorrect Assumption Reasoning	The model invents new assumptions or discourse markers not present in the original text, leading to false reasoning chains.	Supporting veterans means supporting war. War is rape, torture, and pain.	In the assumption step, the model invents the word "all," which is not present in the original argument.
Correct Reasoning but Incorrect Span/Type Mapping	The reasoning is valid, but the model mismatches spans or flaw types with the gold annotation.	No, you are making the claim that 50% of cops are bad. That is 400,000 individuals. You can't make that statement and then proceed to tell me to prove you wrong.	It correctly detects a lack of evidence but mislabels it as "implicit existence."

Table 7: Examples of model reasoning errors categorized by type. Each row shows an error description, the argument where it occurs, and how the model's reasoning deviates from the intended logic.

"Causality Flipped." Mistral-LoRA shows strength in identifying vague language, but GPT-3.5 Turbo and LLaMA2-LoRA do not lead in any category.

Table 6 presents the results for gold and disagreement (All) spans on the GPT-4 and GPT-3.5 dev sets under the 1-shot setting. Since arguments can support multiple interpretations, we include disagreement spans to ensure fairer evaluation. This improves GPT-4's average Jaccard score from 0.156 to 0.194, though the model still struggles to consistently identify appropriate spans.

Overall, all models still lag far behind human performance. Human annotators consistently agree on span selections, thanks to clear guidelines and their ability to understand subtle meanings. In contrast, models often struggle with categories that require deeper reasoning, sometimes failing entirely. This shows that while current models can extract simpler spans, they are not yet reliable for more nuanced understanding.

4.5 Error Analysis

We analyzed model behavior using Chain-of-Thought (CoT) decomposition for argument diagnosis with GPT-4, as shown in Table 7, which breaks down each argument into reasoning steps. On a sample of 50 test instances, this setup improved performance, increasing the macro-F1 score from 0.1464 (GPT-4, 5-shot) to 0.1612. We identified four major error patterns (Table 7): (1) Overgeneralization Bias (22%), (2) Redundant or Missed Clarification (30%), (3) Incorrect Assumption Reasoning (24%), and (4) Correct Reasoning but Wrong Mapping (24%).

As shown in Table 23, all models exhibit a clear gap between macro and micro scores due to label imbalance. Models such as Llama-2 and Llama-3 achieve high micro-F1 (≈0.48–0.53) but low macro-F1 (≈0.10–0.15), showing bias toward frequent labels. In contrast, Qwen3 and GPT-4 (1-shot, 5-shot) show smaller gaps (macro-F1 ≈0.13–0.19; micro-F1 ≈0.40–0.43), suggesting better balance across rare reasoning types.

Overall, this gap highlights that instruction-tuned models often overfit dominant reasoning patterns, inflating micro-level scores while limiting generalization. Models with smaller macro–micro gaps, such as GPT-4, demonstrate stronger robustness and a better understanding of diverse reasoning structures.

5 Generalization to Other Datasets

We assessed domain transfer by sampling 50 sentences from the 1k-sentence LogicClimate corpus (climate-news passages) and annotating them using our rubric, resulting in 250 label decisions. In this setting, we used the same annotator as for the FOCUS test set. Despite the shift in domain, the annotator still achieved high reliability, confirming that the task design generalizes to real-world climate discourse, which includes 13 fallacy types (e.g., false causality, *ad hominem*).

5.1 LogicClimate Annotation Result

Inter-annotator agreement on LogicClimate is strong overall (mean OA = 0.88). "Other Stakeholder Perspective" and "Questionable Cause–Effect" stand out (OA $\geq$ 0.94, $\kappa \approx$ 0.84),

Metric	Alternative Viewpoint		Assumption		Clarification		Implication and Consequences		Reason and Evidence	
	Other Stakeholder Perspective	*Temporal Contrast*	*Implicit Assumption*	*Bias and Subjectivity*	*Vague or Ambiguity*	*Overgeneralized Statement*	*Questionable Cause-Effect*	*Causality Flipped*	*Lacks Evidence*	*Weak Evidence*
Observed Agreement	0.960	0.780	0.880	0.760	0.860	0.820	0.940	1.000	0.880	0.920
Cohen's Kappa	0.834	0.505	0.629	0.407	0.455	0.584	0.854	1.000	0.760	0.836
PABAK	0.920	0.560	0.760	0.520	0.720	0.640	0.880	1.000	0.760	0.840

Table 8: Inter-annotator agreement for each Focused Socratic Question (FSQ) type in the LogicClimate corpus, reported as Observed Agreement (OA), Cohen's κ, and prevalence-adjusted bias-adjusted κ (PABAK). Causality Flipped? shows perfect agreement; missing κ is due to zero variance across raters.

Metric	Alternative Viewpoint		Assumption		Clarification		Implication and Consequences		Reason and Evidence	
	Other Stakeholder Perspective	*Temporal Contrast*	*Implicit Assumption*	*Bias and Subjectivity*	*Vague or Ambiguity*	*Overgeneralized Statement*	*Questionable Cause-Effect*	*Causality Flipped*	*Lacks Evidence*	*Weak Evidence*
Jaccard	0.756	0.601	0.741	0.576	0.703	0.627	0.705	0.000	0.665	0.815
BERTScore	0.852	0.659	0.797	0.650	0.756	0.681	0.784	0.000	0.735	0.862
ROUGE	0.805	0.622	0.770	0.606	0.703	0.668	0.729	0.000	0.699	0.839

Table 9: Span agreement metrics (Jaccard, BERTScore, ROUGE) by Socratic question types in LogicClimate.

suggesting these cues are easy for raters to spot. Reliability is moderate for "Implicit Existence" and "Overgeneralizing Statement" ($\kappa \approx 0.6$) and lowest for "Bias and Subjectivity" and "Vague/Ambiguous Terms" ($\kappa \approx$ 0.4-0.46). Table 9 reports span-level agreement (Jaccard, BERTScore, ROUGE). "Weak Evidence" and "Other Stakeholder Perspective" achieve the highest similarity scores, indicating clear, consistent spans, whereas "Bias and Subjectivity" and "Temporal Contrast" perform worst, reflecting their context-dependent nature. These findings confirm the robustness of our rubric in consistently capturing the intended weakness types across annotators and datasets.

5.2 LogicClimate Analysis

Raters excel on explicit cues. "Weak Evidence" and "Other Stakeholder Perspective" top span metrics (Jaccard, BERTScore, ROUGE) and show the highest κ ($\approx$ 0.84). By contrast, "Temporal Contrast" and "Bias & Subjectivity" score lowest for spans and κ ($\approx$ 0.4–0.46), reflecting their context-dependent nature. "Causality Flipped" is perfectly reliable for humans.

Table 11 illustrates how our FSQType maps onto various fallacy types—flaws in reasoning—in the LogicClimate sample. Our FSQType framework captures these reasoning flaws by categorizing them into pedagogically relevant weakness types. Each fallacy exhibits distinct associations with specific FSQ categories. For instance, appeal to emotion frequently aligns with Bias and Subjectivity and Questionable Cause-Effect, while fallacy of extension and fallacy of relevance often co-occur with Vague Terms and Overgeneralization. This alignment suggests that our typology not only provides fine-grained types but also mirrors how fallacious reasoning manifests in natural discourse. These patterns demonstrate that our typology meaningfully captures the underlying dimensions of argumentative flaws across diverse fallacy categories.

6 Conclusion

We present FOCUS, the first benchmark linking Socratic questions to fine-grained weakness types and the exact spans they reference. It includes an 11-type typology, 440 high-agreement instances, and an explainable task that requires models to output both a label and supporting evidence. Baselines with GPT-4, GPT-3.5, LLaMA-2/3, Mistral, OLMo-2, and Qwen still trail human reliability, especially on context-dependent cues such as "Bias & Subjectivity" and "Temporal Contrast," leaving ample room for progress. A transfer test on 50 climate-news sentences (LogicClimate) shows that the rubric generalizes well: annotators maintained a mean agreement of 0.88 and high span overlap on explicit cues like "Weak Evidence," yet both humans and models struggled with rare or implicit patterns (e.g., reverse causation). This points to two priorities: clearer guidelines for subjective types and discourse-aware models that can handle temporal and causal shifts. All data, guidelines, and code are released to spur research on explainable, learning-oriented Socratic question generation.

Limitations

Our study has several limitations. First, all arguments are in English, which may constrain the cross-linguistic generalizability of our framework. Extending the dataset to other languages would broaden its applicability. Second, some arguments contain multiple claims or supporting points, leading to multiple valid interpretations. While we addressed this by including both gold and disagreement spans, this introduces ambiguity in span-level evaluation and can challenge model training. Third, although we demonstrate the generalizability of our framework on a new dataset (LogicClimate), future work should examine how models interact with this generalized dataset to assess robustness across domains and argument styles.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 22H00524. Ana Brassard's effort was supported by a RIKEN Incentive Research Project FY2024. We thank Nguyen Minh Phuong, Hinata Tezuka for their support. This work was written independently, with minor phrasing assistance from a large language model (ChatGPT).

References

Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan, and Mohsen Dorodchi. 2023. Socratic questioning of novice debuggers: A benchmark dataset and preliminary evaluations. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 709–726, Toronto, Canada. Association for Computational Linguistics.

Beng Heng Ang, Sujatha Das Gollapalli, and See-Kiong Ng. 2023. Socratic question generation: A novel dataset, models, and evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 147–165, Dubrovnik, Croatia. Association for Computational Linguistics.

Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 27–37.

Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 105–116, Miami, FL, USA. Association for Computational Linguistics.

Sie Wai Chew, I-Hsiu Lin, and Nian-Shing Chen. 2019. Using socratic questioning strategy to enhance critical thinking skill of elementary school students. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, volume 2161, pages 290–294. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. ChainCQG: Flow-aware conversational question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2061–2070, Online. Association for Computational Linguistics.

Camélia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. 2023. Teach me how to argue: A survey on nlp feedback systems in argumentation. In *Proceedings of the 10th Workshop on Argument Mining*, pages 19–34.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Yueh-Ren Ho, Bao-Yu Chen, and Chien-Ming Li. 2023. Thinking more wisely: using the socratic method to develop critical thinking skills amongst healthcare students. *BMC medical education*, 23(1):173.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Alison King. 1994. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American educational research journal*, 31(2):338–368.

Nischal Ashok Kumar and Andrew Lan. 2024. Improving socratic question generation using data augmentation and preference optimization. *arXiv preprint arXiv:2403.00199*.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. Qed: A framework and dataset for explanations in question answering. *Transactions of the Association for computational Linguistics*, 9:790–806.

Ming Li, Qian Gao, and Tianfei Yu. 2023. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. *BMC cancer*, 23(1):799.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text summarization branches out*.

Shoichi Naito, Shintaro Sawada, Chihiro Nakagawa, Naoya Inoue, Kenshi Yamaguchi, Iori Shimizu, Farjana Sultana Mim, Keshav Singh, and Kentaro Inui. 2022. TYPIC: A corpus of template-based diagnostic comments on argumentation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5916–5928, Marseille, France. European Language Resources Association.

Shoichi Naito, Wenzhi Wang, Paul Reisert, Naoya Inoue, Camélia Guerraoui, Kenshi Yamaguchi, Jungmin Choi, Irfan Robbani, Surawat Pothong, and Kentaro Inui. 2024. Designing logic pattern templates for counter-argument logical structure analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11313–11331, Miami, Florida, USA. Association for Computational Linguistics.

David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218.

OpenAI. 2023. Gpt-4 technical report. https://openai.com/research/gpt-4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Richard Paul and AJA Binker. 1990. Socratic questioning. *Critical thinking: What every person needs to survive in a rapidly changing world*, pages 269–298.

Richard Paul and Linda Elder. 2007. Critical thinking: The art of socratic questioning. *Journal of developmental education*, 31(1):36.

Richard Paul and Linda Elder. 2019. *The thinker's guide to Socratic questioning*. Rowman & Littlefield.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55.

Irfan Robbani, Paul Reisert, Surawat Pothong, Naoya Inoue, Camélia Guerraoui, Wenzhi Wang, Shoichi Naito, Jungmin Choi, and Kentaro Inui. 2024. Flee the flaw: Annotating the underlying logic of fallacious arguments through templates and slot-filling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20524–20540, Miami, Florida, USA. Association for Computational Linguistics.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. *arXiv preprint arXiv:2211.12835*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

DN Walton. 2008. *Argumentation schemes*. Cambridge University Press.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Ethical Considerations

This study exclusively uses publicly available and anonymized datasets, including SoQG2023 and LogicClimate. No personally identifiable information is present in the data, and no new human data was collected. All annotations were conducted by trained researchers with expertise in argumentation and critical thinking, including the authors of this paper, who participated voluntarily. The study complies with the terms of use of all datasets and does not raise any privacy or data protection concerns.

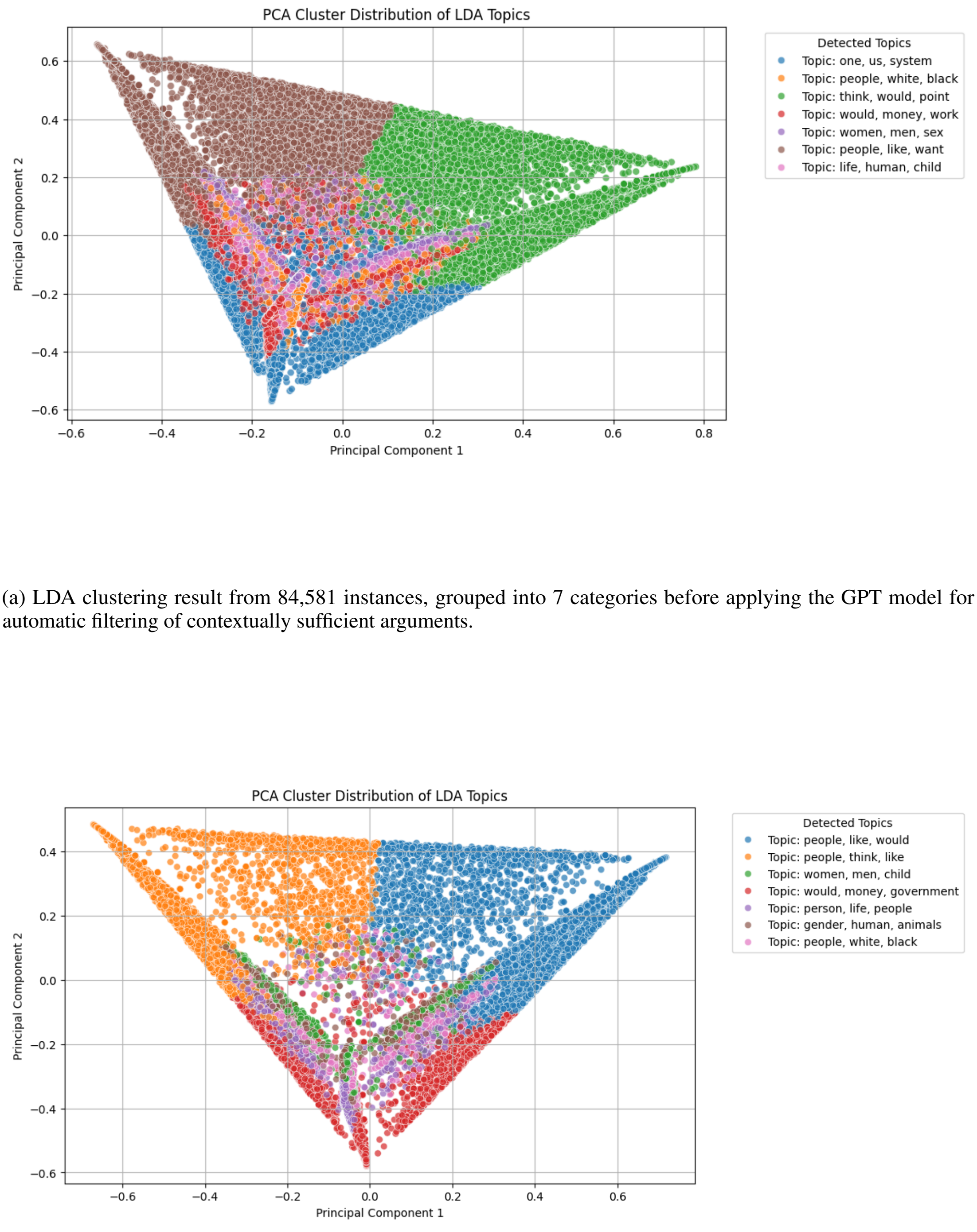
A.2 Experimental Setting

We fine-tuned Llama-2-13B-hf, Llama-2-7B-hf, Meta-Llama-3-8B-Instruct, Mistral-7B-Instruct, OLMo-2-1124-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-7B-Instruct, Qwen3-14B, and Qwen3-8B for Socratic question generation (FOCUS), with sequence packing enabled and a maximum sequence length of 1024 tokens. Training was conducted for 200 optimization steps using a linear learning rate schedule with a 3% warmup. The learning rate was set to 2×10^{-4}, and optimization employed AdamW (fused) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay $= 0.0$. The maximum gradient norm was 0.3. Precision was handled with bfloat16 and TF32, and both gradient checkpointing and length-based batching were enabled. The per-device batch size was 4, with gradient accumulation of 4 steps, yielding an effective batch size of 16 sequences per optimization step. We applied LoRA with a rank of $r = 32$, while all other LoRA hyperparameters followed the defaults in our implementation. Evaluation was performed every $\sim$10% of training steps ($\approx$20 steps) with a 20-step evaluation delay. We used zero-shot prompting (`prompting_type="zeroshot"`, `kshot=0`) for evaluation on the development set (`do_eval_dev=True`), while test-set evaluation was disabled for this run. Random seeds were fixed at 42, 43, and 44 to ensure result stability across runs. All experiments were executed on a single NVIDIA GPU (`_n_gpu=1`, Device: cuda). We used fixed values (learning rate $= 2 \times 10^{-4}$, schedule = linear, warmup = 3%, LoRA $r = 32$, max steps = 200, effective batch = 16) based on prior tuning heuristics for 7B instruction-tuned LLMs using PEFT. For GPT-based baselines, we used the OpenAI Python SDK (version $\geq$ 1.0.0) to interface with the GPT-3.5 and GPT-4 APIs. Inference was performed using the models `gpt-3.5-turbo-instruct`, `gpt-3.5-turbo`, `text-davinci-003`, and `gpt-4`. All API calls were made via OpenAI's `chat.completions.create` and `completions.create` endpoints. Inference parameters were held constant across runs: temperature $= 0.7$, max_tokens $= 512$, top_p $= 1.0$, frequency_penalty $= 0$, and presence_penalty $= 0$. Prompting was standardized using a unified system prompt file concatenated with user-supplied argument text. Preprocessing, CSV handling, and output parsing were implemented in Python 3.10 using pandas and regular expressions for label extraction. All API-based experiments were executed in Google Colab. All experiments were conducted under identical settings across the three random seeds (42, 43, 44) to ensure reproducibility and statistical stability. All models were fine-tuned and evaluated under the same hyperparameter configuration described above, ensuring consistent training conditions and comparability across model families.

A.3 Observed Agreement and Applicability of FSQtype

To calculate observed agreement, each item was evaluated based on the level of consensus among the three annotators. A score of 1 was assigned when all three annotators provided the same label (full agreement), a score of 0.5 when two annotators agreed and one differed (partial agreement), and a score of 0 when all annotators provided different labels (no agreement). The final observed agreement was obtained by averaging these scores across all annotated items.

To compute applicability, the annotations for each subtype were first flattened into a single list—for example, nested lists such as `[[1,1,0], [0,1,1]]` were converted to `[1,1,0,0,1,1]`. The number of 1s (Applicable) and 0s (Not Applicable) was then counted for each subtype. These counts were used to determine the height of the bars in the plot. The applicability percentage per type was calculated by dividing the number of 1s by the total number of annotations (i.e., the sum of 1s and 0s) for that subtype. This approach reflects the distribution of individual judgments rather than majority-vote decisions.



(a) LDA clustering result from 84,581 instances, grouped into 7 categories before applying the GPT model for automatic filtering of contextually sufficient arguments.

(b) LDA clustering result from 23,599 instances, grouped into 7 categories after applying the GPT model for automatic filtering of contextually sufficient arguments.

Figure 3: LDA clustering results before and after applying the GPT model for automatic filtering of contextually sufficient arguments. The similar distributions across both plots indicate that the core topical structure of the dataset was preserved after filtering.

Prompt for Filtering Contextually Sufficient Arguments

Role: You are an expert annotator in argument mining.

Task: Analyze arguments from the `input` column in a CSV file and determine whether they are contextually sufficient based on the definition below. Collect only those that meet this criterion and save them into a new CSV file for download.

Definition: An argument is contextually sufficient if it provides all necessary background information, allowing the reader to fully understand and evaluate the claim without requiring external knowledge.

Examples of Contextually Sufficient Arguments:

- Example 1: *"Because deer and humans are not of the same species. If all life is not equivalent in value, then the lives of some species must be worth more or less than others."*
- Example 2: *"There are no long-term studies on the COVID vaccine because it hasn't existed for a long time. 'Long-term' in this context could mean 5, 10, or even 15+ years. While the vaccine is effective now, its long-term effects on the body remain uncertain."*
- Example 3: *"There are different theories about oppression and its causes. Second-wave feminists often argue that, because we live in a patriarchy, a woman's sexuality cannot be separated from the oppressive male-centric culture."*
- Example 4: *"Taking or doing whatever you want violates the rights of others, making you susceptible to law enforcement. Protecting individual rights through law enforcement is essential to maintaining a free society."*
- Example 5: *"I believe I am extremely unattractive and that no woman will love me for who I am. My hobbies are the only things I have, but because they are extremely common, I am not considered unique and therefore seen as boring."*

Instructions for Annotation:

- Include an argument if it contains enough context to be understood and evaluated independently.
- Exclude an argument if it lacks necessary background information or relies on external knowledge to be fully understood.

Table 10: Prompt format used for filtering contextually sufficient arguments prior to dataset construction. The full prompt is available in our project repository on GitHub.

A.4 Annotator Background

The annotation process involved three individuals with varying academic and linguistic backgrounds, all selected for their relevance to the task. One annotator is a faculty member who holds a Ph.D. and specializes in argument mining; they possess strong expertise in discourse-level analysis and have sufficient proficiency in English for academic annotation tasks. The second annotator is a postdoctoral researcher and a native English speaker from the United States, also specializing in argument mining and natural language understanding. The third annotator is a graduate-level computer science student with sufficient English proficiency and prior experience in data annotation. All annotators had foundational knowledge of Socratic Questioning and were provided with detailed guidelines, exemplars, and iterative feedback to ensure high-quality annotations.

A.5 License and Use of Artifacts

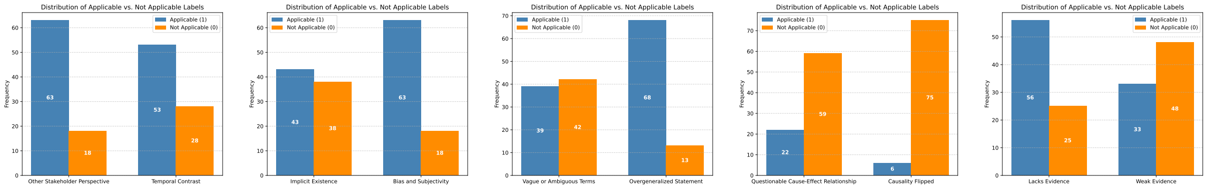
The GPT-3.5 and GPT-4 models were accessed through OpenAI's API under the standard research-compatible terms of service applicable at the time of use.[3] These models were utilized exclusively within the controlled research environment of our institution, following OpenAI's usage policies and documentation to ensure ethical and responsible deployment. All experimental interactions with the models were performed via secure API endpoints, and model configurations were kept consistent across evaluation sessions to maintain comparability and reproducibility.

No model weights were downloaded, stored, or modified during the study. All responses were generated through API-based inference, ensuring that no proprietary model parameters or internal representations were accessed. This approach guaranteed compliance with OpenAI's usage restrictions and preserved the integrity of the closed-weight model setting. The use of these models was consis-

[3] https://openai.com/policies/terms-of-use

Fallacy Type	**OSP**	**TC**	**VAT**	**OG**	**IE**	**BS**	**QCE**	**CF**	**LE**	**WE**
Ad Hominem	3	3	3	3	3	3	0	0	1	2
Ad Populum	1	0	1	2	2	1	0	0	2	0
Appeal to Emotion	5	4	4	2	4	5	3	0	3	2
Circular Reasoning	1	1	0	0	0	1	0	0	1	0
Fallacy of Credibility	4	5	4	4	5	4	1	0	1	3
Fallacy of Extension	5	4	5	5	5	5	1	0	3	1
Fallacy of Logic	2	2	2	2	2	1	1	0	1	1
Fallacy of Relevance	4	3	5	2	3	5	1	0	2	3
False Causality	0	0	0	0	1	1	1	0	0	1
False Dilemma	1	1	0	1	1	1	1	0	1	0
Faulty Generalization	4	4	4	4	4	3	1	0	1	3
Intentional	**12**	**12**	**13**	**12**	**12**	**12**	**5**	0	**10**	**5**

Table 11: Distribution of FSQ Types by Fallacy Type in the LogicClimate Sample (50 instances). FSQ acronyms: OSP = Other Stakeholder Perspective, TC = Temporal Contrast, VAT = Vague or Ambiguous Terms, OG = Overgeneralized Statement, IE = Implicit Existence, BS = Bias and Subjectivity, QCE = Questionable Cause-Effect Relationship, CF = Causality Flipped, LE = Lacks Evidence, WE = Weak Evidence. Bold values indicate the most prominent associations.



(a) Alternative Viewpoint

(b) Assumption

(c) Clarification

(d) Implication & Consequence

(e) Reason & Evidence

Figure 4: The distribution of annotator judgments for each FSQType subtype on dev set is shown, categorized as Applicable (1) or Not Applicable (0). Each subfigure (a–e) corresponds to two of the ten FSQ types—namely, Other Stakeholder Perspective, Temporal Contrast, Implicit Existence, Bias and Subjectivity, Vague or Ambiguous Terms, Overgeneralized Statement, Questionable Cause-Effect Relationship, Causality Flipped, Lacks Evidence, and Weak Evidence—ordered from left to right. The bars represent the total number of 1s and 0s across all annotations for each subtype (three annotations per instance). Blue bars indicate the number of annotations labeled as Applicable, while orange bars represent Not Applicable annotations. This visualization illustrates the perceived applicability of each subtype, with certain types (e.g., Causality Flipped and Weak Evidence) showing a lower proportion of applicable judgments.

tent with their intended purpose—strictly limited to prompt-based inference and benchmarking within a non-commercial academic context. No generated outputs were redistributed or repurposed beyond the scope of reproducible research.

Furthermore, the FOCUS dataset—including annotations, guidelines, and evaluation scripts—will be released publicly under the Creative Commons Attribution–NonCommercial 4.0 International License (CC BY-NC 4.0), which permits use, sharing, and adaptation for academic research purposes only. All source content, including Reddit-derived data and the SoQG-2023 dataset, was handled in accordance with public data-sharing policies and respective platform terms. The FOCUS benchmark is intended solely for research and educational purposes, with no commercial use permitted.

A.6 Dataset Consent and Source Information

The SoQG dataset used in this study, released by Ang et al. (2023), is publicly available at https://github.com/NUS-IDS/eacl23_soqg. It was constructed from Reddit's r/ChangeMyView forum using content from the Pushshift API. A subset of 3,600 instances was manually annotated via Mechanical Turk; the rest were labeled using a BERT-based classifier. All data is anonymized and shared under the CC BY-NC 4.0 license, in compliance with Reddit's content policies. In addition, we use a 50-instance sample from the LogicClimate dataset, a publicly released corpus of annotated fallacious arguments collected from the Kialo debate platform. The dataset is available at https://github.com/FOCUSSocratic2025/focus-socratic-question, and is shared under

Focus of the Question	**Exemplar**	**Example**	**Description**
Other Stakeholder Perspective	What would other stakeholders with an opposing stance say about ______?	*America has the best educational system, including MIT.*	Find something that other stakeholders might disagree with, based on different viewpoints like occupation or demographic background.
Temporal Contrast	What might happen if ______ changes in time?	*Electric cars are environmentally friendly.*	Look for something the author expresses an opinion about that could change over time (e.g., the weak Japanese yen or future environmental impact of batteries).
Vague or Ambiguous Terms	Do you think the general public would understand what you mean by ______?	*2nd wave feminist*	Find a phrase that might be difficult for the public to understand, is subjective, or requires more context.
Overgeneralized Statement	Why do you think ______ is true in all situations?	*1. no woman will love me for me* *2. I am not considered unique and therefore considered boring.*	Identify where a small sample is used to make a broad conclusion or where explicit generalization is made.
Implicit Existence	Are you implicitly assuming that ______ is always the case?	*Most people agree that stricter gun laws make society safer.*	Spot phrases where something is assumed to always be true, without being clearly stated.
Bias and Subjectivity	Are you assuming ______ because of personal experience or preference? Opinion	*I knew democracy is corrupted; I think communist is better.*	Look for statements based on bias, emotion, or subjective interpretation rather than facts.
Questionable Cause-Effect Relationship	What makes you think ______ is the main cause, rather than just a coincidence?	*Switzerland consumes chocolate. Hence, most Nobel Prize winners.*	Find where a correlation is wrongly treated as causation.
Causality Flipped	Could it be that ______ is actually the result, not the cause?	*I studied because I passed the exam.*	Identify phrases where the cause and effect are reversed.
Lacks Evidence	Could you point to any data or examples that back up ______? [Component: claim]	*Smoking is bad!*	Find a claim made without supporting data or examples.
Weak Evidence	What other kinds of evidence might be stronger than ______?	*Smoking is bad. It smells!*	Spot arguments supported by weak or irrelevant evidence.
None of the above	–	–	Does not fit any of the specific categories listed above.

Table 12: Taxonomy of Socratic Question Focus Types with exemplars, examples, and brief descriptions.

the CC BY 4.0 license. All examples are de-identified and used in accordance with Kialo's public data access policy.

Socratic Question Background
Socratic Question is a disciplined method of inquiry that systematically directs thought in multiple directions to achieve various intellectual goals. It is used to explore complex ideas, uncover truths, expose problems, reveal underlying assumptions, analyze concepts, distinguish between what is known and unknown, and trace the logical implications of one's thinking. Unlike ordinary questioning, Socratic Ques employs a deliberate and structured approach to deeply investigate the reasoning behind a claim.

Focus of the Question
This task is designed to identify which part of an argument should be the focus when generating a Socratic question. Each focus type reflects a different way of probing weaknesses, assumptions, or ambiguities within the argument.

Task Objective
Your goal is to determine whether an argument requires probing through one or more aspects of Socratic questioning.

Instructions

- Read the argument along with the exemplar associated with each focus type.
- For a given type (e.g., *Other Stakeholder Perspective*), perform binary classification:
  - Yes: The argument can be probed using this Socratic question type.
  - No: The argument is not relevant to this type.
- If "Yes," select the specific span of text that should be the focus of the Socratic question.
  - The span should be as short as possible while still capturing the necessary meaning.
  - If there are multiple possible spans, choose the one that appears first in the text.
  - The selected span should lend itself to generating a thought-provoking and meaningful Socratic question.

FSQ Types and Criteria

1. **Other Stakeholder Perspective**: Viewpoints from stakeholders who may disagree with the main argument.
2. **Temporal Contrast**: Highlights how criteria, contexts, or norms may shift over time.
3. **Vague or Ambiguous Terms**: Contains imprecise or unclear language.
4. **Overgeneralized Statement**: Applies a broad claim universally without justification.
5. **Implicit Existence**: Assumes something is true without stating it explicitly.
6. **Bias and Subjectivity**: Based on personal belief or emotion rather than fact.
7. **Lacks Evidence**: Claims made without supporting data.
8. **Weak Evidence**: Evidence provided is insufficient or loosely related.
9. **Questionable Cause-Effect Relationship**: Assumes causation from correlation.
10. **Causality Flipped**: Mistakes the effect as the cause.
11. **None of the Above**: No applicable FSQ type; use span = `"Null"`.

Table 13: Prompt Format. The following format was used as input to the GPT model for conducting the baseline experiment. The full prompt is available in our project repository on GitHub.

Input Format

`Argument: <ARGUMENT TEXT>`

Expected Output Format

`Classification:` List of one or two FSQ types (e.g., ["Implicit Existence", "Bias and Subjectivity"])
`Span:` For each FSQ type, the corresponding span from the argument. If "None of the above," return ["Null"].

Example

`Input Argument:` So I know this is a really, really sensitive subject. I'm not trying to stir up conflict; I'm just having what I believe challenged. Here is why I believe this: When I was still in the womb, the doctors recommended that I be aborted because I seemed to be developing abnormally.

```
Output:
Classification: ["Bias and Subjectivity", "Lacks Evidence"]
Span: ["Here is why I believe this", "the doctors recommended that I be aborted because I seemed
to be developing abnormally"]
```

Table 14: Prompt format. The following format was used as input to the GPT model for the baseline experiment. The full prompt is available in our project repository on GitHub Part 2.

Metric	SQ1-Jacc	SQ2-Jacc	SQ1-BiJacc	SQ2-BiJacc	SQ1-Edit	SQ2-Edit	SQ1-BERT	SQ2-BERT	SQ1-ROUGE	SQ2-ROUGE
Alternative$_{\text{PM}}$	0.605	0.511	0.565	0.455	0.608	0.580	0.773	0.692	0.679	0.580
Assumption$_{\text{PM}}$	0.521	0.501	0.490	0.467	0.587	0.525	0.696	0.715	0.572	0.573
Clarity$_{\text{PM}}$	0.512	0.604	0.427	0.569	0.535	0.617	0.669	0.742	0.579	0.656
Implication$_{\text{PM}}$	0.674	1.000	0.633	1.000	0.660	1.000	0.796	1.000	0.730	1.000
Reason$_{\text{PM}}$	0.597	0.539	0.570	0.482	0.621	0.543	0.718	0.703	0.650	0.639

Table 15: Average Similarity Scores from our PM analysis on dev set. SQ1 represents majority spans (F), and SQ2 represents minority spans (G). Subscript "PM" indicates scores based on post-majority disagreement analysis.

Subtype	**SQ1_2_Fleiss**	**SQ1_2_Gwet**	**SQ1_2_Kripp**
Alternative	-0.002	0.295	0.004
Assumption	0.290	0.305	0.295
Clarity	0.312	0.304	0.316
Implication	0.439	0.270	**0.442**
Reason	**0.501**	**0.313**	0.504

Table 16: Inter-annotator agreement scores across both subtypes based on combined span sets (FG) in the dev set. Bold indicates the highest score per subtype.

Subtype	**SQ1_Fleiss**	**SQ1_Gwet**	**SQ1_Kripp**	**SQ2_Fleiss**	**SQ2_Gwet**	**SQ2_Kripp**
Alternative	-0.143	0.143	-0.129	0.072	**0.304**	0.004
Assumption	0.157	0.368	0.168	**0.357**	**0.518**	**0.365**
Clarity	0.258	**0.444**	0.267	0.084	0.313	0.095
Implication	**0.501**	**0.626**	**0.507**	0.100	0.325	0.111
Reason	0.421	0.566	0.429	0.489	**0.616**	**0.495**

Table 17: Inter-annotator agreement scores by Socratic Question subtype in the dev set. Metrics include Fleiss' kappa, Gwet's AC1, and Krippendorff's alpha for majority spans (F), minority spans (G), and their combination (FG, not shown here for brevity). Bold indicates the highest score per row.

Subtype	**SQ1_A1_vs_A2**	**SQ1_A2_vs_A3**	**SQ1_A1_vs_A3**	**SQ2_A1_vs_A3**	**SQ2_A1_vs_A3**	**SQ2_A1_vs_A3**
Alternative	0.000	0.000	-0.011	**0.456**	-0.027	-0.011
Assumption	0.467	0.175	0.175	**0.697**	0.203	0.150
Clarity	**0.682**	0.236	0.108	0.121	-0.110	0.216
Implication	0.667	**0.516**	0.323	0.341	-0.052	-0.059
Reason	**0.743**	0.229	0.289	0.710	0.349	0.438

Table 18: Pairwise Cohen's Kappa scores by subtype and annotator pair in the dev set. Scores are shown separately for majority spans (F) and minority spans (G). Bold indicates the highest value per row.

Subtype	Disagree Argument	Disagree Span
Other Stakeholder Perspective	White people, at their best, have no ill will towards Black folks, but their refusal to acknowledge or ignorance of systemic racism is not helping. Other minorities seem to want to be as white as possible, which is sad for them because white is NOT the standard.	White people
Temporal Contrast	Because he has consistently been topping the polls. As for how he fares against Hillary, those numbers have been fluctuating quite a bit. One week it's Trump, another it's Hillary.	One week it's Trump
Vague or Ambiguous Terms	I'd like to start with the definition of "wet": consisting of, containing, covered with, or soaked with liquid (such as water). For something to be wet, it needs to be soaked in a liquid. Since water cannot be soaked in itself, it can never attain the state of being wet.	For something to be wet
Overgeneralized Statement	White people, at their best, have no ill will towards Black folks, but their refusal to acknowledge or ignorance of systemic racism is not helping. Other minorities seem to want to be as white as possible, which is sad for them because white is NOT the standard.	at their best
Implicit Existence	I live in America if that makes any difference to what the response might be. The reason I believe what I said above is: Because of your basic rights as an American, you are able to vote and help make a difference.	Because of your basic rights as an American

Table 19: Examples of disagreements between the annotated ground truth and the model-generated spans, drawn from a random sample of 50 instances in the experiment.

Argument	FSQ Type	Major Interpretation	Minor Interpretation
I'm pro choice. I don't support abortion, but I support a woman's right to do as she pleases with her body. It's not my place to decide. With this said, a common argument I see against pro-lifers is that it is sexist to outlaw abortions, because it's wanting control of a woman's body.	Other Stakeholder Perspective	support a woman's right to do as she pleases with her body.	it is sexist to outlaw abortions, because it's wanting control of a woman's body.
You act like my only reason for disliking the movie is because it didn't surprise me, but then you proceed to question one of my other reasons. Also I don't care what does or doesn't surprise you. We're two different people. The shootouts reminded me of almost every other shootout I've seen in film.	Temporal Contrast	disliking the movie is because it didn't surprise me.	The shootouts reminded me of almost every other shootout I've seen in film.
I get it. Stereotyping is bad. But what the fuck is with every little thing being labelled cultural appropriation? A white friend of mine has dreadlocks, and the one and only reason he has dreadlocks is because he likes the style.	Bias and Subjectivity	every little thing being labelled cultural appropriation?	Stereotyping is bad.

Table 20: Examples of arguments with annotated FSQ types and corresponding major and minor interpretations.

Argument	Conclusion Span	Main Claim	Premises	Assumptions	Reasoning Flaw
Yes, but in prison, you can, in theory, learn from your actions and become a better person. In hell, you will never redeem yourself, and are tortured forever.	In hell, you will never redeem yourself, and are tortured forever.	Hell is a place of eternal punishment without redemption.	In prison, one can learn from their actions and improve.	Redemption is possible in prison but not in hell. The nature of hell is absolute and unchanging.	Assumes that hell is a fixed state without the possibility of change (implicit existence). Overgeneralizes about punishment in hell.
Theft is immoral by definition, though. You can't establish theft as moral.	You can't establish theft as moral.	Theft cannot be morally justified.	Theft is immoral by definition.	The definition of theft inherently includes immorality; there are no circumstances where theft could be moral.	Assumes definitions are universally agreed upon and static; overgeneralizes without considering differing moral frameworks.
Supporting veterans means supporting war. War is rape, torture, and pain.	Supporting veterans means supporting war.	Supporting veterans equates to endorsing war.	Supporting veterans involves endorsing the consequences of war.	Supporting veterans inherently means supporting the actions of war.	Overgeneralization that equates veteran support with endorsement of war, ignoring alternative perspectives.
No, you are making the claim that 50% of cops are bad. That is 400,000 individuals. You can't make that statement and then proceed to tell me to prove you wrong.	You can't make that statement and then proceed to tell me to prove you wrong.	The claim that 50% of cops are bad is unfounded.	Claiming 50% of cops are bad implies many individuals are bad.	The percentage claim is based on misunderstanding of evidence; burden of proof lies with the claimant.	Assumes the claim is false without addressing supporting evidence (implicit existence).

Table 21: Examples of argument structure diagnosis generated using the Chain of Thought (CoT) framework. Each row represents one of four major error patterns: (1) Overgeneralization Bias, (2) Missed Clarification Probing, (3) Incorrect Assumption Reasoning, and (4) Correct Reasoning but Incorrect Span/Type Mapping. The error segments are highlighted in *italic*.

FSQ Type	A1 vs. A2	A1 vs. A3	A2 vs. A3
Other Stakeholder Perspective	0.8214	0.4643	0.5000
Temporal Contrast	0.7857	0.4643	0.4643
Implicit Assumption	0.7857	0.5000	0.5000
Bias and Subjectivity	0.8929	0.6786	0.7143
Vague and Ambiguous Term	0.8571	0.5000	0.5714
Overgeneralization	0.7500	0.7143	0.7500
Questionable Cause–Effect Relationship	0.8571	0.7143	0.7857
Causality Flipped	0.8929	0.8571	0.8929
Lack of Evidence	0.8929	0.7143	0.6786
Weak Evidence	0.8571	0.6786	0.6786

Table 22: Inter-annotator agreement scores across different FSQ types.

Model (3 seeds)	**Micro P**	**Macro P**	**Micro R**	**Macro R**	**Micro F1**	**Macro F1**
Llama-2-13b-hf	0.2268 ± 0.0082	0.1681 ± 0.0111	0.2476 ± 0.0109	0.1864 ± 0.0184	0.2367 ± 0.0082	0.1544 ± 0.0109
Llama-2-7b-hf	0.3137 ± 0.0039	0.2055 ± 0.0035	0.3677 ± 0.0037	0.1893 ± 0.0293	0.3386 ± 0.0039	0.1475 ± 0.0393
Meta-Llama-3-8B-Instruct	0.4029 ± 0.0324	0.1058 ± 0.0178	0.6231 ± 0.1276	0.1406 ± 0.0665	0.4880 ± 0.0643	0.1018 ± 0.0278
Mistral-7B-Instruct	0.2893 ± 0.0689	0.1701 ± 0.0240	0.3314 ± 0.0762	0.2030 ± 0.0352	0.3089 ± 0.0724	0.1662 ± 0.0321
OLMo-2-1124-7B-Instruct	0.4004 ± 0.0255	0.1342 ± 0.0462	0.5615 ± 0.1383	0.2010 ± 0.0031	0.4645 ± 0.0613	0.1275 ± 0.0387
Qwen2.5-14B-Instruct	**0.4275 ± 0.0038**	0.0919 ± 0.0012	**0.7093 ± 0.0111**	0.1140 ± 0.0456	**0.5334 ± 0.0026**	0.0798 ± 0.0038
Qwen2.5-7B-Instruct	0.4053 ± 0.0083	0.1624 ± 0.0293	0.4551 ± 0.0201	0.1780 ± 0.0128	0.4287 ± 0.0136	0.1289 ± 0.0058
Qwen3-14B	0.4049 ± 0.0166	0.1873 ± 0.0482	0.4427 ± 0.0182	0.2096 ± 0.0577	0.4226 ± 0.0096	0.1383 ± 0.0145
Qwen3-8B	0.3211 ± 0.0519	**0.2354 ± 0.0540**	0.3739 ± 0.0590	0.2214 ± 0.0217	0.3455 ± 0.0552	**0.1720 ± 0.0190**
GPT4-1shot	0.3902 ± 0.0144	0.1632 ± 0.0173	0.4944 ± 0.0199	0.1825 ± 0.0074	0.4329 ± 0.0166	0.1302 ± 0.0044
GPT4-5shot	0.3771 ± 0.0162	0.1950 ± 0.0189	0.4239 ± 0.0130	**0.2030 ± 0.0135**	0.3990 ± 0.0141	0.1464 ± 0.0103

Table 23: Performance of models averaged across 3 random seeds. Values are shown as mean ± standard deviation. The best micro and macro values for Precision, Recall, and F1 are highlighted in bold.