Assessing Cognitive Alignment between Large Language Models and Humans in Generating Socratic Questions

Surawat Pothong[1] Paul Reisert[2] Naoya Inoue[1,3] Machi Shimmei[4] Wenzhi Wang[1,3]
Shoichi Naito[1,3,5] 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
{wang.wenzhi.r7, naito.shoichi.t1,}@dc.tohoku.ac.jp, jungmin.choi@riken.jp, kentaro.inui@mbzuai.ac.ae

Abstract

We propose CognitiveQG, a benchmark for evaluating human–model cognitive alignment in critical-thinking–based question generation for argumentative text. CognitiveQG includes 20 annotated instances grounded in Facione's critical-thinking framework, which decomposes the thinking process into several key cognitive skills and is annotated with expert cognitive processes at the individual level. Our results show that while humans and models align in the interpretation of a given argument, divergence appears in deeper stages: models struggle to infer alternative viewpoints and exhibit overconfidence during evaluation. In generated questions, although models can target vague terms and evidence gaps in the argument, they consistently miss counterargument- and bias-focused questions. These findings highlight cognitive misalignment between models and humans, which may pose risks in educational settings.

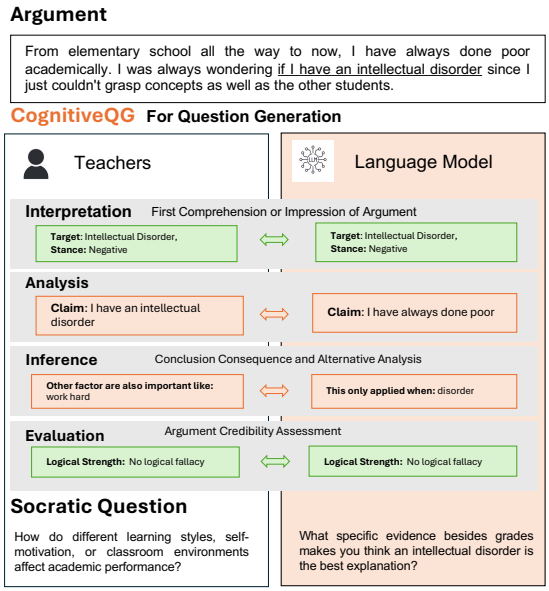

Figure 1: Overview of the CognitiveQG framework for assessing cognitive alignment between humans and LLMs in Focused Socratic Question Generation, motivated by Facione's critical thinking.

1 Introduction

Cognition includes mental processes such as perceiving, remembering, and reasoning. Facione's critical-thinking framework organizes these processes into six skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation [1, 2]. Critically examining argumentative text—including the main claim and its inferences—requires the coordinated use of these six skills. In pedagogical settings that teach critical-thinking skills, a common strategy is to have students write an argument and then have instructors use Socratic questioning to help students reflect on their arguments. When instructors craft these questions, they exercise all six cognitive skills. Notably, human cognitive processes at each stage can vary depending on the downstream task (e.g., question generation) and educational level, even for the same argument. To ensure students receive expert-like thinking through questioning, we argue that LLMs should employ cognitive processes similar to those of humans, yet this alignment remains underexplored.

We introduce CognitiveQG, a benchmark for evaluating cognitive alignment between humans and large language models (LLMs) for Focused Socratic Question Generation [3, 4, 5], annotated with highly educated expert cognition. CognitiveQG builds on insights from argument mining, alignment research, and question generation [6, 7].

This work is guided by the question: How does Facione's critical thinking framework align with LLM cognition in question generation, and where does it misalign across cognitive phases and question types? To address this, CognitiveQG introduces an annotation scheme that operationalizes human cognition across four stages of critical thinking, with questions as output to compare the similarities and dissimilarities between human-generated and model-generated questions by grouping them into FOCUS types to identify patterns in human and model question generation. Figure 1 shows that the process begins with human interpretation of an argument and proceeds through structured subtasks aligned with each cognitive skill. As shown in Figure 1, human annotators exhibit consistent thinking patterns when generating Socratic questions, such as grounding questions in core assumptions and attending to linguistic cues (e.g., conditional markers like if that indicate dependence on specific circumstances).

Our preliminary dataset comprises 20 arguments with 361 annotation steps produced by annotators. We show moderate inter-annotator agreement. Then, we evaluate model alignment by prompting LLMs with the CognitiveQG scheme and comparing their outputs against human annotations. Our findings show that models and humans demonstrate similar cognition in the interpretation phase, but as the process moves to deeper stages, cognitive divergence emerges, particularly in the inference phase, where models show severely limited perspective-taking capability (5.6% "misunderstands" usage for GPT-4 vs. 15.0% for humans), and in the evaluation phase, where models exhibit overconfidence bias (94.4% high-strength ratings vs. 70.0% for humans) and prioritize bias detection over expertise assessment contrasting with humans' balanced approach, revealing systematic misalignment in critical thinking compared to humans.

In summary, we make three key contributions:

1. A **novel schema** for assessing cognitive alignment in question generation;
2. A **preliminary dataset** of 20 annotated instances extending FOCUS, with moderate inter-annotator agreement and partial-match similarity;
3. A **preliminary analysis** comparing model outputs with human cognition, highlighting both similarities and current gaps.

Annotator	Filled / Total	Completion Rate (%)
Annotator1	330 / 361	91.41
Annotator2	331 / 361	91.69
GPT-4.1	292 / 361	80.89
GPT-5.2	328 / 361	90.86

Table 1: Annotation coverage (completion rate) for each annotator and model. Completion Rate = (Filled Fields / Total Required Fields) × 100.

2 Annotation Guideline

2.1 Design Principles for the Scheme

Reflect human and model cognition The scheme captures cognitive processes—not just final outputs—by aligning each step with Facione's critical thinking skills through open-ended and structured subtasks.

Allow easy comparison and mapping Human and model annotations share the same fields and subtasks, enabling direct comparison through alignment metrics such as span overlap, label agreement, and similarity scores.

Provide deeper and structured analysis of cognition By decomposing reasoning into cognitively motivated subtasks, the scheme reveals where human and model thinking diverge, allowing fine-grained analysis of error patterns and thought breakdowns.

2.2 Annotation Scheme and Sub tasks

After establishing the design principles, we developed the annotation scheme based on Facione's critical-thinking framework. Each component aligns with Facione's cognitive skills and is guided by "questions that trigger critical-thinking skills." Table 2 shows that the scheme balances open-ended writing with structured tasks such as label and span selection.

2.3 CognitiveQG

We use the FOCUSdataset as our baseline because it pairs arguments with Socratic questions and span-level weaknesses; each CognitiveQGinstance consists of an argument and its FOCUSquestion. Expert annotators (PhD candidates and postdoctoral researchers in argument mining) label the data using a web interface and a discussion forum to clarify the guideline. We assess quality via coverage and inter-annotator agreement: coverage is high and comparable (about 91%; Table 1), indicating the scheme

Annotation Field	Metric	Score	Metric	Score	Metric	Score	Metric	Score
Initial Understanding	Jaccard	0.373	ROUGE-L	0.571	-	-	BERT	0.755
Stance Target	C'Kappa	0.544	PABAK	0.368	OA	0.684	Gwet's AC	0.593
Paraphrase of core claim	Jaccard	0.251	ROUGE-L	0.483	-	-	BERT	0.767
Knowledge Domain	C'Kappa	0.000	PABAK	0.474	OA	0.737	Gwet's AC	0.701
Core Claim	Jaccard	0.606	ROUGE-L	0.720	-	-	BERT	0.755
Minor Claim	Jaccard	0.347	ROUGE-L	0.456	-	-	BERT	0.516
Premise	Jaccard	0.339	ROUGE-L	0.417	-	-	BERT	0.493
Has Assumption	C'Kappa	-0.048	PABAK	0.158	OA	0.579	Gwet's AC	0.474
Missing Component	Jaccard	0.000	ROUGE-L	0.267	-	-	BERT	0.753
Inductive or Deductive	C'Kappa	0.431	PABAK	0.474	OA	0.737	Gwet's AC	0.521
Consequence	C'Kappa	0.296	PABAK	0.053	OA	0.526	Gwet's AC	0.301
Primary Domain Affected	C'Kappa	0.240	PABAK	-0.053	OA	0.474	Gwet's AC	0.366
Alternative Type	C'Kappa	0.568	PABAK	0.474	OA	0.737	Gwet's AC	0.622
Alternative Keyword	Jaccard	0.065	ROUGE-L	0.228	-	-	BERT	0.492
Inference Score	C'Kappa	0.000	PABAK	-0.158	OA	0.421	Gwet's AC	-0.163
Credibility Factors	C'Kappa	0.166	Exact Match	0.316	Jaccard	0.474	Gwet's AC	0.525
Logical Fallacy	C'Kappa	0.563	PABAK	0.368	OA	0.684	Gwet's AC	0.633
Trustworthiness	C'Kappa	0.373	Exact Match	0.474	Jaccard	0.640	Gwet's AC	0.691
Trust Explanation	Jaccard	0.108	ROUGE-L	0.209	-	-	BERT	0.499

Table 2: Inter-annotator agreement. Text-span similarity uses **non-empty** pairs; categorical fields report Cohen's κ, PABAK, observed agreement (OA), and Gwet's AC.

is feasible. Agreement is measured with Jaccard/ROUGE-L and BERTscore for spans, and Cohen's κ, PABAK [8], OA, and Gwet's AC for categorical fields; Table 2 shows moderate agreement, higher for spans than higher-level categories. We therefore use BERT/Sent2Vec and Gwet's AC as primary metrics, reporting the others as supplementary diagnostics.

3 Baseline Experiment

The baseline experiments are conducted using GPT-5 and GPT-4, both run as standard models with a temperature of 0.3, using the same prompt inspired by the annotation guideline with a one-shot example setting [9].

Table 3 shows the baseline results from GPT-4 and GPT-5, reported field by field. We motivate the choice of evaluation metrics as follows: for span-based evaluation, we use BERTScore [10], ROUGE-L [11], and Jaccard similarity [12, 13]; for categorical classification, we use Gwet's AC [14], macro F1, and micro F1. Gwet's AC is used because it provides a stable and reliable measure of categorical agreement that is robust to class imbalance.

3.1 Alignment Analysis

First, we analyze coverage rates across the annotation scheme. Table 1 shows no significant gap, except GPT4, between humans and models, indicating similar completion rates.

Second, we examine cognitive alignment in each process, starting with interpretation. Both humans and models show similar alignment for interpretation, with BERT scores ranging from 0.582-0.695 for target identification and paraphrasing. However, misalignment occurs in domain knowledge classification, where model performance drops. Investigation reveals that humans recognize when arguments discuss general topics requiring no specialized expertise, while models force domain-specific predictions.

Third, in the analysis process, core claim and premise alignment appear similar based on BERT scores (0.525-0.656), showing comparable capability. However, GPT struggles with minor claim detection. Significant misalignment occurs in assumption detection, where models only predict that arguments lack assumptions without providing explanations. Additionally, models confuse inductive and deductive reasoning patterns that humans distinguish clearly.

Fourth, in the inference phase, GPT4 outperforms humans in consequence prediction, with different inter-annotator agreement scores (Gwet's AC: 0.301 vs 0.317), suggesting models align with one annotator's inference approach. For primary domain prediction, models and humans show similar patterns, though models predict "social" slightly more frequently. Critically, GPT-4 shows severe cognitive bias with 77.8% preference for other-factors" versus humans' balanced 50.0%. The severely limited use of

Annotation Field	Metric 1	GPT-4	GPT-5	Metric 2	GPT-4	GPT-5	Metric 3	GPT-4	GPT-5
Initial Understanding	BERT	0.647	0.582	ROUGE-L	0.398	0.388	Jaccard	0.314	0.320
Stance Target	Gwet's AC	0.257	0.326	F1 Macro	0.283	0.354	F1 Micro	0.450	0.550
Paraphrase of core claim	BERT	0.695	0.700	ROUGE-L	0.315	0.294	Jaccard	0.175	0.155
Knowledge Domain	Gwet's AC	0.005	-0.073	F1 Macro	0.027	0.014	F1 Micro	0.100	0.050
Core Claim	BERT	0.634	0.656	ROUGE-L	0.492	0.590	Jaccard	0.407	0.532
Minor Claim	BERT	0.401	0.376	ROUGE-L	0.294	0.248	Jaccard	0.247	0.209
Premise	BERT	0.525	0.587	ROUGE-L	0.450	0.567	Jaccard	0.396	0.488
has Assumption	Gwet's AC	0.563	0.659	F1 Macro	0.250	0.275	F1 Micro	0.600	0.700
Missing Component	BERT	0.000	0.000	ROUGE-L	0.000	0.000	Jaccard	0.000	0.000
Inductive or Deductive	Gwet's AC	0.276	0.063	F1 Macro	0.263	0.270	F1 Micro	0.500	0.550
Consequence	Gwet's AC	0.317	0.137	F1 Macro	0.399	0.296	F1 Micro	0.500	0.400
Primary Domain Affected	Gwet's AC	0.479	0.507	F1 Macro	0.335	0.293	F1 Micro	0.541	0.600
AlternativeType	Gwet's AC	0.282	0.024	F1 Macro	0.246	0.178	F1 Micro	0.450	0.350
Alternative Keyword	BERT	0.293	0.273	ROUGE-L	0.051	0.000	Jaccard	0.010	0.000
Inference Score	Gwet's AC	-0.248	0.527	F1 Macro	0.000	0.249	F1 Micro	0.000	0.650
credibility Factors	Gwet's AC	0.014	0.167	Exact Match	0.053	0.263	Jaccard	0.175	0.351
logicalFallacy	Gwet's AC	0.470	0.328	F1 Macro	0.205	0.198	F1 Micro	0.500	0.400
trustworthiness	Gwet's AC	0.120	0.053	Exact Match	0.211	0.158	Jaccard	0.281	0.386
trustExplanation	BERT	0.426	0.432	ROUGE-L	0.120	0.122	Jaccard	0.053	0.066

Table 3: GPT-4 vs GPT-5. Metric 1 uses BERT for span fields and Gwet's AC for categorical fields. Metrics 2/3 are ROUGE-L and Jaccard for span fields, and F1 Macro/Micro (human-vs-GPT classification) for categorical fields. Highlighting applies only to **BERT** and **Gwet's AC**: green for > 0.7 and red for < 0.25.

misunderstands" categorization (5.6% versus 15.0% for humans) reveals diminished perspective-taking capability. While both use the same framework and generate consistent keywords, humans use "misunderstands" (15.0%) to model opponent viewpoints nearly three times more frequently than GPT-4, indicating that models struggle to authentically adopt alternative perspectives.

In the evaluation phase, humans and models diverge strongly. Inference strength agreement ranges from -0.248 to 0.527, and GPT-4 tends to overrate arguments (94.4% rated as "high" vs. 70.0% for humans), missing subtle weaknesses. For credibility factors, humans more often cite expertise limitations than bias, whereas GPT-4 focuses almost entirely on bias and never flags expertise issues. For trustworthiness, humans remain more skeptical (8 logical, 8 untrusted), while GPT-4 is more lenient (12 logical, 3 untrusted). Trust explanations reflect this pattern: humans highlight expertise and domain limitations, while GPT-4 emphasizes bias and fallacies.

3.2 Question Alignment Analysis

To investigate the Socratic questions generated from critical thinking, we compare the outputs of both the models and humans. The prompt instructs both humans and the models to create five priority questions that should be asked for each argument.

The distributions in Figure 2 show a serious mismatch in what the models prioritize. Humans ask more Other Stakeholder Perspective questions, but the models undergenerate them, which is problematic for critical-thinking instruction. GPT-5 also over-focuses on Vague/Ambiguous Terms (30% vs. 18% for humans) and cause–effect probes, suggesting a bias toward surface issues over deeper reasoning. Both models also under-produce Bias/Subjectivity questions.

4 Conclusion

We evaluate cognitive alignment between humans and models in the critical-thinking process for generating Socratic questions. Our findings reveal that models share similar cognition with humans in the interpretation phase, but their ability to identify hidden assumptions, generate alternative perspectives, and perform evaluation is largely absent. Although models and humans align in their priorities when crafting questions, models tend to omit counterargument- and bias-focused questions derived from the thinking process. We provide insights for educational AI by highlighting where model thinking aligns with humans and where users should be cautious when using model-generated questions.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H00524 and the Nakajima Foundation.

References

[1] Peter A Facione. Critical thinking: What it is and why it counts. millbrae. **California Academic Press. Haziran**, Vol. 13, p. 2009, 1998.

[2] Peter A Facione, et al. Critical thinking: What it is and why it counts. **Insight assessment**, Vol. 1, No. 1, pp. 1–23, 2011.

[3] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 7180–7198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[4] Beng Heng Ang, Sujatha Das Gollapalli, and See-Kiong Ng. Socratic question generation: A novel dataset, models, and evaluation. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 147–165, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[5] Thomas Huber and Christina Niklaus. LLMs meet bloom's taxonomy: A cognitive view on large language model evaluations. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 5211–5246, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[6] Daiki Shiono, Ana Brassard, Yukiko Ishizuki, and Jun Suzuki. Evaluating model alignment with human perception: A study on shitsukan in LLMs and LVLMs. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 11428–11444, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[7] Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In Noam Slonim and Ranit Aharonov, editors, **Proceedings of the 5th Workshop on Argument Mining**, pp. 79–89, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[8] Guanmin Chen, Peter Faris, Brenda Hemmelgarn, Robin L Walker, and Hude Quan. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. **BMC medical research methodology**, Vol. 9, No. 1, p. 5, 2009.

[9] OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023.

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

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

[12] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. **Bull Soc Vaudoise Sci Nat**, Vol. 37, pp. 547–579, 1901.

[13] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. **arXiv preprint arXiv:2006.14799**, 2020.

[14] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. A comparison of cohen' s kappa and gwet' s ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. **BMC medical research methodology**, Vol. 13, No. 1, p. 61, 2013.

A Appendix

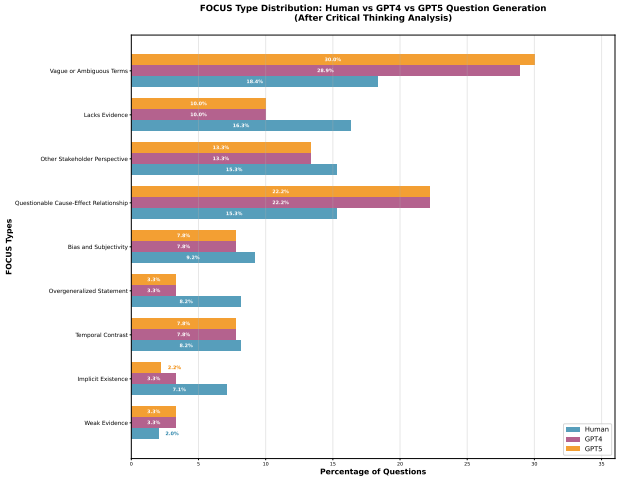

Figure 2: FOCUS Type Distribution: Human vs GPT4 vs GPT5 Question Generation (After Critical Thinking Analysis). The distribution reveals systematic differences in questioning priorities, with models over-emphasizing vague term clarification while under-generating stakeholder perspective and bias-focused questions compared to human experts.