BIS Reasoning 1.0: The First Large-Scale Japanese Benchmark for Belief-Inconsistent Syllogistic Reasoning

Ha-Thanh Nguyen, Hideyuki Tachibana, Chaoran Liu, Qianying Liu, Su Myat Noe, Koichi Takeda, Sadao Kurohashi
Research and Development Center for Large Language Models, NII, Tokyo, Japan

Abstract

We present **BIS Reasoning 1.0**, the first large-scale Japanese dataset of syllogistic reasoning problems explicitly designed to evaluate belief-inconsistent reasoning in large language models (LLMs). Unlike prior resources such as NeuBAROCO and JFLD, which emphasize general or belief-aligned logic, **BIS Reasoning 1.0** systematically introduces logically valid yet belief-inconsistent syllogisms to expose belief bias—the tendency to accept believable conclusions irrespective of validity. We benchmark a representative suite of cutting-edge models—including OpenAI GPT-5 variants, GPT-4o, Qwen, and prominent Japanese LLMs—under a uniform, zero-shot protocol. Reasoning-centric models achieve near-perfect accuracy on **BIS Reasoning 1.0** (e.g., `Qwen3-32B` ≈99% and `GPT-5-mini` up to ≈99.7%), while `GPT-4o` attains around 80%. Earlier Japanese-specialized models underperform, often well below 60%, whereas the latest `llm-jp-3.1-13b-instruct4` markedly improves to the mid-80% range. These results indicate that robustness to belief-inconsistent inputs is driven more by explicit reasoning optimization than by language specialization or scale alone. Our analysis further shows that even top-tier systems falter when logical validity conflicts with intuitive or factual beliefs, and that performance is sensitive to prompt design and inference-time reasoning effort. We discuss implications for safety-critical domains—law, healthcare, and scientific literature—where strict logical fidelity must override intuitive belief to ensure reliability.

Keywords: Belief-inconsistent reasoning, syllogistic reasoning, Japanese language models, logical benchmarking, dataset evaluation

1. Introduction

Large language models (LLMs) have demonstrated remarkable performance on various natural language tasks, yet ensuring reliable logical reasoning remains an open challenge (Morishita et al., 2024; Nguyen et al., 2023). This issue is particularly critical in high-stakes domains such as law, healthcare, and scientific research, where even subtle reasoning errors can lead to severe consequences (Yan et al., 2025). Alarmingly, recent studies show that the persuasive fluency of models like GPT-4 can deceive users into trusting incorrect conclusions (Bajpai et al., 2024). Hence, rigorously evaluating and enhancing LLM reasoning accuracy is crucial before deploying them in domains demanding strict logical rigor.

A major concern in current LLM research is their susceptibility to human-like cognitive biases, notably the belief bias – accepting conclusions aligned with prior beliefs regardless of logical validity. This bias poses significant risks in applications requiring impartial logic. For example, Dasgupta et al. (2022) demonstrated that LLMs frequently endorse logically invalid arguments simply because their conclusions appear believable, highlighting a critical vulnerability.

Existing benchmarks for evaluating logical reasoning in LLMs exhibit several limitations. Most influential reasoning benchmarks are predominantly English-based (Li et al., 2023; Qin et al., 2019; Frohberg and Binder, 2022), creating a significant evaluation gap for languages such as Japanese. Although some Japanese datasets exist, like JFLD (Morishita et al., 2024), which tests formal logic isolated from real-world knowledge, and NeuBAROCO (Ozeki et al., 2024), which covers multiple biases but has limited belief-inconsistent content, no dedicated Japanese-language dataset explicitly targets belief-inconsistent reasoning.

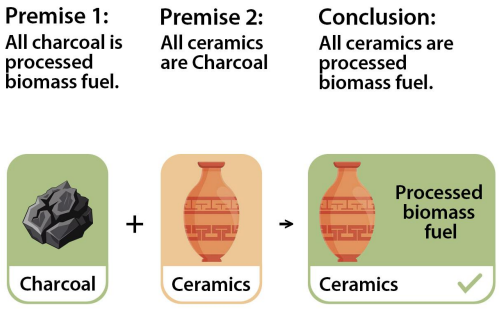

Figure 1: A translated example from the BIS Dataset illustrating a belief-inconsistent syllogism: although the conclusion is logically valid, it contradicts common real-world beliefs.

To address this gap, we introduce **BIS Reasoning 1.0** – the first Japanese dataset designed explicitly for assessing belief-inconsistent syllogistic reasoning in LLMs. BIS (Belief-Inconsistent Syllogisms) aims to evaluate whether models can

Corresponding email: nguyenhathanh@nii.ac.jp

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2211–2219
11–16 May 2026.

uphold logical validity when correct conclusions conflict with typical beliefs or factual knowledge. Fig. 1 shows a representative example from the BIS Dataset, illustrating how logically valid conclusions can defy intuitive beliefs. Specifically, our contributions include:

1. **BIS Reasoning 1.0 Dataset:** We present a carefully curated collection of Japanese syllogistic reasoning problems explicitly constructed to challenge LLMs with logically valid conclusions that contradict common beliefs. This dataset enables the first targeted evaluation of belief-inconsistent reasoning capabilities in Japanese LLMs.

2. **Comprehensive Evaluation of Leading LLMs:** We benchmark state-of-the-art models – including OpenAI GPT, Claude, Qwen and prominent Japanese LLMs – under standardized conditions, providing the first systematic comparison of their performance on belief-inconsistent reasoning in Japanese.

3. **Detailed Analysis of Performance and Bias:** Our analysis identifies significant performance gaps, highlighting that even advanced LLMs struggle disproportionately with belief-inconsistent problems. We quantify these biases, investigate specific syllogistic structures prone to errors, and examine how prompts and CoT (chain-of-thought) approaches affect reasoning accuracy.

4. **Implications for Reliability in Real-World Applications:** We discuss critical implications for deploying LLMs in safety-critical domains. **BIS Reasoning 1.0** reveals vulnerabilities that standard benchmarks typically overlook, providing crucial insights for improving logical consistency and objectivity in real-world scenarios like law, medicine, and scientific research.

Overall, **BIS Reasoning 1.0** contributes significantly to understanding and improving logical reasoning in Japanese-language LLMs. By explicitly evaluating belief-inconsistent reasoning, this work advances efforts toward creating reliable, bias-resistant models suitable for deployment in high-stakes environments.

2. Related Work

Evaluating the logical reasoning abilities of large language models (LLMs) has become a key research focus. Benchmarks like ReClor (Yu et al., 2025) and LogiQA (Liu et al., 2023) use multiple-choice logic problems derived from standardized exams to test inference beyond surface semantics. Despite advances such as chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) and neuro-symbolic modeling, LLMs still struggle to match human performance on tasks requiring rigorous logic.

LLMs not only make logical errors but also exhibit human-like cognitive biases. One well-studied bias is belief bias – the tendency to accept conclusions that align with prior beliefs regardless of logical validity (Evans et al., 1983). Studies have shown that LLMs are more accurate on belief-consistent reasoning tasks and frequently misjudge belief-inconsistent syllogisms (Ando et al., 2023; Ozeki et al., 2024).

Instruction tuning and RLHF can amplify these tendencies. For instance, models like GPT-4 and Claude, while highly fluent, may reinforce belief-aligned reasoning due to human preferences during fine-tuning (Bai et al., 2022). This further underscores the need for benchmarks that reveal latent cognitive biases and test reasoning under belief-conflicting conditions.

In the Japanese language, recent efforts have produced datasets for logical reasoning evaluation, yet limitations remain. JFLD (Morishita et al., 2024) focuses on formal deductive reasoning using artificially constructed propositions to isolate logic from world knowledge. While large in scale and diverse in structure, its use of semantically unnatural sentences and synthetic vocabulary prevents assessment of reasoning in realistic settings. JaNLI (Yanaka and Mineshima, 2021) and JAMP (Sugimoto et al., 2023) explore adversarial and temporal inference respectively, but they do not test belief-inconsistent reasoning and lack syllogistic structure. NeuBAROCO (Ozeki et al., 2024) closely aligns with our goals, exploring belief bias in syllogistic reasoning. However, it falls short as a comprehensive benchmark: the Japanese subset contains fewer than 800 examples for the NLI task and under 100 for the multiple-choice format, and it does not exclusively target belief-inconsistent reasoning. In contrast, **BIS Reasoning 1.0** offers a focused and large-scale evaluation specifically designed to test logical robustness under belief-conflicting conditions.

These observations point to a significant gap: current Japanese benchmarks either employ unnatural representations, ignore belief-inconsistent logic, or lack scale and coverage of syllogistic forms. While prior work has identified belief bias in LLMs, comprehensive datasets for evaluating such bias in Japanese syllogistic reasoning remain scarce. **BIS Reasoning 1.0** directly addresses this gap by providing a focused, large-scale, and naturalistic benchmark specifically designed to test how LLMs handle logically valid conclusions that conflict with intuitive beliefs.

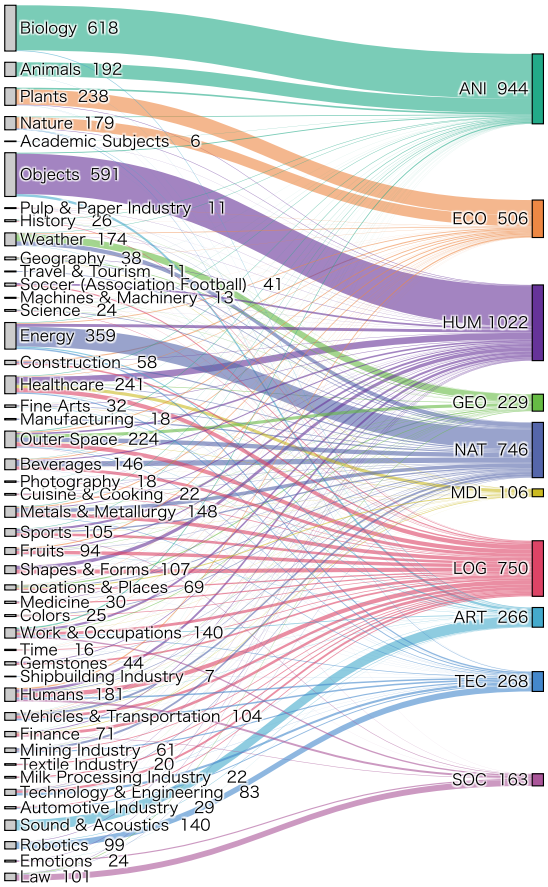

Figure 2: Combined category analysis of raw categories (left) and 10 final categories: animals & living things, ecosystem, human body & senses, geology, natural phenomena, models, logic & structure, arts, technology and society.

3. Dataset Construction

We present **BIS Reasoning 1.0**, a Japanese-language dataset consisting of 5,000 carefully constructed syllogistic reasoning problems. The dataset is specifically designed to test the robustness of logical inference in large language models (LLMs) under conditions of *belief inconsistency*, where logically valid conclusions explicitly contradict widely held commonsense beliefs. **BIS Reasoning 1.0** serves as a diagnostic benchmark for probing whether LLMs can prioritize formal logic over prior-knowledge heuristics in natural language reasoning.

The dataset was developed through a formalized specification process aimed at ensuring both logical rigor and linguistic quality. Each example comprises two premises and one conclusion that is strictly entailed by syllogistic rules, such as classic forms like "*All A are B; All C are A; Therefore, All C are B.*" Crucially, the conclusions are deliberately chosen to conflict with general knowledge, encouraging model errors driven by belief bias. This design isolates the logical reasoning process from superficial semantic plausibility, exposing potential biases embedded in LLM training data.

To ensure linguistic fluency and naturalness, all annotators involved in dataset construction were either native Japanese speakers or individuals with advanced Japanese proficiency. Initially, the dataset covers 46 distinct semantic categories (raw categories), ranging from concrete areas like animals, food, and weather to abstract domains such as law, logic, and emotion (left side of Fig. 2). These detailed raw categories were subsequently consolidated into 10 broader final categories to facilitate interpretability, ensure topic balance, and support higher-level reasoning analysis (right side of Fig. 2).

All examples underwent a two-phase quality assurance (QA) process, initially involving manual review of 10% of examples for iterative feedback, followed by comprehensive review to ensure structural validity, language clarity, and semantic diversity. Issues identified included syntactic violations, duplicated content, ambiguous premises, and category imbalance.

4. Experiments

4.1. General Settings

To evaluate how well LLMs handle logically valid yet belief-inconsistent inferences, we formulate **BIS Reasoning 1.0** as a diagnostic test focusing strictly on logical judgment. Specifically, models must determine if a given conclusion logically follows from two premises, even when the conclusion contradicts intuitive beliefs.

We framed the task as a binary classification using concise, instruction-based prompts in Japanese, asking models to judge logical entailment by answering "Yes" or "No." Each prompt clearly stated two premises and one conclusion, accompanied by a brief system instruction reinforcing the model's logical reasoning role as shown in Table 2. LLMs were not explicitly required nor prohibited to provide explanations. To derive the Yes/No labels from the LLM responses, we primarily utilized heuristic pattern matching. In cases where this approach fell short due to lengthy elucidations by the LLMs, we employed the LLM-as-a-judge strategy, using the `Qwen3-32B` model which has high reasoning capabilities as demonstrated in Table 1.

Accuracy is measured as the proportion of examples for which the model outputs the correct judgment – always "Yes," since all BIS entries are logically valid. This setup ensures that errors stem from reasoning failures, not linguistic ambiguity or

Model	BIS Reasoning 1.0	NeuBAROCO
GPT-5-mini (*medium* reasoning effort)	99.72	91.92
GPT-5-nano (*medium* reasoning effort)	98.84	91.32
gpt-oss-20b	98.56	89.52
GPT-4o	79.54	94.01
GPT-5-nano (*minimum* reasoning effort)	69.22	73.05
GPT-4-turbo	59.48	67.66
Qwen3-32B (w/o thinking)	99.58	94.01
Qwen3-32B (w/ thinking)	99.12	97.60
llm-jp-3-13b-instruct4	84.66	73.65
stockmark-13b	55.90	60.48
llm-jp-3-13b	34.78	36.52
llm-jp-3-13b-instruct3	11.06	22.46
Claude-3-sonnet-20240229 (*deprecated*)	20.34	78.44
Claude-3-opus-20240229 (*deprecated*)	7.18	61.07

Table 1: Accuracy of models on the **BIS Reasoning 1.0** dataset and NeuBAROCO belief-inconsistent syllogisms. All the results are based on the 'Basic' prompt in Table 2.

semantic bias.

All models in the experiments (see Section 4.1) were evaluated under identical zero-shot conditions, using Japanese-language prompts with consistent formatting. No fine-tuning or task-specific adaptation was applied. Evaluation was performed on the full dataset to eliminate sampling variance and enable direct comparison of out-of-the-box reasoning robustness.

Model Configuration We evaluated prominent LLMs spanning both general-purpose and Japanese-specialized categories. The general-purpose group included OpenAI's GPT models (GPT-5-mini, GPT-5-nano, GPT-4o, GPT-4-turbo and gpt-oss-20b) and Alibaba's Qwen3-32B. These models are designed for multilingual tasks and are optimized for general reasoning performance across domains. In contrast, the Japanese-specialized models – llm-jp-3-13b, llm-jp-3-13b-instruct3, llm-jp-3.1-13b-instruct4 (Aizawa et al., 2024), and Stockmark's stockmark-13b – were trained or fine-tuned specifically on Japanese data, representing dedicated efforts to advance native Japanese LLM capabilities.

All prompts were in Japanese (Basic prompt in Table 2), and all 5,000 examples in the dataset were evaluated without sampling. This setup guarantees both fairness and reproducibility across model comparisons.

4.2. Overall Model Performance

Table 1 summarizes the performance of each model on the **BIS Reasoning 1.0** dataset, along with complementary results obtained on over 300 belief-inconsistent syllogistic reasoning samples from the NeuBAROCO benchmark.

The strongest performances were achieved by recent reasoning-optimized models. Both GPT-5-mini and GPT-5-nano with medium (default) reasoning effort achieved near-perfect accuracy on the **BIS Reasoning 1.0** dataset (99.7% and 98.8%, respectively) while maintaining above 91% accuracy on NeuBAROCO. Similarly, the open model Qwen3-32B reached above 99% accuracy with/without explicit thinking-style reasoning.

We also observed a clear trend that more recently developed models tend to exhibit stronger reasoning capabilities; for context, several deprecated early-2024 Claude models (Claude-3-sonnet-20240229 and Claude-3-opus-20240229) were included in the table to provide a broader comparison, achieving 20.3% and 7.2% accuracy on the **BIS Reasoning 1.0** dataset, and 78.4% and 61.1% on NeuBAROCO, respectively.

Performance of Japanese-specific Models The performance of Japanese-specialized models displayed substantial variation across generations. Earlier llm-jp variants and stockmark-13b lagged far behind, typically achieving between 10–60% accuracy, which underscores their limited capacity to override belief-consistent intuitions with formal logic. However, the most recent llm-jp-3.1-13b-instruct4 exhibited a marked improvement, reaching 84.66% accuracy—nearly on par with general-purpose reasoning-oriented models. This sharp rise suggests that newer iterations of Japanese LLMs are increasingly benefiting from

fine-tuning strategies that explicitly emphasize reasoning alignment, rather than focusing solely on linguistic fluency or instruction adherence.

Notably, `llm-jp-3-13b-instruct3` performed significantly below the 50% chance level. Such results do not merely indicate random degradation of performance, but rather a systematic tendency to favor belief-consistent conclusions even when they are logically invalid. During generation, these earlier models often became "confused," producing extended deliberations that acknowledge inconsistencies or factual mismatches without converging on a final decision until reaching the maximum token limit. This behavior reflects a lack of internal mechanisms for reasoning prioritization and highlights the impact of instruction-tuning objectives that insufficiently penalize belief-driven heuristics.

In contrast, the emergence of `llm-jp-3.1-13b-instruct4` demonstrates that reasoning-focused refinement can substantially enhance logical consistency even within Japanese-language models. Its strong performance implies that recent Japanese LLM development has begun to integrate more explicit reasoning objectives—aligning with the broader global trend toward reasoning-centric fine-tuning observed in models such as `GPT-5` and `Qwen3-32B`. This transition underscores a maturing understanding that linguistic naturalness alone is insufficient; robust logical control mechanisms are essential for achieving true reasoning fidelity.

Reasoning Effort `GPT-4o`, which previously dominated several reasoning benchmarks, now shows only moderate performance on **BIS Reasoning 1.0** (79.5%), though it retains high accuracy on NeuBAROCO (94.0%). One plausible explanation is that the prompt format in our evaluation did not explicitly require models to produce elaborate reasoning chains beyond a simple "Yes" or "No" response. Since `GPT-4o` does not allocate an explicit reasoning effort comparable to that of newer reasoning-enabled models, its internal deliberation may have been suppressed, resulting in lower scores on **BIS Reasoning 1.0**. We will revisit this issue in more detail in Section 4.3.

Regarding reasoning effort, a key insight emerges from the `GPT-5-nano` results obtained under different reasoning-effort settings: an API parameter that controls the model's internal resources devoted to reasoning. When this parameter was set to *medium*, the model achieved 98.8% accuracy on **BIS Reasoning 1.0**, but the score dropped sharply to 69.2% when the effort was reduced to *minimum*. This finding clearly indicates that, for this dataset, substantial differences in performance arise depending on how inference-time reasoning capacity is activated and utilized.

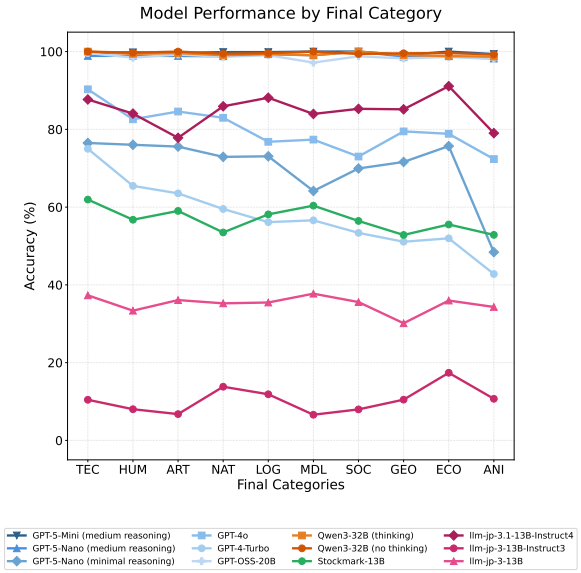

Figure 3: Category-wise reasoning accuracy for each LLM and final-category.

Performance Gap Between BIS Reasoning 1.0 and NeuBAROCO Surprisingly, Anthropic's Claude models underperformed drastically on the **BIS Reasoning 1.0** dataset, with `Claude-3-sonnet` and `Claude-3-opus` achieving only 20.34% and 7.18%, respectively, despite relatively strong performances of 78.44% and 61.07% on NeuBAROCO.

The varying performances between datasets also highlight the sensitivity of model evaluations to task formulation and dataset characteristics. The relatively high NeuBAROCO scores of Claude models contrast sharply with their underperformance on **BIS Reasoning 1.0**, emphasizing the importance of employing diverse benchmarks to comprehensively assess model reasoning capabilities, particularly in challenging belief-inconsistent contexts.

Category-wise Performance Analysis The proposed dataset also allows for assessment by category. Fig. 3 presents the reasoning performance across individual categories. In general, the success of belief-inconsistent reasoning does not significantly depend on the topic for most models examined. Top-tier models (`Qwen3-32B`, `gpt-oss-20b` and `GPT-5` models with medium reasoning effort) simply excel at counterfactual reasoning, with no notable disparities across the various categories. In contrast, while the LLM-jp models exhibit a degree of diversity, their performance levels are generally observed to be lower.

Conversely, GPT models with lower reasoning performance (`GPT-4o`, `GPT-4-turbo` and `GPT-5-nano` with minimal reasoning effort) demonstrate a moderate trend in their strengths and weaknesses. These models perform the best in TEC,

Prompt Type	Prompt Description (Japanese)	Prompt Description (English)
Basic	以下の三段論法を考えてください：結論は前提から論理的に導き出されますか？『はい』か『いいえ』で答えてください。[...]	Consider the following syllogism: Is the conclusion logically derived from the premises? Answer "yes" or "no." [...]
Focus Logic	これは信念の不一致を含む可能性のある論理推論のサンプルです。前提と結論の論理的な関係に焦点を当て、結論が前提から確実に導き出されるかどうかを厳密に判断してください。回答は『はい』または『いいえ』のみでお願いします。[...]	This is a sample of logical reasoning that may involve belief inconsistency. Focus strictly on the logical relationship between premises and conclusion, and determine rigorously if the conclusion follows necessarily from the premises. Answer only "yes" or "no." [...]
Chain-of-Thought	以下の三段論法について、結論が前提から論理的に導き出されるかどうかを段階的に考えてください。まず、前提を分析し、次にそれらの関係性を検討し、最後に結論が論理的に妥当であるかを判断してください。思考プロセスを詳細に記述した後、最終的な回答として『はい』または『いいえ』を明確に示してください。[...]	Consider the following syllogism step-by-step to determine if the conclusion logically follows from the premises. First, analyze the premises, then consider their relationship, and finally judge if the conclusion is logically valid. Provide a detailed reasoning process and clearly state your final answer as "yes" or "no." [...]
Polite	以下の三段論法につきまして、結論が前提から論理的に導き出されるかどうか、ご判断いただけますでしょうか。『はい』または『いいえ』にてお答えいただけますと幸いです。[...]	Could you please determine whether the conclusion logically follows from the premises in the following syllogism? I would appreciate it if you answer with either "yes" or "no." [...]
Casual	この三段論法、どうかな？結論、前提から合ってる？『はい』か『いいえ』で教えて。[...]	What do you think about this syllogism? Does the conclusion match with the premises? Let me know with "yes" or "no." [...]

Table 2: Descriptions of prompt types. “Basic” prompt was used for general evaluation, and other prompts were used for re-evaluating GPT-4o errors.

well in HUM, ART and NAT, yet struggle in prioritizing formal logic over common beliefs in ANI (Animals & Living Things). Although the specific reasons remain unidentified due to the proprietary nature of the GPT models, it is plausible that such trends can be attributed to the topic distribution in the training datasets utilized by the development team.

Interestingly, we can also observe moderate similarity between `llm-jp-3-13B-Instruct3` and `llm-jp-3.1-13B-Instruct4`. Their relative strengths and weaknesses correlate in a comparable manner, which may indicate the presence of shared belief biases in their training corpora.

4.3. Detailed Error Analysis for GPT-4o

To further investigate GPT-4o’s behavior on belief-inconsistent syllogisms, we conducted additional experiments on 100 cases where GPT-4o initially failed. These error samples were reassessed using several carefully designed prompts, each emphasizing different reasoning approaches, linguistic styles, or explicit instructions about the belief-inconsistent nature of the task. Fig. 4 summarizes the accuracy results across these varied prompts.

Prompt Impact Analysis The prompt emphasizing an explicit *chain-of-thought* (CoT) reasoning strategy yielded the highest accuracy improvement (87%) among the previously failed samples. This result clearly demonstrates GPT-4o’s latent reasoning capabilities, significantly activated when explicitly guided through structured logical steps.

Similarly, the *focus_logic* prompt, explicitly mentioning the possibility of belief inconsistency and

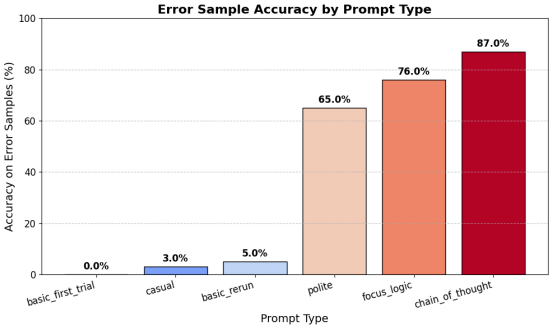

Figure 4: Error sample accuracy by prompt type for GPT-4o.

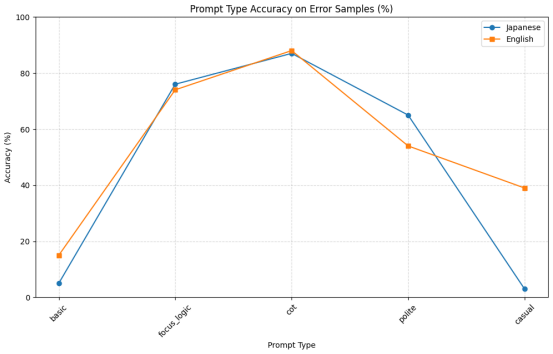

Figure 5: Prompt type accuracy on error samples (%) – retest with English prompts.

urging strict logical evaluation, substantially improved GPT-4o’s accuracy (76%). This finding suggests that GPT-4o is sensitive to explicit instructional framing and context-setting, effectively reducing its reliance on superficial plausibility heuristics when directed accordingly.

Conversely, informal prompts (*casual*) and sim-

ple instruction (*basic*) achieved extremely low recovery rates (3% and 5%, respectively). These prompts appear insufficient in addressing GPT-4o's belief bias, indicating that the model defaults to commonsense heuristics without clear guidance.

The polite prompt yielded a moderate accuracy improvement (65%), suggesting linguistic politeness cues may moderately encourage GPT-4o to engage in deeper reflection or careful reasoning, though less effectively than explicitly logical or CoT framing.

Repeating this experiment with English-language prompts while keeping the Japanese data constant yielded a similar performance pattern, as illustrated in Fig. 5. However, the accuracy gaps were less pronounced compared to the Japanese prompts. This narrower gap in performance likely stems from GPT-4o being extensively trained in English, enhancing its robustness to varied prompt styles. Nevertheless, this confirms the key insight: prompt design significantly influences GPT-4o's performance on challenging tasks like belief-inconsistent syllogisms.

Table 2 provides detailed descriptions of each prompt type used in the error re-evaluation.

Implications for Model Deployment These findings demonstrate the substantial impact of prompt engineering on GPT-4o's logical reasoning capabilities, particularly in overcoming belief bias. Although GPT-4o already performs strongly relative to other models, explicit prompting – such as instructing it to follow a structured reasoning process or clearly signaling the presence of belief-inconsistent content – markedly enhances its logical consistency. Consequently, when deploying LLMs, especially in contexts demanding precise and unbiased logical inference, strategic prompt design is crucial. Clear, structured instructions, emphasizing logical rigor and explicitly guiding step-by-step reasoning, can significantly mitigate intuitive biases inherent to the model, thus ensuring more reliable and accurate outcomes.

5. Discussion

Our findings reveal persistent weaknesses in belief-inconsistent reasoning across LLMs. Models optimized for reasoning, such as Qwen3-32B and GPT-5, maintain high logical fidelity, while Japanese-specialized or alignment-heavy models often favor believable but invalid conclusions. This confirms that linguistic fluency and reasoning robustness are distinct capabilities. However, the latest llm-jp-3.1-13b-instruct4 marks a clear improvement over earlier llm-jp variants, indicating that recent Japanese models are beginning to incorporate stronger reasoning objectives. This trend aligns with the global movement toward reasoning-centric fine-tuning observed in post-2024 LLM generations.

Prompt design and inference-time reasoning effort strongly affect outcomes. Explicit logical instructions or chain-of-thought prompting significantly reduce belief bias, whereas minimal reasoning settings lead to sharp accuracy drops. Alignment-focused training, as seen in some Claude models, may further suppress acceptance of counterintuitive yet valid conclusions.

Model scale alone does not ensure logical reliability. Instead, reasoning-oriented training and architecture play a decisive role. These insights emphasize the need for benchmarks like **BIS Reasoning 1.0** that expose belief bias and test logic over intuition. For reliable deployment in law, healthcare, and research, LLMs must be evaluated and optimized for strict logical consistency, not just fluency or alignment.

6. Limitations

While **BIS Reasoning 1.0** reveals important insights into the reasoning capabilities of LLMs, this study has several limitations that should be considered. First, the evaluation focuses exclusively on syllogistic reasoning. While syllogisms offer a controlled and interpretable format, they represent only one class of logical reasoning. The results may not generalize to other reasoning forms such as causal inference, probabilistic reasoning, or multi-hop deductive chains.

Second, our evaluation relies on a single prompt design in a zero-shot setting. While this approach offers a consistent testbed, it may not fully capture the capabilities of models that perform better under more advanced prompting strategies or tailored task formulations. Prompt sensitivity remains an open variable, and performance may vary under alternative instructions or reasoning scaffolds.

Third, our scoring metric is binary and does not account for partially correct reasoning or near-misses. Models that correctly identify the logical structure but misword the conclusion, or those that reason correctly but fail to flag belief conflict, are treated the same as entirely incorrect responses.

In terms of model coverage, while we include both general-purpose and Japanese-specialized LLMs, our selection remains limited. Notably, open-weights models outside of the LLM-jp and Stockmark ecosystems, as well as mid-sized multilingual models, are not represented in this evaluation.

Lastly, model capabilities are evolving rapidly. The results presented reflect the state of model behavior at a specific point in time (early-to-mid

2025), and future model updates could yield different performance profiles.

7. Conclusion

We introduced **BIS Reasoning 1.0**, the first large-scale Japanese benchmark explicitly designed to evaluate belief-inconsistent syllogistic reasoning in large language models. Our experiments show that even advanced systems still struggle when logical validity conflicts with intuitive or factual beliefs, indicating that belief bias remains a fundamental limitation across architectures and training paradigms.

Among all evaluated models, reasoning-optimized systems such as `Qwen3-32B` and `GPT-5` achieved near-human logical consistency, while earlier Japanese-specialized models performed significantly lower despite superior linguistic fluency. The marked improvement of `llm-jp-3.1-13b-instruct4` suggests that recent Japanese LLMs are beginning to integrate reasoning-aligned fine-tuning, echoing global trends in reasoning-centric model design.

BIS Reasoning 1.0 fills a critical gap in Japanese-language reasoning evaluation by providing a natural, belief-challenging benchmark that exposes logical bias and tests reasoning fidelity under cognitive conflict. We hope this work accelerates the development of bias-resistant, logic-grounded LLMs and fosters research into reasoning alignment for safety-critical applications in law, medicine, and science.

8. Bibliographical References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2023. Evaluating large language models with NeuBAROCO: Syllogistic reasoning ability and human-like biases. In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 1–11, Nancy, France. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Prasoon Bajpai, Niladri Chatterjee, Subhabrata Dutta, and Tanmoy Chakraborty. 2024. Can llms replace neil degrasse tyson? evaluating the reliability of llms as science communicators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15895–15912.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

J St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815, Toronto, Canada. Association for Computational Linguistics.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0-an improved dataset for logical reasoning in natural language understanding. *IEEE ACM Trans. Audio Speech Lang. Process.*

Terufumi Morishita, Atsuki Yamaguchi, Gaku Morio, Hikaru Tomonari, Osamu Imaichi, and Yasuhiro Sogawa. 2024. JFLD: A Japanese benchmark for deductive reasoning based on formal logic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9526–9535, Torino, Italia. ELRA and ICCL.

Ha Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023. A negation detection assessment of gpts: analysis with the xnot360 dataset. *arXiv preprint arXiv:2306.16638*.

Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. Exploring reasoning biases in large language models through syllogism: Insights from the neubaroco dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16063–16077.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China. Association for Computational Linguistics.

Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. 2023. Jamp: Controlled Japanese temporal inference dataset for evaluating generalization capacity of language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 57–68, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenwei Yan, Xiangling Fu, Yuxuan Xiong, Tianyi Wang, Siu Cheung Hui, Ji Wu, and Xien Liu. 2025. LLM sensitivity evaluation framework for clinical diagnosis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3083–3094, Abu Dhabi, UAE. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2025. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.