Can Large Language Models Robustly Perform Natural Language Inference for Japanese Comparatives?

Yosuke Mikami[1,2] **Daiki Matsuoka**[1,2] **Hitomi Yanaka**[1,2]
[1]The University of Tokyo
[2]Riken
{ymikami, daiki.matsuoka, hyanaka}@is.s.u-tokyo.ac.jp

Abstract

Large Language Models (LLMs) perform remarkably well in Natural Language Inference (NLI). However, NLI involving numerical and logical expressions remains challenging. Comparatives are a key linguistic phenomenon related to such inference, but the robustness of LLMs in handling them, especially in languages that are not dominant in the models' training data, such as Japanese, has not been sufficiently explored. To address this gap, we construct a Japanese NLI dataset that focuses on comparatives and evaluate various LLMs in zero-shot and few-shot settings. Our results show that the performance of the models is sensitive to the prompt formats in the zero-shot setting and influenced by the gold labels in the few-shot examples. The LLMs also struggle to handle linguistic phenomena unique to Japanese. Furthermore, we observe that prompts containing logical semantic representations help the models predict the correct labels for inference problems that they struggle to solve even with few-shot examples.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated high performance across a wide range of tasks, including Natural Language Inference (NLI; Bowman et al. 2015). However, inference with numerical and logical expressions remains challenging for LLMs (She et al. 2023, Liu et al. 2023a, Parmar et al. 2024). In particular, NLI involving comparatives is important, as it requires a proper understanding of such expressions. Indeed, there are English benchmarks focusing on comparatives for pre-trained models and inference systems (Haruta et al. 2022, Liu et al. 2023b).

However, it has not been thoroughly investigated how *robust* LLMs are in handling various types of inference involving comparatives, regardless of the prompt formats or the few-shot example selection. Moreover, there is growing attention to analyzing the robustness of inference in languages that are not dominant in the pre-training data.

Given these motivations, we construct an NLI dataset focusing on Japanese comparatives by creating templates from an existing Japanese NLI dataset and filling in them with words.[1] Using this dataset, we evaluate five LLMs, including both open and commercial models. We analyze how robustly LLMs can perform inference on comparatives regardless of the way prompts are given in zero-shot and few-shot settings. We also compare LLMs with ccg-jcomp[2] (Mikami et al. 2025), a logical inference system for Japanese comparatives.

The experimental results suggest that the prompt formats impact the model behavior in the zero-shot settings, and that the few-shot performance is influenced by the gold labels in the few-shot examples. In addition, prompts with semantic representations from ccg-jcomp can improve model accuracy on problems that remain difficult even with standard few-shot settings.

2 Related Work

In this section, we describe existing datasets that contain inference problems involving comparatives. JSeM (Kawazoe et al. 2017) is a Japanese NLI dataset, constructed from the English NLI dataset FraCaS (Cooper et al. 1996) with some additional problems that cover inference unique to Japanese. The problems are divided into sections based on semantic phenomena, including comparatives, which allows us to evaluate the strengths and weaknesses of models with respect to individual phenomena. However, since JSeM is limited in vocabulary and small in scale, we create templates from the dataset

[1] Our dataset is available on https://github.com/ynklab/comparativeNLI_dataset

[2] https://github.com/ynklab/ccg-jcomp

Proceedings of the 16th International Conference on Computational Semantics, pages 137–146
September 22-23, 2025, Licensed under the Creative Commons Attribution 4.0 International License

ID	Category	Template		Example	Label
jsem-570	basic comparative	P	X-wa Y-yori A. X-TOP Y-than A (X is more A than Y)	Taro-wa Hanako-yori omoi. Taro-TOP Hanako-than heavy (Taro is heavier than Hanako)	*unk*
		H	X-wa A. X-TOP A (X is A)	Taro-wa omoi. Taro-TOP heavy (Taro is heavy)	
jsem-577	equative	P	X-wa Y-to onaji-kurai-no $\mathrm{N_A}$-da. X-TOP Y-COM as $\mathrm{N_A}$-COP (X is as A as Y)	Taro-wa Jiro-to onaji-kurai-no omosa-da. Taro-TOP Jiro-COM as weight-COP (Taro is as heavy as Jiro)	*unk*
		H	X-wa Y-yori A. X-TOP Y-than A (X is more A than Y)	Taro-wa Jiro-yori omoi. Taro-TOP Jiro-than heavy (Taro is heavier than Jiro)	
jsem-620	presupposition	P	X-wa Y izyoo-ni A. X-TOP Y than A (X is more A than Y)	Taro-wa Hanako izyoo-ni omoi. Taro-TOP Hanako than heavy (Taro is heavier than Hanako)	*yes*
		H	Y-wa A. Y-TOP A (Y is A)	Hanako-wa omoi. Hanako-TOP heavy (Hanako is heavy)	

Table 1: Examples of categories and their corresponding templates. P and H denote the premise and the hypothesis, respectively. X (Y), A, and $\mathrm{N_A}$ are a proper noun, an adjective, and the noun form of an adjective, respectively. ID indicates the ID in the original JSeM dataset. *unk* stands for the *unknown* label.

and generate new problems by filling in the templates with various words.

CAD (Haruta et al. 2022) is a dataset on English adjectives, comparatives, adverbs, and quantifiers. The authors chose inference examples from linguistic papers and constructed new problems by applying transformations such as adding negation and replacing words. Adjective Scale Probe (Liu et al. 2023b) is a dataset designed to investigate how well language models understand degree semantics. It is semi-automatically generated based on templates. While these studies evaluate the extent to which pre-trained language models perform inference involving comparatives in fine-tuned settings, they do not specifically focus on the robustness of the inference in in-context learning settings. To address this gap, we provide a scalable NLI dataset involving Japanese comparatives based on templates created from existing hand-crafted NLI problems.

3 Dataset Creation

To analyze the extent to which LLMs robustly perform inference involving Japanese comparatives, we create an NLI dataset based on the comparatives section of JSeM. Our dataset construction process is composed of (i) template creation based on JSeM and (ii) problem creation using the templates.

3.1 Template Creation

First, for each problem in JSeM, we manually construct a template containing blanks for adjectives, verbs, numerals, and nouns. Each template has at least one premise and one hypothesis. The gold labels are *yes*, *no*, and *unknown*, corresponding to entailment, contradiction, and neutral, respectively.

The templates are classified into ten categories based on JSeM: basic comparative, equative, clausal comparative, numerical, ambiguous, temporal, quantifier, absolute adjective, presupposition, and superlative. One problem may have multiple categories.

Table 1 shows some examples of categories and their corresponding templates. In what follows, we will refer to a template with its original ID in JSeM, which is shown in the leftmost column. First, jsem-570 involves a basic comparative expression *yori*. Second, jsem-577 targets the equative construction, with its premise meaning that the degree of property A is almost the same for X and Y. Since the premise does not specify which degree is greater, its gold label is *unknown*. Third, jsem-620 is one of the problems focusing on the fact that some Japanese comparative expressions trigger a presupposition (Kubota 2012, Hayashishita 2007). Here, the phrase "izyoo-ni" makes the premise presuppose that Y is A, as a result of which the premise entails the hypothesis.

3.2 Problem Creation

We create new problems by filling in the templates with words corresponding to each part of speech, in order to see whether the models can consistently capture the inference patterns independently of specific content words. The words to be inserted into the templates are carefully chosen by the authors, who are native speakers of Japanese, for their naturalness. In what follows, we detail the concrete procedure for word insertion.

As for a placeholder for an adjective, we insert gradable adjectives in a way that the gold label remains unchanged. More specifically, we avoid using a certain class of adjectives called *absolute adjectives* (Kennedy and McNally 2005), which allow inference from “X is more A than Y” to “Y is A” (e.g., “wet”). Since this property may lead to undesirable changes to the gold label in some templates, we make sure that the inserted word is not an absolute adjective.

In addition, we adopt different strategies depending on whether the placeholder involves the predicative or attributive use. With the predicative use, we insert only adjectives that can take a person as their subject. When the placeholder for an adjective involves the attributive use, in which case the whole template also contains placeholders for a noun and a verb, we construct and apply a list of plausible adjective-noun-verb combinations. More concretely, we first input the template into GPT-4o to generate some adjective-noun-verb combinations. Then, we manually select natural ones from them. To illustrate, consider the template “Taro [verb] a more [adjective] [noun] than Jiro” (for expository purposes, we write the template in English). If the LLM produces the combinations *expensive-car-bought* and *expensive-backpack-drank*, we choose the first output but not the second, since only the first combination results in a semantically-natural sentence when inserted.

Finally, for templates involving numerals, we set a natural range of numerical values compatible with the lexical item for each problem and select the numbers to fill in the templates within that range. For instance, in the template “Taro ate [number] apples,” we choose numbers less than 5.

With these strategies, we generate approximately 60 problems from each template. As a result, the total number of problems is 4304, and the distribution of the gold labels is (*yes*/*no*/*unknown*) = (2524/466/1314).

4 Evaluation of Zero-shot NLI

First, we analyze how consistent the performance of the LLMs is regardless of the prompts in the zero-shot prompt setting, compared with a logical inference system.

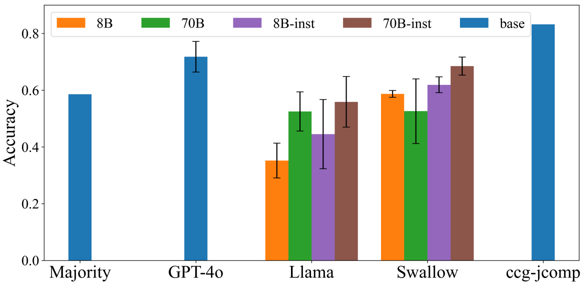

Figure 1: Accuracies on our dataset in the zero-shot setting (average and standard deviation of nine prompts). “Majority” indicates the accuracy achieved by answering *yes*, the most frequent label in the dataset, for all problems.

4.1 Experimental Setting

Models We evaluate five LLMs: GPT-4o[3], Llama-3.1-8B/70B[4] (Llama8B/70B), instruction-tuned Llama-3.1-8B/70B (Grattafiori et al. 2024), Llama-3.1-Swallow-8B/70B[5] (Swallow8B/70B), and instruction-tuned Llama-3.1-Swallow-8B/70B (Fujii et al. 2024). Llama 8B/70B are open-source and multilingual models but do not officially support Japanese. Swallow is a model obtained by performing continual pre-training on Llama with a large Japanese corpus to enhance Japanese language capabilities.

Prompts We conduct experiments using nine different prompts.[6] We create the prompts based on the templates in the FLAN collection (Longpre et al. 2023), which compiles instruction tuning data and methods. The templates contain multiple evaluation instructions, so we use them to examine the models’ robustness to prompts. The details of the prompts are shown in Appendix A.

Logical Inference System We also evaluate ccg-jcomp (Mikami et al. 2025), a logical inference system for Japanese comparatives. This system derives semantic representations of the input sentences and performs theorem proving to judge the entailment relation.

4.2 Results and Discussion

Figure 1 presents the accuracy of each system. As shown, GPT-4o demonstrated the best performance

[3] https://openai.com/index/gpt-4o-system-card/

[4] https://huggingface.co/collections/meta-llama/llama-31-669fc079a0c406a149a5738f

[5] https://huggingface.co/collections/tokyotech-llm/llama-31-swallow-66fd4f7da32705cadd1d5bc6

[6] The experiments were conducted in May and June 2025.

of all the LLMs. Among the open-source models, Swallow, which specifically targets Japanese, outperformed Llama. In addition, larger models performed better, and instruction-tuned models outperformed their non-tuned counterparts of the same size. All models had variations depending on the prompt, and these variations were particularly large for Llama8B-inst and Swallow70B.

LLMs tended to produce incorrect answers even for relatively simple problems. For instance, they often incorrectly answered *yes* to the problems generated from jsem-570 in Table 1, possibly due to the lexical overlap between the premise and the hypothesis. Previous studies have suggested that there are lexical overlap heuristics or order-preserving subset heuristics in pre-trained models performing NLI tasks (McCoy et al. 2019, Yanaka and Mineshima 2021). The experimental result indicates that such heuristics may also be present in LLMs.

We also highlight that the LLMs struggled to handle linguistic phenomena that exist in Japanese but not in English. GPT-4o failed to correctly answer the problems related to presupposition (e.g., jsem-620), which is unique to Japanese comparatives. About Llama and Swallow, they tended to incorrectly answer *yes* to problems such as (1), in which (1a) is the premise and (1b) is the hypothesis.

(1) a. Taro-wa Jiro ka Saburo-yori omoi.
Taro-TOP Jiro or Saburo-than heavy
"Taro is heavier than Jiro or Saburo."

b. Taro-wa Jiro-yori omoi.
Taro-TOP Jiro-than heavy
"Taro is heavier than Jiro."

Here, the gold label is *unknown* because the disjunction in (1a) cannot have narrow scope below *than*. In contrast, its English counterpart does allow such a reading (i.e., Taro is heavier than both Jiro and Saburo), making the label *yes*. It is possible that the errors of the models are due to this difference between the two languages.

5 Evaluation of Few-shot NLI

Next, we analyze the extent to which model predictions change depending on how few-shot examples related to the problem category are given.

5.1 Experimental Setting

For GPT-4o, Llama70B-inst, and Swallow70B-inst, we conduct two types of few-shot experiments with the prompt that showed the highest accuracy in Section 4.

Few_normal For each problem, we give the models one few-shot example generated from the same template. For instance, we show an example generated from jsem-570 to a model, and then evaluate it on a modified version where at least one of X, Y, and A is replaced with a different word.

Few_adversarial For each problem, we give the models an example that is closely related to the problem but has a different gold label. For example, when evaluating a model on jsem-577, we give it an example whose premise is augmented with "Y-wa A" (Y is A). This revision changes the gold label to *yes*. Note that we conduct this experiment only for categories with more than one kind of gold label.

5.2 Results and Discussion

Figure 2 shows the accuracies of the three models in each setting. In FEW_NORMAL, all the models showed improved accuracy compared to the zero-shot setting. In particular, Swallow70B-inst exhibited a significantly larger improvement than the other two. In FEW_ADVERSARIAL, the accuracy of GPT-4o showed a slight improvement, whereas Llama70B-inst and Swallow70B-inst exhibited performance degradation, which was especially notable in Swallow70B-inst.

The results of the two experiments indicate that Swallow70B-inst is highly susceptible to the gold labels of few-shot examples. The other two models effectively leveraged the few-shot examples with the same label, and also were not greatly affected when given examples with a different label.

Although the models avoided many of the errors in the zero-shot experiment with the prompts in FEW_NORMAL, the accuracy did not improve sufficiently in some cases. For example, GPT-4o still failed to correctly answer the problems that require an understanding of presuppositions. In addition, the accuracy of Llama70B-inst for the problems such as (1) was zero.

5.3 Analysis with Semantic Representation Prompts

Inspired by Ozeki et al. (2024), we construct few-shot prompts with not only example problems, but also their semantic representations obtained via ccg-jcomp (see Appendix D for details). We instruct LLMs to generate semantic representations of sentences and then infer the entailment label. We conduct experiments on problems with which each model showed low accuracy even with the

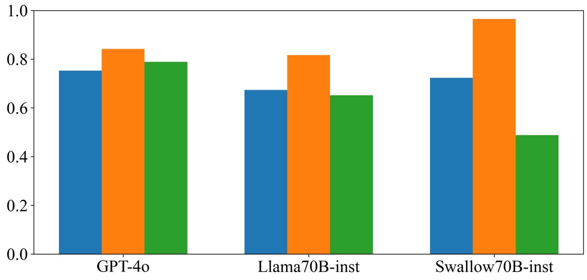

Figure 2: Accuracies of three LLMs in each experimental setting (blue: zero-shot; orange: FEW_NORMAL; green: FEW_ADVERSARIAL)

FEW_NORMAL prompt: namely, presupposition (e.g., jsem-620) for GPT-4o and disjunctive sentences (e.g., (1)) for Llama70B-inst. As a result, the accuracy of GPT-4o and Llama70B-inst increased from 0.049 to 0.230 and from 0.0 to 0.148, respectively. This result suggests that providing semantic representations can improve model performance.

6 Conclusion

In this study, we constructed an NLI dataset focusing on Japanese comparatives, and analyzed how robustly LLMs can perform inference involving comparatives in zero-shot and few-shot settings. The zero-shot experiment revealed that the models' performance varies depending on the prompts, and each model exhibited a distinctive pattern of errors. In the few-shot experiments, we observed that some models, such as Swallow70B-inst, showed a decrease in accuracy when given adversarially designed examples. This observation suggests that some models may be overly sensitive to the specific labels included in the few-shot examples. For problems that the models struggled to solve in the few-shot settings, we found that the accuracy can be improved by making the models predict the semantic representations of the sentences.

Acknowledgments

We thank the three anonymous reviewers for their helpful comments and feedback. This work was partially supported by the Institute for AI and Beyond of the University of Tokyo, and JSPS KAKENHI grant number JP24H00809.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2022. Implementing natural language inference for comparatives. *Journal of Language Modelling*, 10(1):139–191.

J-R Hayashishita. 2007. Izyoo (ni)-and gurai-comparatives: Comparisons of deviation in japanese. *GENGO KENKYU (Journal of the Linguistic Society of Japan)*, 132:77–109.

Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, and Daisuke Bekki. 2017. An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In *New Frontiers in Artificial Intelligence*, pages 58–65, Cham. Springer International Publishing.

Christopher Kennedy and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81:345 – 381.

Yusuke Kubota. 2012. The presuppositional nature of izyoo (-ni) and gurai comparatives: A note on hayashishita (2007). *GENGO KENKYU (Journal of the Linguistic Society of Japan)*, 141:33–47.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023a. Evaluating the logical reasoning ability of chatgpt and gpt-4. *Preprint*, arXiv:2304.03439.

Wei Liu, Ming Xiang, and Nai Ding. 2023b. Adjective scale probe: can language models encode formal semantics information? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13282–13290.

S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yosuke Mikami, Daiki Matsuoka, and Hitomi Yanaka. 2025. Implementing a logical inference system for japanese comparatives. In *Proceedings of the 5th Natural Logic Meets Machine Learning Workshop*. Association for Computational Linguistics. To appear.

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, Bangkok, Thailand. Association for Computational Linguistics.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

Jingyuan Selena She, Christopher Potts, Samuel R Bowman, and Atticus Geiger. 2023. Scone: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. *arXiv preprint arXiv:2305.19426*.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Prompt Templates

Table 2 shows the prompt templates used in Sections 4 and 5. They are translations of the templates in FLAN related to NLI.

B Results by Category in Zero-shot Experiments

Figure 3 shows the accuracies of each LLM and ccg-jcomp across categories.

C Errors of LLMs in the Zero-shot Experiments

In addition to the errors described in Section 4.2, the LLMs also failed to correctly answer the problems related to equatives such as jsem-577-1 in Table 1. They tended to answer *no*, which suggests that they interpret the premise as meaning that the degrees of the two people are exactly equal.

D Details of the Experiment with Semantic Representation Prompts

Table 3 shows the instruction and a few-shot example used in Section 5.3. It provides the semantic representations adopted in ccg-jcomp.

As for the experimental results, although the accuracy of Llama 70B Instruct was still low compared to other models, the semantic representations it predicted were correct in most problems. Most of the errors stemmed from the reasoning step. Table 4 is an example of reasoning errors. The semantic representations are correct; the model successfully interpreted the premise as "Taro is kinder than Jiro, or Taro is kinder than Saburo." However, it incorrectly concluded that the hypothesis follows the premise.

Template	Translation
{premises} 選択肢付きの質問です：上記の段落に基づいて「{hypothesis}」と結論付けることはできますか。 選択肢：含意、矛盾、中立 回答：	({premises} Question with options: Based on the paragraph above can we conclude that “{hypothesis}”? options: entailment, contradiction, neutral answer:)
{premises} この段落に基づいて、下の文が真であると結論付けることはできますか。 {hypothesis} 選択肢：含意、矛盾、中立 回答：	({premises} Based on that paragraph can we conclude that the sentence below is true? {hypothesis} options: entailment, contradiction, neutral answer:)
{premises} 選択肢付きの質問です：以下の結論を導くことはできますか。 {hypothesis} 選択肢：含意、矛盾、中立 回答：	({premises} Q with options: Can we draw the following conclusion? {hypothesis} options: entailment, contradiction, neutral answer:)
{premises} 前の文が与えられたとき、この次の文は従いますか。 {hypothesis} 選択肢：含意、矛盾、中立 回答：	({premises} Does this next sentence follow, given the preceding text? {hypothesis} options: entailment, contradiction, neutral answer:)
{premises} 選択肢：含意、矛盾、中立 問題：次の文を推論できますか。 {hypothesis} 回答：	({premises} options: entailment, contradiction, neutral Question: Can we infer the following? {hypothesis} answer:)
次の段落を読んで仮説が真かどうかを決定してください。最後の選択肢の中から選んでください： {premises} 仮説：hypothesis 選択肢：含意、矛盾、中立 回答は	(Read the following paragraph and determine if the hypothesis is true. Select from options at the end: {premise} Hypothesis: {hypothesis} options: entailment, contradiction, neutral answer:)
テキストを読んで文が真かどうかを決定してください： {premises} 文：{hypothesis} 選択肢：含意、矛盾、中立 回答：	(Read the text and determine if the sentence is true: {premises} Sentence: {hypothesis} options: entailment, contradiction, neutral answer:)
選択肢付きの質問です：以下の文脈から仮説を導くことはできますか。 文脈： {premises} 仮説：{hypothesis} 選択肢：含意、矛盾、中立 回答：	(Question with options: can we draw the following hypothesis from the context? Context: {premises} Hypothesis: {hypothesis} options: entailment, contradiction, neutral answer:)
次の文が真かどうかをその下のテキストに基づいて決定してください。選択肢から選んでください。 {hypothesis} {premises} 選択肢：含意、矛盾、中立 回答：	(Determine if the sentence is true based on the text below. Choose from options. {hypothesis} {premises} options: entailment, contradiction, neutral answer:)

Table 2: Prompt templates used in Section 4

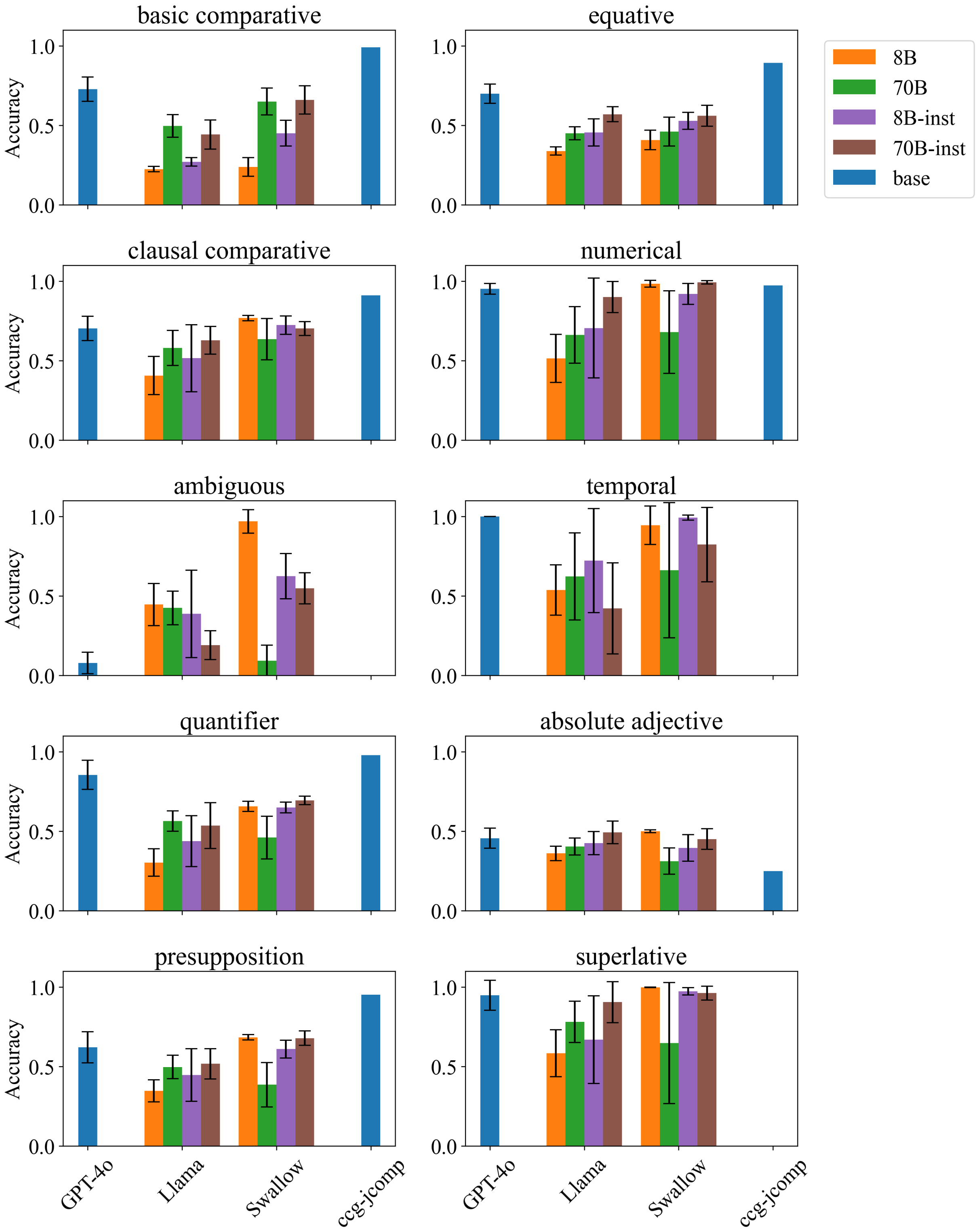

Figure 3: Accuracies of each model and system across categories.

```
与えられた前提と仮説の間の正しい論理関係を決定してください。
- 仮説が前提から論理的に導かれる場合は「含意」と答えてください。
- 前提と仮説が論理的に両立しない場合は「矛盾」と答えてください。
- 「含意」でも「矛盾」でもない場合は「中立」と答えてください。

## 入力
前提：太郎は次郎か三郎より明るい。
仮説：太郎は次郎より明るい。

## 述語論理への翻訳
前提：∃d (明るい(太郎, d) ∧¬明るい(次郎, d)) ∨∃d (明るい(太郎, d) ∧¬明るい(三郎, d))
仮説：∃d (明るい(太郎, d) ∧¬明るい(次郎, d))

## 推論
[その答えに対する理由を説明してください]

## 回答
[含意、矛盾、中立のいずれかを答えてください]

(Determine the correct logical relationship between the given premises and the hypothesis.
- Answer “entailment” if the hypothesis follows logically from the premises.
- Answer “contradiction” if the premises and the hypothesis are logically incompatible with each other.
- Answer “neutral” if the relationship is neither “entailment” nor “contradiction.”

## Input
Premise: Taro is more cheerful than Jiro or Saburo.
Hypothesis: Taro is more cheerful than Jiro.

## Translation into predicate logic
Premise: ∃d (cheerful(taro, d) ∧¬cheerful(jiro, d)) ∨∃d (cheerful(taro, d) ∧¬cheerful(saburo, d))
Hypothesis: ∃d (cheerful(taro, d) ∧¬cheerful(jiro, d))

## Reasoning
[Explain your reasoning for the answer ]

## Answer
[Answer entailment, contradiction, or neutral ])
```

Table 3: Instruction and a few-shot example used in the experiment in Section 5.3. The lower part is an English translation.

Input	前提：太郎は次郎か三郎より優しい。 仮説：太郎は次郎より優しい。 (Premise: Taro is kinder than Jiro or Saburo. Hypothesis: Taro is kinder than Jiro.)
Semantic Representations	前提：∃d(優しい(太郎, d) ∧¬優しい(次郎, d)) ∨∃d(優しい(太郎, d) ∧¬優しい(三郎, d)) 仮説：∃d(優しい(太郎, d) ∧¬優しい(次郎, d)) (Premise: ∃d(kind(taro, d) ∧¬kind(jiro, d)) ∨∃d(kind(taro, d) ∧ ¬kind(saburo, d)) Hypothesis: ∃d(kind(taro, d) ∧¬kind(jiro, d)))
Reasoning	前提が真であれば、太郎は次郎より優しい、または太郎は三郎より優しい。仮説が真であれば、太郎は次郎より優しい。したがって、仮説は前提から論理的に導かれる。 (If the premise is true, then Taro is kinder than Jiro, or Taro is kinder than Saburo. If the hypothesis is true, then Taro is kinder than Jiro. Therefore, the hypothesis is logically derived from the premise.)

Table 4: Example of reasoning errors of Llama70B-inst. Semantic representations and reasoning are the output.