LLMs Struggle with NLI for Perfect Aspect: A Cross-Linguistic Study in Chinese and Japanese

Jie Lu[1] **Du Jin**[1] **Hitomi Yanaka**[1,2]
[1]the University of Tokyo [2]RIKEN
{lujie2001yoshino, dujin728}@gmail.com
hyanaka@g.ecc.u-tokyo.ac.jp

Abstract

Unlike English, which uses distinct forms (e.g., had, has, will have) to mark the perfect aspect across tenses, Chinese and Japanese lack separate grammatical forms for tense within the perfect aspect, which complicates Natural Language Inference (NLI). Focusing on the perfect aspect in these languages, we construct a linguistically motivated, template-based NLI dataset (1,350 pairs per language). Experiments reveal that even advanced LLMs struggle with temporal inference, particularly in detecting subtle tense and reference-time shifts. These findings highlight model limitations and underscore the need for cross-linguistic evaluation in temporal semantics. Our dataset is available at https://github.com/Lujie2001/CrossNLI.

1 Introduction

Recent advances in large language models (LLMs) have raised important questions about the depth and limits of their language understanding. While these models perform well on many standardized benchmarks, most such evaluations are heavily centered on English and often overlook linguistic features that are specific to other languages.

This paper focuses on whether LLMs have human-like understanding of the perfect aspect of punctual verbs in Chinese and Japanese. Although both languages exhibit features that differ from English (See Section 2.1), there has been no systematic investigation of how the perfect aspect is represented or interpreted in these languages within the NLI framework.

To address this gap, we construct a challenging dataset targeting the interpretation of the perfect aspect with punctual verbs (e.g., *die*) in Chinese and Japanese. Our dataset is linguistically motivated, template-based, and contains 1,350 sentence pairs per language.

Our contributions are as follows:

1. We construct a bilingual NLI dataset focused on perfect aspect in Chinese and Japanese.
2. Our analysis reveals that even the current state-of-the-art LLMs repeatedly fail on specific types of problems in our dataset, indicating that they have not fully acquired a robust or generalizable understanding of the perfect aspect in Chinese and Japanese.

2 Background

2.1 Perfect Aspect in Chinese and Japanese

Following Reichenbach (1947), we analyze the temporal interpretation of the perfect aspect by appealing to a three-way temporal distinction: Speech Time (S), Event Time (E), and Reference Time (R). In Reichenbach's framework, different tenses can be interpreted as different relations between S, E, and R. In the past, R occurs before S; in the present, R and S are simultaneous; in the future, R is after S. Furthermore, in the perfect aspect, E always occurs before R, regardless of tense. In Example (1), E ("Hanako graduates") precedes R ("Taro gets PhD"), thus the overall temporal relation of the sentence is ($S < E < R$). Here, $A < B$ signifies that A takes place before B.

(1) When Taro gets his PhD next year, Hanako will have graduated from college.

In addition, the time interval between E and R is specified by adding temporal adverbs in the main clause (e.g., "When Taro gets his PhD next year, Hanako will have graduated from college *3 months ago*").

In English, the perfect aspect is marked differently depending on tense (e.g., had, has, will have). However, Chinese and Japanese do not morphologically vary aspect markers across tenses. Chinese typically uses the marker "*-le*(了)"(Klein et al., 2000; Mochizuki, 1997) to indicate the perfect aspect regardless of tense and relies on temporal adverbs or context to convey temporal information. Japanese expresses the perfect aspect using the auxiliary "*-tei-*(-てい-)"(Kudo, 1995; Iori, 2001), combined with either the past "*-tei-ta*(-てい-た)" or non-past "*-tei-ru*(-てい-る)" form, reflecting its binary tense system.[1]

These aspect markers are also used in other contexts and are not exclusively used to express the perfect aspect. For example, Chinese "*le*" may also serve as a modal particle to express urgency or emotional emphasis (e.g., "太好了!" means "great!"). Because such non-perfect uses dominate everyday usage, we hypothesized that LLMs may struggle to generalize the meaning of the perfect aspect in these languages.

2.2 Temporal NLI Datasets

There are already some NLI datasets that focus on aspect (Kober et al., 2019; Pruś et al., 2024). Kober et al. (2019) introduced a carefully curated NLI dataset with a specific focus on tense and aspect. However, these studies focus only on English.

Several studies (Hu et al., 2020; Yanaka and Mineshima, 2021, 2022; Sugimoto et al., 2024) have addressed NLI tasks involving challenging linguistic phenomena in Japanese and Chinese, but they rarely involve NLI tasks focusing on the perfect aspect. OCNLI (Hu et al., 2020) is a Chinese NLI dataset, and JaNLI (Yanaka and Mineshima, 2021) and JSICK (Yanaka and Mineshima, 2022) are Japanese NLI datasets. However, they scarcely address temporal inference. Jamp_sp (Sugimoto et al., 2024) is a Japanese temporal inference dataset, but it does not systematically investigate inference tasks concerning the perfect aspect.

3 Dataset

Based on tense (past (Pst), present (Pres), future (Fut)) and the presence (t) or absence (None) of a temporal adverb in the main clause discussed in Section 2.1, we designed six Japanese sentence templates based on linguistic literature (Kudo, 1995) and created corresponding Chinese templates. By using these sentence templates as premises and hypotheses, we constructed 30 premise–hypothesis pairs (P, H) of NLI problems for Japanese and Chinese, respectively. Since the perfect aspect with punctual verbs expresses a stable temporal relation in sentences, each (P, H) pair is theoretically expected to have a unique correct label (*entailment* or *non-entailment*) under various punctual verb phrases (See a and b in Example (2)). This enables us to generate a large number of (P, H) pairs with entailment labels by inserting different lexical items semi-automatically.

(2) a. Pres(t): Hanako has already **been dead** for 3 months.
⇒ Pres: Hanako has already **been dead**.

b. Pres(t): Hanako has already **graduated from college** for 3 months.
⇒ Pres: Hanako has already **graduated from college**.

The examples of sentence templates with labels for Chinese are shown in Table 1. Full examples of (P, H) pairs (Table 5) and sentence templates in Chinese and Japanese (Tables 6 and Table 7) can be found in Appendix B.

We manually collected 45 sets of common lexical items (nouns and punctual verbs) and clauses to fill our templates. To minimize semantic influence, the items were designed to maintain one-to-one semantic correspondence between Chinese and Japanese. In total, we generated 1,350 (P, H) pairs for each language, comprising 405 instances labeled as entailment and 945 instances labeled as non-entailment.

Some studies have noted that uncertainty may arise in NLI tasks when temporality is involved (Kober et al., 2019; Pavlick and Kwiatkowski, 2019). To address this issue, we limited the verb types to punctual verbs that denote irreversible changes (e.g., *die*).

To validate the reliability of the sentences, all instances in the dataset underwent rigorous review and were refined by native speakers. Additionally, to ensure labeling reliability, multiple native speakers independently annotated 30 different (P, H) pairs. Under a majority voting scheme, their judgments consistently matched the gold labels, demonstrating high inter-annotator agreement.[2]

[1] Other markers such as "guo" (Chinese), "zhe" (Chinese), and "-ta" (Japanese) may express perfect meanings; however, this paper primarily focuses on the prototypical "-le" and "-tei-".

[2] We collected answers from seven native Chinese speakers and three native Japanese speakers. The average match rate between the Chinese responses and the golden label is 94%, while Japanese is 100%.

Categories	Template Example
P: Pst(t) (E < R < S)	[Event-Past] 的时候, [NP] 已经 [VP] [TIME] 了. 太郎去年取得博士学位 的时候, 花子 已经 死 三个月 了. “When Taro got his PhD last year, Hanako *had* already been dead for 3 months.”
⇒ H_1: Pst (E < R < S)	[Event-Past] 的时候, [NP] 已经 [VP] 了. 太郎去年取得博士学位 的时候, 花子 已经 死 了. “When Taro got his PhD last year, Hanako *had* already been dead.”
⇏ H_2: Pres(t) (E < S = R)	[NP] 已经 [VP] [TIME] 了. 花子 已经 死 三个月 了. “Hanako *has* already been dead for 3 months.”
⇒ H_3: Pres (E < S = R)	[NP] 已经 [VP] 了. 花子 已经 死 了. “Hanako *has* already been dead.”
⇏ H_4: Fut(t) (S < E < R)	[Event-Future] 的时候, [NP] 已经 [VP] [TIME] 了. 太郎明年取得博士学位 的时候, 花子 已经 死 三个月 了. “When Taro gets his PhD next year, Hanako *will have* already been dead for 3 months.”
⇒ H_5: Fut (S < E < R)	[Event-Future] 的时候, [NP] 已经 [VP] 了. 太郎明年取得博士学位 的时候, 花子 已经 死 了. “When Taro gets his PhD next year, Hanako *will have* already been dead.”

Table 1: Template examples of premise and hypothesis sentences in Chinese. In category column, the symbol (t) indicates the presence of a temporal adverb in the main clause. The slot [Event-Past] and [Event-Future] is a subordinate clause containing a temporal expression referring to the past or future, such as “太郎去年取得博士学位” (“Taro got his PhD last year”). ⇒ indicates *entailment* and ⇏ indicates *non-entailment*.

4 Experimental Setup

We conducted experiments on multilingual LLMs and LLMs with enhanced monolingual capability with varying parameter scales. The multilingual models we used include GPT-4 (gpt-4-0613), Claude 3.5 (claude-3-5-sonnet-20241022), Deepseek-V3 (deepseek-chat), and Llama3.1[3] (8B and 70B). The LLMs with enhanced monolingual capability include the Chinese models Qwen3[4] (8B and 32B) and the Japanese models Swallow[5] (9B and 27B). These models cover both multilingual and language-specialized types.

Each model received every premise–hypothesis pair in the corresponding language, together with an instructional prompt that introduces the NLI task and asks whether the premise entails the hypothesis. Model predictions were then compared with gold labels to compute classification accuracy. All experiments were conducted in a zero-shot setting. Our Japanese prompts were adapted from (Sugimoto et al., 2024) and then translated into Chinese by native speakers. The full Chinese and Japanese prompts are provided in Appendix A.

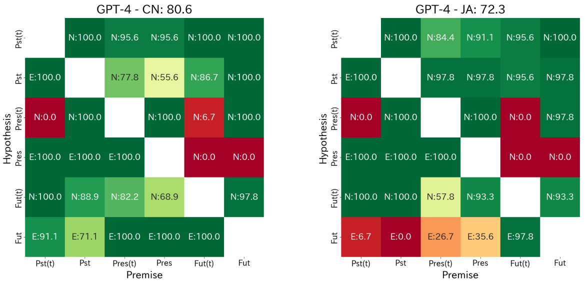

Figure 1: Detailed results from GPT-4 in Chinese and Japanese.The overall accuracy is shown in the title. E/N:number in cells shows the gold label and the accuracy for each (P, H) pair.

5 Results and Discussion

Table 4 shows the average accuracy of tested models on our dataset. Figure 1 shows the detailed results of GPT-4. See Appendix C for detailed results of other models.

Comparison between models As shown in Table 4, Claude 3.5 achieved the best overall performance, outperforming GPT-4—the second-best model—by over 10% in both Chinese and Japanese.

Most models performed similarly on Chinese and Japanese, with accuracy differing by less than 5%. However, Llama-8B was a notable outlier, showing a large performance gap of 26.2% (Chinese: 37.3%, Japanese: 65.6%). Notably, Llama-

[3] hf.co/collections/meta-llama/llama-31-669fc079a0c406a149a5738f

[4] hf.co/collections/Qwen/qwen3-67dd247413f0e2e4f653967f

[5] hf.co/collections/tokyotech-llm/gemma-2-swallow-67f2bdf95f03b9e278264241

Tense of (P, H)	Label	GPT-4	Claude3.5	Deepseek-v3	Llama-8B	Llama-70B	Qwen3-8B	Qwen3-32B	Swallow-9B	Swallow-27B
(Pst(t), Pres(t))	**N**	**0.0/0.0**	**77.8/2.2**	**0.0/0.0**	**0.0/17.8**	**0.0/0.0**	**0.0/0.0**	**0.0/0.0**	**0.0/0.0**	**0.0/0.0**
(Pst(t), Pres)	E	100.0/100.0	100.0/97.8	100.0/100.0	100.0/57.8	100.0/95.6	100.0/100.0	100.0/100.0	95.6/62.2	100.0/62.2
(Pst, Pres(t))	N	100.0/100.0	100.0/100.0	100.0/100.0	0.0/100.0	93.3/80.0	91.1/62.2	51.1/75.6	93.3/62.2	15.6/62.2
(Pst, Pres)	**E**	**100.0/100.0**	**95.6/84.4**	**100.0/100.0**	**100.0/62.2**	**100.0/88.9**	**100.0/100.0**	**100.0/100.0**	**100.0/60.0**	**100.0/60.0**
(Fut(t), Pres(t))	**N**	**6.7/0.0**	**91.1/24.4**	**0.0/0.0**	**0.0/51.1**	**8.9/8.9**	**0.0/0.0**	**0.0/2.2**	**0.0/0.0**	**0.0/0.0**
(Fut(t), Pres)	**N**	**0.0/0.0**	**53.3/20.0**	**0.0/0.0**	**2.2/62.2**	**0.0/37.8**	**0.0/0.0**	**2.2/2.2**	**13.3/4.4**	**6.7/8.9**
(Fut, Pres(t))	N	100.0/97.8	100.0/100.0	100.0/100.0	11.1/95.6	97.8/93.3	97.8/46.7	48.9/82.2	97.8/62.2	51.1/60.0
(Fut, Pres)	**N**	**0.0/0.0**	**86.7/42.2**	**2.2/0.0**	**0.0/51.1**	**0.0/73.3**	**0.0/0.0**	**0.0/33.3**	**0.0/24.4**	**2.2/15.6**

Table 2: Model accuracy (%) when the premise is in the past or future, and the hypothesis is in the present tense. Left side of "/" shows accuracy in Chinese cases, and the right side shows Japanese cases. E indicates entailment labels and N indicates non-entailment labels. The rows in boldface indicate the questions with lexical overlap.

Model	Language	Accuracy (E / N)
Llama-8B	CN	**92.6% / 13.5%**
	JA	45.2% / 71.3%
Qwen3-8B	CN	44.6% / 74.6%
	JA	62.7% / 71.4%
Swallow-9B	CN	46.9% / 80.2%
	JA	32.6% / 48.4%

Table 3: The differences in accuracy between *entailment* and *non-entailment* cases for Llama-8B, Qwen3-8B and Swallow-9B.

Model	Accuracy (CN / JA)
GPT-4	80.6% / 72.3%
Claude3.5	91.5% / 76.7%
Deepseek-v3	77.3% / 70.1%
Llama-8B	37.3% / 65.6%
Llama-70B	75.8% / 72.3%
Qwen3-8B	74.2% / 68.8%
Qwen3-32B	51.4% / 56.6%
Swallow-9B	70.2% / 43.6%
Swallow-27B	54.9% / 42.7%

Table 4: Overall accuracy of each model on our dataset.

8B shows an accuracy gap of nearly 80% between instances labeled as entailment and those labeled as non-entailment (See Table 3). Given that the contexts in which the perfect aspect appears in Chinese are more homogeneous, this result suggests that multilingual models with smaller parameter sizes may struggle to generalize the meaning of the perfect aspect in Chinese.

Furthermore, LLMs with enhanced monolingual capability (Qwen3 and Swallow) exhibit a negative correlation between accuracy and model size. We aim to explore this phenomenon in greater depth in future studies.

Comparison based on linguistic phenomena When the tense of the premise and the hypothesis is the same, models with parameter sizes over 32 billion achieve near-perfect accuracy, while those with lower parameter sizes still struggle with it. Example (3) shows a case of (P: Pst(t), H: Pst).

(3) Pst(t): 太郎上周回到家的时候，花子已经死三天了。
"When Taro came home last week, Hanako had already been dead for 3 days."
$\not\Rightarrow$ Pst: 太郎上周回到家的时候，花子已经死了。
"When Taro came home last week, Hanako had already been dead."

This demonstrates that models with larger parameter sizes can capture the semantic nuances introduced by temporal adverbs.

However, when the tense of the premise and the hypothesis differ, the situation becomes more complex. In cases where the premise is the past or future and the hypothesis is the present (e.g., (P: Fut, H: Pres), we found all models except Claude3.5 consistently predict *entailment* (See Table 2). One possible reason is that the models rely on lexical overlap heuristics mentioned in (McCoy et al., 2019) to solve these problems. In Chinese, since the aspect marker "*le*" applies across all tenses, lexical overlap naturally occurs. In Japanese, sentence pairs where both the premise and the hypothesis use the same perfect aspect marker (e.g., (Fut, Pres)) involve lexical overlap. Examples (4) and (5) illustrate cases where lexical overlap occurs.

(4) Fut: 太郎明年大学毕业的时候，花子已经辞职了。
"When Taro graduates from college next year, Hanako will have already quit her job."
$\not\Rightarrow$ Pres: 花子已经辞职了。
"Hanako has already quit her job."

(5) Fut: 太郎が来年大学を卒業するとき、花子はとっくに会社を辞めている。
"When Taro graduates from college next year, Hanako will already have quit her job."
$\not\Rightarrow$ Pres: 花子は会社を辞めている。
"Hanako has already quit her job."

To our surprise, in Japanese cases where the premise and hypothesis use different tense markers, models still tend to incorrectly predict *entailment*, as illustrated by Example (6), in which "*-tei-ta*" is used in the premise and "*-tei-ru*" in the hypothesis. This result may suggest that the models' low accuracy in handling the perfect aspect in both Chinese and Japanese is not merely a consequence of heuristic biases, but also reflects their incomplete understanding of the semantic distinction between the Japanese perfect aspect marker "*-tei-ta*" and the simple past marker "*-tei-ru*".

(6) Pst(t): 太郎が先週に家に帰ったとき、花子は既に三日前に死んでいた。
"When Taro came home last week, Hanako had already been dead for 3 days."
$\nRightarrow$ Pres(t): 花子は三日前に死んでいる。
"Hanako has already been dead for 3 days."

6 Conclusion

In this study, we presented a bilingual NLI dataset targeting the interpretation of the perfect aspect with punctual verbs in Japanese and Chinese. Our results show that even state-of-the-art LLMs often fail to capture the correct temporal relations, especially when tense and reference times differ between sentences. Our findings highlight the need for evaluation benchmarks that are both linguistically diverse and sensitive to temporal inference.

7 Limitation and Future Work

One limitation of this study is that our experiments deliberately include only punctual, irreversible verbs (e.g., die) to avoid truth-conditional ambiguities. Consequently, our findings do not yet generalize to verbs that occur in perfect-progressive constructions. Extending coverage to such verb classes is left for future work.

Another limitation is that our experiments are only performed in a zero-shot setting. We plan to expand the range of prompt formats used in future experiments.

Finally, some phenomena highlighted in Section 5 remain speculative, most notably the negative scaling trend observed for the Qwen3 series and the Swallow series. We will design additional controlled experiments to validate or refute these hypotheses.

8 Acknowledgement

We would like to express our heartfelt gratitude to Koki Shibata, Tomoki Doi, Taiga Someya, Daiki Matsuoka, and Izumi Konishi for their generous support and invaluable assistance in both the discussions and the writing process of this research. Without them, this work would not have been possible. We also 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

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Larry Moss. 2020. Ocnli: Original chinese natural language inference. In *Findings of EMNLP*.

Isao Iori. 2001. Teirukei, teitakei no imi no toraekata ni kansuru hitoshian (a proposal for the proper understanding of the meaning of the aspectual morpheme -tei in japanese). *Hitotsubashi Daigaku Ryuugakusei Sentaa Kiyoo (Bulletin of the Center for International Education, Hitotsubashi University)*, 4:75–94.

W. Klein, Ping Li, and H. Hendriks. 2000. Aspect and assertion in mandarin chinese. *Natural Language and Linguistic Theory*, 18(4):723–770.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Mayumi Kudo. 1995. *Aspect and Tense System and Text: Temporal Expressions in Modern Japanese*. Japanese Language Research Series, 2nd Series, Vol. 7. Hitsuji Shobo. Originally in Japanese.

R. Thomas 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.

Keiko Mochizuki. 1997. Chuugokugo no paafekuto sou (the perfect aspect in chinese). *Toukyou Gaikokugo Daigaku Ronshuu (Tokyo University of Foreign Studies Journal)*, 55:55–71.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Katarzyna Pruś, Mark Steedman, and Adam Lopez. 2024. Human temporal inferences go beyond aspectual class. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1923, St. Julian' s, Malta. Association for Computational Linguistics.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. Macmillan.

Tomoki Sugimoto, Yasumasa Onoe, and Hitomi Yanaka. 2024. Jamp_sp : A controlled japanese temporal inference dataset considering aspect. *Journal of Natural Language Processing*, 31(2):637–679.

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.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

A Prompts

Chinese:
指示: 从 entailment, non-entailment 中回答前提和假设的关系.不需要给出解释.
限制：
- 如果能够通过逻辑知识或常识性知识从前提推导出假设，则输出 entailment.
- 如果前提成立无法保证假设成立,则输出 non-entailment.
- 前提和假设中没有省略任何时间成分.
- 前提和假设的发话时点为现在.
前提: {premise}
假设: {hypothesis}
答案:

Japanese:
指示: 前提と仮説の関係を entailment,non-entailment の中から回答してください.説明は不要です.
制約：
- 前提から仮説が,論理的知識や常識的知識を用いて導出可能である場合は entailment と出力
- 前提が成り立つとしても仮説が必ずしも成り立たない場合は non-entailment と出力
- 前提と仮説には,時間的な成分を省略していない
- 前提と仮説の発話時を現在とする
前提: {premise}
仮説: {hypothesis}
答え:

English translation:
Instruction: Answer the relationship between the premise and the hypothesis with one of the following: entailment or non-entailment. No explanation is needed.
Constraints:
- If the hypothesis can be deduced from the premise through logical reasoning or common sense knowledge, output entailment.
- If the truth of the premise does not guarantee the truth of the hypothesis, output non-entailment. - There is no omission of any temporal information in both the premise and hypothesis.
- The utterance time for both the premise and hypothesis is the present.
Premise: {premise}
Hypothesis: {hypothesis}
Answer:

B Templates

Table 5 shows all (P, H) templates and their labels in our dataset. Table 6 and Table 7 show Chinese and Japanese sentence templates used to create our dataset.

C Detailed Results

Figure 2 and Figure 3 show detailed results of all models under our dataset.

Premise	Hypothesis	Example	Label
Pst(t)		When Taro got his PhD last year, Hanako had already been dead for 3 months.	
	Pst	When Taro got his PhD last year, Hanako had already been dead.	Entailment
	Pres(t)	Hanako has already been dead for 3 months.	Non-Entailment
	Pres	Hanako has already been dead.	Entailment
	Fut(t)	When Taro gets his PhD next year, Hanako will have already been dead for 3 months.	Non-Entailment
	Fut	When Taro gets his PhD next year, Hanako will have already been dead.	Entailment
Pst		When Taro got his PhD last year, Hanako had already been dead.	
	Pst(t)	When Taro got his PhD last year, Hanako had already been dead for 3 months.	Non-Entailment
	Pres(t)	Hanako has already been dead for 3 months.	Non-Entailment
	Pres	Hanako has already been dead.	Entailment
	Fut(t)	When Taro gets his PhD next year, Hanako will have already been dead for 3 months.	Non-Entailment
	Fut	When Taro gets his PhD next year, Hanako will have already been dead.	Entailment
Pres(t)		Hanako has already been dead for 3 months.	
	Pst(t)	When Taro got his PhD last year, Hanako had already been dead for 3 months.	Non-Entailment
	Pst	When Taro got his PhD last year, Hanako had already been dead.	Non-Entailment
	Pres	Hanako has already been dead.	Entailment
	Fut(t)	When Taro gets his PhD next year, Hanako will have already been dead for 3 months.	Non-Entailment
	Fut	When Taro gets his PhD next year, Hanako will have already been dead.	Entailment
Pres		Hanako has already been dead.	
	Pst(t)	When Taro got his PhD last year, Hanako had already been dead for 3 months.	Non-Entailment
	Pst	When Taro got his PhD last year, Hanako had already been dead.	Non-Entailment
	Pres(t)	Hanako has already been dead for 3 months.	Non-Entailment
	Fut(t)	When Taro gets his PhD next year, Hanako will have already been dead for 3 months.	Non-Entailment
	Fut	When Taro gets his PhD next year, Hanako will have already been dead.	Entailment
Fut(t)		When Taro gets his PhD next year, Hanako will have already been dead for 3 months.	
	Pst(t)	When Taro got his PhD last year, Hanako had already been dead for 3 months.	Non-Entailment
	Pst	When Taro got his PhD last year, Hanako had already been dead.	Non-Entailment
	Pres(t)	Hanako has already been dead for 3 months.	Non-Entailment
	Pres	Hanako has already been dead.	Non-Entailment
	Fut	When Taro gets his PhD next year, Hanako will have already been dead.	Entailment
Fut		When Taro gets his PhD next year, Hanako will have already been dead.	
	Pst(t)	When Taro got his PhD last year, Hanako had already been dead for 3 months.	Non-Entailment
	Pst	When Taro got his PhD last year, Hanako had already been dead.	Non-Entailment
	Pres(t)	Hanako has already been dead for 3 months.	Non-Entailment
	Pres	Hanako has already been dead.	Non-Entailment
	Fut(t)	When Taro gets his PhD next year, Hanako will have already been dead for 3 months.	Non-Entailment

Table 5: All (P, H) templates and their labels. Here, we only present the English translation of one example to illustrate the correspondence between the (P, H) pair and their label in our dataset. As mentioned in Section 3, the label remains unchanged even when different punctual verbs are used.

Category	Template	Example
Pst(t)	[Event-Past]的时候，[NP]已经[VP][TIME]了	田中上周搬家的时候，山本已经合格大学一周了
Pst	[Event-Past]的时候，[NP]已经[VP]了	田中上周搬家的时候，山本已经合格大学了
Pres(t)	[NP]已经[VP][TIME]了	山本已经合格大学一周了
Pres	[NP]已经[VP]了	山本已经合格大学了
Fut(t)	[Event-Future]的时候，[NP]已经[VP][TIME]了	佐藤下个月工作的时候，山本已经合格大学一周了
Fut	[Event-Future]的时候，[NP]已经[VP]了	佐藤下个月工作的时候，山本已经合格大学了

Table 6: Sentence Templates for Chinese.

Category	Template	Example
Pst(t)	[Event-Past]とき、[NP]は[TIME]前にすでに[V-teita]	田中が先月引っ越したとき、山本は一週間前にすでに大学に合格していた
Pst	[Event-Past]とき、[NP]はすでに[V-teita]	田中が先月引っ越したとき、山本はすでに大学に合格していた
Pres(t)	[NP]は[TIME]前に[V-teiru]	山本は一週間前に大学に合格している
Pres	[NP]は[V-teiru]	山本は大学に合格している
Fut(t)	[Event-Future]とき、[NP]は[TIME]前に[V-teiru]	佐藤が来月転職するとき、山本は一週間前に大学に合格している
Fut	[Event-Future]とき、[NP]はとっくに[V-teiru]	佐藤が来月転職するとき、山本はとっくに大学に合格している

Table 7: Sentence Templates for Japanese.

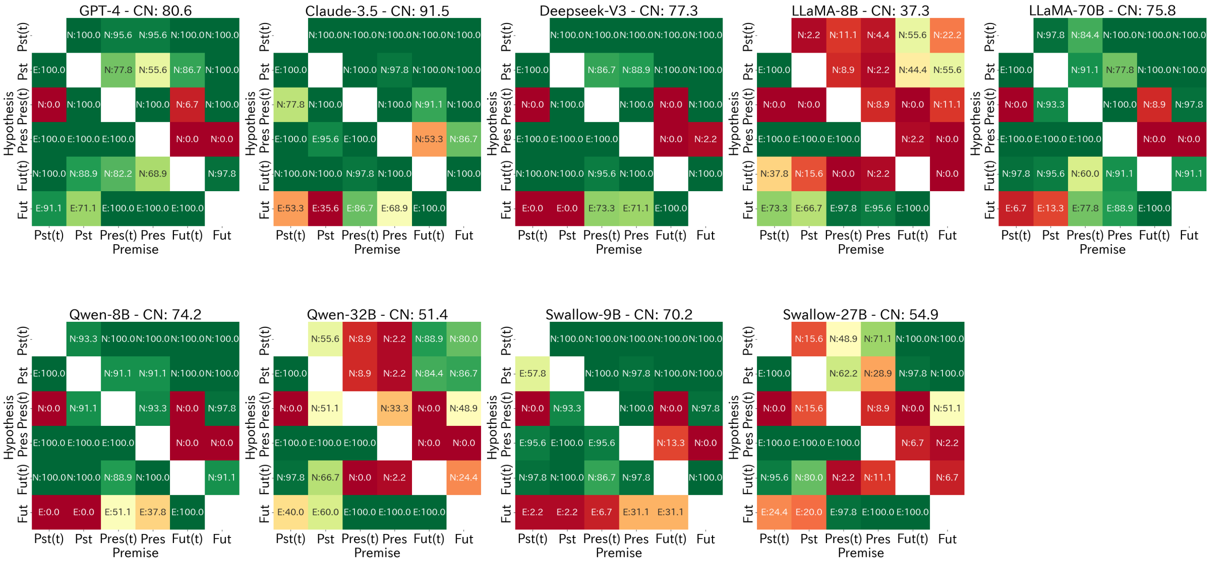

Figure 2: Results on our Chinese dataset. The overall accuracy is shown in the title. E/N:number in cells shows the gold label and the accuracy for each (P, H) pair.

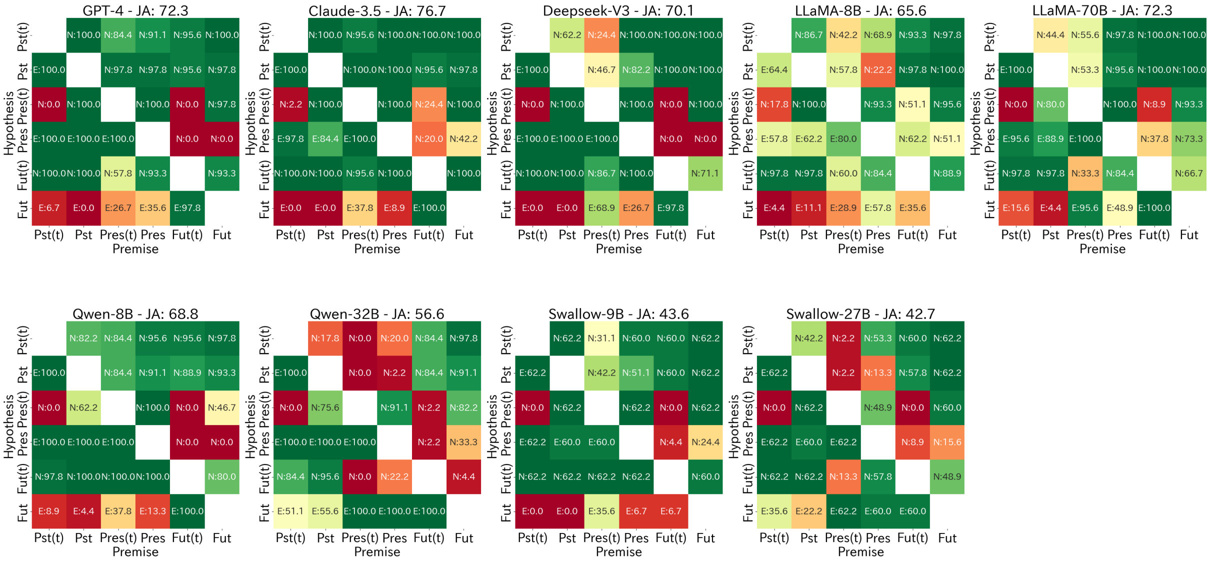

Figure 3: Results on our Japanese dataset. The overall accuracy is shown in the title. E/N:number in cells shows the gold label and the accuracy for each (P, H) pair.