Implementing a Logical Inference System 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

Natural Language Inference (NLI) involving comparatives is challenging because it requires understanding quantities and comparative relations expressed by sentences. While some approaches leverage Large Language Models (LLMs), we focus on logic-based approaches grounded in compositional semantics, which are promising for robust handling of numerical and logical expressions. Previous studies along these lines have proposed logical inference systems for English comparatives. However, it has been pointed out that there are several morphological and semantic differences between Japanese and English comparatives. These differences make it difficult to apply such systems directly to Japanese comparatives. To address this gap, this study proposes ccg-jcomp, a logical inference system for Japanese comparatives based on compositional semantics. We evaluate the proposed system on a Japanese NLI dataset containing comparative expressions. We demonstrate the effectiveness of our system by comparing its accuracy with that of existing LLMs.

1 Introduction

Natural Language Inference (NLI) (Bowman et al. 2015) is the task of determining the entailment relation between premise and hypothesis sentences. In particular, this paper focuses on inferences involving comparative expressions (e.g., *heavier*, where the comparative morpheme *-er* is attached). In (1), for example, the premise (1a) and (1b) entail the hypothesis (1c).

(1) a. John is heavier than Bob.
b. Bob is heavier than 70 kg.
c. John is heavier than 70 kg.
(entailment)

Inferences involving comparatives like (1) are challenging to an NLI system because the system needs to correctly understand the meaning of the quantity expression “70 kg” and the comparative relation between John’s and Bob’s weights.

There are two main approaches to NLI. One is a deep learning (DL)-based approach. Large Language Models (LLMs), such as GPT-4o,[1] have been performing accurately in various tasks, including NLI. However, recent works (She et al. 2023, Liu et al. 2023, Parmar et al. 2024) have pointed out that even such models have difficulties in handling problems involving logical connectives such as negation and quantification. This fact indicates that DL-based models still have room for improvement.

The other approach to NLI is a logic-based approach (Abzianidze 2015, Mineshima et al. 2015, Bernardy and Chatzikyriakidis 2017, Hu et al. 2020, Bernardy and Chatzikyriakidis 2021), in which mathematical logic is utilized to perform NLI involving various logical expressions robustly. In particular, inference systems based on compositional semantics have achieved high performance on NLI problems composed of lexical, syntactic, and semantic phenomena. As for comparatives, Haruta et al. (2022) proposed a logical inference system for English comparatives based on *Combinatory Categorial Grammar* (CCG, Steedman 2000) and *degree semantics* (Cresswell 1976, Klein 1980). However, we cannot apply the system directly to Japanese comparatives because of morphological and semantic differences between Japanese and English comparatives, which we will describe in detail in Section 4.

In this study, we aim to develop a logical inference system for Japanese comparatives based on CCG and degree semantics. Inspired by the logical inference system for English comparatives proposed by Haruta et al. (2022), our system, named ccg-jcomp, compositionally derives the semantic

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

Sentence	Semantic Representation
John is heavy.	$\mathsf{heavy}(\mathsf{john}, \theta)$
John is heavier than 70 kg.	$\exists d.\,(\mathsf{heavy}(\mathsf{john}, d) \wedge d > 70\mathsf{kg})$
John is heavier than all the student.	$\forall x.(\mathsf{student}(x) \to \exists d.(\mathsf{heavy}(\mathsf{john}, d) \wedge \neg\mathsf{heavy}(x, d)))$

Table 1: Basic semantic representations for comparatives

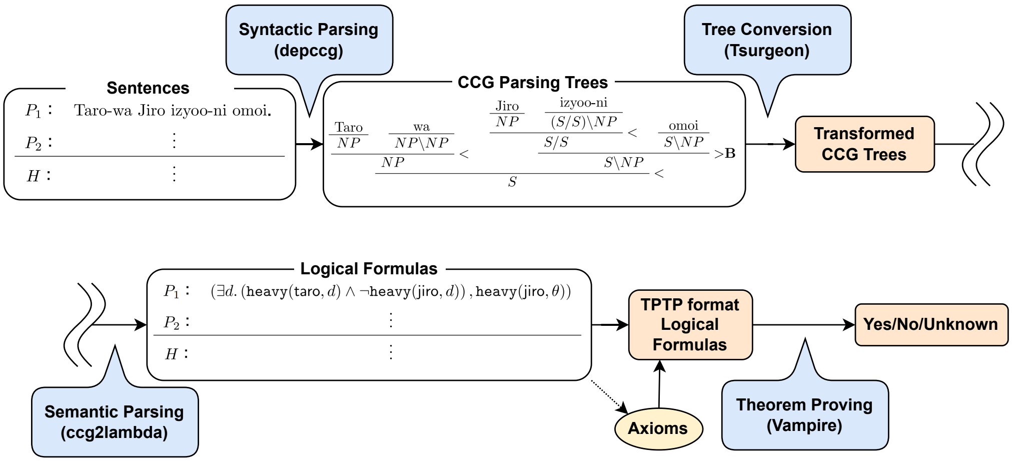
Syntactic Parsing (depccg)
Tree Conversion (Tsurgeon)
Sentences
P_1 : Taro-wa Jiro izyoo-ni omoi.
P_2 :
H :
CCG Parsing Trees
Taro NP
wa $NP\backslash NP$
Jiro NP
izyoo-ni $(S/S)\backslash NP$
omoi $S\backslash NP$
S/S
NP
$S\backslash NP$
S
<
>B
Transformed CCG Trees
Logical Formulas
P_1 : $(\exists d.\,(\mathtt{heavy}(\mathsf{taro}, d) \wedge \neg\mathtt{heavy}(\mathsf{jiro}, d))\,, \mathtt{heavy}(\mathsf{jiro}, \theta))$
P_2 :
H :
TPTP format Logical Formulas
Yes/No/Unknown
Axioms
Semantic Parsing (ccg2lambda)
Theorem Proving (Vampire)

Figure 1: Overview of the proposed system

representations (i.e., the logical formulas representing the sentence meanings) of Japanese sentences through syntactic and semantic parsing and judges the entailment relation using a theorem prover. Further, we implement syntactic and semantic analyses to systematically handle some phenomena specific to Japanese comparatives.

We experiment with JSeM (Kawazoe et al. 2017), a Japanese NLI dataset containing problems involving comparatives. We compare the performance of our system with GPT-4o and some Japanese LLMs. Our experiment shows that our proposed system outperforms all of them in accuracy on the dataset.

Our contributions are as follows:

1. We compositionally derive the semantic representations of Japanese sentences containing some comparative expressions based on CCG and degree semantics.

2. We implement ccg-jcomp, a logical inference system for Japanese comparatives.[2]

3. We demonstrate the effectiveness of our proposed system through experiments on a Japanese NLI dataset involving comparatives.

2 Degree Semantics

In our study, we adopt a theoretical framework called degree semantics, which allows us to analyze the meanings of gradable adjectives and comparatives formally. Its basic idea is to treat a gradable adjective as a binary predicate that takes an entity and a degree as arguments. For instance, "John is d feet tall" can be represented as $\mathsf{tall}(\mathsf{john}, d)$ (for simplicity, we omit units such as "feet").

We handle comparatives following the so-called A-not-A analysis (Seuren 1973, Klein 1982) in degree semantics. According to this analysis, (2a) can be represented as (2b), which means that there exists a degree d such that John's weight is more than or equal to d and Bob's weight is not.

(2) a. John is heavier than Bob.
b. $\exists d.\,(\mathsf{heavy}(\mathsf{john}, d) \wedge \neg\mathsf{heavy}(\mathsf{bob}, d))$

Table 1 shows some other examples of basic constructions involving comparatives and their semantic representations.

3 System Overview

Figure 1 shows the overview of ccg-jcomp, our proposed system. The overall system flow follows Haruta et al. (2022): CCG syntactic parsing, tree conversion, semantic parsing, and theorem proving. In what follows, we describe the details of each

[2] Our system is available for research use at https://github.com/ynklab/ccg-jcomp

Category	Word Type	Semantic Template
NP	common noun	$\lambda E\ N\ F.\exists x.\,(N(E, x) \wedge F(x))$
$S\backslash NP$	positive adjective	$\lambda E\ Q\ N.\ Q(\lambda I.I, \lambda x.N(E, \lambda d.d, \lambda d.d, \lambda t.t, x))$
$S\backslash NP$	negative adjective	$\lambda E\ Q\ N.Q(\lambda I.I, \lambda x.N(E, \lambda d.d, \lambda d.-d, \lambda t.\neg t, x))$
$(S/S)\backslash NP$	yori	$\lambda E\ Q\ V.V(\lambda A\ x.Q(\lambda I.I,\ \lambda y.\exists d.(A(x, d) \wedge \neg A(y, d))))$
$(S/S)\backslash NP$	yori (measure phrase)	$\lambda E\ Q\ V.V(\lambda A\ F\ x.Q(\lambda I.I,\ \lambda y.\exists d.(A(x, d) \wedge F(y < d))))$

Table 2: Examples of basic semantic templates

step, deferring the explanation of the specifics of Japanese comparatives until section 4.

3.1 Syntactic Parsing

First, a tokenizer tokenizes the input sentences, and a CCG parser converts them into CCG trees. CCG is a grammar formalism that assigns a syntactic category to each grammatical expression. The set of syntactic categories is defined recursively as follows: (i) atomic categories: NP (noun phrase), S (sentence), etc., (ii) functional categories: X/Y, $X\backslash Y$ (where X and Y are syntactic categories). Both X/Y and $X\backslash Y$ take the category Y as an argument and return the category X. "/" and "\" indicate that the argument is taken from the right and left, respectively.

CCG parsers generally use CCGbank (Hockenmaier and Steedman 2007) or its modified versions for training, which are not necessarily compatible with comparatives. Thus, the output CCG trees are not always the ones we expect at this point. To deal with this issue, we modify the CCG trees if necessary. Another possible way to modify CCG trees is to revise the CCG parser itself. However, this method is costly because it requires re-training or fine-tuning the CCG parser. Thus, we leave this approach for future work.

3.2 Semantic Parsing

In this step, we assign a semantic representation to each lexical item of the CCG tree based on the semantic templates. Then, the semantic representation of the whole sentence is composed according to the CCG rules. To illustrate, we show two rules below. Some other rules are provided in Appendix A.

- Forward functional application rule

$$\frac{X/Y : f \qquad Y : a}{X : f\,a} >$$

- Backward functional application rule

$$\frac{Y : a \qquad X\backslash Y : f}{X : f\,a} <$$

We set up the semantic templates in order to give semantic representations to the lexical items. Table 2 shows the semantic templates for basic comparative expressions.[3]

Let us proceed to some details of the templates in Table 2, focusing on the function N that appears in the templates for positive/negative adjectives, which we have newly added to handle comparatives.[4] N has five arguments, the first one E being the base form of the adjective, and the fifth one x being the subject of the adjective. Turning to the second argument $\lambda d.d$, it is introduced for the differential comparatives. Consider (3a) for example. In the semantic composition process, this argument becomes $\lambda d.(d + 5)$ as a result of the combination of "5 kg" and the adjective "omoi" (heavy), which leads to the intended semantic representation (3b).

(3) a. Taro-wa Jiro yori 5 kg omoi. (Taro is 5 kg heavier than Jiro.)

b. $\forall d.(\mathsf{heavy}(\mathsf{jiro}, d) \to \mathsf{heavy}(\mathsf{taro}, d + 5))$

The third argument, $\lambda d.d$ (or $\lambda d.-d$), indicates whether the adjective is positive or negative. This allows us to distinguish between (3a) and (4a),

[3] For expository purposes, the semantic templates listed here are simplified from the original ones, which are more complicated in order to handle various expressions.

[4] E (resp. Q) represents the surface form of the word (resp. the generalized quantifier (Barwise and Cooper 1981)).

which contain adjectives of the opposite polarity. For instance, by assuming that “karui” (light) is a negative adjective, we can derive the semantic representation (4b) for (4a), where the argument $\lambda d.{-d}$ corresponds to -5.

(4) a. Taro-wa Jiro yori 5 kg karui. (Taro is 5 kg lighter than Jiro.)

b. $\forall d.\,(\mathsf{light}(\mathsf{jiro}, d) \to \mathsf{light}(\mathsf{taro}, d-5))$

Similarly, the fourth argument, $\lambda t.t$ (or $\lambda t.\neg t$), makes a distinction about the polarity of the adjectives in comparatives with measure phrases. Taking (5) and (6) for example, the arguments $\lambda t.t$ and $\lambda t.\neg t$ correspond to $d > 70$ and $\neg(d > 70)$ in the semantic representations, respectively.

(5) a. Taro-wa 70 kg yori omoi. (Taro is heavier than 70 kg.)

b. $\exists d.\,(\mathsf{heavy}(\mathsf{taro}, d) \wedge d > 70)$

(6) a. Taro-wa 70 kg yori karui. (Taro is lighter than 70 kg.)

b. $\exists d.\,(\mathsf{light}(\mathsf{taro}, d) \wedge \neg(d > 70))$

3.3 Theorem Proving

In this step, we input the logical formulas of the premises and hypothesis obtained in the previous step into an automated theorem prover and judge their entailment relation.

Axioms In order to prove entailment relations, we introduce some axioms. To illustrate, we describe one of the axioms, (CP), which is shown below (here, **A** is an adjective). It corresponds to a basic axiom in degree semantics called Consistency Postulate (Klein 1980).

(CP) $\forall x\, y.\, ((\exists d.\, (\mathbf{A}(x, d) \wedge \neg\mathbf{A}(y, d)))$
$\to \forall d.\, (\mathbf{A}(y, d) \to \mathbf{A}(x, d)))$

Intuitively, this axiom requires that **A** be a predicate such that if the degree of x is greater than the degree of y, then the degree of x is greater than or equal to the degree of y. Using this axiom, we can make inferences such as (1). We give the details of the proof in Appendix B, where we also explain other axioms.

Implementation First, we choose some axioms based on the adjectives in the input sentences and add them as premises. Then, we input the logical formulas of the premises and hypothesis into the automated theorem prover. Given the premises and axioms $P_1, \ldots P_n$ and the hypothesis H, the system output is *yes* (entailment) when $P_1 \wedge \ldots \wedge P_n \to H$ is proven, *no* (contradiction) when $P_1 \wedge \ldots \wedge P_n \to \neg H$ is proven, and *unknown* (neutral) when neither is proven.

4 Challenges in Handling Japanese Comparatives

In this section, we explain some linguistic phenomena specific to Japanese comparatives and how we treat them in this study.

4.1 Absence of Overt Comparative Morphemes

English has overt comparative morphemes, such as *more* and *-er*. On the other hand, Japanese has no such morphemes. The examples (7a) and (7b) illustrate that the adjective “omoi” has the same surface form whether it is used for comparison or not.

(7) a. Taro-wa Jiro yori <u>omoi</u>.
Taro-TOP Jiro than heavy
“Taro is heavier than Jiro.”

b. Taro-wa <u>omoi</u>.
Taro-TOP heavy
“Taro is heavy.”

Although it is possible to give different semantic representations to “omoi” in both sentences, we assign the same semantic representation to simplify the semantic parsing process. Accordingly, we introduce an unpronounced symbol (empty category) to distinguish the semantic representations of the two sentences. Specifically, when there is no comparative expression such as “...yori” and “...izyoo-ni,” we insert an empty category *cmp* of category S/S instead. We introduce the aforementioned comparison criterion θ by assigning the following semantic representation (8) to this empty category.

(8) $\lambda S.S(\lambda A\, x.A(x, \theta))$

This inserted operator also plays a role of matching the types of the semantic representations of “Jiro yori omoi” and “*cmp* omoi” (Figures 2 and 3).

4.2 Equatives

English equative sentences such as (9a) are interpreted as indicating “... is at least as heavy as ... .” Thus, Haruta et al. (2022) represented (9a) as (9b).

(9) a. John is as heavy as Bob.

b. $\forall d.\,(\mathsf{heavy}(\mathsf{bob}, d) \to \mathsf{heavy}(\mathsf{john}, d))$

$$\dfrac{\dfrac{\dfrac{\text{Jiro}}{\begin{array}{c}NP\\ :\lambda P.P(\mathsf{jiro})\end{array}} \qquad \dfrac{\text{yori (than)}}{\begin{array}{c}(S/S)\backslash NP\\ :\lambda Q\,S.S(\lambda A\,x.Q(\lambda y.\exists d.(A(x,d)\wedge\neg A(y,d))))\end{array}}}{\begin{array}{c}S/S\\ :\lambda S.S(\lambda A\,x.\exists d.(A(x,d)\wedge\neg A(\mathsf{jiro},d)))\end{array}}< \qquad \dfrac{\text{omoi (heavy)}}{\begin{array}{c}S\backslash NP\\ :\lambda Q\,N.Q(\lambda x.N(\mathsf{heavy},x))\end{array}}}{\begin{array}{c}S\backslash NP\\ :\lambda Q.Q(\lambda x.\exists d.(\mathsf{heavy}(x,d)\wedge\neg\mathsf{heavy}(\mathsf{jiro},d)))\end{array}}>\mathbf{B}_{\times}$$

Figure 2: A part of semantic composition of (7a)

$$\dfrac{\dfrac{\textit{cmp}}{\begin{array}{c}S/S\\ :\lambda S.S(\lambda A\,x.A(x,\theta))\end{array}} \qquad \dfrac{\text{omoi (heavy)}}{\begin{array}{c}S\backslash NP\\ :\lambda Q\,N.Q(\lambda x.N(\mathsf{heavy},x))\end{array}}}{\begin{array}{c}S\backslash NP\\ :\lambda Q.Q(\lambda x.\mathsf{heavy}(x,\theta))\end{array}}>\mathbf{B}_{\times}$$

Figure 3: A part of semantic composition of (7b)

On the other hand, Japanese equatives merely express that the degrees are close to each other. For instance, (10a) can be true even when Taro's weight is slightly less than Jiro's.

(10) a. Taro-wa Jiro to onaji kurai-no
Taro-TOP Jiro same as-GEN
omosa-da.
weight-COP
"Taro is as heavy as Jiro."

b. Jiro-wa omoi.
Jiro-TOP heavy
"Jiro is heavy."

c. Taro-wa omoi. (entailment)
Taro-TOP heavy
"Taro is heavy."

To handle the meaning of equatives, we propose the following representation (11) for (10a). This intuitively indicates that the difference in weight between Taro and Jiro is less than the constant δ.

$$\begin{aligned}(11)\quad &\forall d_1 d_2.\,((\neg\,(\mathsf{heavy}(\mathsf{taro},d_1)\\ &\leftrightarrow \mathsf{heavy}(\mathsf{jiro},d_1))\\ &\wedge\neg\,(\mathsf{heavy}(\mathsf{taro},d_2)\leftrightarrow\mathsf{heavy}(\mathsf{jiro},d_2)))\\ &\to |d_1-d_2|<\delta)\end{aligned}$$

We also introduce the following axiom (12), which prescribes the relation between θ and δ. Intuitively, this axiom indicates that δ is so small that the truth value of the predicate heavy does not change within the range of δ from θ.

$$(12)\quad \forall x.\,(\mathsf{heavy}(x,\theta-\delta)\leftrightarrow\mathsf{heavy}(x,\theta+\delta))$$

We can make inferences such as (10) using this axiom together with (UP) and (DOWN) (see Appendix B for details).

4.3 Clausal Comparatives

Clausal comparatives are comparatives with subordinate clauses. (13a) is an example of a clausal comparative. We also deal with related sentences such as (13b) and (13c).

(13) a. Taro-wa Hanako-ga katta yori
Taro-TOP Hanako-NOM bought than
takai hon-o katta.
expensive book-ACC bought
"Taro bought a more expensive book than Hanako bought."

b. Taro-wa Hanako-ga katta no yori
Taro-TOP Hanako-NOM bought NO than
takai hon-o katta.
expensive book-ACC bought
"Taro bought a more expensive book than what Hanako bought."

c. Taro-wa Hanako yori takai
Taro-TOP Hanako than expensive
hon-o katta.
book-ACC bought
"Taro bought a more expensive book than Hanako."

We assign the same semantic representation (14) to the three sentences in (13).

$$\begin{aligned}(14)\quad &\exists d.\,(\exists x.\,(\mathsf{book}(x)\wedge\mathsf{expensive}(x,d)\\ &\wedge\exists e.\,(\mathsf{bought}(e)\wedge(\mathsf{Nom}(e)=\mathsf{taro})\\ &\wedge\ (\mathsf{Acc}(e)=x)))\\ &\wedge\neg\exists x.\,(\mathsf{book}(x)\wedge\mathsf{expensive}(x,d)\\ &\wedge\exists e.\,(\mathsf{bought}(e)\wedge(\mathsf{Nom}(e)=\mathsf{hanako})\\ &\wedge\ (\mathsf{Acc}(e)=x))))\end{aligned}$$

Category	Word	Semantic Template
$((NP/NP)/(NP/NP))\backslash(S\backslash NP)$	yori (13a)	$\lambda E\,V\,M.V(\lambda G.\exists d.(M(\lambda A\,x.A(x,d))$ $\wedge\neg M(\lambda A\,x.(A(x,d)\wedge G(x)))))$
$((NP/NP)/(NP/NP))\backslash NP$	yori (13b)	$\lambda E\,Q\,M.\exists d.(M(\lambda A\,x.A(x,d))$ $\wedge\neg M(\lambda A\,x.(A(x,d)\wedge Q(\lambda y.(x=y)))))$
$((NP/NP)/(NP/NP))\backslash NP$	yori (13c)	$\lambda E\,Q\,M\,F\,x.Q(\lambda y.$ $(\exists d.M(\lambda A\,z.(A(z,d)\wedge F(x,z)))$ $\wedge\neg M(\lambda A\,z.(A(z,d)\wedge F(y,z)))))$

Table 3: Semantic templates for clausal comparatives

In order to obtain this semantic representation, we assign different semantic representations to "yori" in each sentence, which are listed in Table 3. Note that the template in the second row includes $\lambda y.(x=y)$, which is necessary to consider the fact that the pronominal "no" is identified with "hon" (book) in (13b).

4.4 Presupposition

Some Japanese comparative expressions have a special semantic content called a presupposition (Kubota 2012, Hayashishita 2007). A presupposition is a type of meaning not affected by entailment-canceling operators such as negation and modals (cf. Potts (2015)). The predicate "know" is an example of a presupposition trigger (i.e., an expression or a construction causing presuppositions). In (15a), the presupposition is that Bob ran. This can be confirmed by the fact that the negated sentence (15b) also implies that Bob ran.

(15) a. John knows that Bob ran.

b. John does not know that Bob ran.

We list some Japanese comparative sentences with a presupposition in (16), where the trigger is underlined. Here, the presupposition is that the comparative standard has the property expressed by the predicate. That is, the three sentences in (16) all presuppose that Jiro is heavy.

(16) a. Taro-wa Jiro izyoo-ni omoi.
Taro-TOP Jiro than heavy
"Taro is heavier than Jiro."

b. Taro-wa Jiro to onaji kurai omoi.
Taro-TOP Jiro as same as heavy
"Taro is as heavy as Jiro."

c. Taro-wa Jiro hodo omoku nai.
Taro-TOP Jiro hodo heavy not
"Taro is not as heavy as Jiro."

In formally analyzing presuppositions, it is not adequate to simply conjoin the presupposition with other parts of the sentence. For example, suppose we represent the meaning of (15a) as a conjunction of the semantic representations of "John knows that Bob ran" and "Bob ran," as shown below.

(17) $\mathsf{know}(\mathsf{john},\mathsf{ran}(\mathsf{bob}))\wedge\mathsf{ran}(\mathsf{bob})$

The negation of this formula, which is shown in (18), does not entail $\mathsf{ran}(\mathsf{bob})$, failing to capture the fact that the presupposition is not subject to the negation (cf. (15b)).

(18) $\neg\left(\mathsf{know}(\mathsf{john},\mathsf{ran}(\mathsf{bob}))\wedge\mathsf{ran}(\mathsf{bob})\right)$
$\Leftrightarrow\neg\mathsf{know}(\mathsf{john},\mathsf{ran}(\mathsf{bob}))\vee\neg\mathsf{ran}(\mathsf{bob})$

In order to correctly handle presuppositions, we use a framework called *multidimensional semantics* (Karttunen and Peters 1979). In this framework, the semantic representation of an entire sentence is represented by a pair of semantic representations. The first element is for the central content conveyed by the sentence (the *at-issue* content), and the second one is for the presupposition. For example, the semantic representation of the sentence (16a) is shown in (19).

(19) $\langle\exists d.\left(\mathsf{heavy}(\mathsf{taro},d)\wedge\neg\mathsf{heavy}(\mathsf{jiro},d)\right),$
$\mathsf{heavy}(\mathsf{jiro},\theta)\rangle$

When the sentence is negated, we only negate the semantic representation of the at-issue content in the semantic composition, and the semantic representation of the entire sentence is (20).

(20) $\langle\neg\exists d.\left(\mathsf{heavy}(\mathsf{taro},d)\wedge\neg\mathsf{heavy}(\mathsf{jiro},d)\right),$
$\mathsf{heavy}(\mathsf{jiro},\theta)\rangle$

In the theorem proving step, we conjoin the semantic representations for the at-issue content and for the presupposition with $\wedge$.

5 Experiment

5.1 Settings

In this section, we describe the implementation settings of the proposed system.

Syntactic Parsing We use a Japanese tokenizer Janome.[5] As a CCG parser, we use depccg (Yoshikawa et al. 2017), the best-performing model provided for Japanese. We use Tsurgeon (Levy and Andrew 2006) to modify CCG parsing trees and insert empty categories. Our modification processes are as follows:

- We add rules to merge some multiword expressions. For instance, "izyoo ni" is converted to "izyoo-ni," "yori mo" to "yori-mo," and "to onaji kurai no" to "to-onaji-kurai-no."
- We insert the empty category *cmp* (cf. Section 4.1).
- We add a new syntactic feature to "yori" in phrasal comparatives related to clausal comparatives[6] in order to distinguish it from "yori" in ordinary phrasal comparatives.
- We add a new syntactic feature to "yori" in comparatives with a measure phrase in order to distinguish it from "yori" in ordinary phrasal comparatives.

In total, we make 60 entries in the Tsurgeon script for these processes.

Semantic Parsing For semantic composition, we use ccg2lambda (Martínez-Gómez et al. 2016), which supports Japanese as well as English. It uses λ-calculus to derive semantic representations. We extend the semantic templates to introduce the semantic representations based on degree semantics. We create two templates, one with multidimensional semantics and one without. The total number of lexical entries in each semantic template file is 222. We newly add 58 entries for words related to comparatives.

Theorem Proving We use Vampire 4.9 (Kovács and Voronkov 2013), a resolution-based automated theorem prover, for theorem proving. Vampire uses the Thousand of Problems for Theorem Provers (TPTP, Sutcliffe 2017) format to describe logical formulas. For this reason, we convert the output of ccg2lambda into first-order predicate logic formulas in the TPTP format. At this point, we add the axioms described in Section 3.3. In this step, we use the CASC mode, the fastest mode in Vampire. We try to prove $P_1 \wedge P_2 \wedge \ldots \wedge P_n \rightarrow H$ and $P_1 \wedge P_2 \wedge \ldots \wedge P_n \rightarrow \neg H$ for up to 20 seconds each to determine the system output.

5.2 Dataset

We use the comparatives section of the JSeM dataset (Kawazoe et al. 2017) for evaluation of our inference system. This NLI dataset contains Japanese counterparts of the FraCaS test suite (Cooper et al. 1996). It also contains newly added problems that involve phenomena FraCaS does not address or phenomena unique to Japanese.

In this study, we do not address tense and aspect, so we eliminated problems involving them. We do not address modality as well. With regard to modality, JSeM only has problems involving the property that modals do not affect the presupposition. Thus, we replaced modals with negation on these problems. As a result, the number of problems in the dataset is 71. The distribution of the gold answer labels is (*yes*/*no*/*unknown*) = (42/8/21). Table 4 shows some problems in the dataset.

jsem-569, Gold answer: yes	
P1	PC-6082-wa ITEL-XZ yori hayai. (PC-6082 is faster than ITEL-XZ.)
P2	ITEL-XZ-wa hayai. (ITEl-XZ is fast.)
H	PC-6082-wa hayai. (PC-6082 is fast.)
jsem-576, Gold answer: no	
P1	PC-6082-wa ITEL-XZ to onaji kurai-no hayasa-da. (PC-6082 is as fast as ITEL-XZ.)
P2	PC-6082-wa osoi. (PC-6082 is slow.)
H	ITEL-XZ-wa hayai. (ITEL-XZ is fast.)

Table 4: Examples of the problems in JSeM. P and H stand for "premise" and "hypothesis," respectively.

5.3 Evaluation Method

We use accuracy as an evaluation metric. When an error occurs in the proposed system, we treat it as an incorrect answer. As a baseline, we adopt

[5] https://github.com/mocobeta/janome

[6] For example, "Taro-wa Hanako yori takai hon-o katta. (Taro bought a more expensive book than Hanako.)"

GPT-4o and Swallow 8B[7]/70B[8] (S-8B/S-70B), the latter being competitive open Japanese LLMs. We conduct experiments using six different prompts for these models and calculate the accuracy as the average across these prompts.[9] The details of the prompts are shown in Appendix D.

6 Results and Discussion

6.1 Results

Table 5 shows the accuracy on the JSeM dataset. The table shows that our system outperformed all baseline models in terms of accuracy. The detailed results are shown in Appendix E.

Majority	GPT-4o	S-8B	S-70B	Ours
.592	.774	.549	.712	**.845**

Table 5: Accuracy on the JSeM dataset. "Majority" indicates the accuracy achieved when "yes," the most common label in the dataset, is answered for all the problems.

jsem-570, Gold answer: unknown GPT-4o: yes, Ours: unknown	
P	PC-6082-wa ITEL-XZ yori hayai. (PC-6082 is faster than ITEL-XZ.)
H	PC-6082-wa hayai. (PC-6082 is fast.)
jsem-620, Gold answer: yes GPT-4o: unknown, Ours: yes	
P	Taro-wa Hanako izyoo-ni hayaoki-da. (Taro is an earlier riser than Hanako.)
H	Hanako-wa hayaoki-da. (Hanako is an early riser.)

Table 6: Examples of problems that GPT-4o did not answer correctly but ours did

Table 6 shows some examples of problems that our system could predict correct answers while GPT-4o could not. GPT-4o incorrectly answered some of the relatively simple problems, such as jsem-570. The possible reason is that GPT-4o inferred "X is fast" from "X is faster."

Notably, GPT-4o failed to answer correctly some problems with presupposition triggers, such as jsem-620. In order to perform this inference, it is necessary to infer the presupposition that Hanako is an early riser from the premise. GPT-4o was rarely able to solve such problems. On the other hand, our proposed system correctly predicted the entailment relation, thanks to multidimensional semantics.

6.2 Error Analysis

Table 7 shows two cases where our system failed to obtain correct semantic representations, but GPT-4o gave correct answers. In jsem-589, we can interpret "APCOM-no keiyaku" either as the contracts that APCOM won or as the contracts that ITEL won from APCOM. To handle this kind of ambiguity, we need to (i) add a new semantic representation of "yori," and (ii) implement a system for distinguishing between the two interpretations based on syntactic information.

Jsem-606 is another case where our system failed to make a correct prediction. The verb "magaru" (bend) behaves like an adjective when combined with "te-i-ru." However, our system treats the resultant predicate "magatte-i-ru" as a verb, so its semantic type does not match the one required for the argument of "yori," causing an error in semantic parsing. To handle this error, we need to give an exceptional semantic representation to "te-i-ru" when it forms an adjective-like predicate with certain verbs like "magaru."

jsem-589, Gold answer: yes GPT-4o: yes, Ours: error	
P	ITEL-wa APCOM-no keiyaku yori ooku-no chuumon-o kakutoku-sita. (ITEL won more orders than the APCOM contract.)
H	ITEL-wa APCOM-no chuumon-o kakutoku-shita. (ITEL won the APCOM contract.)
jsem-606, Gold answer: yes GPT-4o: yes, Ours: error	
P	Kono boo-wa ano boo yori magatte-i-ru. (This stick is more bent than that one.)
H	Kono boo-wa magatte-i-ru. (This stick is bent.)

Table 7: Examples of problems our system answered incorrectly

7 Conclusion

In this study, we have proposed ccg-jcomp, a logical inference system for Japanese comparatives

[7] tokyotech-llm/Llama-3.1-Swallow-8B-v0.1

[8] tokyotech-llm/Llama-3.1-Swallow-70B-v0.1

[9] The model inferences were conducted in May 2025.

based on CCG, degree semantics, and some analyses of phenomena unique to Japanese comparatives. In our experiments with the Japanese NLI dataset that involves comparatives, we demonstrated that our proposed system achieved higher accuracy than several LLMs.

In future work, we are considering handling the ambiguity of certain sentences and the behavior of the adjective-like verbs discussed in Section 6.2. Additionally, it would be desirable to address adverbial comparatives, which are not covered in JSeM.

Limitations

Few-shot Learning In this study, we did not compare methods using few-shot learning as a baseline. It may improve the performance of the baseline models. For example, the LLMs may correctly answer jsem-620 in Section 6.1 by looking at some example inferences with a presupposition and learning the inference patterns. However, we do not have a sufficient number of problems involving Japanese comparatives to carry out and evaluate few-shot learning. Therefore, we conducted all experiments in a zero-shot setting for all models.

Scalability In addition to comparatives, JSeM has sections on other linguistic phenomena, such as anaphora. However, since our proposed system focuses only on Japanese comparatives, it cannot be used as is to handle these phenomena. To address them, we need to introduce the mechanism employed by some specific frameworks (e.g., *dynamic semantics* (Groenendijk and Stokhof 1991) for anaphora) in a manner consistent with degree semantics, which is not trivial. Hence, we leave for future work the development of a unified system that can handle these phenomena together with comparatives.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments, which helped us improve this paper. This work was supported by the Institute for AI and Beyond of the University of Tokyo and JSPS KAKENHI Grant Number JP24H00809, Japan.

References

Lasha Abzianidze. 2015. A tableau prover for natural logic and language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502, Lisbon, Portugal. Association for Computational Linguistics.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence: Resources for processing natural language*, pages 241–301. Springer.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2017. A type-theoretical system for the FraCaS test suite: Grammatical framework meets coq. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Long papers*.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2021. Applied temporal analysis: A complete run of the FraCaS test suite. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 11–20, Groningen, The Netherlands (online). Association for Computational Linguistics.

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.

Max J Cresswell. 1976. The semantics of degree. In *Montague grammar*, pages 261–292. Elsevier.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100.

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.

Julia Hockenmaier and Mark Steedman. 2007. Ccgbank: a corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3):355–396.

Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S. Moss, and Sandra Kuebler. 2020. MonaLog: a lightweight system for natural language inference based on monotonicity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 334–344, New York, New York. Association for Computational Linguistics.

Lauri Karttunen and Stanley Peters. 1979. Conventional Implicature. In *Presupposition*, pages 1–56. Brill.

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.

Ewan Klein. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4:1–45.

Ewan Klein. 1982. The interpretation of adjectival comparatives1. *Journal of Linguistics*, 18(1):113–136.

Laura Kovács and Andrei Voronkov. 2013. First-order theorem proving and vampire. In *International Conference on Computer Aided Verification*, pages 1–35. Springer.

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.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

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

Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. 2016. ccg2lambda: A compositional semantics system. In *Proceedings of ACL-2016 System Demonstrations*, pages 85–90.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061.

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.

Christopher Potts. 2015. Presupposition and implicature. *The Handbook of Contemporary Semantic Theory*, pages 168–202.

Pieter A. M. Seuren. 1973. *The Comparative*, pages 528–564. Springer Netherlands, Dordrecht.

Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. 2023. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1803–1821, Toronto, Canada. Association for Computational Linguistics.

Mark Steedman. 2000. *The Syntactic Process*. MIT press.

Geoff Sutcliffe. 2017. The TPTP problem library and associated infrastructure: from CNF to TH0, TPTP v6. 4.0. *Journal of Automated Reasoning*, 59(4):483–502.

Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. 2017. A* CCG parsing with a supertag and dependency factored model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 277–287. Association for Computational Linguistics.

A Combinatory Rules of CCG

We show some combinatory rules of CCG below (see Steedman (2000) for details).

- Forward functional application rule

$$\frac{X/Y : f \qquad Y : a}{X : f\,a} >$$

- Backward functional application rule

$$\frac{Y : a \qquad X\backslash Y : f}{X : f\,a} <$$

- Forward functional composition rule

$$\frac{X/Y : f \qquad Y/Z : g}{X/Z : \lambda x.f(g\,x)} >\mathbf{B}$$

- Backward functional composition rule

$$\frac{Y\backslash Z : g \qquad X\backslash Y : f}{X\backslash Z : \lambda x.f(g\,x)} <\mathbf{B}$$

- Forward functional crossed composition rule

$$\frac{X/Y : f \qquad Y\backslash Z : g}{X\backslash Z : \lambda x.f(g\,x)} >\mathbf{B}_{\times}$$

- Backward functional crossed composition rule

$$\frac{Y/Z : g \qquad X\backslash Y : f}{X/Z : \lambda x.f(g\,x)} >\mathbf{B}_{\times}$$

B Details of Axioms

Table 8 shows the axioms employed in our system. (CP) is the axiom we already introduced in Section 3.3. We can make the following inferences using this axiom. (21a) and (21b) are the premises, and (21c) is the hypothesis.

Name	Logical Formula
CP	$\forall x\, y.\, ((\exists d.\, (\mathbf{A}(x,d) \wedge \neg\mathbf{A}(y,d))) \rightarrow \forall d.\, (\mathbf{A}(y,d) \rightarrow \mathbf{A}(x,d)))$
ANT	$\forall x\, d.\, (\mathbf{P}(x,d) \leftrightarrow \neg\mathbf{N}(x,d))$
UP	$\forall x\, d.\, (\mathbf{P}(x,d) \rightarrow \forall d'.\, (d' \leq d \rightarrow \mathbf{P}(x,d')))$
DOWN	$\forall x\, d.\, (\mathbf{N}(x,d) \rightarrow \forall d'.\, (d' \geq d \rightarrow \mathbf{N}(x,d')))$
DELTA	$\forall x.\, (\mathbf{A}(x,\theta-\delta) \leftrightarrow \mathbf{A}(x,\theta+\delta))$

Table 8: Axioms for Japanese comparatives. $\mathbf{A}$ denotes adjectives, $\mathbf{P}$ denotes positive adjectives such as *long*, and $\mathbf{N}$ denotes negative adjectives such as *short*.

(21) a. Taro-wa Jiro yori omoi.
Taro-TOP Jiro than heavy
"Taro is heavier than Jiro."

b. Jiro-wa 70 kg yori omoi.
Jiro-TOP 70 kg than heavy
"Jiro is heavier than 70 kg."

c. Taro-wa 70 kg yori omoi.
Taro-TOP 70 kg than heavy
"Taro is heavier than 70 kg."

(entailment)

Concretely, from (CP) and (21a), we obtain $\mathsf{heavy}(\mathsf{jiro}, 70) \rightarrow \mathsf{heavy}(\mathsf{taro}, 70)$. Then, from this formula and (21b), we can derive $\mathsf{heavy}(\mathsf{taro}, 70)$.

(ANT) indicates the antonymy relation between positive and negative adjectives. The following is an example of an inference using this axiom. (22a) is the premise and (22b) is the hypothesis.

(22) a. Taro-wa Jiro yori omoi.
Taro-TOP Jiro than heavy
"Taro is heavier than Jiro."

b. Taro-wa Jiro yori karui.
Taro-TOP Jiro than light
"Taro is lighter than Jiro."

(contradiction)

(UP) and (DOWN) are axioms that indicate the monotonicity of positive and negative adjectives, respectively. (DELTA) is an axiom about equatives. Using these axioms, we can prove the entailment relation in (10) as follows. (23a), (23b), and (23c) are the semantic representations of (10a), (10b), and (10c), respectively.

(23) a. $\forall d_1\, d_2.\, ((\neg\, (\mathsf{heavy}(\mathsf{taro}, d_1)$
$\leftrightarrow\ \mathsf{heavy}(\mathsf{jiro}, d_1))$
$\wedge\, \neg\, (\mathsf{heavy}(\mathsf{taro}, d_2)$
$\leftrightarrow\ \mathsf{heavy}(\mathsf{jiro}, d_2)))$
$\rightarrow |d_1 - d_2| < \delta)$

b. $\mathsf{heavy}(\mathsf{jiro}, \theta)$

c. $\mathsf{heavy}(\mathsf{taro}, \theta)$

First, from (23b), (UP), and (DELTA), we can derive $\mathsf{heavy}(\mathsf{jiro}, \theta)$ and $\mathsf{heavy}(\mathsf{jiro}, \theta + \delta)$. Then, by replacing d_1 (resp. d_2) in (27a) with $\theta + \delta$ (resp. δ), and by contraposition, we obtain either $\mathsf{heavy}(\mathsf{taro}, \theta + \delta) \leftrightarrow \mathsf{heavy}(\mathsf{jiro}, \theta + \delta)$ or $\mathsf{heavy}(\mathsf{taro}, \theta) \leftrightarrow \mathsf{heavy}(\mathsf{jiro}, \theta)$. In both cases, $\mathsf{heavy}(\mathsf{taro}, \theta)$ is true since we have $\mathsf{heavy}(\mathsf{jiro}, \theta + \delta)$ and $\mathsf{heavy}(\mathsf{jiro}, \theta)$.

In the implementation, (CP) and (DELTA) are added for all gradable adjectives. (ANT) is added for adjectives that have an antonym. (UP) and (DOWN) are added for positive adjectives and negative adjectives, respectively.

C Problem Replacement

Table 9 shows an example of the problems in JSeM to which we applied the replacement we discussed in section 5.2. The original problem uses the property that the presupposition "Hanako is an early riser" is not affected by the modal "kamo-sire-nai." We did not implement the semantic representation of modals, so we replaced them with a negation "to-iu-wake-de-wa-nai." Since presuppositions are unaffected by negation (as well as by modals), this replacement does not alter the purpose of the problem—namely, to test whether the model understands that presuppositions are not influenced by entailment-canceling operators.

D Prompts for the Baseline Models

Table 10 and Table 11 show examples of prompts for GPT-4o and Swallow, respectively.

jsem-621 (**original**), Gold answer: yes	
Premise	Taro-wa Hanako izyoo-ni hayaoki kamo-sire-nai. (Taro may be an earlier riser than Hanako.)
Hypothesis	Hanako-wa Hayaoki-da. (Hanako is an early riser.)
jsem-621 (**replaced**), Gold answer: yes	
Premise	Taro-wa Hanako izyoo-ni hayaoki toiu-wake-de-wa-nai. (Taro is not an earlier riser than Hanako.)
Hypothesis	Hanako-wa Hayaoki-da. (Hanako is an early riser.)

Table 9: An example of the problems in which we replaced a modal with a negation

system	前提文と仮説文が与えられます。 前提文が仮説文を含意しているか答えてください。 「含意」、「矛盾」、「中立」のいずれかで答えてください。 (You are given premises and a hypothesis. Answer whether the premises entail the hypothesis. Answer with “entailment”, “contradiction”, or “neutral.”)
user	前提１：PC-6082はITEL-XZより速い。 前提２：ITEL-XZは速い。 仮説：PC-6082は速い。 (Premise 1: PC-6082 is faster than ITEL-XZ. Premise 2: ITEL-XZ is fast. Hypothesis: PC-6082 is fast.)

Table 10: Example of the prompt for GPT-4o

前提文と仮説文が与えられます。 前提文が仮説文を含意しているか答えてください。 「含意」、「矛盾」、「中立」のいずれかで答えてください。 前提１：PC-6082はITEL-XZより速い。 前提２：ITEL-XZは速い。 仮説：PC-6082は速い。 回答：

Table 11: Example of the prompt for Swallow

E Detailed Results

Table 12, Table 13, and Table 14 show the detailed evaluation results of each baseline model. E, C, and N represent entailment, contradiction, and neutral, respectively. “Prompt Type” indicates the order of the words 含意 (entailment), 矛盾 (contradiction), and 中立 (neutral) as they appear in each prompt. For example, Table 10 and Table 11 show prompts of the E-C-N (含意-矛盾-中立) type.

Swallow 8B tended to output *yes* when 含意 or 中立 appeared first in the prompt, resulting in substantially lower F1 scores for contradiction and neutral compared to entailment. Conversely, when 矛盾 was presented first, the number of *no* responses increased.

In contrast, Swallow 70B and GPT-4o produced more balanced outputs, achieving higher F1 scores than Swallow 8B.

Prompt Type	Accuracy	Gold Label	Precision	Recall	F1 Score
E-C-N	0.619	E	0.62	1.00	0.76
		C	0.00	0.00	0.00
		N	0.67	0.10	0.17
E-N-C	0.605	E	0.61	1.00	0.76
		C	0.50	0.12	0.20
		N	0.00	0.00	0.00
C-E-N	0.225	E	0.80	0.19	0.31
		C	0.13	1.00	0.23
		N	0.00	0.00	0.00
C-N-E	0.591	E	0.77	0.81	0.79
		C	0.30	1.00	0.46
		N	0.00	0.00	0.00
N-E-C	0.605	E	0.60	1.00	0.75
		C	1.00	0.12	0.22
		N	0.00	0.00	0.00
N-C-E	0.647	E	0.64	1.00	0.78
		C	0.75	0.38	0.50
		N	1.00	0.05	0.09

Table 12: Evaluation results of Swallow 8B on each prompt

Prompt Type	Accuracy	Gold Label	Precision	Recall	F1 Score
E-C-N	0.647	E	0.80	0.86	0.83
		C	0.36	1.00	0.53
		N	0.50	0.10	0.16
E-N-C	0.690	E	0.81	0.83	0.82
		C	0.55	0.75	0.63
		N	0.47	0.38	0.42
C-E-N	0.676	E	0.80	0.86	0.83
		C	0.40	1.00	0.57
		N	0.67	0.19	0.30
C-N-E	0.661	E	0.80	0.86	0.83
		C	0.42	1.00	0.59
		N	0.43	0.14	0.21
N-E-C	0.760	E	0.88	0.83	0.85
		C	0.62	1.00	0.76
		N	0.61	0.52	0.56
N-C-E	0.788	E	0.88	0.86	0.87
		C	0.62	1.00	0.76
		N	0.71	0.57	0.63

Table 13: Evaluation results of Swallow 70B on each prompt

Prompt Type	Accuracy	Gold Label	Precision	Recall	F1 Score
E-C-N	0.746	E	0.83	0.81	0.82
		C	0.75	0.75	0.75
		N	0.59	0.62	0.60
E-N-C	0.774	E	0.84	0.86	0.85
		C	0.75	0.75	0.75
		N	0.65	0.62	0.63
C-E-N	0.774	E	0.85	0.83	0.84
		C	0.78	0.88	0.82
		N	0.62	0.62	0.62
C-N-E	0.760	E	0.85	0.81	0.83
		C	0.78	0.88	0.82
		N	0.59	0.62	0.60
N-E-C	0.788	E	0.86	0.86	0.86
		C	0.75	0.75	0.75
		N	0.67	0.67	0.67
N-C-E	0.788	E	0.86	0.86	0.86
		C	0.78	0.88	0.82
		N	0.65	0.62	0.63

Table 14: Evaluation results of GPT-4o on each prompt