JSTS-Neg: Japanese Semantic Textual Similarity Dataset for Evaluating Negation Understanding Ability

Reiko Yuasa[1], **Yoshihide Kato**[2], **Shigeki Matsubara**[1,2]
[1]Graduate School of Informatics, Nagoya University
[2]Information & Communications, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
yuasa.reiko.k5@s.mail.nagoya-u.ac.jp

Abstract

Negation is a common linguistic phenomenon in natural language. Thus, datasets and benchmarks focused on negation are being constructed to evaluate the negation understanding abilities of language models. Negation is especially crucial when estimating the semantic similarity between sentences because it inverses their meaning. Although semantic textual similarity (STS) is one of the useful tasks to evaluate the abilities of large language models (LLMs), few STS datasets focus on negation. In this research, we introduce JSTS-Neg, a new Japanese STS dataset focusing on negation. Most instances in JSTS-Neg include negations and they are composed of both clausal and sub-clausal negations to reflect a variety of negation types. Moreover, JSTS-Neg consists of negation minimal pairs that only differ in the presence or absence of a negation cue. We evaluate the performance of existing LLMs on JSTS-Neg using negation minimal pairs to explore their abilities and limitations in understanding negation. LLMs tend to predict the similarity of two sentences ignoring negation cues in specific settings.

Keywords: Negation, Minimal Pair, Large Language Models, Semantic Textual Similarity

1. Introduction

Negation is a common linguistic phenomenon in natural language. Processing negation correctly is crucial for natural language processing systems because it inverses the meaning of sentences, phrases, words, and other linguistic elements. Recent studies have shown that existing large language models (LLMs) struggle with negation by evaluation on negation-focused datasets and benchmarks (Hossain et al., 2022; Truong et al., 2023; García-Ferrero et al., 2023, inter alia).

Various datasets focusing on negation have been created for natural language inference (NLI) (Hossain et al., 2020; Hartmann et al., 2021), acceptability judgment (Someya and Oseki, 2023; Taktasheva et al., 2024), and question answering (QA) (Ravichander et al., 2022; García-Ferrero et al., 2023). However, few datasets target semantic textual similarity (STS), despite being a fundamental task in natural language understanding to evaluate the abilities of LLMs. STS demands deep understanding of the meaning of sentences and thus allows to compare, analyze, and evaluate the factuality of LLM responses (Wang et al., 2024). Thus, STS datasets focusing on negation should be constructed. Besides, regarding target languages, most negation-focused datasets are in English, while datasets in Japanese are found in few studies, such as Matsuyoshi et al. (2014); Uchida and Nanjo (2024); Yoshida et al. (2025).

We introduce JSTS-Neg, a new Japanese STS dataset focusing on negation[1]. To construct JSTS-Neg, we adopt a similar approach to that developed by Yoshida et al. (2025), which creates NLI instances with a negation cue from an existing one without that. JSTS-Neg includes sufficient number and variety of negations, as approximately 85% of instances in JSTS-Neg include negation cues involving both clausal and sub-clausal negations. Moreover, JSTS-Neg consists of negation minimal pairs that only differ in the presence or absence of a negation cue. Evaluation based on negation minimal pairs eliminates factors other than negation and unveils the negation understanding ability solely by comparing the outputs between these pairs.

We also evaluate the negation understanding ability of existing LLMs on JSTS-Neg to explore their current performance and limitations. The evaluation results show that LLMs tend to predict the similarity of two sentences ignoring negation cues in specific settings.

2. Related Work

This section describes existing datasets focusing on negation. To clarify our contributions, we focus particularly on Japanese and STS datasets.

[1]Our dataset and source code are publicly available at https://github.com/reiko-y/JSTS-Neg.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 12031–12041
11–16 May 2026.

Table 1: Datasets focusing on negation

Dataset	Task	Minimal pair	Type of negation	Language(s)
(Hossain et al., 2020)	NLI	✓	Clausal	EN
(Hartmann et al., 2021)	NLI	✓	Clausal and sub-clausal	EN, BG, DE, FR, ZH
RuBLiMP (Taktasheva et al., 2024)	Acceptability judgment	✓	Clausal and sub-clausal	RU
CONDAQA (Ravichander et al., 2022)	QA		Clausal and sub-clausal	EN
(García-Ferrero et al., 2023)	QA	✓ (partly)	Clausal and sub-clausal	EN
N-JSNLI (Uchida and Nanjo, 2023)	NLI	✓ (partly)	Clausal	JA
JNLI-Neg (Yoshida et al., 2025)	NLI	✓	Clausal and sub-clausal	JA
JBLiMP (Someya and Oseki, 2023)	Acceptability judgment	✓	Clausal	JA
N-JSTS (Uchida and Nanjo, 2024)	STS		Clausal	JA
JSTS-Neg (Ours)	STS	✓	Clausal and sub-clausal	JA

2.1. Datasets Focusing on Negation

Several datasets focusing on negation have been constructed to evaluate the negation understanding ability of language models, as listed in Table 1. For example, Hossain et al. (2020) constructed a benchmark for NLI in which negation played a critical role. Hartmann et al. (2021) created a multilingual (i.e., English, Bulgarian, German, French, and Chinese) negation-focused benchmark collection for NLI. Someya and Oseki (2023); Taktasheva et al. (2024) developed acceptability judgment datasets focusing on linguistic minimal pairs including negation. Ravichander et al. (2022) constructed an English reading comprehension dataset requiring reasoning about the implications of negated statements in paragraphs. García-Ferrero et al. (2023) semiautomatically generated a large QA dataset about commonsense knowledge, in which negation appeared in approximately two-thirds of the corpus in different forms. However, few STS datasets focus on negation.

2.2. Japanese Datasets Focusing on Negation

Few Japanese datasets focus on negation. Representative Japanese datasets, such as JaNLI (Yanaka and Mineshima, 2021), JGLUE (Kurihara et al., 2022), and JSICK (Yanaka and Mineshima, 2022) remain insufficient to evaluate the negation understanding ability of language models because they include less negation (Uchida and Nanjo, 2023; Yuasa et al., 2025). Although N-JSNLI (Uchida and Nanjo, 2023), JNLI-Neg (Yoshida et al., 2025), and JBLiMP (Someya and Oseki, 2023) address this issue, they are not intended to evaluate STS.

2.3. Japanese STS Datasets Including Negation

Uchida and Nanjo (2024) automatically constructed N-JSTS, a Japanese STS evaluation dataset including negation from JSTS (Kurihara et al., 2022). New STS instances including negation were generated by inserting a negation cue to the end of both sentences of STS instances. However, N-JSTS did not include negation minimal pairs because the negation cues were added to both sentences of every instance. Thus, the negation understanding ability was difficult to solely evaluate using this dataset.

In addition, the negation cue only appeared in the main clause, and negation was assumed to be clausal (in Japanese, the end of sentence is always

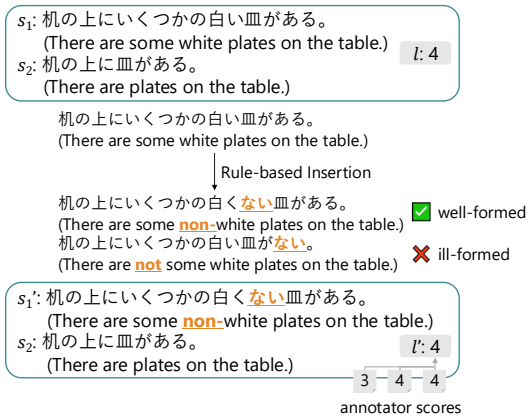

Figure 1: Dataset construction flow

user: 次の日本語が文法的・意味的に正しいかどうかを判定してください。 正しい場合は「正しい」、そうでない場合は「正しくない」と出力してください。 「正しい」「正しくない」のいずれかのみを出力することを遵守してください。
user: {Original JSTS sentence}
assistant: 正しい
user: {Generated sentence}

user: Check the grammatical and semantical well-formedness of the following Japanese sentence. If the sentence is well-formed, output "Well-formed", otherwise output "Ill-formed". Please strictly follow the instruction to output either "Well-formed" or "Ill-formed".
user: {Original JSTS sentence}
assistant: Well-formed
user: {Generated sentence}

Figure 2: Well-formedness prompt (The original prompt is shown above, and its translation is provided below.)

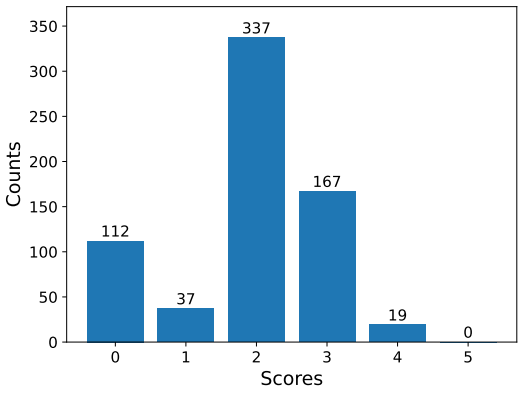

Figure 3: Score distribution of the training portion of D_{orig}

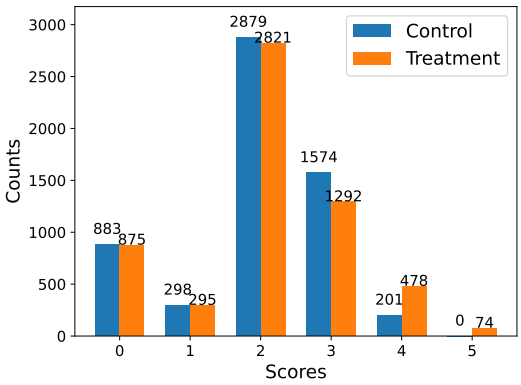

Figure 4: Score distribution of between the control and treatment groups of the training portion of minimal pairs

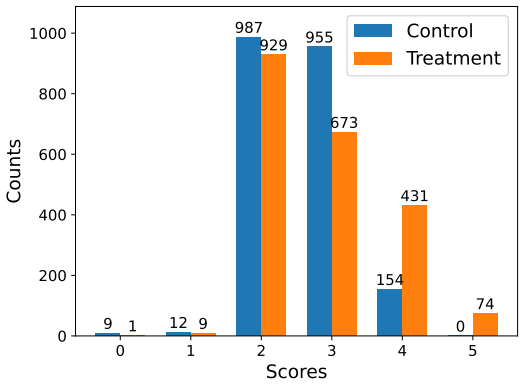

Figure 5: Score distribution of the control and treatment groups of the training portion of negation minimal pairs with *important* negation cues

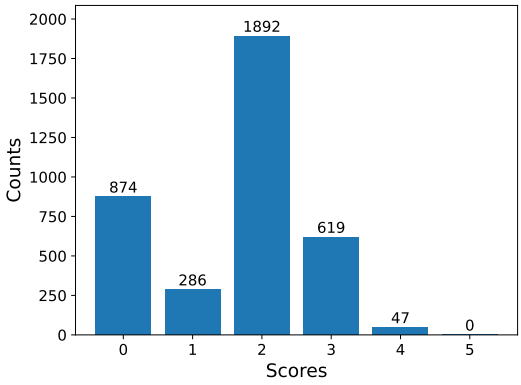

Figure 6: Score distribution of the training portion of negation minimal pairs with *unimportant* negation cues[2]

the verb of the main clause). Consequently, the variety of negation types was insufficient.

3. Constructed JSTS-Neg

We construct JSTS-Neg from JSTS (Kurihara et al., 2022), as detailed in this section. We adopt the method proposed by Yoshida et al. (2025), in which new instances consisting of negation minimal pairs are generated from an existing Japanese NLI dataset. Two Japanese representative negative morphemes, namely, "ない" and "ず", are the target negation cues, and we do not consider neither the affixal negation nor double negation.

3.1. Dataset Requirements

JSTS-Neg is intended to satisfy the following requirements:

1. All instances consist of negation minimal pairs.
2. The dataset includes clausal and sub-clausal negations.
3. Sentences are constructed originally in Japanese (i.e., they are not translations from other languages).

These requirements agree with those imposed by Yoshida et al. (2025). Requirements 1 and 2 are satisfied during dataset construction. Requirement 3 is satisfied by using JSTS (Kurihara et al., 2022) as the source dataset, whose sentences are originally constructed in Japanese.

3.2. Dataset Construction Procedure

JSTS-Neg was constructed following two main processes, and its construction flow is illustrated in Figure 1.

1. Insert a negation cue into a sentence.
   (a) Rule-based insertion: Insert a negative morpheme, either "ない" or "ず", into a verb, adjective, or adjectival noun.

[2]The gold similarity scores remain unchanged between the control and treatment groups of negation minimal pairs with *unimportant* negation cues.

Table 2: Evaluated open LLMs

Language	Model	Name on Hugging Face [3]
Japanese	llm-jp-3.1-1.8B-instruct4 (Aizawa et al., 2024)	llm-jp/llm-jp-3.1-1.8b-instruct4
	llm-jp-3.1-13B-instruct4 (Aizawa et al., 2024)	llm-jp/llm-jp-3.1-13b-instruct4
	Llama 3.1 Swallow 8B Instruct v0.5 (Fujii et al., 2024; Okazaki et al., 2024)	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
	Gemma-2-Llama Swallow 2B IT v0.1 (Fujii et al., 2024; Okazaki et al., 2024)	tokyotech-llm/Gemma-2-Llama-Swallow-2b-it-v0.1
	Gemma-2-Llama Swallow 9B IT v0.1 (Fujii et al., 2024; Okazaki et al., 2024)	tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1
	Swallow-MS 7B instruct v0.1 (Fujii et al., 2024; Okazaki et al., 2024)	tokyotech-llm/Swallow-MS-7b-v0.1
	Llama 3 Youko 8B Instruct (Sawada et al., 2024)	rinna/llama-3-youko-8b-instruct
	Gemma 2 Baku 2B Instruct (Sawada et al., 2024)	rinna/gemma-2-baku-2b-it
multilingual	Llama 3.1 8B Instruct	meta-llama/Llama-3.1-8B-Instruct
	Llama 3 8B Instruct	meta-llama/Meta-Llama-3-8B-Instruct
	Mistral 7B Instruct v0.3 (Jiang et al., 2023)	mistralai/Mistral-7B-Instruct-v0.3
	Ministral 8B Instruct	mistralai/Ministral-8B-Instruct-2410
	Mistral Nemo Instruct	mistralai/Mistral-Nemo-Instruct-2407
	Gemma 3n E2B instruct	google/gemma-3n-E2B-it
	Gemma 3n E4B instruct	google/gemma-3n-E4B-it
	Gemma 3 270M instruct (Kamath et al., 2025)	google/gemma-3-270m-it
	Gemma 3 1B instruct (Kamath et al., 2025)	google/gemma-3-1b-it
	Gemma 3 4B instruct (Kamath et al., 2025)	google/gemma-3-4b-it
	Gemma 2 2B Instruct (Rivière et al., 2024)	google/gemma-2-2b-it
	Gemma 2 9B Instruct (Rivière et al., 2024)	google/gemma-2-9b-it

(b) Check the grammatical and semantical well-formedness of the obtained sentences using an LLM and discard the sentences judged as ill-formed.

2. Create a new STS instance from another STS instance.

(a) Make a sentence pair from an STS instance either by keeping one sentence unchanged and inserting a negation cue to the other or by inserting a negation cue to both.

(b) Manually assign a similarity score to the sentence pair.

We represent similarity score l between sentences s_1 and s_2 as (s_1, s_2, l) and call it an STS instance.

3.2.1. Negation Cue Addition

We create sentences with a negation cue according to the method proposed by Yoshida et al. (2025). In detail, we append negation cue "ない" or "ず" to a verb, adjective, or adjectival noun and inflect surrounding morphemes if necessary. If a sentence has multiple candidates to insert a negation cue, the sentence is duplicated for each candidate, and a single negation cue is added per candidate. That is, if a sentence has n candidates to insert a negation cue, n sentences with a negation cue are created. Note that every sentence with a negation cue and its original sentence form a negation minimal pair.

After cue addition, the well-formedness of the generated sentences is automatically checked to remove ill-formed sentences. Semantically ill-formed sentences are sometimes created because rule-based insertion only considers morphological information. For example, the following sentence (1), which is created from sentence (2), is semantically ill-formed:

(1) 群衆がいて混雑しない。
(There are crowds, and it is **not** congested.)

(2) 群衆がいて混雑する。
(There are crowds, and it is congested.)

Checking the well-formedness of sentences is based on in-context learning of an LLM[4]. The prompt shown in Figure 2 follows Yoshida et al. (2025). Here, we give the original sentence as an example of a well-formed sentence (one-shot) to make the model consider only the effect of the added negation cue. Only sentences judged as well-formed are used, while those judged as ill-formed are discarded from the dataset.

3.2.2. New STS Instances with Negation Cue

For an STS instance $i = (s_1, s_2, l)$, we create new instances by adding negation cues. Let S'_1 and S'_2 be the sets of sentences obtained by adding a negation cue to s_1 and s_2, respectively. If i satisfies $|S'_1| > 0$, $|S'_2| > 0$, $neg(s_1) = 0$, and $neg(s_2) = 0$, then create a set of STS instances, $D_{\text{neg}}(i)$, where $neg(s)$ denotes the number of negation cues in s[5]. $D_{\text{neg}}(i)$ is defined as

$$
\begin{aligned}
D_{\text{neg}}(i) &= D_1(i) \cup D_2(i) \cup D_{1,2}(i) \ , \\
D_1(i) &= \{(s'_1, s_2, l') | s'_1 \in S'_1\} \ , \\
D_2(i) &= \{(s_1, s'_2, l') | s'_2 \in S'_2\} \ , \\
D_{1,2}(i) &= \{(s'_1, s'_2, l') | s'_1 \in S'_1 \wedge s'_2 \in S'_2\} \ .
\end{aligned}
$$

In addition, l' is the gold similarity score obtained from annotators' scores. Three annotators assigned a score to each sentence pair according to

[3] https://huggingface.co/

[4] The verification model is `gpt-4.1-2025-04-14` by OpenAI API (https://openai.com/index/openai-api/).

[5] We use the negation cue detector made by Yuasa et al. (2025) to calculate $neg(s)$.

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示:
２つの文の類似度を0~5のいずれかで答えてください。
数字が大きいほど２つの文は似ています。

以下の指標を参考に回答してください。

【類似度： 5】
２つの文は意味が完全に一致している。

【類似度： 4】
２つの文はほぼ意味が同じだが、重要でない部分で意味が異なる。

【類似度： 3】
２つの文はある程度同じ意味だが、重要な部分で意味が異なる。

【類似度： 2】
２つの文は異なる意味であるが、細かい要素が共通している。

【類似度： 1】
２つの文は異なる意味であるが、話題に共通点がある。

【類似度： 0】
２つの文は全く異なる意味であり、話題も単語も共通していない。

We provide instructions describing the task, paired with contextual input. A response should adequately fulfills the requirement.

Instructions:
Rate the similarity between the two sentences on a scale from 0 to 5.
A higher value indicates more similarity between the sentences.

The following indicators are intended to support scoring:

[Similarity: 5]
The two sentences completely match in meaning.

[Similarity: 4]
The two sentences are almost identical in meaning but differ in minor details.

[Similarity: 3]
The two sentences are somewhat similar in meaning but differ in important details.

[Similarity: 2]
The two sentences have different meanings but have some commonalities.

[Similarity: 1]
The two sentences have different meanings but share a common topic.

[Similarity: 0]
The two sentences have completely different meanings, sharing neither topic nor vocabulary.

Figure 7: Task guideline prompt (The original prompt is shown on the left, and its translation is provided on the right.)

the guideline of JSTS, and the median across their three scores was used as the gold score.

3.3. Dataset Construction

We created STS instances with negation by applying the procedure described in Section 3.2 to the JSTS dataset.

First, we randomly reordered instances in the training, validation, and test sets of JSTS and created new instances from the top. We repeated instance creation until the number of created instances exceeded 4,000, 1,000, and 1,000 in the training, validation, and test sets, respectively. For the sampled original instances (672, 167, and 170 instances from the training, validation, and test sets, respectively), the similarity scores were reassigned by the same annotators[6]. We calculated Fleiss' kappa (Fleiss, 1971) using the set consisting of both the sampled original instances and newly created ones. The kappa score of the three annotators was 0.59, indicateing moderate agreement.

[6] In STS, the similarity scores are ordinal rather than interval scales. However, JSTS only includes average scores being treated as interval scales. As the scores are regarded as ordinal scale values, we decided to reassign them.

Table 3: AccChg results in zero-shot setting

Model	M			M_{i}			M_{u}		
	Acc	Acc'	AccChg^*	Acc	Acc'	AccChg^*	Acc	Acc'	AccChg^*
llm-jp-3.1-1.8B-instruct4	50.58	51.27	0.69	40.75	41.29	0.54	53.97	54.70	0.74
llm-jp-3.1-13B-instruct4	47.22	48.04	0.82	29.49	38.34	8.85	53.32	51.38	**-1.94**
Llama 3.1 Swallow 8B Instruct v0.5	30.89	37.61	6.73	43.16	52.28	9.12	26.66	32.56	5.90
Gemma-2-Llama Swallow 2B IT v0.1	41.46	40.01	**-1.44**	31.37	20.64	**-10.72**	44.93	46.68	1.75
Gemma-2-Llama Swallow 9B IT v0.1	46.47	54.56	8.10	63.81	64.08	0.27	40.50	51.29	10.79
Swallow-MS 7B instruct v0.1	55.18	49.62	**-5.56**	58.71	37.00	**-21.72**	53.97	53.97	0.00
Llama 3 Youko 8B Instruct	37.61	44.61	7.00	56.57	48.79	**-7.77**	31.09	43.17	12.08
Gemma 2 Baku 2B Instruct	63.56	59.57	**-3.98**	56.57	35.66	**-20.91**	65.96	67.80	1.85
Llama 3.1 8B Instruct	45.37	50.79	5.42	56.84	58.98	2.14	41.42	47.97	6.55
Llama 3 8B Instruct	1.85	3.02	1.17	0.54	3.75	3.22	2.31	2.77	0.46
Mistral 7B Instruct v0.3	40.97	47.15	6.18	45.04	52.01	6.97	39.58	45.48	5.90
Ministral 8B Instruct	58.54	47.01	**-11.53**	35.39	18.77	**-16.62**	66.51	56.73	**-9.78**
Mistral Nemo Instruct	62.73	62.18	**-0.55**	64.08	45.31	**-18.77**	62.27	67.99	5.72
Gemma 3n E2B instruct	35.83	40.91	5.08	48.79	50.67	1.88	31.37	37.55	6.18
Gemma 3n E4B instruct	36.51	40.01	3.50	46.92	46.92	0.00	32.93	37.64	4.70
Gemma 3 270M instruct	55.18	49.62	**-5.56**	58.71	37.00	**-21.72**	53.97	53.97	0.00
Gemma 3 1B instruct	3.84	5.63	1.78	10.99	16.09	5.09	1.38	2.03	0.65
Gemma 3 4B instruct	38.02	39.74	1.72	32.71	39.95	7.24	39.85	39.67	**-0.18**
Gemma 2 2B Instruct	52.85	51.27	**-1.58**	31.37	19.57	**-11.80**	60.24	62.18	1.94
Gemma 2 9B Instruct	24.43	29.92	5.49	32.44	40.21	7.77	21.68	26.38	4.70
GPT-5-nano	31.09	35.42	4.32	30.03	41.82	11.80	31.46	33.21	1.75
GPT-5-mini	42.35	51.61	9.27	44.77	49.87	5.09	41.51	52.21	10.70
GPT-5	54.08	62.94	8.85	62.47	54.69	**-7.77**	51.20	65.77	14.58
GPT-4.1-nano	52.78	54.98	2.20	41.55	39.14	**-2.41**	56.64	60.42	3.78
GPT-4.1-mini	50.31	51.20	0.89	46.92	37.00	**-9.92**	51.48	56.09	4.61
GPT-4.1	48.80	51.54	2.75	58.18	45.31	**-12.87**	45.57	53.69	8.12

Both Acc and Acc' are presented as percentages, and AccChg is presented as percent points.
* Values in **bold** are negative.

JSTS-Neg, or $D_{\text{JSTS-Neg}}$, is defined as

$$D_{\text{JSTS-Neg}} = D_{\text{orig}} \cup D_{\text{neg}} ,$$
$$D_{\text{neg}} = \bigcup_{i \in D_{\text{orig}}} D_{\text{neg}}(i) .$$

where D_{orig} denotes the set of sampled JSTS instances. JSTS-Neg consists of a set of negation minimal pairs M defined as

$$M = M_{\text{single}} \cup M_{\text{both}} ,$$
$$M_{\text{single}} = \{(i, i') | i \in D_{\text{orig}} \wedge i' \in D_1(i) \cup D_2(i)\} ,$$
$$M_{\text{both}} = \{(i', i'') | \exists i \in D_{\text{orig}}((i' \in D_1(i) \wedge i'' \in D_2(i')) \vee (i' \in D_2(i) \wedge i'' \in D_1(i')))\} .$$

M_{single} denotes the set of negation minimal pairs consist of i from D_{orig} and i' made by inserting a negation cue to either s_1 or s_2 of i. M_{both} denotes the set of negation minimal pairs consist of i' made by inserting a negation cue to s_1 of i and i'' made by inserting a negation cue to s_2 of i', and vice versa.

For minimal pair $m = (i, i') \in M$, we call i and i' control and treatment instances of m, respectively. The treatment instance is a version obtained by adding a negation cue to one sentence of the control instance.

M can be divided into sets M_{i} and M_{u}, with M_{i} and M_{u} being sets of negation minimal pairs constructed using *important* and *unimportant* negation cues, respectively. The definition of this division is based on Hossain et al. (2022) as follows:

$$M_{\mathrm{i}} = \{((s_1, s_2, l), (s_1', s_2', l')) \in M | l \neq l'\} ,$$
$$M_{\mathrm{u}} = \{((s_1, s_2, l), (s_1', s_2', l')) \in M | l = l'\} .$$

3.4. Score Distribution

Figure 3 shows the STS score distribution of the training set of D_{orig}. As the training set of D_{orig} is randomly sampled from the JSTS dataset, its statistical properties are retained. The distribution shows the imbalance of JSTS.

Figure 4 shows the STS score distribution between the control and treatment groups of negation minimal pairs. Figures 5 and 6 show the STS score distribution between negation minimal pairs with *important* and *unimportant* negation cues, respectively. According to Figures 4 and 5, the STS score distributions are similar between the control and treatment groups.

Table 4: AccChg results in 4-shot setting

Model	M			M_{i}			M_{u}		
	Acc	Acc'	AccChg^*	Acc	Acc'	AccChg^*	Acc	Acc'	AccChg^*
llm-jp-3.1-1.8B-instruct4	44.49	44.21	**-0.27**	29.65	31.69	2.04	49.59	48.52	**-1.07**
llm-jp-3.1-13B-instruct4	47.59	46.07	**-1.52**	27.35	29.60	2.25	54.56	51.73	**-2.82**
Llama 3.1 Swallow 8B Instruct v0.5	31.93	31.43	**-0.49**	23.32	25.79	2.47	34.89	33.38	**-1.51**
Gemma-2-Llama Swallow 2B IT v0.1	52.86	49.62	**-3.24**	46.60	34.42	**-12.17**	55.02	54.85	**-0.17**
Gemma-2-Llama Swallow 9B IT v0.1	37.30	40.56	3.27	30.94	34.26	3.32	39.48	42.73	3.25
Swallow-MS 7B instruct v0.1	32.13	30.87	**-1.26**	23.22	19.95	**-3.27**	35.20	34.63	**-0.57**
Llama 3 Youko 8B Instruct	39.85	36.25	**-3.60**	27.40	20.91	**-6.49**	44.13	41.53	**-2.60**
Gemma 2 Baku 2B Instruct	43.83	44.57	0.74	34.80	33.46	**-1.34**	46.94	48.39	1.46
Llama 3.1 8B Instruct	28.51	26.84	**-1.67**	17.91	18.02	0.11	32.16	29.87	**-2.29**
Llama 3 8B Instruct	22.25	23.23	0.97	10.78	13.08	2.31	26.20	26.72	0.52
Mistral 7B Instruct v0.3	29.86	32.34	2.48	20.97	24.77	3.81	32.92	34.94	2.03
Ministral 8B Instruct	41.22	40.85	**-0.37**	27.67	24.72	**-2.95**	45.89	46.40	0.52
Mistral Nemo Instruct	34.47	34.85	0.38	27.83	28.69	0.86	36.75	36.97	0.22
Gemma 3n E2B instruct	33.99	36.21	2.22	23.59	27.83	4.24	37.56	39.10	1.53
Gemma 3n E4B instruct	37.68	38.52	0.84	30.67	35.66	4.99	40.09	39.50	**-0.59**
Gemma 3 270M instruct	21.66	20.99	**-0.67**	12.65	8.74	**-3.91**	24.76	25.20	0.44
Gemma 3 1B instruct	14.74	13.33	**-1.41**	20.00	18.02	**-1.98**	12.93	11.72	**-1.22**
Gemma 3 4B instruct	39.81	40.80	0.99	21.72	21.88	0.16	46.03	47.31	1.27
Gemma 2 2B Instruct	47.01	46.55	**-0.47**	30.99	27.51	**-3.49**	52.53	53.10	0.57
Gemma 2 9B Instruct	22.03	23.90	1.87	18.12	28.47	10.35	23.38	22.32	**-1.05**
GPT-5-nano	40.82	46.71	5.89	38.82	41.72	2.90	41.51	48.43	6.92
GPT-5-mini	46.45	54.07	7.62	47.40	48.42	1.02	46.13	56.01	9.89
GPT-5	54.14	61.88	7.74	58.61	53.24	**-5.36**	52.60	64.85	12.25
GPT-4.1-nano	47.06	47.03	**-0.03**	32.87	29.44	**-3.43**	51.94	53.08	1.14
GPT-4.1-mini	43.54	42.79	**-0.75**	40.70	29.33	**-11.37**	44.52	47.42	2.90
GPT-4.1	53.21	56.16	2.95	56.41	49.22	**-7.18**	52.10	58.54	6.44

Both Acc and Acc' are presented as percentages, and AccChg is presented as percent points.
The averages across five trials with different random seeds are shown.
* Values in **bold** are negative.

4. Experiment

We evaluated a wide range of LLMs on JSTS-Neg to explore their ability to correctly understand negation.

4.1. Experimental Setup

We evaluated Japanese and Japanese-supporting multilingual LLMs. The first eight models listed in Table 2 are Japanese LLMs, while the rest are multilingual LLMs. We also evaluated generative pretrained transformer (GPT) models via the OpenAI API[7].

Following the approach by Han et al. (2024), we provided prompts including a task guideline to the evaluated LLMs and let them solve STS. The experiment involved zero- and few-shot settings. In the zero-shot setting, only a task guideline was given as a prompt to every LLM. The prompt shown in Figure 7 follows llm-jp-eval[8], except for the instructions. The instruction adhered to the task guideline of JSTS (Kurihara et al., 2022) to ensure the same conditions used for with human annotation. In addition, we used four- and 11-shot prompts. In the four-shot setting, we randomly provided two examples as shots from the training sets of D_{orig} and D_{neg}. In the 11-shot setting, we provided five examples from the training set of D_{orig} and six examples from that of D_{neg} as shots. The shots from one dataset had different gold similarity scores (i.e., the prompt always contained STS instances with all gold similarity scores ranging from 0 to 5.)[9]. In both few-shot settings, we evaluated five patterns of different random seeds and calculated their average as the final evaluation result. The examples sampled in one trial were fixed for all instances. The test set of JSTS-Neg was then used for evaluation in every experimental setting.

[7] We used `gpt-5-nano-2025-08-07`, `gpt-5-mini-2025-08-07`, `gpt-5-2025-08-07`, `gpt-4.1-nano-2025-04-14`, `gpt-4.1-mini-2025-04-14`, and `gpt-4.1-2025-04-14` available at https://openai.com/index/openai-api/.

[8] https://github.com/llm-jp/llm-jp-eval

[9] We did not evaluate a 12-shot setting with six examples per dataset because no STS instance with gold similarity score is 5 appeared in the training set of D_{orig}.

Table 5: AccChg results in 11-shot setting

Model	M			M_{i}			M_{u}		
	Acc	Acc'	AccChg^{*}	Acc	Acc'	AccChg^{*}	Acc	Acc'	AccChg^{*}
llm-jp-3.1-1.8B-instruct4	47.14	47.14	0.00	38.23	39.03	0.80	50.20	49.93	**-0.28**
llm-jp-3.1-13B-instruct4	54.29	54.93	0.65	43.06	45.52	2.47	58.15	58.17	0.02
Llama 3.1 Swallow 8B Instruct v0.5	47.43	48.84	1.41	36.89	33.03	**-3.86**	51.05	54.28	3.23
Gemma-2-Llama Swallow 2B IT v0.1	55.32	55.29	**-0.03**	52.49	40.00	**-12.49**	56.29	60.55	4.26
Gemma-2-Llama Swallow 9B IT v0.1	51.42	55.88	4.46	50.29	46.11	**-4.18**	51.81	59.24	7.44
Swallow-MS 7B instruct v0.1	37.25	35.70	**-1.55**	25.90	18.87	**-7.02**	41.16	41.49	0.33
Llama 3 Youko 8B Instruct	40.41	45.93	5.52	31.47	35.39	3.91	43.49	49.56	6.07
Gemma 2 Baku 2B Instruct	43.20	42.53	**-0.67**	39.62	34.10	**-5.52**	44.43	45.42	1.00
Llama 3.1 8B Instruct	35.87	32.81	**-3.06**	17.27	14.85	**-2.41**	42.27	38.99	**-3.28**
Llama 3 8B Instruct	37.63	37.08	**-0.55**	17.80	15.07	**-2.73**	44.45	44.65	0.20
Mistral 7B Instruct v0.3	34.73	35.13	0.40	21.82	25.15	3.32	39.17	38.56	**-0.61**
Ministral 8B Instruct	42.80	44.32	1.52	32.06	29.44	**-2.63**	46.49	49.45	2.95
Mistral Nemo Instruct	45.23	51.89	6.66	36.84	33.89	**-2.95**	48.12	58.08	9.96
Gemma 3n E2B instruct	25.05	26.05	1.00	15.07	16.73	1.66	28.49	29.26	0.77
Gemma 3n E4B instruct	34.82	34.98	0.15	27.56	26.06	**-1.50**	37.32	38.04	0.72
Gemma 3 270M instruct	18.41	18.61	0.21	18.66	16.46	**-2.20**	18.32	19.35	1.03
Gemma 3 1B instruct	21.95	20.56	**-1.39**	37.59	35.28	**-2.31**	16.57	15.50	**-1.07**
Gemma 3 4B instruct	40.78	40.48	**-0.30**	29.81	28.69	**-1.13**	44.56	44.54	**-0.02**
Gemma 2 2B Instruct	48.62	47.39	**-1.24**	34.75	27.88	**-6.86**	53.39	54.10	0.70
Gemma 2 9B Instruct	47.80	47.91	0.11	46.70	44.66	**-2.04**	48.17	49.02	0.85
GPT-5-nano	34.02	39.09	5.08	28.26	36.46	8.20	36.00	40.00	4.00
GPT-5-mini	46.64	53.40	6.75	48.63	48.31	**-0.32**	45.96	55.15	9.19
GPT-5	56.12	64.34	8.22	59.84	57.96	**-1.88**	54.83	66.53	11.70
GPT-4.1-nano	44.16	44.91	0.75	30.83	26.17	**-4.66**	48.75	51.37	2.62
GPT-4.1-mini	45.93	45.37	**-0.56**	46.43	33.89	**-12.55**	45.76	49.32	3.56
GPT-4.1	54.55	54.67	0.12	62.57	50.40	**-12.17**	51.79	56.14	4.35

Both Acc and Acc' are presented as percentages, and AccChg is presented as percent points.
The averages across five trials with different random seeds are shown.
* Values in **bold** are negative.

4.2. Evaluation Metrics

As we defined STS as a six-value classification task, we adopted the accuracy as the main evaluation metric. To evaluate the performance toward negation, we used the accuracy change (AccChg) that measures the change in accuracy among negation minimal pairs as follows:

$$\mathrm{AccChg} = \mathrm{Acc}' - \mathrm{Acc}\ ,$$
$$\mathrm{Acc} = \frac{1}{|M|} \sum_{((s_1,s_2,l),(s'_1,s'_2,l'))\in M} \mathbf{1}[\hat{l} = l]\ ,$$
$$\mathrm{Acc}' = \frac{1}{|M|} \sum_{((s_1,s_2,l),(s'_1,s'_2,l'))\in M} \mathbf{1}[\hat{l'} = l']\ .$$

where $\hat{l}$ and $\hat{l'}$ are the predicted similarity scores of STS instances (s_1, s_2, l) and (s'_1, s'_2, l'), respectively. AccChg is the subtraction of Acc' and Acc, which are the accuracies for the treatment and control groups, respectively. The change in performance toward negation can be evaluated using this metric. AccChg ranges from -1 to 1, with a value closer to 0 indicating a smaller change in performance owing to negation and a value closer to -1 indicating a poorer performance on the treatment group than on the control group.

4.3. Results

The experimental results are listed in Tables 3, 4 and 5.

In the zero- and four-shot settings, AccChg varies across models. In the 11-shot setting, many models show negative AccChg values on M_{i}, indicating that these models cannot capture the change in gold similarity score caused by a negation cue. In contrast, AccChg on M_{u} tends to be positive in the same setting. This suggests that many models ignore a negation cue that does not change the gold similarity score. In particular, the GPT series show this tendency regardless of the settings. These results indicate that LLMs that are given many shots or closed LLMs predict similarity ignoring negation cues.

5. Conclusion and Future Work

In this study, we developed JSTS-Neg, a Japanese STS dataset for evaluating the negation understanding ability of language models. We evaluated various LLMs on JSTS-Neg to analyze their performances to negation. Closed LLMs or LLMs with many shots tended to ignore negation cues in STS.

In future work, we will evaluate the performance of models more extensively on JSTS-Neg. In addition, we plan to expand JSTS-Neg with affixal negation cues and double negation, which are not covered in its current version.

6. Acknowledgements

We would like to thank Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata for providing the JSTS dataset and its annotation guideline used in this research. The computation of experiments was partly carried out on supercomputer "Flow" at Information Technology Center, Nagoya University.

7. Bibliographical References

Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. *Computing Research Repository (CoRR)*, abs/2407.03963.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

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. In *Proceedings of the First Conference on Language Modeling (COLM)*, University of Pennsylvania, USA.

Namgi Han, Nobuhiro Ueda, Masatoshi Otake, Satoshi Katsumata, Keisuke Kamata, Hirokazu Kiyomaru, Takashi Kodama, Saku Sugawara, Bowen Chen, Hiroshi Matsuda, Yusuke Miyao, Yugo Murawaki, and Koki Ryu. 2024. llm-jp-eval: Automatic evaluation tool for Japanese large language models. In *Proceedings of the Thirtieth Annual Meeting of the Association for Natural Language Processing*. In Japanese.

Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. An analysis of negation in natural language understanding corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (Short Papers)*, volume 2.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Computing Research Repository (CoRR)*, abs/2310.06825.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey,

Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, and Ivan Nardini. 2025. Gemma 3 technical report. *Computing Research Repository (CoRR)*, abs/2503.19786.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large Japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling (COLM)*, University of Pennsylvania, USA.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *Computing Research Repository (CoRR)*, abs/2408.00118.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*.

Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Rethinking STS and NLI in large language models. In *Findings of the Association for Computational Linguistics (EACL 2024)*, pages 965–982, St. Julian's, Malta. Association for Computational Linguistics.

Reiko Yuasa, Asahi Yoshida, Yoshihide Kato, and Shigeki Matsubara. 2025. Evaluation of japanese language understanding benchmarks from the perspective of negation. In *Proceedings of the Thirty-first Annual Meeting of the Association for Natural Language Processing*. In Japanese.

8. Language Resource References

Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. *This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models*. Association for Computational Linguistics (ACL).

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. *A Multilingual Benchmark for Probing Negation-Awareness with Minimal Pairs*. Association for Computational Linguistics (ACL).

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. *An Analysis of Natural Language Inference Benchmarks through the Lens of Negation*. Association for Computational Linguistics (ACL).

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. *JGLUE: Japanese General Language Understanding Evaluation*. European Language Resources Association (ELRA).

Suguru Matsuyoshi, Ryo Otsuki, and Fumiyo Fukumoto. 2014. *Annotating the Focus of Negation in Japanese Text*. European Language Resources Association (ELRA).

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. 2022. *CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation*. Association for Computational Linguistics (ACL).

Taiga Someya and Yohei Oseki. 2023. *JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs*. Association for Computational Linguistics (ACL).

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. *RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs*. Association for Computational Linguistics (ACL).

Takumi Uchida and Hiroaki Nanjo. 2023. *Data Augmentation by Contrapositives for Recognizing Entailment in Sentences with Negation Expressions*. In Japanese.

Takumi Uchida and Hiroaki Nanjo. 2024. *Performance verification of natural language understanding in sentences with negation expressions*. In Japanese.

Hitomi Yanaka and Koji Mineshima. 2021. *Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference*. Association for Computational Linguistics (ACL).

Hitomi Yanaka and Koji Mineshima. 2022. *Compositional Evaluation on Japanese Textual Entailment and Similarity*. MIT Press.

Asahi Yoshida, Yoshihide Kato, Yasuhiro Ogawa, and Shigeki Matsubara. 2025. *Construction of a Japanese Language Inference Dataset for Evaluating Negation Understanding Ability*. In Japanese.