HOTATE: A Japanese Dialogue Corpus Annotated with Responses of Private Thoughts and Public Statements

Yuko Toda[1] **Daisuke Maekawa**[1] **Kota Manabe**[1] **Eito Yoneyama**[1]
Kanade Nonomura[1] **Yuki Fujiwara**[1] **Tomoyuki Kajiwara**[1,2]
[1]Ehime University [2]The University of Osaka
toda@ai.cs.ehime-u.ac.jp kajiwara@cs.ehime-u.ac.jp

Abstract

This study aims to reveal how accurately Large Language Models (LLMs) can deal with a speaker's actual utterances and their true feelings behind them in Japanese dialogue. Speakers use not only ***private thoughts*** which express one's true feelings and intentions, but also ***public statements*** which convey their intentions while considering the interlocutor's feelings and social status. While *public statements* help to maintain interpersonal relationships, they can obscure the speaker's true intention, potentially leading to misunderstandings. We extended existing Japanese dialogue corpora by annotating *public statements* and *private thoughts* responses for each dialogue in the corpora, and then evaluated LLMs' ability to classify and generate between these two types of expressions. The results of the classification task revealed that the current LLMs do not understand those expressions at all, and that training with our corpus can significantly improve the recognition performance. Furthermore, the results of the generation task demonstrated that generating *private thoughts* is more difficult than generating *public statements*, according to both automatic and human evaluations. We release our corpus, which contains 7,964 human-annotated dialogues.

Keywords: Social Relationships, Private Thoughts and Public Statements, Japanese Dialogue

1. Introduction

In daily dialogues, speakers strategically alternate their utterances depending on contextual factors and interpersonal relationships to maintain societal harmony (Melansyah and Haristiani, 2020). While ***private thoughts*** convey genuine emotions and intentions, ***public statements*** serve as socially adaptive expressions aimed at preserving harmonious relationships. However, it is often difficult to determine whether an utterance reflects the speaker's *private thoughts* or not. Suppose a scenario in which a manager asks a member to take on additional work, as shown in Figure 1. The member positively agrees to the request (*public statements*), while internally believing that the manager should handle it insted (*private thoughts*). Consequently, the manager may act against the member's true intention, resulting in frustration.

The gap between *public statements* and *private thoughts* is a crucial issue in achieving smooth communication. Related studies include those on direct and indirect responses (Takayama et al., 2021), sarcasm (Wilson, 2006; Oprea and Magdy, 2019; Abu Farha et al., 2022), and style transfer in terms of politeness (Srinivasan and Choi, 2022). As will be described in Section 2, *public statements* and *private thoughts* are contrastive to them in terms of their meanings and the purposes of use. It is important for maintaining social harmony to understand and switch *public statements* and *private thoughts* in dialogue. However, no dataset explicitly addresses *public statements* and *private thoughts*, and it remains unclear to what extent natural language processing models, including Large Language Models (LLMs), can handle them.

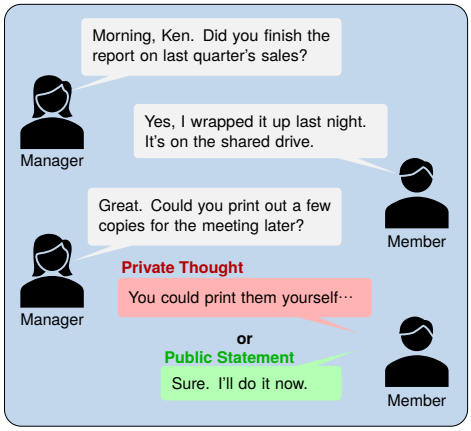


Figure 1: A dialogue illustrating an example situation of *private thoughts* and *public statements*.

To address this issue, we extended existing Japanese dialogue corpora by annotating 3,982 dialogues with *public statements* and *private thoughts* by human annotators, and constructed HOTATE[1] corpus. We investigated the extent to which LLMs can understand *public statements* and *private thoughts* through two tasks: (1) classification of

[1]Dialogue corpus consisting of pairs of *private thoughts* (**HO**nne in Japanese) and *public statements* (**TATE**mae in Japanese). It is available at https://github.com/EhimeNLP/HOTATE

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2987–2995
11–16 May 2026.

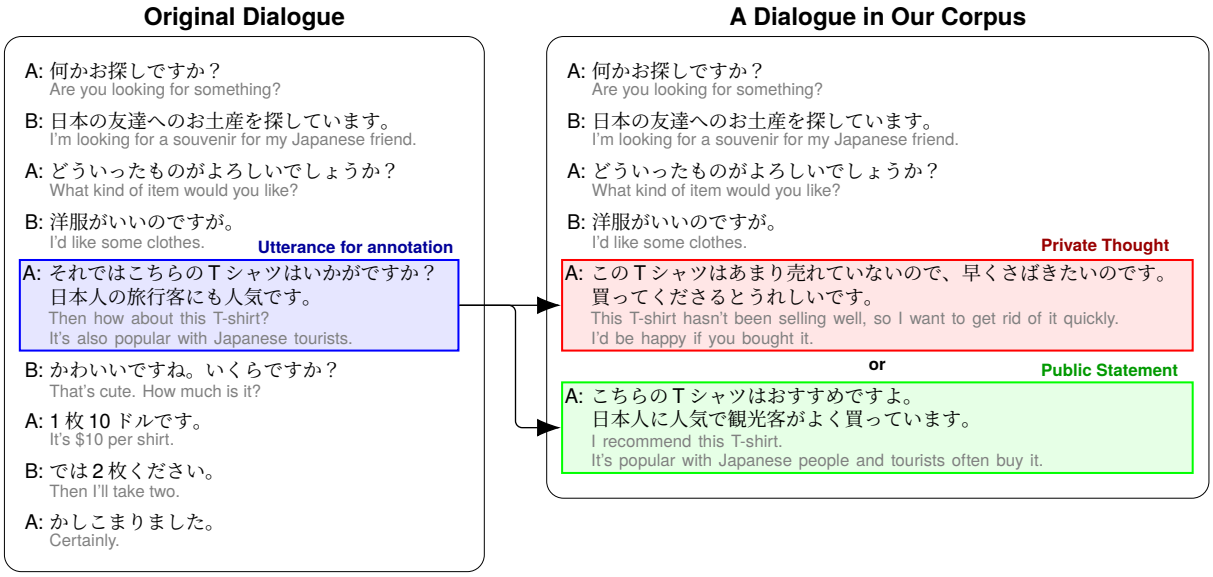


Figure 2: An overview of the annotating operation. The original dialogue is interrupted at a certain point and replaced with the *private thoughts* and *public statements* responses.

whether the final utterance in a dialogue represents *public statements* or *private thoughts*, and (2) transforming *public statements* and *private thoughts*. Experimental results revealed that current Japanese LLMs completely failed to classify *public statements* and *private thoughts* correctly. Furthermore, it was found that transforming *public statements* into *private thoughts* was more challenging than the reverse direction.

2. Related Works

The concept of *public statements* is related to indirect, sarcastic, and polite expressions. In the dialogue, speakers often do not express their requests or intentions directly, but instead convey them through indirect utterances that carry implied meanings (Brown and Levinson, 1987). Prior work on direct and indirect expressions in dialogue (Takayama et al., 2021) showed that transforming indirect utterances into direct expressions improves LLMs' understanding and enhances dialogue quality. Here, it is assumed that indirect utterances convey the same semantic contents as their direct counterparts. On the other hand, *public statements* convey different semantic contents from *private thoughts*, making them contrastive to direct/indirect expressions. Similarly, sarcasm (Wilson, 2006; Oprea and Magdy, 2019; Abu Farha et al., 2022) can be seen as expressions that conceal the speaker's genuine intention. However, sarcasm is used to express negative or critical sentiments toward the interlocutor, making them contrastive in terms of a social function to *public statements*, which aims to maintain smooth interpersonal relationships.

Another research topic closely related to *public statements* is the study of politeness. For example, the multilingual politeness dataset TyDiP (Srinivasan and Choi, 2022) assigns politeness scores to individual utterances independently of the dialogue flow, thereby facilitating text generation that maintains appropriate levels of politeness and contributes to smoother interpersonal communication through language use.

While politeness primarily affects the style of an utterance, *public statements* alter the content itself, contributing to societal harmony. Moreover, since the politeness score in TyDiP was annotated at the utterance level, it does not take into account the contextual flow of dialogue.

Although previous studies have investigated aspects such as indirect expressions, sarcasm, and polite speech, which share partial similarities with *public statements*, they do not consider the unique sociocultural background underlying them. In this study, we construct a corpus focusing explicitly on *private thoughts* and *public statements* to examine how well LLMs can understand and express human intentions.

3. Corpus Construction

In this study, we extended the existing two Japanese dialogue corpora by adding *public statements* and *private thoughts*, thereby constructing a new corpus. This section describes the annotation design, statistical information, and analyses of the constructed corpus.

Dataset	Topic	# Dialogues	Per Dialogue		Per Utterance	Vocabulary Size
			Avg. # Utterances	Avg. # Words	Avg. # Words	
JDD	Dailylife	1,648	5.93	90.35	15.30	4,727
	School	1,692	5.82	99.24	17.12	4,798
	Travel	1,400	5.50	91.66	16.83	4,382
	Health	1,160	6.08	97.59	16.34	3,804
	Entertainment	1,140	5.53	96.82	17.53	3,964
BSD	Business	924	18.57	238.82	12.88	6,102
	Total	7,964	7.26	111.68	15.46	14,808

Table 1: Statistical information of the dialogues, utterances, and vocabulary for each topic in the corpus.

3.1. Source Dialogue Corpora

The use of *public statements* and *private thoughts* can appear not only in business conversations at workplaces and meetings, but also in everyday life's dialogues. To address this, we utilized a Japanese corpus containing dialogue data whose topics are business and daily life.

We used Business Scene Dialogue corpus (BSD)[2] (Rikters et al., 2019), which contains a variety of business situations such as meetings, negotiations, and casual conversation, as a source corpus for the business domain. We also used the Japanese Daily Dialogue (JDD)[3], which consists of dialogues with five topics: *Dailylife, School, Travel, Health, Entertainment*, as a source corpus for the daily conversation domain. With these corpora, we construct our corpus, which spans both business and daily life by adding annotated utterances of *public statements* and *private thoughts* responses to each dialogue.

3.2. Public Statements and Private Thoughts Annotation

We hired 33 annotators via crowdsourcing service *Lancers* [4]. To ensure annotation quality and consistency, we exclusively hired the certified workers with top-tier ranking on *Lancers*. Those certified workers are skilled crowdworkers who meet criteria such as a history of consistently high ratings, a deadline compliance rate of 90% or higher, and high trust scores from clients, thus ensuring more stable work quality compared to general workers. The 33 workers hired were all certified workers. An hourly wage of ¥1,400 was paid for annotation work. This amount exceeds the minimum compensation level on *Prolific* ($8, roughly ¥1,220)[5], a representative crowdsourcing service, and is considered a sufficient payment level for the task contents and quality requirements.

Figure 2 shows an overview of the annotation. As shown in the figure, we asked annotators to select an utterance in a dialogue and replace it with the response of *private thoughts* and *public statements*. Therefore, each dialogue in the annotated data was constructed with the original dialogue history up to the point where the utterance was replaced, adding the two annotated responses. When conducting annotation, the following guidelines were provided to annotators to ensure consistent quality of outputs.

1. Ensure that all annotations are natural and adequate as speaker utterances and intentions in the context of the dialogue.
2. Do not introduce any new background information or contextual setting not derived from the previous dialogue.
3. Clearly distinguish meaning and intent in pairs, and avoid similar expressions between them.

Moreover, we applied a two-stage quality check process to ensure corpus quality. First, the authors manually reviewed each annotator's output. Authors requested annotators to re-annotate for data which deemed inconsistent with the guidelines. After this step, authors manually reviewed and revised all data, ensuring contextual consistency and appropriate wording. This two-stage quality assessment enabled high-quality, consistent annotations while preserving the flow of the dialogue.

3.3. Statistical Information of the Corpus

This corpus obtained 7,964 dialogues by assigning two entries-*public statements* and *private thoughts*-to each of the original 3,982 dialogues. Statistical information about the constructed corpus is shown in Table 1. Since Japanese is a language that does not use word separators such as spaces, we used the Japanese morphological analyzer [6] (Takaoka

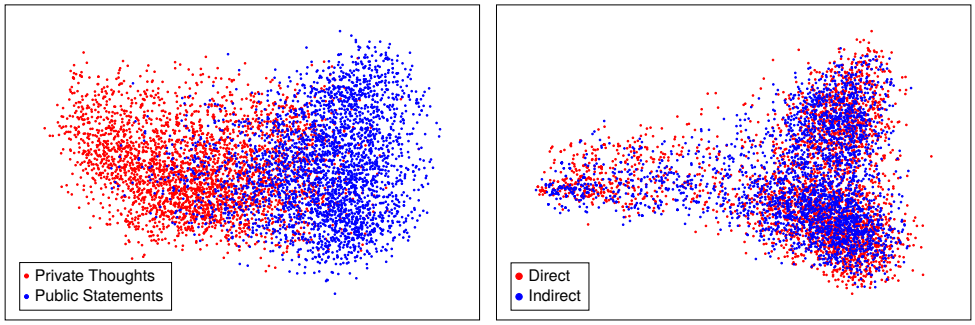
[2] https://github.com/tsuruoka-lab/BSD

[3] https://github.com/jqk09a/japanese-daily-dialogue

[4] https://www.lancers.jp

[5] https://www.prolific.com/pricing

[6] https://github.com/WorksApplications/SudachiPy



(a) Our corpus (b) Direct and indirect corpus

Figure 3: Visualization of embeddings from 3,000 utterance pairs in our corpus and direct/indirect corpus.

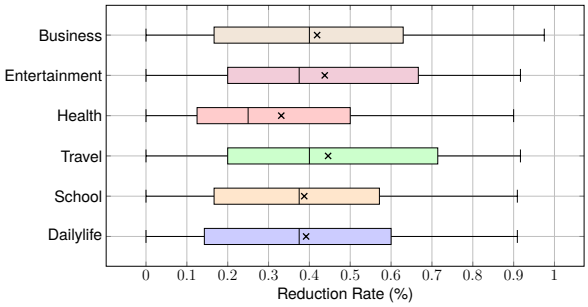


Figure 4: The reduction rate of the annotating operation. Marks of × represent the mean value.

et al., 2018) for word segmentation [7] during measurement. Regarding topic-specific tendencies, the number of utterances in the business domain tended to be higher than in other topics. This is attributed to the fact that the source BSD corpus contains more utterances than the JDD corpus.

Furthermore, as shown in Figure 2, annotators interrupt the original dialogs at the point of the replacement. This operation reduces partial utterances of the original dialogues. Figure 4 shows the information on the reduction rate by topic. For the *Health* topic, the reduction rate was relatively low, confirming a tendency for annotators to assign *public statements* and *private thoughts* to the latter part of the dialogue. In contrast, the median reduction rate was approximately 40% for other topics, showing a relatively stable trend. This suggests that cases where the dialogue is interrupted in the first half are rare, and that annotators assigned the *public statements* and *private thoughts* based on the context of the original dialogue.

3.4. Semantic Difference

We analyzed the semantic difference between *public statements* and *private thoughts* using embedding representation from a Japanese text embedding model[8]. Figure 3 shows the sentence embeddings for randomly sampled 3,000 sentence pairs from our corpus and direct/indirect responses corpus (Takayama et al., 2021), where dimensions were reduced by principal component analysis (Maćkiewicz and Ratajczak, 1993). Embeddings from our corpus formed clusters overall, while partial overlap exists between *public statements* and *private thoughts* as shown in Figure 3a. This suggests that there exists a certain degree of difference in semantic features between the two. On the other hand, embeddings from direct/indirect corpus formed a large overlap as shown in Figure 3b, suggesting that they show high semantic textual similarity. These results indicate that our corpus captures semantic features of utterances from a different perspective of the existing direct/indirect responses corpus.

3.5. Sentimental Difference

Lexical-level Analysis To analyze lexical feature differences between *public statements* and *private thoughts*, we examined word frequencies in both utterances. We used the same setting as in Section 3.3 for measurement. Additionally, we excluded words common to both with frequencies of 30 or more occurrences in order to clarify the difference between them.

The top-10 frequently occurring words are shown in Table 2. In utterances of *private thoughts*, negative words such as ”troublesome” (面倒), ”impossible” (無理), ”difficult” (難しい), and ”not good at” (苦手) appeared frequently. In contrast, positive words such as ”look forward to” (楽しみ), ”interesting” (面白い), and ”work hard” (頑張る) were widely used in utterances of *public statements*. These observations suggest that while *public statements* tend to expand dialogue turns through positive expressions, *private thoughts* tend to constrain dialogue turns through expressions of difficulty or negation.

[7] https://github.com/WorksApplications/Sudachi

[8] https://huggingface.co/cl-nagoya/ruri-v3-310m

Rank	*private thoughts*	*public statements*
1	面倒 (Troublesome)	頑張る (Work Hard)
2	高い (Expensive, High)	楽しみ (Look Forward to)
3	必要 (Necessary)	願う (Wishing)
4	苦手 (Not Good at)	楽しい (Enjoyable)
5	無理 (Impossible)	嬉しい (Happy)
6	調べる (Check)	面白い (Interesting)
7	難しい (Difficult)	今度 (Next Time)
8	悪い (Bad)	美味しい (Delicious)
9	一人 (Alone)	仕方無い (Can't Be Helped)
10	他 (Else)	楽しむ (Enjoy)

Table 2: Frequent words of *public statements* and *private thoughts* (English follows Japanese).

Sentence-level Sentiment Analysis To reveal the emotional tendencies of *public statements* and *private thoughts* utterances, we conducted sentence-level sentiment analysis. We employed a 5-point scale (strong positive, positive, neutral, negative, strong negative) to assess sentiment features of *public statements* and *private thoughts* utterances. For this analysis, we implemented a sentiment classifier by fine-tuning a Japanese version model of ModernBERT (Warner et al., 2025) using a Japanese sentiment polarity dataset [9](Suzuki et al., 2022).

Figure 5 shows the results of sentiment analysis. When comparing *public statements* and *private thoughts*, the former tended to include more positive utterances, while the latter tended to be more negative, consistent with the findings from the lexical-level analysis. In particular, utterances of *public statements* showed a relatively strong positive tendency, with few utterances classified as negative.

4. Experiments

To examine how accurately LLMs can handle *public statements* and *private thoughts*, we conducted evaluation experiments on both classification and generation tasks.

4.1. Tasks

We train LLMs on our dataset constructed in Section 3, and conduct two experiments to evaluate whether the models' ability to classify and generate *public statements* and *private thoughts* can be improved. The first one is a classification task where a dialogue is input into the model and it predicts whether the final utterance is *public statements* or *private thoughts*. The second one is a generation task where a dialogue is input, like the first one, and the model transforms the final utterance of *public statements* into *private thoughts*, and vice versa. In the experiments, we compared the performance of

[9] https://github.com/ids-cv/wrime

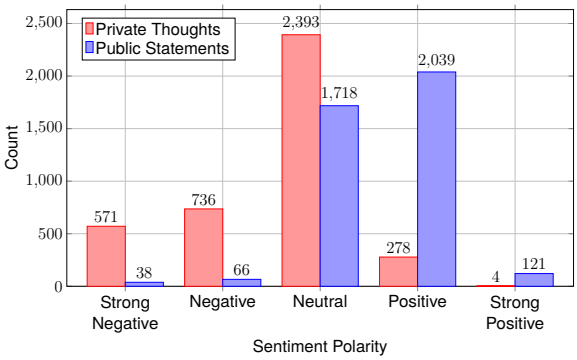


Figure 5: Sentence-level sentiment analysis of *public statements* and *private thoughts* utterances.

untrained LLMs (0-shot) as the baseline against the method that provides multiple examples from the train set of the dataset (few-shot) and instruction-tuning using the train set.

For the classification task, we employed few-shot inference and instruction-tuning to generate responses of labels of *public statements* or *private thoughts*, evaluating accuracy. For the generation task, we employed instruction-tuning to generate responses of transformed utterances as *public statements* or *private thoughts*, evaluating BLEU[10] (Papineni et al., 2002; Post, 2018) and Sentence-BERT[11] (Reimers and Gurevych, 2019, 2020) (SBERT) to the reference sentences.

The dataset was split into training, validation, and test sets at a ratio of 8:1:1. For few-shot learning, we randomly selected 10 examples from the training set, ensuring each example came from a different dialogue for both *public statements* and *private thoughts*.

To assess the quality of the generated sentences and the adequacy as *public statements* and *private thoughts*, we conducted a human evaluation for 50 sentences randomly sampled from the test set. Grammaticality (gram.), contextual consistency as a dialogue(cons.), and contrast clarity between *public statements* and *private thoughts* (cont.) were evaluated on a 3-point scale. Human Evaluation was conducted by three Japanese university students whose native language was Japanese.

4.2. Models

The prompts used for classification and generation tasks are shown in Figures 6a and 6b. For the experiments, we employed prompts that format the model's output as JSON to ensure output stability.

For instruction tuning, we used Hugging Face's

[10] https://github.com/mjpost/sacrebleu

[11] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

```
以下の対話における最後の発言は「本音」か「建
前」かを判定してください。
出力はJSON形式で、キーは "answer" とし、値
は "本音" または "建前" としてください。
なお、理由は出力に含めないでください。

### 対話:

<dialogue>
```

(a) Prompt for classification task

```
以下の対話における最後の発言は<本音/建前>です。
その発言を、<建前/本音>の表現に書き換えてください。
出力はJSON形式で、キーは <"tatemae"/"honne"> と
してください。
なお、理由は出力に含めないでください。

### 対話:

<dialogue>
```

(b) Prompt for generation task

```
Determine whether the last statement
in the following dialogue is "private
 thoughts" or "public statements".
Output should be in JSON format, with
 the key set to “answer” and the
value set to either "private thoughts
" or "public statements".
Do not include the reasoning in the
output.

### Dialogue:

<dialogue>
```

```
The last statement in the following
dialogue is <private thoughts/public
statements>.
Rewrite that statement to express it as <
public statements/private thoughts>.
Output should be in JSON format, with the
key <"public_statements"/"private_thoughts
">.
Do not include the reasoning in the output.

### Dialogue:

<dialogue>
```

Figure 6: Prompts used for classification and generation tasks. English follows Japanese.

SFTTrainer[12] with a learning rate of `2e-4` and the optimizer `AdamW_8bit`. We utilized a single NVIDIA RTX PRO 6000 Blackwell GPU for both training and inference.

We conducted experiments on classification and generation tasks using three types of models. The models we used are as follows:

- **gpt-oss-20b**[13] (OpenAI, 2025) with 20 billion parameters. We used gpt-oss, a reasoning model, with the configuration parameter `reasoning_effort` set to `medium` to control the degree of reasoning applied.
- **llm-jp-13b**[14] (LLM-jp, 2024) with 13 bilion parameters. LLM-jp is a model trained primarily on Japanese, English, and source code, achieving strong performance in Japanese.
- **swallow-8b**[15] (Fujii et al., 2024). Swallow was built through continual pre-training on the Meta Llama 3.1 models, enhancing the Japanese language capabilities of the original model.

[12] https://huggingface.co/docs/trl/sft_trainer

[13] https://huggingface.co/openai/gpt-oss-20b

[14] https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4

[15] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

4.3. Results

Classification Task Table 3 shows the accuracy scores of each model under the 0-shot, 10-shot, and instruction-tuning settings. In the 0-shot setting, all models showed limited accuracies around 50%, which is nearly equivalent to random guessing. Although performance improved in the 10-shot setting, it still remained insufficient for a binary classification task. After instruction-tuning, all models showed further improvement, with swallow-8b achieving the highest accuracy of 92.98%. These results suggest that training with our constructed corpus is effective for recognizing the distinction between *public statements* and *private thoughts*.

Generation Task Table 4 shows the results of the automatic and human evaluations for each model. In both automatic and human evaluations, generating ”*private thoughts* → *public statements*” consistently scored higher than its inverse generation. For the automatic evaluation metrics, swallow-8b achieved the highest scores for both BLEU and SBERT. On the other hand, gpt-oss-20b scored generally lower than the other models, particularly for BLEU. In human evaluation, for ”*private thoughts* → *public statements*” generation, all models achieved scores of around 2.6 to 3.0 on average in all three metrics, with relatively high scores in grammaticality. On the other hand, for ”*public statements* → *private thoughts*” generation, scores

Model	Accuracy (%)		
	0-shot	10-shot	Instruction
gpt-oss-20b	**55.51**	55.39	89.72
llm-jp-13b	45.61	65.66	91.85
swallow-8b	50.38	**78.82**	**92.98**

Table 3: Accuracy scores for each model in the classification task.

Model	Automatic		Human		
	BLEU	SBERT	gram.	cons.	cont.
private thoughts → public statements					
gpt-oss-20b	10.59	0.69	2.98	2.76	2.61
llm-jp-13b	12.14	0.70	2.93	2.71	2.59
swallow-8b	13.84	0.70	2.91	2.76	2.73
public statements → private thoughts					
gpt-oss-20b	8.02	0.67	2.57	2.24	2.00
llm-jp-13b	9.60	0.68	2.91	2.61	2.42
swallow-8b	8.73	0.68	2.87	2.51	2.47

Table 4: Evaluation scores for each model in the generation task. In columns, "gram." stands for grammar, "cons." for consistency, and "cont." for contrast clarity.

decreased for all models compared to the inverse generation. In particular, scores of the contrast clarity between *public statements* and *private thoughts* only achieved scores of around 2.0 to 2.5. Overall, while "*private thoughts* → *public statements*" generation demonstrated consistently high performance in both automatic and human evaluations, "*public statements* → *private thoughts*" generation revealed issues with the consistency and contrast clarity between them.

4.4. Analysis

Training Data Size Analysis To investigate the effect of training data size on classification performance, we evaluated the accuracy changes as the amount of training data increased. For training the classification models, we used the first $N \in \{300, 500, 1000, 2000, 4000, 6000\}$ samples from the training set. The validation and test sets were the same as those described in Section 4.1. All other experimental settings followed those described in Section 4.2.

Figure 7 shows how the classification performance changes with different amounts of training data. Overall, we observed that accuracy improved as the number of training data increased. However, while performance increased up to around $N = 1,000$, it plateaued thereafter and converged around 90%. When comparing across models, gpt-oss-20b achieved relatively high accuracy with a small data size, whereas swallow-8b and llm-jp-13b showed a rapid performance improvement as the training data increased. In particular, swallow-8b consistently outperformed the other models regardless of the amount of training data.

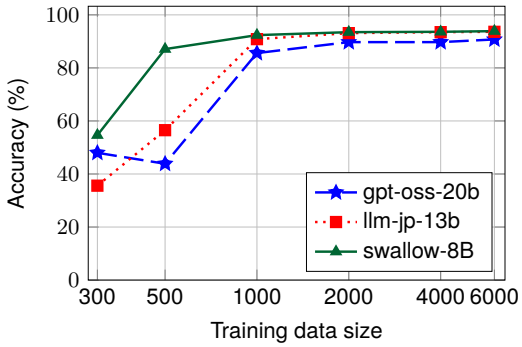


Figure 7: Models classification performance across different training data sizes.

Dialogue Topics Analysis Our corpus was built based on two existing Japanese dialogue corpora, JDD and BSD. JDD corpus consists of daily dialogues across multiple topics: *School*, *Dailylife*, *Travel*, *Entertainment*, and *Health*, and BSD is a corpus for the business domain (Rikters et al., 2019). To analyze the impact of dialogue topics on LLMs' performance in classifying *public statements* and *private thoughts*, we evaluated their few-shot performance for each topic. In this experiment, the data of each topic was split into training, validation, and test sets at a ratio of 8:1:1. Few-shot examples were randomly sampled from the training set. For each topic, we ensured an equal number of public and private examples, and prohibited sampling both types from the same dialogue. We used the models' default generation parameters, and accuracy on the test set was used for evaluation.

Table 5 shows the accuracy of each model by topic. Overall, topics such as *Health* and *Travel* showed relatively high accuracy for most models, suggesting that such dialogues in these topics are easier to handle. In the 0-shot setting, all models achieved around 50% accuracy, indicating a clear limitation when no task-specific content is provided. Among the models, we found that gpt-oss-20b consistently achieved the highest performance across all topics in the 0-shot setting, and each model has its own specific strong topics. In the 10-shot setting, accuracy improved for all models except gpt-oss-20b. In particular, swallow-8b improved significantly across all topics, achieving the highest accuracy around 85% in the *Health* topic. In certain topics such as *Travel* and *Health*, llm-jp-13b also demonstrated notable improvement, confirming the effectiveness of few-shot examples for this classifi-

Model	Accuracy (%)					
	School	Dailylife	Travel	Entertainment	Health	Business
			0-shot			
gpt-oss-20b	55.29	52.41	54.29	54.39	61.21	64.89
llm-jp-13b	46.47	46.39	44.29	45.61	47.41	46.81
swallow-8b	52.94	47.59	44.29	52.63	46.55	44.68
			10-shot			
gpt-oss-20b	54.12	57.23	47.86	55.26	61.21	52.13
llm-jp-13b	60.00	55.42	71.43	54.39	75.00	52.13
swallow-8b	74.12	75.90	76.43	76.32	85.34	68.09

Table 5: Few-shot performance comparison across different topics. Scores are reported in %.

cation task. In contrast, gpt-oss-20b showed only limited improvement from the 0-shot setting, suggesting that its performance saturates even without additional in-context examples. These results indicate clear topic-dependent effects in model performance and demonstrate that few-shot prompting is generally effective for improving classification accuracy across most topics.

5. Conclusion

In this study, we constructed the HOTATE corpus by annotating 3,982 dialogues from the existing Japanese dialogue corpora, JDD and BSD, with human-annotated response pairs of *private thoughts* and *public statements*, totaling 7,964 annotations. Using this corpus, we evaluated the ability of large language models (LLMs) to understand *private thoughts* and *public statements* in dialogue. We designed two tasks for this evaluation, which consist of classifying whether the final utterance in a dialogue expresses a *private thought* or a *public statement*, and converting between *private thoughts* and *public statements*. The experimental results revealed that current Japanese LLMs perform poorly in classifying *private thoughts* and *public statements*. Converting from *public statements* to *private thoughts* was found to be particularly challenging compared to the reverse. Furthermore, the training with our corpus resulted in a significant improvement in classification performance, demonstrating the effectiveness of the corpus for recognizing and distinguishing between *public statements* and *private thoughts*.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24K20840.

Bibliographical References

Ibrahim Abu Farha, Steven Wilson, Silviu Oprea, and Walid Magdy. 2022. Sarcasm Detection Is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5284–5295.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics.

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*.

LLM-jp. 2024. Llm-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese Llms. *arXiv:2407.03963*.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal Components Analysis (Pca). *Computers & Geosciences*, 19(3):303–342.

Raden Regine Melansyah and Nuria Haristiani. 2020. Analysis of Japanese Refusal Speech Acts to an Invitation as a Tatemae. In *Proceedings of the 3rd International Conference on Language, Literature, Culture, and Education (ICOLLITE 2019)*, pages 112–115.

OpenAI. 2025. Gpt-oss-120b & Gpt-oss-20b Model Card. *arXiv:2508.10925*.

Silviu Oprea and Walid Magdy. 2019. Exploring Author Context for Detecting Intended Vs Perceived Sarcasm. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, pages 2854–2859.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A Call for Clarity in Reporting Bleu Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence Embeddings Using Siamese Bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the Business Conversation Corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61.

Anirudh Srinivasan and Eunsol Choi. 2022. TYDIP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738.

Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2022. A JApanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7022–7028.

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. 2018. Sudachi: A Japanese Tokenizer for Business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. Direct: Direct and Indirect Responses in Conversational Text Corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Deirdre Wilson. 2006. The Pragmatics of Verbal Irony: Echo or Pretence? *Lingua*, 116(10):1722–1743. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.