

Evaluating LLaMA-2’s Adaptation to Social Context in Japanese Emails via Fine-Tuning

Muxuan Liu^{1,2} Tatsuya Ishigaki²

Yusuke Miyao^{3,2} Hiroya Takamura² Ichiro Kobayashi^{1,2}

¹ Ochanomizu University

² National Institute of Advanced Industrial Science and Technology

³ The University of Tokyo

{liu.muxuan, koba}@is.ocha.ac.jp

{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp

yusuke@is.s.u-tokyo.ac.jp

Abstract

We explore the capability of the LLaMA-2 models in generating Japanese business emails that accurately reflect social contexts. The current issue is that the unmodified LLaMA-2 model struggles to produce emails suitable for various social situations in Japanese culture. To address this problem, we fine-tuned the model using a business email corpus. Our objective is to identify the additional information (annotation labels) necessary to improve the model’s performance in generating contextually appropriate emails. By training the model with annotation labels representing different social statuses and positions, we investigate the effective input information for incorporating these social contexts into the generated text. Through ablation experiments and manual evaluation, we identify the necessary annotation labels to enhance the accuracy of text generation that reflects social contexts. Additionally, we evaluate the generated emails using two common GPT-based evaluation methods.

1 Introduction

LLMs (Large Language Models) have made remarkable advances in the field of deep learning, playing a crucial role in natural language generation. Recent studies have increased focus on how LLMs process and adapt to specific knowledge. In this paper, we explore the capabilities of LLMs in generating Japanese business emails, with a particular focus on the automatic generation of language expressions considering social status and cultural elements. In Japanese business emails, the use of honorifics and language expressions according to social status is important. These elements deeply affect the content and context of emails and are essential for ensuring appropriate communication.

Table 1 provides examples of Japanese business emails, illustrating how expressions change based on the social status of the sender and the receiver. The examples include the original Japanese text

From a subordinate to a superior: In Japanese: XX部長 (Honorific title: Indicates respect towards the superior) いつもお世話になっております。 (Set phrase: Expresses gratitude and appreciation) 下記のプロジェクトに関する報告書を添付いたしました。 (Formal expression: Uses keigo “いたしました” to show respect) ご確認のほどよろしくお願ひ申し上げます。 (Formal request: Uses keigo “申し上げます” to show respect) 山田太郎 (Sender’s name)
Translation: Dear Manager, Thank you for your continued support. I have attached the report regarding the project below. I would appreciate it if you could review it. Sincerely, Taro Yamada
From a superior to a subordinate: In Japanese: 山田さん (Name with san: A respectful but less formal way to address a subordinate) お疲れ様です。 (Set phrase: Acknowledges the hard work of the subordinate) 以下のプロジェクトに関する報告書を添付しました。 (Direct expression: Uses direct form “しました” indicating less formality) ご確認のほどよろしくお願ひします。 (Request: Uses standard polite form “お願ひします”) 佐藤一郎 (Sender’s name)
Translation: Dear Yamada, You did a good job today. I have attached the report regarding the project below. Please review it. Sincerely, Ichiro Sato

Table 1: Examples of Japanese Business Emails with Annotations

and their English translations. The first example shows an email from a subordinate to a superior. The language used in this email is formal and re-

spectful, utilizing honorifics and polite expressions appropriate for addressing someone of higher status. The second example is an email from a superior to a subordinate, where the language is less formal, reflecting the superior’s higher status. These examples illustrate that even when intending to convey the same message, the way emails are expressed can differ due to the unique social hierarchy and cultural norms in Japanese business communication. To improve LLMs understanding of social relationships in Japanese business emails, we conducted experiments using a Japanese business email dataset and the LLaMa-2-7B model developed by Meta AI¹, fine-tuned the model based on annotation labels related to the social status of the receivers and senders to automatically generate Japanese emails. We performed ablation experiments to evaluate the impact of each annotation label on the quality of generated emails. By systematically removing individual labels and observing the effects on email generation, we were able to identify which specific labels are essential for improving contextual accuracy. Additionally, we assessed the effectiveness of two GPT-based evaluation methods: few-shot prompting and chain-of-thought (CoT) prompting. These methods were used to determine how well different annotation labels and prompting techniques capture and reflect social contexts in the generated emails. By analyzing the results, we aim to provide a clearer understanding of the necessary inputs and methods to enhance the contextual appropriateness and overall quality of automatically generated Japanese business emails.

2 Related Work

Recent studies have advanced our understanding of how LLMs process knowledge and adapt to different cultural and social contexts. For example, Farquhar et al. (2023) analyzed LLMs in an unsupervised environment, discussing key challenges related to data preprocessing, model interpretability, and the accuracy and reliability of knowledge discovery. Kovač et al. (2023) evaluated how LLMs reflect different cultural perspectives, personal values, and personality traits. They used psychological questionnaires to analyze the controllability of LLMs’ perspectives, exploring methods to reflect personal and cultural values and personality traits

in LLMs. Masoud et al. (2023) quantitatively analyzed how well LLMs can adapt to different cultural values using a framework of cultural congruence. They assessed the extent to which LLMs reflect cultural values and personality traits based on Hofstede et al. (2010)’s cultural dimensions. Nguyen et al. (2023) reported on the development and utilization of a multilingual dataset supporting 167 languages. This dataset provides a foundation for LLMs to learn diverse linguistic cultures and adapt to different cultural contexts. Salewski et al. (2023) evaluated how accurately LLMs can mimic individuals with different attributes such as age, profession, gender, and skin color, revealing how LLMs reflect social characteristics and biases. These studies shed light on various aspects of LLMs’ knowledge processing and social adaptability, examining their ability to understand and represent diverse perspectives.

In addition to these studies, several works have focused on the evaluation of text generated by LLMs. One of the key challenges in evaluating natural language generation (NLG) models is the development of reliable and valid evaluation metrics. Traditionally, automatic metrics such as BLEU, ROUGE, and METEOR have been used to assess the quality of generated text by comparing it to reference texts. However, these metrics often fail to capture the nuanced aspects of human communication, such as style, coherence, and context appropriateness. Recent developments in evaluation methodologies have started to leverage the capabilities of LLMs as evaluators themselves. Hackl et al. (2023) introduced the concept of using GPT-based models for evaluating the stylistic quality of generated text, demonstrating that these models can provide more human-like assessments compared to traditional metrics. This approach leverages the inherent language understanding capabilities of LLMs to perform nuanced evaluations. Another promising direction is the use of chain-of-thought (CoT) prompting, which guides the evaluation process by explicitly modeling the reasoning steps taken by humans during evaluation. Building on the insights from Wei et al. (2022), who demonstrated that CoT prompting significantly improves the performance of LLMs in complex reasoning tasks, Liu et al. (2023b) proposed the G-Eval method. This method utilizes GPT models for comprehensive evaluation of generated text, focusing on various dimensions such as fluency, relevance, and coherence. G-Eval incorporates chain-

¹<https://huggingface.co/meta-LLaMa/LLaMa-2-7b-hf>

of-thought prompting and a form-filling paradigm to systematically assess multiple aspects of the text, achieving high correlation with human judgments. The method has demonstrated significant improvements in alignment with human evaluations compared to traditional metrics, particularly in tasks requiring high levels of creativity and contextual understanding.

Our study builds on these advancements by employing both few-shot prompting and CoT prompting to evaluate the generated Japanese business emails. We aim to assess the effectiveness of different annotation labels in incorporating social contexts into the text and to determine which evaluation method better captures the stylistic and contextual appropriateness of the emails. This dual evaluation approach not only provides a more comprehensive assessment of the generated emails but also contributes to the ongoing research on the evaluation methodologies for NLG tasks.

3 Corpus Annotation

In the experiments, we used a Japanese business email corpus reflecting social contexts (Liu et al., 2023a). This corpus was constructed to analyze the impact of social contexts, such as the social status and intimacy between speakers, on the use of Japanese. As shown in Table 2, the corpus includes business emails that clearly indicate social status, annotated with tags that denote the roles and hierarchical relationships of the speakers. The annotations leverage contextual information from Systemic Functional Linguistics (SFL) (Halliday and Matthiessen, 2014.), which considers the establishment of linguistic systems with respect to social contexts. This forms a corpus that emphasizes information related to social roles. As shown in Table 3, the Japanese business email corpus includes 770 situations corresponding to various sender actions, each containing emails written by five different workers. For a comprehensive description of the corpus and Systemic Functional Linguistics, please refer to the Appendix.

4 Experiments

4.1 Methodology

The experiments were conducted based on the ablation settings shown in Table 4. The objective was to enhance the model’s ability to generate texts considering social contexts by fine-tuning LLaMA-2 model using “situation,” “text,” and “labels” data

Situation You are under the care of department A of your client. Please write a year-end greeting email to all members of department A at your client.	
Text Subject: Greetings for the End of the Year To all members of department A at XX Corporation, I am writing to express my gratitude for your continuous support throughout the year. My name is XX from XX Corporation. As the year-end approaches, there is only a little time left in this year. I would like to express my sincere appreciation for your significant cooperation during this fiscal year. We will continue to do our best in our business as much as possible in the coming years, so we would appreciate your continued support. Finally, I would like to express my best wishes for your further prosperity. I hope you have a wonderful new year. From XX at XX Corporation	
Labels (Participants)	
Superiority relationship (receiver)	Superior
Superiority relationship (sender)	Subordinate
Sender’s role	Employee
receiver’s role	All members of a department in a client company
Internal/External	External
Number of senders	Individual
Number of receivers	Multiple
Labels (Speech function)	
Sender’s action	Assertion
Sender’s detailed action	Greeting
Exchange role	Giving
Exchange item	Information

Table 2: Example corpus: Email text and its labels for an employee greeting all members of a department in a client company (adapted from (Liu et al., 2023a))

Sender’s Action	Number of situations	Percentage of situations	Number of Emails
Refusal	70	0.09	350
Request	100	0.13	500
Apology	100	0.13	500
Reminder	100	0.13	500
Gratitude	100	0.13	500
Greeting	100	0.13	500
Notification	100	0.13	500
Inquiry	100	0.13	500
Total	770	1	3850

Table 3: Statistics Showing Characteristics of the Corpus (Modified from (Liu et al., 2023a))

extracted from the corpus, as shown in the example in Table 2. Specifically, using 11 types of labels indicating social relationships included in the corpus (e.g., hierarchical relationships, status, internal-external relations), we conducted ablation experiments to examine the impact of these labels on the generated texts. For the ablation experiments,

Model	Situation & Text	SR_R	SR_S	SR	RR	IE	NS	NR	SA & SDA	ER & EI
Model-0	✓									
Model-1	✓	✓								
Model-2	✓	✓	✓							
Model-3	✓	✓	✓	✓						
Model-4	✓	✓	✓	✓	✓					
Model-5	✓	✓	✓	✓	✓	✓				
Model-6	✓	✓	✓	✓	✓	✓	✓			
Model-7	✓	✓	✓	✓	✓	✓	✓	✓		
Model-8	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Model-9	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 4: Details of Ablation Experiments. The abbreviations are: SR_R (Superiority relationship (receiver)), SR_S (Superiority relationship (sender)), SR (Sender’s role), RR (Receiver’s role), IE (Internal/External), NS (Number of senders), NR (Number of receivers), SA (Sender’s action), SDA (Sender’s detailed action), ER (Exchange role), EI (Exchange item).

the parameters were set with a learning rate of $1e-4$, 100 epochs, a batch size of 4 per training step, and a gradient accumulation step count of 2. To optimize the model’s memory usage and computational efficiency, we utilized automatic device mapping along with BF16 precision. We randomly selected 3,080 emails from our dataset for training purposes, using these to adjust and fine-tune our model. Following the training phase, we employed another set of 770 emails to validate the model’s performance, ensuring that it generalized well across different but unseen data points. After training, the output limit for each fine-tuned model was set to 300 tokens, and new emails were generated. After validation, we evaluated the model’s text generation capabilities using 80 distinct situations. We extracted 10 situations from each of eight different sender actions, resulting in a diverse set of 80 situations. Each model then generated one email per situation. This approach ensured a balanced representation of various business email behaviors and offered a comprehensive assessment of the model’s performance across different communication styles. Additionally, to compare the quality of the generated emails, LLaMa-2-7B model was also used to generate emails for the same situations, and compared its results with those of emails generated by models set with different parameters previously.

4.2 Evaluation Method

We evaluated the generated emails based on two aspects: (1) **Stylistic Evaluation**: Assessing whether the generated emails conform to the standard style of Japanese emails, and (2) **Label Evaluation**: Determining whether the generated emails are appropriate for the labels, meaning whether the content and structure of the emails accurately reflect the

social context and roles indicated by the labels that should be present in the corpus. For the stylistic evaluation, two human reviewers manually scored 30 emails randomly selected from the 80 emails. We then applied the same criteria to have GPT-4 score these emails using two different methods: few-shot prompting (refer to Section 4.2.1) and chain-of-thought (CoT) (refer to Section 4.2.2). The effectiveness of these methods was compared by calculating the kappa coefficient. For the label evaluation, all emails were manually scored, and the results were statistically analyzed (refer to Section 4.2.3).

4.2.1 Automatic Evaluation Using GPT-4 with Few-shot Prompting

We utilize GPT-4 and few-shot prompting (Brown et al., 2020; Wang et al., 2020; Song et al., 2023) to evaluate the email texts generated by each model. Few-shot prompting is a technique in which the model is given a few examples of the task it needs to perform, which significantly enhances the model’s ability to generalize and perform well on the task without extensive fine-tuning. By leveraging this capability, the model can learn from a small number of examples to generate appropriate responses or predictions. In our evaluation, the texts are input into GPT-4 following a set of rules using the Few-shot prompting method, to observe the characteristics of the topics output by each generation model. We aim to obtain scores for the content of the emails generated by each model and the reasons for those scores. Regarding the uncertainty of scoring by LLMs, it has been revealed that LLMs are sensitive to the order of inputs (Wang et al., 2023). Specifically, it has been pointed out that the order of results can lead to completely opposite conclusions. LLMs

tend to be biased towards responses at certain positions, a phenomenon recognized as “Positional Bias”. When the quality difference between evaluated candidates is significant, positional bias is less impactful. To address this issue, it has been suggested to take multiple scores and average them, or to change the input order multiple times and average the scores. Therefore, in this paper, we take the scores three times and calculate their average.

4.2.2 Automatic Evaluation Using GPT-4 with Chain-of-Thought (CoT) Reasoning

Several studies (Amatriain, 2024; Hsieh et al., 2023; Zhou et al., 2022; Li et al., 2024) show that LLMs have a significant advantage in prompt generation, often surpassing human-written prompts in various natural language processing tasks. This advantage is particularly evident in tasks requiring nuanced understanding and contextual adaptation, where LLMs can generate more effective and precise prompts. Building on this foundation, we evaluated the content of Japanese emails by referencing the G-Eval method (Liu et al., 2023b) and incorporating the Chain-of-Thought (CoT) prompting technique to ensure thoroughness and accuracy in the scoring process. By leveraging these advanced methods, we aim to enhance the evaluation process, making it more reliable and consistent. This approach highlights the practical applications of LLM-generated prompts in improving the accuracy and efficiency of automated assessments.

We first used an initial prompt to guide the model in generating a detailed prompt, as shown below:

Based on the following labels and definitions, please generate a detailed prompt to evaluate the quality of the email content.

The labels are as follows: [Subject], [Salutation], [Self-introduction], [Content and Purpose], [Closing Greeting], [Signature].

The definitions for each label are as follows:

[Subject]: The email subject should specifically and clearly indicate the main content of the email.

[Salutation]: At the beginning of the email, use an appropriate salutation for the receiver or receiver group.

[Self-introduction]: The email should start with the sender’s self-introduction. For example, introducing oneself as “I am XX.”

[Content and Purpose]: The email body should explain the purpose of the email (refusal, request,

apology, reminder, thanks, greeting, notice, inquiry) and the relevant details.

[Closing Greeting]: The email should conclude with a polite closing greeting expressing respect and gratitude to the receiver. For example, ending with “Thank you.”

[Signature]: At the end of the email, include the sender’s signature so that the receiver knows who the email is from.

Evaluate whether the above labels are included, and assign a score (1 or 0) for each label.

Subsequently, we utilized the prompt generated by GPT-4 and made slight modifications to the scoring criteria to align with human standards. The final prompt used for scoring is as follows:

This is a task to evaluate email content. Based on the following email content, please assign a score (1 or 0) for each label.

Email content: (omitted)

Evaluation process:

- 1. Subject:*** First, check the subject. Evaluate if the email subject is appropriate.
- 2. Salutation:*** Next, assess if the greeting is appropriate. After the subject, is there an appropriate greeting for individual receivers (e.g., “Mr. XX,” “Ms. XX”) and for multiple receivers (e.g., “Everyone,” “Dear all”)?
- 3. Self-intro:*** Then, check if there is a self-introduction. Is there a self-introduction of the sender at the beginning of the email?
- 4. Content and Purpose:*** Evaluate if the details related to the purpose are explained in detail in the body of the email.
- 5. Closing Remarks:*** Lastly, check if there is a closing greeting at the end of the email.
- 6. Signature:*** Confirm if the sender’s signature is included at the end of the email.

******The evaluation criteria are as follows:***

Subject: Evaluation: Is the subject of the email indicated? Score: 1 (appropriate) / 0 (lack of)

Salutation: Evaluation: After the subject, is there an appropriate greeting for the receiver (e.g., “Mr. XX,” “Ms. XX”) Score: 1 (appropriate) / 0 (inappropriate or lack of)

Self-introduction: Evaluation: Is there a self-introduction of the sender at the beginning of the email? Score: 1 (appropriate) / 0 (lack of)

Content and Purpose: Evaluation: In the body of the email, are there explanations related to the purpose such as clarification, request, apology, reminder, gratitude, greeting, notice, or inquiry? Score: 1 (even if not entirely clear or somewhat confusing, as long as the intention is somewhat understood) / 0 (no meaning understood at all)

Closing Remarks: Evaluation: Is there a closing greeting at the end of the email? Score: 1 (appropriate) / 0 (lack of)

Signature: Evaluation: Is there a sender’s signature at the end of the email, such as XX?

Score: 1 (appropriate) / 0 (lack of)

*****Please output the evaluation results in the following format:**

Subject: Score

Salutation: Score

Introduction: Score

Content and Purpose: Score

Closing Remarks: Score

Signature: Score

With this detailed prompt, the model can think step-by-step and provide scoring. Please note that the original prompt were provided in Japanese. For readability, the content is presented in English in this paper. For the original Japanese version, please refer to the Appendix A.

4.2.3 Manual Evaluation Based on Social Context Labels

We manually evaluate the extent to which the emails generated by each model reflect those labels. Additionally, we analyze the presence of specific words or phrases in the emails generated by each model to verify if they are included in a manner that meets our expectations. Furthermore, we focus on cross-comparing the results generated by each model to evaluate performance differences between the models.

5 Result

5.1 Few-shot prompting

To evaluate the details of the generated emails, we used GPT-4 to score the same set of emails that were scored by two human reviewers, as introduced in Section 4.2. As shown in Figure 1, Few-shot prompting was employed, allowing the model to learn from three examples and six scoring criteria. Each time the generated emails violated any of these rules, one point was deducted, with a perfect score being 6 points. GPT-4 output scores based on these rules, enabling a comparative evaluation of the quality of emails generated by different models.

As shown in Appendix Figure 5, the Few-shot Prompting scoring approach results in the highest average scores for the fine-tuned Models 6 and 7, while in contrast, the performance of the untuned LLaMa-2-7B is significantly lower. As shown in the top half of Appendix Figure 6, many of the emails generated by LLaMa-2-7B contain repetitive

Good Example 1: Subject: Regarding the School Festival
Dear Students,
This is XX from the School Festival Executive Committee. The dates for this year's school festival have been decided as from X Month X Day to X Day. Clubs that wish to participate should fill out the necessary information in the attached file and contact us by email by X Month X Day. Thank you for your cooperation.

XX Executive Committee XX

Good Example 2: Subject: To All Employees in Charge of Pamphlet Creation
Dear Sales Department,
Thank you for your hard work. This is XX, Company A, our client, has requested samples of food-related pamphlets that we have produced so far. Therefore, those who have been involved in their production, please reply to me, XX, by next Wednesday.

XX Sales Department XX

Good Example 3: Subject: Thank You
Dear Mr./Ms. A,
Thank you for your hard work. This is XX. Thank you very much for covering for me when I was absent due to a cold. It was very helpful. I look forward to your continued support.

XX

Scoring Rules:
Clear Subject: The email subject should clearly and specifically indicate the main content of the email.
Appropriate Salutation: At the beginning of the email, an appropriate salutation should be used to address the recipient or recipient group.
Self-introduction: The email should begin with a self-introduction of the sender. For example, introducing oneself as "This is XX. (XXです)"
Specific Content and Purpose: The email body should clearly explain the purpose of the email and the relevant details.
Polite Closing Remarks: The email should end with polite closing remarks that express respect and gratitude to the recipient. For example, ending with "Thank you for your cooperation. (よろしくお願いたします)"
Clear Signature: The email should include the sender's signature at the end so that the recipient can identify who the email is from.
For each violation of the above rules, 1 point will be deducted, with a maximum score of 6 points.
Based on the above rules, calculate the score for the following emails (+ Each model-generated email for the same scenario)

Figure 1: Few-shot prompting

Label	R1 vs. R2	R1 vs. GPT-4	R2 vs. GPT-4
Subject	1.000	0.423	0.423
Salutation	1.000	0.216	0.216
Self-intro	1.000	0.420	0.420
Content and Purpose	0.911	0.152	0.262
Closing Remarks	1.000	0.524	0.524
Signature	0.923	0.286	0.250

Table 5: Cohen's Kappa Values of Few-shot prompting scores compared to human ratings. R1: Reviewer 1, R2: Reviewer 2

sentences, making it difficult to generate appropriate email content. However, as shown by the Kappa scores in Table 5, there is a high level of agreement between human reviewers 1 and 2, but a significantly lower level of agreement between the reviewers and the predictions generated by GPT-4. This suggests that the Few-shot Prompting scoring approach is less accurate.

5.2 Chain-of-Thought (CoT) Reasoning

We used the same comparison method as in the previous table 5 to compute the kappa values shown in Table 6. It is evident that the GPT-4 model demonstrates a high level of agreement with human raters across most dimensions, as indicated by the Kappa values approaching or equal to 1. For instance, in aspects such as "Subject," "Salutation," "Self-intro," "Closing Remarks," and "Signature," the agreement between human raters and the model is

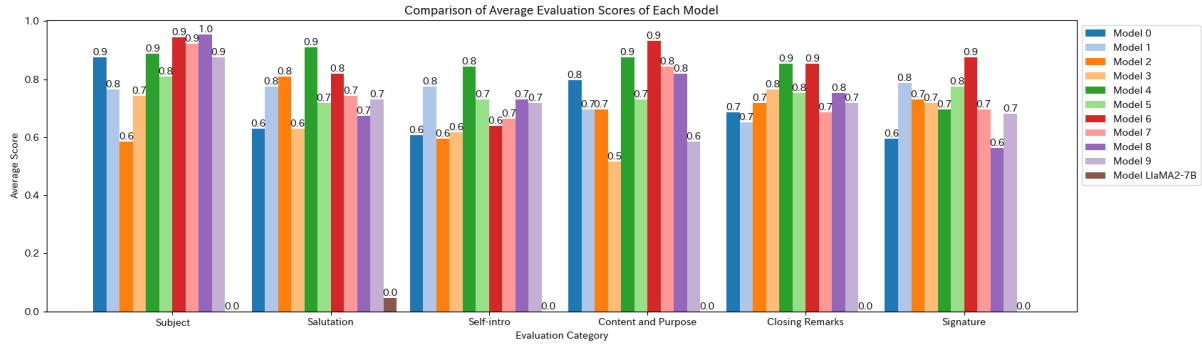


Figure 2: Comparison of Average Evaluation Scores of Each Model

Label	R1 vs. R2	R1 vs. GPT-4	R2 vs. GPT-4
Subject	1.000	1.000	1.000
Salutation	1.000	0.918	0.918
Self-intro	1.000	0.862	0.862
Content and Purpose	0.911	0.734	0.830
Closing Remarks	1.000	0.889	0.889
Signature	0.923	1.000	0.923

Table 6: Cohen’s Kappa Values of CoT reasoning scores compared to human ratings. R1: Reviewer 1, R2: Reviewer 2

nearly perfect. This suggests that GPT-4 effectively mimics human scoring in these areas. However, there are slight variations observed in certain aspects, such as “Content and Purpose,” where the agreement is relatively lower compared to other dimensions. Despite these variances, the overall trend indicates that GPT-4 is proficient at emulating human scoring across a range of text evaluation criteria. This validation supports the efficacy of the CoT approach in leveraging automated scoring models like GPT-4 for reliable and efficient text evaluation. We evaluated all generated emails using the CoT method, and Figure 2 displays the average performance of different models across several categories of email assessment. Each category represents a key component of an email, including the subject, salutation, self-introduction, content and purpose, closing remarks, and signature. The scores for each category were determined by assessing whether the emails met the criteria in that category (1 for meeting the criteria, 0 for not meeting), and then calculating the average score. Here’s a detailed progression through each model:

- **Model 0** set the baseline using only email content and situation information, achieving moderate scores across the board.
- **Model 1** added the “superiority relationship (receiver)”, which led to notable improvements in salutations and

self-introductions, showcasing how adaptation to the receiver’s status can refine greetings and introductory remarks.

- **Model 2** incorporated “superiority relationship (sender)”, improving salutations slightly further and enhancing signatures, suggesting that understanding both parties’ social statuses helps in tailoring the email’s formal aspects appropriately.
- **Model 3** included the “sender’s role”, which did not show improvement in performance, especially in content and purpose, indicating potential challenges in integrating this identity information effectively.
- **Model 4** added “receiver’s role”, significantly improving self-introductions and salutations by adapting more personally to the receiver’s specifics. This model managed to elevate the self-introduction scores and maintained high performance in subsequent models.
- **Model 5** introduced “internal and external” relationship details, which slightly decreased performance, possibly due to the complexity added by these relational dynamics.
- **Model 6** further added “number of senders”. This label significantly improved the performance in ‘content and purpose’ from 0.7 to 0.9, highlighting the importance of this information in emails involving discussions or announcements.
- **Model 7** added “number of receivers”, where the scores in “content and purpose” and “closing remarks” slightly decreased, suggesting that handling emails with multiple receivers introduced additional complexity.
- **Model 8** included “sender’s action” and “sender’s detailed action”, which enhanced the “content and purpose” significantly, showing that understanding the sender’s specific actions is crucial for accurately crafting the core message of the email.
- **Model 9**, despite utilizing all labels, did not always yield the highest scores.

Above analysis shows that Models 4 to 6 performed relatively well, indicating that these models effectively balanced the amount of contextual information used. While the additional context from new labels generally improved the performance of subsequent models, the integration of all labels

in the final model did not necessarily achieve the highest scores across all categories. This outcome suggests that there may be an optimal amount of information, beyond which the inclusion of more details does not continue to benefit, and might even hinder, model performance.

5.3 Manual Evaluation

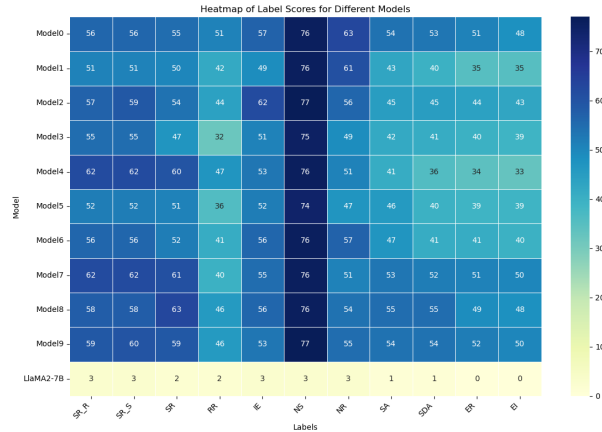


Figure 3: Comparison of Label Scores for Different Models. The abbreviations are: SR_R (Superiority relationship (receiver)), SR_S (Superiority relationship (sender)), SR (Sender’s role), RR (Receiver’s role), IE (Internal/External), NS (Number of senders), NR (Number of receivers), SA (Sender’s action), SDA (Sender’s detailed action), ER (Exchange role), EI (Exchange item).

Additionally, we analyzed the frequency of specific labels in the email content generated by each model, as detailed in Section 4.2.3. As shown in Figure 3, we observed significant variations in scores across different models for various tags. The **LlaMA2-7B** model exhibited very low scores across all tags, with most tag scores being 0 or 1, indicating poor performance. In contrast, **Model0** showed high scores in most tags, particularly in the *NS* and *NR* tags, demonstrating outstanding performance. **Model1** had high scores in the *NS* tag, similar to Model0, but relatively lower scores in other tags such as *ER* and *EI*. **Model2** achieved high scores in the *IE* and *NS* tags, showcasing strong performance. **Model3** had high scores in the *NS* tag but lower scores in the *RR* tags. **Model4** performed well in the *SR_R*, *SR_S* and *NS* tags but had relatively lower scores in the *ER* and *EI* tags. **Model5** scored highly in the *NS* tag but lower in the *SR_R* and *SR_S* tags. **Model6** had high scores in the *NS* tag and also performed well in the *RR*, *SA* and *NR* tags. **Model7** exceeded 50 scores in most tags, indicating excellent performance. **Model8** showed high scores in the *NS* and *SR* tags, with

overall performance close to Model7. **Model9** had the highest score in the *NS* tag, with overall performance close to Model8.

From these results, it can be concluded that crucial labels contributing to the model’s performance and adaptability include the superiority relationship (receiver/sender), sender’s role, receiver’s role, internal/external, number of senders, and number of receivers. The inclusion of these labels significantly improved the model’s performance and adaptability. Overall, in complex situations, as shown in Appendix Table 7, the models tend to confuse relationships between characters, leading to content that deviates from the intended purpose. Conversely, in simpler situations with straightforward relationships, as shown in Appendix Table 8, the models could focus on limited elements and generate more appropriate content. Additionally, it was observed that the labels *SA*, *SDA*, *ER*, and *EI* were not well-learned by the models. This could be due to several reasons: these labels may overlap with information the model already implicitly understands; or the complexity of these labels may exceed the model’s current understanding capabilities.

6 Conclusion

LlaMA-2 struggles with understanding situations in Japanese emails that are easily comprehended by humans. However, by adding specific labels, such as the receiver’s and sender’s social status and identity, we significantly improved the quality of the generated content, particularly in personalized components like salutations and self-introductions. Our ablation study and tag-based evaluation showed that these labels provided the model with more contextual information, enabling it to simulate the human thought process more accurately. While some labels significantly improved the quality of the generated content, others, like "email response" (*ER*) and "email intent" (*EI*) tags, were less effective, indicating that there is room for improvement in these areas. These findings highlight the importance of carefully selecting and integrating labels to enhance model performance in crafting emails that meet specific communicative goals. Future efforts should focus on optimizing the integration and effectiveness of critical labels to improve the model’s ability to generate contextually accurate and nuanced email communications.

References

- Xavier Amatriain. 2024. Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. 2023. Challenges with unsupervised llm knowledge discovery. *arXiv preprint arXiv:2312.10029*.
- Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4 text ratings. *arXiv preprint arXiv:2308.02575*.
- M. A. K. Halliday and Christian M. I. M. Matthiessen. 2014. *Halliday's introduction to functional grammar* /, 4th ed. edition. Routledge, Abingdon, Oxon .
- M.A.K. Halliday. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Open University set book. Edward Arnold.
- M.A.K. Halliday and Christian Matthiessen. 2006. *Constructing Experience Through Meaning: A Language-Based Approach to Cognition*. Continuum. Illustrated edition, 672 pages.
- Shusuke Hirabayashi and Yumiko Hamada. 1988. *Series of Japanese Example Sentences and Problems for Foreigners 10 Honorifics(gaikokujin no tame no nihongo reibun mondai shiri-zu 10 keigo ,in japanese)*. Aratake.
- Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*, 3rd edition. McGraw-Hill Professional.
- Cho-Jui Hsieh, Si Si, Felix X Yu, and Inderjit S Dhillon. 2023. Automatic engineering of long prompts. *arXiv preprint arXiv:2311.10117*.
- Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.
- Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM on Web Conference 2024*, pages 3367–3378.
- Muxuan Liu, Tatsuya Ishigaki, Yusuke Miyao, Hiroya Takamura, and Ichiro Kobayashi. 2023a. Constructing a japanese business email corpus based on social situations. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 499–509.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.
- Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*.
- Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*.
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Appendix

Below is the Japanese version of the initial prompt mentioned in Section 4.2.2:

以下のラベルと定義に基づいて、メール内容の品質を評価するための詳細なプロンプトを生成してください。

ラベルは次の通りです：【件名】、【呼びかけ】、【自己紹介】、【内容と目的】、【終わりの挨拶】、【署名】。

各ラベルの定義は以下の通りです：

【件名】：メールの件名は、メールの主要な内容を具体的かつ明確に示すべきである。

【呼びかけ】：メールの始めに、受信者あるいは受信者グループに適切な呼びかけを使用する。

【自己紹介】：メールは、送信者の自己紹介で始めるべきである。例えば、「XXです」と自己紹介する。

【内容と目的】：メール本文では、メールの目的(断り、依頼、謝罪、催促、感謝、挨拶、お知らせ、問い合わせ)と関連する詳細を説明する。

【終わりの挨拶】：メールは、受信者への尊敬と感謝を表す礼儀正しい終わりの挨拶で締めくくべきである。例えば、「よろしくお願いたします」という言葉で終わる。

【署名】：メールの最後には、送信者の署名を含めることで、受信者が誰からのメールかを把握できるようにする。

上記のラベルが含まれているかどうかを評価し、各ラベルに対してスコア (1または0) を付けてください。

Below is the Japanese version of the final prompt mentioned in Section 4.2.3:

***これは、メール内容の評価タスクです。

**以下のメール内容に基づいて、各ラベルに対してスコア (1または0) を付けてください。

【メール内容】：(略)

【評価プロセス】：

1. まず、件名を確認します。メールの件名が適切かどうかを評価してください。
2. 次に、呼びかけが適切かどうかを評価します。件名の後に、個人の受信者に対する適切な呼びかけ (XXさん、XX様など)、複数人の受信者には (皆さま、みなさんなど) が使用されていますか？
3. その後、自己紹介が行われているかどうかを確認します。メールの冒頭で送信者の

自己紹介が行われていますか？

4. メール本文で目的と関連する詳細が詳しく説明されているかを評価します。
5. 最後に、メールの終わりに挨拶が含まれているかを確認します。
6. メール最後に送信者の署名が含まれているかを確認します。

***各ステップで確認した内容に基づいて、それぞれのラベルに対してスコアを付けてください。

**評価対象と基準は次の通りです：

【件名】：評価：メールの件名は示されていますか？ スコア：1 (適切) / 0 (欠如)

【呼びかけ】：評価：件名の後に、受信者に対する適切な呼びかけ (XXさん、XX様など) が使用されていますか？ スコア：1 (適切) / 0 (不適切または欠如)

【自己紹介】：評価：メールの冒頭で送信者の自己紹介が行われていますか？ スコア：1 (適切) / 0 (欠如)

【内容と目的】：評価：メール本文で、断り、依頼、謝罪、催促、感謝、挨拶、お知らせ、問い合わせなどの目的と関連する内容が説明されていますか？ スコア：1 (明確でなくてもいい。一部混乱してもいい。意図がある程度わかる) / 0 (意味が全くわからない)

【終わりの挨拶】：評価：メールの終わりに挨拶が含まれていますか？ スコア：1 (適切) / 0 (欠如)

【署名】：評価：XXなど、メールの最後に送信者の署名が含まれていますか？ スコア：1 (適切) / 0 (欠如)

***評価結果を以下の形式で出力してください：

【件名】：スコア

【呼びかけ】：スコア

【自己紹介】：スコア

【内容と目的】：スコア

【終わりの挨拶】：スコア

【署名】：スコア

Below is the Japanese version of the 1 figure:

良い例1:
 件名: 学園祭について¥¥¥¥¥¥学生各位¥¥¥¥¥¥学園実行委員のXXです。
 ¥¥¥¥¥¥今年の学園祭の開催期間はX月X日〜X日となりました。¥¥¥¥参加を希望するサークルは、添付のファイルに必要事項を記入し、X月X日までにメールにてご連絡ください。¥¥¥¥¥¥よろしく申し上げます。¥¥¥¥¥¥-----¥¥¥¥¥¥実行委員 XX¥¥¥¥

良い例2:
 件名: パンフレット制作者の社員へ¥¥¥¥営業部門各位¥¥¥¥¥¥お疲れ様です。
 ¥¥¥¥¥¥取引先A社が、これまでに制作した食品関係のパンフレットのサンプルをご覧になりたいとのことです。¥¥¥¥については、今までに製作に関わった人は、来週の水曜日X/Xまでに私XXまでに返信願います。
 ¥¥¥¥-----¥¥¥¥営業部門 XX

良い例3:
 件名: ありがとうございます¥¥¥¥A殿¥¥¥¥お疲れさまです。XXです。¥¥¥¥風邪をひいて欠勤した際、代わりに出勤いただきありがとうございました。¥¥¥¥非常に助かりました。¥¥¥¥今後ともよろしく申し上げます。
 ¥¥¥¥-----¥¥¥¥XX

スコアリングのルール:
 明確な件名: メールは、メールの件名は、メールの主要な内容を具体的に示すべきである。
 適切な呼びかけ: メールは、メールの始めに、受信者あるいは受信者グループに適切な呼びかけを使用する。
 個人紹介: メールは、送信者の自己紹介で始めるべきである。
 例: 「XXです」と自己紹介する。
 具体的な内容と目的: メール本文では、メールの目的と関連する詳細を詳しく説明する。
 終わりの挨拶と礼儀正しい言葉使い: メールは、受信者への尊敬と感謝を表す礼儀正しい終わりの挨拶で締めくくべきである。
 例: 「よろしく申し上げます」という言葉で終わる。
 明確な署名: メールは、送信者の署名を含めることで、受信者が誰からのメールかを把握できるようにする。

以上の各ルールに違反するたびに1点を差し引き、満点は6点。
 以下のメールに対して、以上のルールに従い、スコアを出力する。
 (+各モデルが生成した同一場面のメール)

Figure 4: Few-shot Prompting

場面: あなたはクライアント企業に向いた際、担当のA様に「あなたの同僚のBさんには以前大変お世話になったので、一度ご挨拶したい」と言われました。Bさんにその旨を伝えるメールを書いて下さい。
model0: "件名: A様とは、お疲れ様です。表題の件ですが、A様は以前大変お世話になったのでこちらですんですが、1度お話いただけるとありがたいものにと存じます。私の都合で2度とこちらでご回答いただけようでしたら一度お挨拶をさせていただけると幸いです。私の都合で返信の件ですが、私の連絡先をお書きしておきてすみません。またの名を言いませんが、何卒よろしくお願い致します。XX会社XX部XX"
model1: "XX社 B様 お世話になりました。XX社のXXです。新鮮nessでご挨拶いたします。A様からのご誘いで、Bさんにご挨拶することになりました。XX社 XX"
model2: "メール名: 同僚BさんについてA様 お世話になります。先日は大変お世話になりました。担当のA様に、Bさんには、以前一度ご挨拶したいとのご要望です。"
model3: "B様いつも大変お世話になっております。本日一度お見舞いさせていただきます。XXです。"
model4: "B部 XX様 お世話になりました。XXです。A社 XXです。私、おかげでとことでも存在しないでしょう。早速お伺い申し上げますが、XX様にとどめがありましたらご連絡をお待ちしております。恐縮ではありますがどうかご対応のほどよろしくお願い申し上げます。"
model5: "件名: 大変お世話になりました。 A様。お世話になりました。XXのXXです。Bさんに感謝しています。皆さんにもご指導のほどお願いいたします。よろしくお願い致します。XX月XX日"
model6: "件名: B様についてA様 お疲れ様です。XXです。Bさんには大変お世話になったので、Bさんにもお会いしたいと思っています。これからXXでは業種の視察や研修などの機会が多く、Bさんには大変助かりそうです。それで、一度BさんにBさんにBさんが大変助かりそうですと思っています。以上、よろしくお願い致します。XX"
model7: "件名: Bさんお世話になりました。ご連絡B様 お疲れ様です。XXです。A様からご連絡を受けております。A様からはBさんには大変お世話になりました。BさんにはA様からご連絡を受けた後の間も悪くなりません。よろしくお願い致します。"
model8: "件名: A様のご要請 XX株式会社 XX部 XX様 いつもお世話になっております。XX株式会社XX部のXXです。A様のご要請をいただきありがとうございます。Bさんに一度ご挨拶したいため、A様のご要請をいただきました。お忙しいところ恐縮ですが、Bさんのご教えの通りやすうに、お担当ご調整いただければと思います。今後の繋がりをおくるため、一度ご挨拶をさせていただければと思いますので、お手数をおかけしてはどうかご理解頂きます。お忙しいところ恐縮ですが、お手数をおかけしてはどうかご連絡の方をお知らせください。どうぞよろしくお願い致します。"
model9: "件名: Bさんとは今後の協力 B様 お疲れ様でございます。XX社のXXです。A様からBさんの宜しくお伝えさせていただきました。ご連絡させて頂きました。Bさんには、大変お世話になったのですが、恐縮ですが、一度ご挨拶できると嬉しくお思いにしております。WHEN、WHERE、TIMEをご連絡していただけますでしょうか。お忙しいところ恐れ入りますが、よろしくお願い致します。XX社XX"
L1aMa-2-7B: "こんにちは、あなたの同僚のBさんには以前大変お世話になったので、一度ご挨拶したいと言われました。(300トークンまで重複)"

Table 7: Example of output from an email indicating difficulty in identifying a large number of characters (some formatting is omitted to save space, only the text is retained)

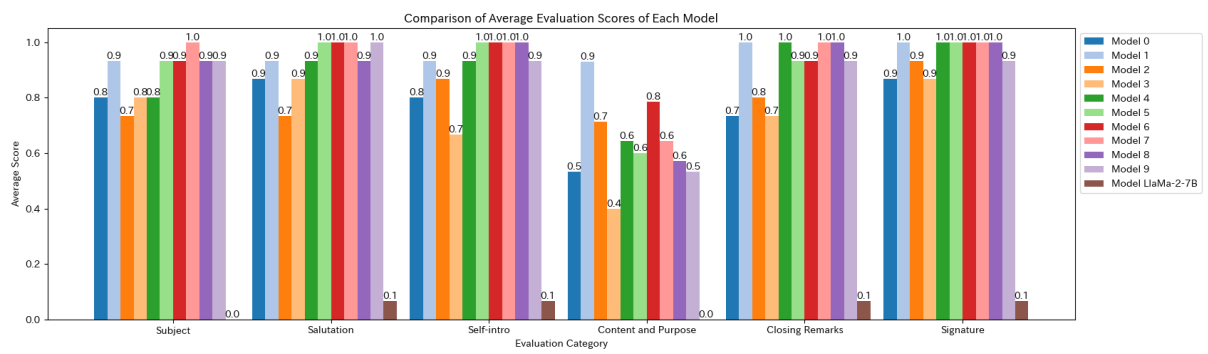


Figure 5: This figure shows the Comparison of Average Evaluation Scores of Each Model when using few-shot prompting. Due to the low kappa values, these results were not adopted in the main text.

<p>場面：あなたはAサークルに所属している学生です。2週間後にクリスマスイベントの飲み会を予定していますが、サークルメンバーの出欠希望がまだ集まっていません。飲み会への出欠について、メンバー全員に問い合わせるメールを考えなさい。</p>
<p>model0:"件名：飲み会の出欠について サークルメンバーへ 2週間後の飲み会ですが、まだ出欠確認が出来ていない方がいます。お手数ですが、期日までに必ずご回答ください。-----XX"</p>
<p>model1:"件名：サークル・Aクリスマスイベントについて Aサークルの皆さんお疲れ様です。XXです。2週間後にサークルのクリスマスイベントを予定しています。つきましては、until thenからuntil afterまでの出欠、費用の負担状況などについて、XX一月XX日までご返信ください。-----XX"</p>
<p>model2:"件名：飲み会(クリスマスイベント)について Aサークルの皆さんご担当者様お疲れ様です。XXです。2週間後に飲み会を予定しているのですが、出欠についてまだ分かっていない人がいるかと思ひます。ご連絡をお待ちしております。よろしくお願いいたします。-----XX"</p>
<p>model3:"件名：飲み会の出欠について サークルメンバーの皆さんへ Aです。2週間後の飲み会の旨いなりについてですが、出欠のことを早め知りたいと思います。となたかご連絡いただけると助かります。よろしくお願ひいたします。"</p>
<p>model4:"件名：飲み会の出欠について Aサークルのメンバーへ XXです。2週間後の飲み会ですが、出欠についてまだ誰も回答をしてくれていません。私都合でも出るかごめんなさい。誰か出欠することがありますら、先に返信してください。もし出欠が必要なら、ご確認ください。XX"</p>
<p>model5:"Aさんです。クリスマスイベントの飲み会の出欠はまだ誰か知りたいです。ご連絡をお待ちしております。よろしくお願いいたします。"</p>
<p>model6:"件名：クリスマスイベントのご連絡 Aサークルのみなさまお疲れ様です。様々な予定が入り、急なお知らせが必要になってきたのですが、2週間後の日にはクリスマスイベントの飲み会を予定しています。現在出欠がございますので、ご返信のほどお願ひいたします。-----XX"</p>
<p>model7:"件名：飲み会の出欠確認の件 Aサークルメンバー各位お疲れ様です。XX学部XXです。2週間後のクリスマスイベントの飲み会の出欠について、皆さんの出欠確認をお願ひいたします。Aさんから連絡を受けていることをお聞きしていただきました。メールにてお願ひいたします。XX"</p>
<p>model8 "件名：飲み会の出欠について Aサークルのみなさん お疲れ様です。XXです。2週間後にクリスマスイベントの飲み会を予定していますが、サークルメンバーの出欠希望はまだ集まっていません。サークル全員で行けるよう出欠を依頼したいと考えています。ご連絡までお待ちしております。-----XX"</p>
<p>model9:"メンバー各位お疲れ様です。XXです。2週間後にサークルでクリスマスイベントをする飲み会をしようと話しています。できれば出欠も予め知りたいので、出欠希望を知らせてください。よろしくお願いいたします。XX"</p>
<p>L1aMa-2-7B:"こんにちは、あなたはAサークルに所属している学生です。2週間後にクリスマスイベントの飲み会を予定していますが、サークルメンバーの出欠希望がまだ集まっていません。飲み会への出欠について、メンバー全員に問い合わせるメールを考えなさい。(300トークンまで重複)</p>

Table 8: Example of mail output showing relatively little difficulty in identifying situations (some formatting omitted to save space, only text retained)

```

model_Llama-7B_output :
こんにちは、
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー全員に急遽メールを書いているところです。
あなたは学園祭の出し物についての希望を聞くためにサークルのメンバー
-----
Model Llama-7B のスコア : 0/6
1. 件名 : 0点
2. 呼びかけ : 0点
3. 自己紹介 : 0点
4. 内容と目的 : 0点
5. 終わりの挨拶 : 0
6. 署名 : 0点

```

Figure 6: Top: mail generated by LLaMa-2-7B; Bottom: evaluation by GPT-4

A.1 Systemic Functional Linguistics (SFL) and Japanese Business Email Corpus

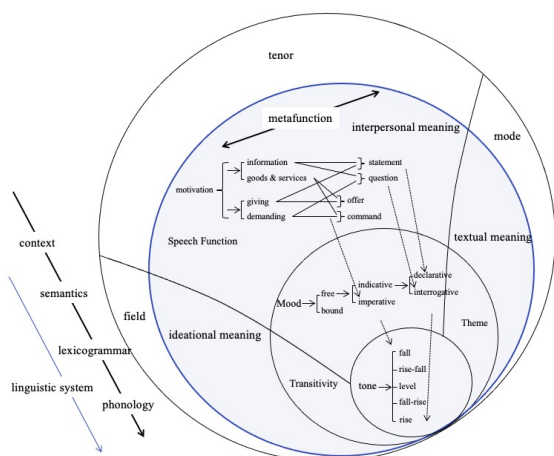


Figure 7: Language systems by systemic functional linguistics adapted from (Halliday and Matthiessen, 2006)

Systemic Functional Linguistics (SFL), founded by M.A.K. Halliday, is essential for understanding the linguistic aspects of social situations, the focus of our research. SFL views linguistic systems as social semiotic systems, emphasizing the interplay between language and social contexts. SFL divides the linguistic system into three semiotic systems: semantic, lexicogrammar, and expression stratum.

Figure 7 outlines SFL’s linguistic system. According to Halliday, situational context is explained through three frameworks: "Field" (what is happening), "Tenor" (who is involved), and "Mode" (how language is used) (Halliday, 1978).

Japanese Business Email Corpus uses SFL to analyze email communication, exploring how it uncovers linguistic knowledge and the relationships between language and social activities. These form a contextually conditioned network of linguistic options for social communication, known as the “system network”. SFL highlights the relationship between situational selection, meaning, and linguistic features like vocabulary and grammar. For example, in an educational context, events like “lecture” and “discussion” occur, and corresponding lexico-grammatical resources such as “present the topic” and “share your thoughts” are selected. The system network represents the process of realizing texts by describing the relationships between different resources (features) and how they are chosen. In terms of “choice,” the system network uses square brackets (‘[’) to indicate the selection of one feature and curly braces (‘{’) for selecting multiple features simultaneously. This framework helps understand how language resources are chosen in the creation of texts (Liu et al., 2023a).

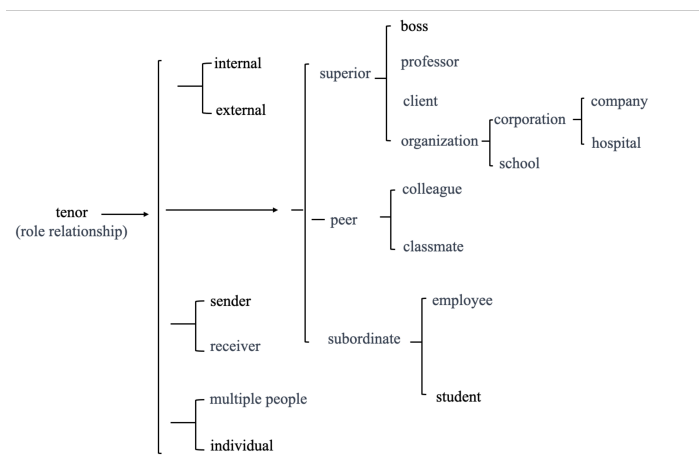


Figure 8: System Network of “Tenor” adapted from (Liu et al., 2023a)

One of the focuses of the Japanese Business E-mail Corpus is examining the “Tenor” relationship in email communication, which refers to the relationship between the sender and receiver. To consider the social standing of participants in typical business email conversations, Japanese Business Email Corpus constructed a selectional system for the tenor relationship. An example of a network system is provided in Figure 8. The attributes of “internal” and “external” represent the internal and external positional relationships of the conversation participants. Generally, “internal” refers to “family, colleagues, or members of the

same group,” while “external” refers to “unfamiliar people, outsiders, people from other companies, or people from other groups” (Hirabayashi and Hamada, 1988). Additionally, to represent the sender’s position, the characters and organizations commonly used in business emails are divided into three attributes: superior, peer, and subordinate (Liu et al., 2023a). The entire corpus is built upon this system.