

# Evaluating Social Intelligence in LLMs via Japanese Honorifics in Email Generation: A Social Semiotic System Perspective

Muxuan Liu<sup>\*,‡</sup>, Tatsuya Ishigaki<sup>‡</sup>, Yusuke Miyao<sup>†,‡</sup>,  
Hiroya Takamura<sup>‡</sup>, Ichiro Kobayashi<sup>†\*,‡</sup>

\* Ochanomizu University † The University of Tokyo

‡ National Institute of Advanced Industrial Science and Technology  
{liu.muxuan, koba}@is.ocha.ac.jp yusuke@is.s.u-tokyo.ac.jp  
{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp

## Abstract

We propose JaSocial, a novel evaluation framework that leverages Japanese emails to comprehensively evaluate large language models' (LLMs) social intelligence across varied social-status relationships. The framework integrates three core components. First, we construct and publicly release a meticulously human-annotated Japanese email dataset covering six distinct social-status contexts, thereby capturing nuanced shifts in social hierarchy and politeness. Second, we adopt Systemic Functional Linguistics (SFL)—a social-semiotic linguistic theory that explicitly models how linguistic choices realize interpersonal relations and hierarchical distinctions—to classify email content in terms of three perspectives: social relationships, speech functions, and honorific expressions. Based on these perspectives, we design an automated evaluation method that assigns each LLM-generated email a contextual appropriateness score, quantifying how well it reflects socially intelligent behavior. Third, we release the full evaluation code to ensure reproducibility and enable fair cross-model comparisons. JaSocial exposes current LLMs' limitations in capturing cultural nuance, while providing an open benchmark for future research.

**Keywords:** Social Intelligence, Japanese Honorifics, Systemic Functional Linguistics

## 1. Introduction

Social intelligence, classically defined as the ability to manage interpersonal relations (Thorndike, 1920; Goleman, 1995), has been increasingly studied through computational models (Guo et al., 2024). In this study, we examine the social intelligence of large language models (LLMs) through the lens of Japanese honorific expressions. These are polite linguistic forms that reflect social roles and interpersonal relationships. In Japanese communication, their appropriate use depends heavily on contextual factors such as the relative status and roles of the speaker and listener, making them an ideal testbed for evaluating context-aware linguistic competence.

We frame this task using Systemic Functional Linguistics (SFL) (Halliday, 1985; Halliday and Matthiessen, 2014), which models language as a stratified system of meaning shaped by social context. For our present purpose, we distill SFL into three strata most relevant to written business email communication, as follows:

- **Stratum 1: Social Context** — In SFL, context is organized through register variables *Field*, *Tenor*, and *Mode*. In our setting, *Field* (business communication) and *Mode* (written email) are fixed, while *Tenor* varies. We operationalize *Tenor* as role relations (e.g., Subordinate → Superior, Student → Professor), reflecting socially conditioned *choices* in the context stratum.

- **Stratum 2: Speech Function** — At the semantic stratum, especially the *interpersonal metafunction*, discourse is modeled as networks of *choices* in the speech-function (mood) system (e.g., statement, question, offer, command). We capture this by annotating sentence-level communicative intent, such as *Apology*, *Inquiry*, or *Reminder*, which realize semantic-level selections.

- **Stratum 3: Honorific Style** — At the lexicogrammatical stratum, meanings from higher levels are *realized* through linguistic resources. In Japanese, honorific morphology and stylistic resources realize interpersonal meanings from Strata 1–2. We categorize these realizations into composite styles such as *Polite*, *Humble+Polite*, or *Honorific+Polite*.

These strata form a realization hierarchy: Social Context constrains Speech Function, which in turn determines the appropriate Honorific Style. This provides the foundation for our JaSocial Evaluation Framework, which consists of three components: (1) a stratally annotated dataset; (2) classifiers trained to infer stratal annotations from generated texts; and (3) a scoring mechanism that compares LLM-generated emails with human-written emails at each stratum.

This framework enables interpretable evaluation of the social intelligence of LLMs. By incorporating SFL, it not only analyzes how LLMs adapt to Social Context, perform communicative functions,

and select appropriate honorific styles, but also assigns stratum-level scores for assessing pragmatic appropriateness (see Figure 1). Our three-layer schema follows the SFL principle of stratification, where meanings are realized across context, semantics, and lexicogrammar. All data and code will be publicly released upon publication at: <https://github.com/ja-email-eval/JaSocial>.

## 2. Related Work

**Automatic evaluation for Japanese LLMs.** Recent years have seen active development of automatic evaluation for Japanese LLMs. A representative effort is *llm-jp-eval*, which integrates multiple Japanese benchmarks and provides a unified pipeline for large-scale assessment (Han et al., 2024).<sup>1</sup> While such suites cover general abilities, they do not target pragmatics central to Japanese business communication, such as sensitivity to SOCIAL CONTEXT (Stratum 1) and appropriateness of HONORIFIC STYLE (Stratum 3).

**Controlling honorifics in generation.** In machine translation (MT), prior work has shown that honorific realization can be controlled at the surface level. Feely et al. (2019) annotate and condition EN→JA Neural MT outputs to control honorific verb morphology, achieving formal style control but not explicitly modeling role relations. More recently, Sekizawa and Yanaka (2023) construct a Japanese honorific conversion task conditioned on social relationships, requiring structure-sensitive transformation of honorific predicates across simple and complex sentences, and show that fine-tuned LLMs outperform prompt-based learning in syntactic generalization.

**Role dynamics, longitudinal change, and learning support.** Empirical studies reveal that honorific usage tracks evolving role identities. Ohashi (2018) conduct a longitudinal study of Japanese email communication, finding that the frequency and type of honorifics shift with participants' changing roles. From a pedagogical perspective, research on Japanese honorific education highlights the importance of social context and interactional settings in learning appropriate honorific usage (Okamoto and Shibamoto-Smith, 2016), while Rahayu (2018) emphasize cultural factors as key determinants of honorific choice.

**Systemic Functional Linguistics (SFL) and honorifics.** SFL offers a stratified account of how linguistic resources encode social relationships and

interpersonal roles. The KeiCO corpus (Liu and Kobayashi, 2022) organizes honorific phenomena with SFL-informed social status annotations. Complementarily, Halliday (1985) shows that linguistic choices such as formality encode interpersonal relations, linking lexicogrammatical forms to social meaning.

**Gap and our contribution.** Existing evaluation toolkits emphasize general abilities; work on honorific generation focuses largely on surface control or task-specific conversion; and empirical studies highlight the importance of social roles and cultural factors but do not provide a direct, interpretable evaluation scaffold for LLM outputs. We address this gap by proposing a stratified, SFL-informed evaluation framework that (i) operationalizes SOCIAL CONTEXT, SPEECH FUNCTION, and HONORIFIC STYLE on a Japanese business email dataset, (ii) trains a classifier to infer these strata from generated texts, and (iii) scores LLM outputs at each stratum against human references.

## 3. Methodology

### 3.1. Overview of the JaSocial Evaluation Framework

Figure 1 presents an overview of the JaSocial Evaluation Framework, which evaluates the social appropriateness of LLM outputs in Japanese honorific email generation. Given a social situation prompt, an LLM generates an email that is automatically analyzed through three interdependent strata grounded in Systemic Functional Linguistics (SFL). Each stratum captures a distinct level of social, semantic, and lexicogrammatical meaning. The outputs are compared with human-written emails, and the resulting similarity scores are combined into a comprehensive final score that quantifies overall social appropriateness.

### 3.2. Dataset Construction and Statistics

We extend the Japanese business email dataset (Liu et al., 2023), which originally used a flat annotation schema, by adding sentence-level stratified annotations across three strata (Social Context → Speech Function → Honorific Style), following the framework of SFL.

The distribution of annotated emails is shown in Table 1, which was used to train, validate, and test the stratal annotation classifiers that power our evaluation framework (Section 3.3). The overall annotation framework is summarized in Table 2.

The original dataset assigned one of eight coarse-grained speech function labels (e.g., gratitude, apology, greeting) at the email level. To en-

<sup>1</sup>The LLM-jp overview paper (LLM-jp et al., 2024) is related but does not itself introduce this evaluation tool.

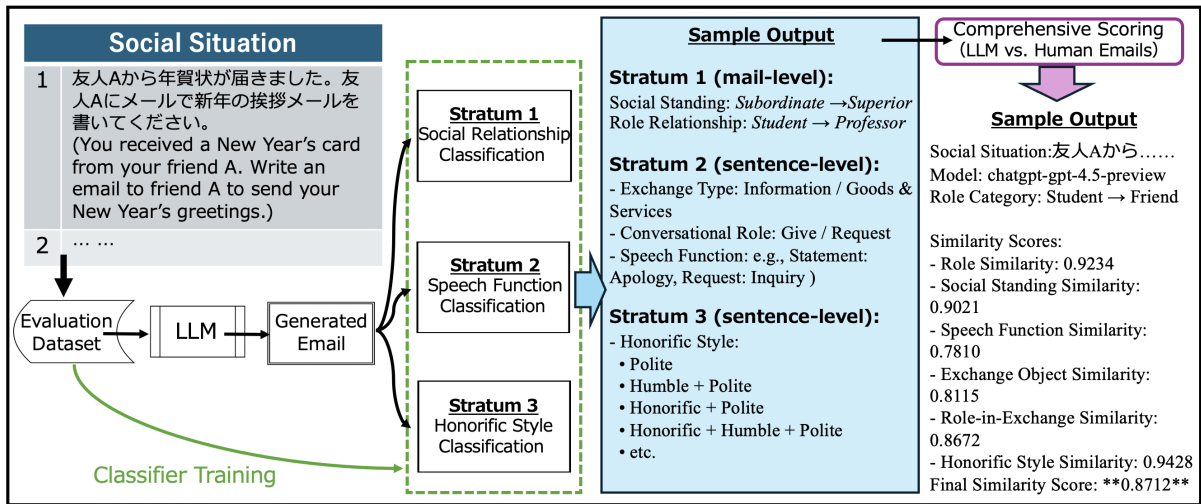


Figure 1: JaSocial Evaluation Framework

Social Standing	Role Relationship	Total	Train (80%)	Val (10%)	Test (10%)
Higher to Lower	Teacher → Student	200	160	20	20
Higher to Lower	Employee → Subordinate	200	160	20	20
Lower to Higher	Student → Professor	200	160	20	20
Lower to Higher	Employee → Supervisor	200	160	20	20
Peer	Student → Friend	200	160	20	20
Peer	Employee → Colleague	200	160	20	20
<b>Total</b>		<b>1200</b>	<b>960</b>	<b>120</b>	<b>120</b>

Table 1: Distribution of the real-email dataset used to train/validate/test the stratal annotation classifiers (Stratum 1 splits). These splits are for classifier training only and are separate from the LLM evaluation benchmark described in Section 3.4.

able fine-grained analysis, we re-annotated 1,200 representative emails at the sentence level. Figure 2 illustrates the framework for sentence-level speech functions (Stratum 2). Each annotation consists of *Object of Exchange*, *Role in Conversation*, and *Sender's Action*. The Sender's Action is structured as *A: B*, where A is a general category (e.g., *Statement*, *Question*, *Offer*, *Request*) and B specifies the functional meaning (e.g., *Apology*, *Inquiry*, *Reminder*).

In total, the dataset contains 5,855 annotated sentences ( $\approx 4.9$  per email). Sentence distributions differ across roles (e.g., Student→Professor: 5.9 vs. Student→Friend: 4.3). Honorific style coverage is also balanced, with Polite (61%), Humble+Polite (31%), and Informal (7%) as the dominant categories.

A compact snapshot of the label families is given in Table 2. Table 3 then presents a fully annotated email example, including Social Context (Stratum 1), subject, recipient name, and sentence-level annotations for Speech Function and Honorific Style (Strata 2 and 3).

Most stratal labels are surface-deterministic and can be reliably inferred from observable linguistic forms: honorific verb morphology (e.g., “いたしま

す” vs. “します”) directly signals style, role relations are given by the scenario, and many speech functions are transparently marked by syntactic constructions or lexical choices. To operationalize this, we first applied a simple rule-based script to pre-assign labels, and then the research team manually reviewed and corrected all annotations. Ambiguous cases accounted for approximately 1% of the sentences and primarily concerned the classification of fixed honorific predicate constructions into composite style categories such as Polite and Honorific+Polite; these were resolved through consensus discussion among all native-speaker co-authors.

Honorific Style is modeled as a single multi-class variable with six composite categories (Informal, Polite, Honorific+Polite, Humble+Polite, Honorific+Humble+Polite, Humble+Informal), which avoids incoherent combinations such as “Honorific+Informal” and ensures both robustness and interpretability.

### 3.3. Classifier Design and Training

We implement a cascaded classifier system to predict Social Context (Stratum 1), Speech Function (Stratum 2), and Honorific Style (Stratum 3)

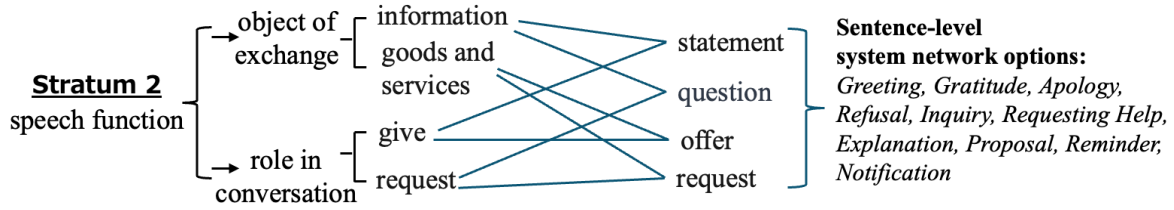


Figure 2: Sentence-Level Speech Function Annotation System (Stratum 2), adapted with modifications from Liu et al. (2023).

Stratum	Annotation Category	Representative Values	$K$ / Granularity
Stratum 1: Social Context	Social Standing Role Relationship	Higher→Lower; Lower→Higher; Peer Teacher→Student; Student→Professor; Student→Friend; Employee→Supervisor/Colleague/ Subordinate	3 / email 6 / email
Stratum 2: Speech Function	Object of Exchange Role in Conversation Sender’s Action	Information; Goods & Services Give; Receive Apology; Inquiry; Request; Reminder; Notification; Greeting; Gratitude; Suggestion; Explanation; Re- fusal; Proposal; Question:Requesting Help	2 / sentence 2 / sentence 12 / sentence
Stratum 3: Honorific Style	Honorific Style	Informal; Polite; Honorific+Polite; Humble+Polite; Honorific+Humble+Polite; Humble+Informal	6 / sentence

Table 2: Consolidated summary of annotation categories across Strata 1–3, including representative values, cardinalities ( $K$ ), and annotation granularity.

from LLM-generated emails. Each classifier is fine-tuned independently using Japanese BERT (*cl-tohoku/bert-base-japanese-v2*<sup>2</sup>), with outputs from higher strata used as auxiliary features for lower ones. An overview of the inputs and targets across strata is summarized in Table 2.

### 3.4. Evaluation Protocol and Situation-Level Analysis

We evaluate LLM generations on 480 case outputs (16 models  $\times$  30 situations across six role relationships). All models were evaluated with identical prompt templates across the 30 social situations, without model-specific tuning, to ensure comparability. For each situation, one email is generated per model and compared against five human-written references crafted under the same situation, yielding per-reference similarities that are averaged into the final score. Evaluation uses the annotation-based scoring protocol introduced in Section 3.4. Given a social situation prompt, we compare each generated email against multiple human-written emails crafted for the same situation. Each sentence or email is annotated with six labels across three strata: Social Standing, Role

Relationship, Object of Exchange, Role in Conversation, Sender’s Action, and Honorific Style. These annotations are represented as category-wise vectors, predicted by trained classifiers introduced in Section 3.3.

To compute the final score, we first concatenate the annotation vectors from all three strata—Social Context (Stratum 1), Speech Function (Stratum 2), and Honorific Style (Stratum 3)—into a single integrated representation  $\mathbf{v}_{\text{all}}^{\text{gen}}$ . This representation is then compared against human-written counterparts  $\mathbf{v}_{\text{all}}^{\text{ref}_i}$  constructed under the same social situation. The final score is defined as the average cosine similarity between the generated and human-written vectors:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^N \cos \left( \mathbf{v}_{\text{all}}^{\text{gen}}, \mathbf{v}_{\text{all}}^{\text{ref}_i} \right) \quad (1)$$

where  $N$  denotes the number of human-written emails available under the same situation. This computation jointly considers all annotation types across the three strata, capturing their interdependence and providing a holistic measure of social appropriateness.

We adopt cosine similarity over concatenated annotation vectors, rather than computing per-metric accuracy, because each generated email is compared against multiple human-written references that often differ in detail but are equally acceptable.

<sup>2</sup>Pre-trained Japanese BERT model provided by the Tohoku NLP Lab: <https://huggingface.co/tohoku-nlp/bert-base-japanese-v2>

Item	Content
Social Situation	<b>You forgot the details of the homework assigned today at school. Moreover, the submission deadline is tomorrow. Consider what kind of email you should send to your friend in order to receive the homework details promptly.</b>
Stratum 1: Social Context	<i>Sender's social standing = Peer Receiver's social standing = Peer Sender's role relationship = Student Receiver's role relationship = Friend In/Out relationship = Internal Number of senders = Individual Number of receivers = Individual</i>
Subject	<b>Homework, Help!!!</b>
Recipient's Name	<i>A-san</i>
<b>Email Body (Sentence-Level) and Annotations</b>	
<b>Sentence 1</b>	<b>Sorry!</b> <i>Annotation: [Stratum 2: Object of Exchange = Information; Role in Conversation = Give; Sender's Action = Statement:Apology; Stratum 3 = Informal]</i>
<b>Sentence 2</b>	<b>Regarding today's homework, I'm sorry, but could you please tell me again?</b> <i>Annotation: [Stratum 2: Object of Exchange = Goods and Services; Role in Conversation = Receive; Sender's Action = Request:Requesting Help; Stratum 3 = Informal]</i>
<b>Sentence 3</b>	<b>Really, I'm sorry!!!</b> <i>Annotation: [Stratum 2: Object of Exchange = Information; Role in Conversation = Give; Sender's Action = Statement:Apology; Stratum 3 = Informal]</i>
<b>Sentence 4</b>	<b>I need your help.</b> <i>Annotation: [Stratum 2: Object of Exchange = Goods and Services; Role in Conversation = Receive; Sender's Action = Request:Requesting Help; Stratum 3 = Informal]</i>

Table 3: A Complete Annotated Email Example with Sentence-Level Stratal Annotations (originally in Japanese, shown as a representative data example)

A strict accuracy measure would assume a single gold label per stratum, penalizing legitimate alternatives in honorific style or speech function. By contrast, our vector-based similarity provides a soft matching mechanism that accommodates natural variation while still capturing socio-pragmatic appropriateness, since our goal is not classification but to use classifier outputs as structured features for evaluating generated texts.

Beyond overall model-level scores, we conduct a situation-level analysis across all 30 scenarios (five situations for each of the six role relationships defined in Table 2). Two robust patterns emerge. First, *upward-facing* contexts (e.g., *Student* → *Professor*, *Employee* → *Supervisor*) are consistently easier across models, with many scores approaching or exceeding 0.90, reflecting clearer socio-pragmatic constraints and stable honorific usage. Second, *peer-level* and *downward-facing* contexts (e.g., *Student* → *Friend*, *Employee* → *Subordinate*) are substantially harder and more variable; some models drop to 0.40–0.50 in certain situations due to overformality or awkward informality. Even within a fixed role relationship, some situations are systematically more difficult (e.g., directive vs. informative styles under *Teacher* → *Student*).

## 4. Experiments

We evaluated the cascaded stratal classifiers with an 80/10/10 email-level split, ensuring no sentence overlap across sets. All models were initialized from *cl-tohoku/bert-base-japanese-v2* and fine-tuned with AdamW (Loshchilov and Hutter, 2019) ( $\text{lr}=2 \times 10^{-5}$ , batch size 16) and early stopping. Across strata, classifiers achieved high reliability (>94% accuracy on average): near-perfect for Social Context (role relations),  $\approx 95\%$  for Speech Function, and  $\approx 97\%$  for Honorific Style. Frequent categories such as *Apology* and *Greeting* were predicted stably, while rare functions (e.g., *Reminder*) remained less reliable due to class imbalance.

Table 4 lists the evaluated models, including both Japanese-tuned and multilingual LLMs. The results highlight strengths such as consistent honorific usage in specialized models, and weaknesses in expressing functions like *Gratitude* or *Apology*.

## 5. Results and Analysis

### 5.1. Stratum-Level Performance Analysis

Figure 3 presents model performance across six annotation strata: **Social Standing, Role Relation-**

Model Name	Developer	Language
[M-LLaMA] Meta-Llama-3.1-8B-Instruct	Meta (2024)	Multilingual
[Swallow] Llama-3.1-Swallow-8B-Instruct-v0.3	Fujii et al. (2024); Okazaki et al. (2024)	Japanese
[ELYZA] Llama-3-ELYZA-JP-8B	Hirakawa et al. (2024)	Japanese
[Youko] llama-3-youko-8b-instruct	Chen et al.; Sawada et al. (2024)	Japanese
[JP-1.8B] llm-jp-3-1.8b-instruct3	National Institute of Informatics (2024a)	Japanese
[JP-3.7B] llm-jp-3-3.7b-instruct3	National Institute of Informatics (2024c)	Japanese
[JP-7.2B] llm-jp-3-7.2b-instruct3	National Institute of Informatics (2024c)	Japanese
[JP-13B] llm-jp-3-13b-instruct3	National Institute of Informatics (2024b)	Japanese
[Qwen-2] Qwen-2-7B-Instruct	Yang et al. (2024)	Multilingual
[Qwen-2.5] Qwen-2.5-7B-Instruct	Team (2024)	Multilingual
[GPT-3.5] chatgpt-gpt-3.5-turbo	Brown et al. (2020)	Multilingual
[GPT-4o-mini] chatgpt-gpt-4o-mini	OpenAI (2024b)	Multilingual
[GPT-4o] chatgpt-gpt-4o	OpenAI (2024c)	Multilingual
[GPT-4.5] chatgpt-gpt-4.5-preview	OpenAI (2024a)	Multilingual
[Fugaku] Fugaku-LLM-13B-instruct	RIKEN and Research System (2023)	Japanese
[Rakuten] RakutenAI-7B-instruct	Rakuten Group, Inc. et al. (2024)	Japanese

Table 4: List of evaluated LLMs and their language capabilities. Model abbreviations in square brackets are used consistently across all figures in this paper.

ship, Object of Exchange, Role in Conversation, Sender’s Action, and Honorific Style. Results reveal substantial variation across models and strata.

**Object of Exchange** and **Role in Conversation** achieved uniformly high scores (mean > 0.94, std < 0.06), suggesting these communicative roles are more easily captured by LLMs. In contrast, socio-pragmatic strata—especially **Social Standing**, **Role Relationship**, and **Honorific Style**—exhibited larger variance, indicating greater modeling difficulty.

Among these, **Role Relationship** was the most unstable (std up to 0.446), with some models failing entirely (score = 0) in certain prompts. **Social Standing** also posed significant challenges; for example, **Meta-Llama-3.1-8B** performed poorly (mean = 0.4809), sometimes responding in English despite being evaluated on Japanese prompts. We also quantified English contamination. English contamination varied across models: Meta-Llama-3.1-8B produced English tokens in 100% of generations (avg. 70% characters), while GPT-4o, Qwen-2, and Swallow showed 0% contamination. Other models ranged between 3–20%.

Interestingly, most models overproduced highly formal expressions, even in informal situations. This leads to inflated scores in upward-facing contexts (e.g., student → professor), but poor performance in peer-level settings. From an SFL perspective, this pattern reflects a breakdown in the realization hierarchy: models capture lexicogrammatical forms of honorifics but fail to condition these choices on contextual variables such as role or social standing. In other words, they have memorized honorific morphology but not the contextual constraints governing register shift. This mirrors the human linguistic phenomenon of overgeneralization, but here it arises because contextual strata

are not explicitly represented in training.

In fact, four of the five lowest-scoring situations involved *student* → *friend* interactions (e.g., apologizing for canceling a study session, mean = 0.660, SD = 0.363), primarily due to excessive use of honorifics. Conversely, upward-facing situations such as *student* → *professor* consistently ranked among the highest (e.g., thanking a professor for thesis guidance, mean = 0.907, SD = 0.199).

**Sender’s Action**, which includes fine-grained functions like *Request: Requesting Help* or *Statement: Greeting*, yielded moderate yet discriminative scores. GPT-based models generally outperformed others (mean > 0.83), suggesting stronger handling of pragmatic intent.

**Honorific Style** showed clear model-dependent gaps. Japanese-specialized models such as **Llama-3-ELYZA-JP-8B** achieved strong performance (mean = 0.8881), while general-purpose models like **Meta-Llama-3.1-8B** underperformed (mean = 0.518). This confirms the effectiveness of language-specific tuning.

These results can also be interpreted through the lens of SFL. Consistent with SFL’s realization hierarchy, we observe that LLMs handle semantic-level communicative functions (Stratum 2: Speech Function) relatively well, but often fail to align higher-level Social Context (Stratum 1) with lower-level realizations in Honorific Style (Stratum 3). For instance, peer-level prompts such as *Student* → *Friend* frequently elicit overly formal styles, indicating a breakdown between strata. This demonstrates how the SFL perspective provides explanatory power for the observed mismatches across strata.

In addition to the annotated strata, we observed stylistic mismatches not captured by the tag system, for example LLaMA-based models frequently generated Western-influenced phrases such as “Dear

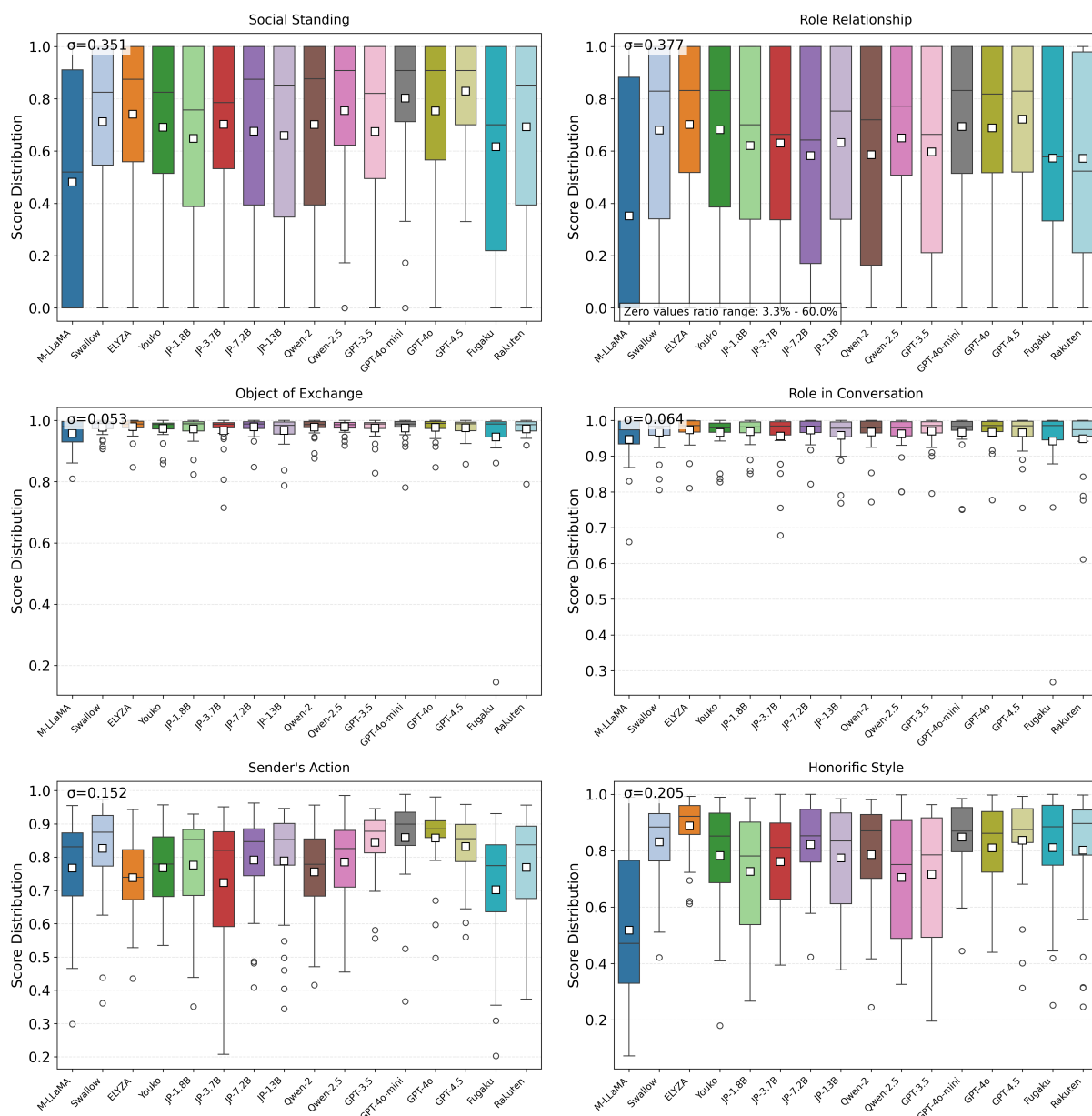


Figure 3: Box plots of LLM performance across six annotation categories.  $\sigma$  denotes the standard deviation of model scores across the 30 social situations, indicating how variable model behavior is within each annotation stratum. Each box shows the score distribution of a model, spanning the middle 50% of values. Median and mean are shown as a black line and white square, respectively. Whiskers and dots represent non-outliers and outliers.

friend” or “Best,” directly translated into Japanese, which reflects pretraining data bias and highlights that while basic communicative roles are tractable, strata involving social inference and cultural norms remain more challenging.

## 5.2. Final Score Rankings

To assess overall social appropriateness, we compute the **Final Score** of each model based on its email generation across 30 situations, covering six role relationships (e.g., *Student* → *Friend*, *Em-*

*ployee* → *Supervisor*) and five situations per relationship. For each situation, the generated email is compared to multiple human-written references, and tag-based similarity is calculated across all annotated dimensions. Final scores reflect alignment with human-like socio-pragmatic behavior and are averaged per model and role category.

Figure 4 shows that Japanese-adapted models—**GPT-4.5-preview** (0.859), **GPT-4o-mini** (0.857), and **ELYZA-JP-8B** (0.839)—achieve the highest scores. **Swallow-8B** (0.835) also performs consistently well, highlighting the value

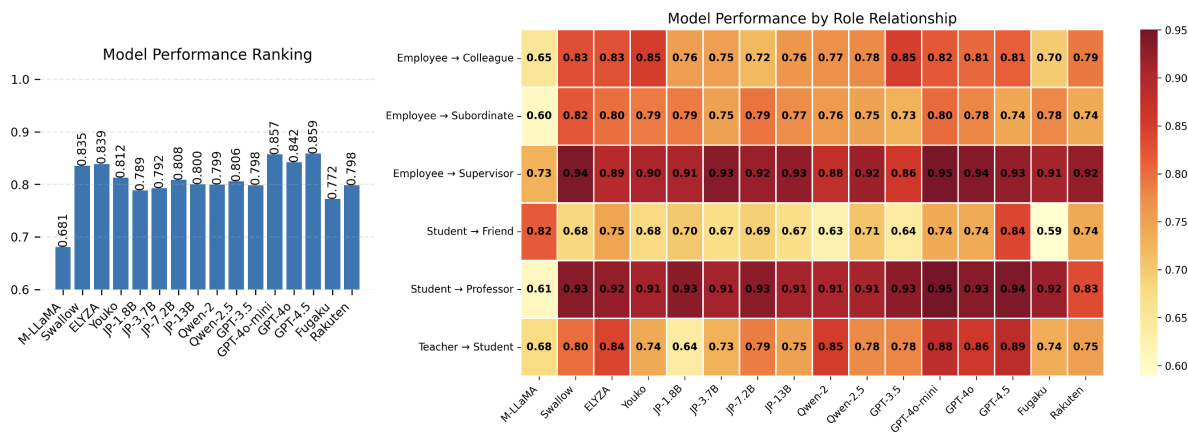


Figure 4: Comparison of model performance across different role relationships. Left: bar plot showing overall average score per model. Right: heatmap showing average performance by role relationship category. Abbreviated model names are used.

of in-language and culturally aligned tuning. In contrast, **Meta-Llama-3.1-8B** obtains the lowest score (0.681), often failing to adapt honorific tone and formality.

To illustrate role-specific performance, Figure 4 (right) presents a heatmap of scores averaged over five situations within each role type. Two key patterns emerge: (1) **Formal situations** (e.g., *Student → Professor*, *Employee → Supervisor*) yield higher scores, suggesting that models handle expected formality more effectively; (2) **Informal or lateral roles** (e.g., *Student → Friend*) pose greater challenges, often leading to excessive use of honorifics and style mismatches. The top-tier models at this time, such as **GPT-4o-mini** and **Swallow-8B** maintain robust performance across all role types, while others—like **JP-1.8B** and **Qwen-2.5**—exhibit greater variability. For example, thanking a professor achieved a high mean score of 0.907, while apologizing to a friend for canceling a study session dropped to 0.660, illustrating the gap between structured hierarchical contexts and casual peer interactions.

Beyond these average trends, we also considered the upper and lower bounds of performance. For the lower bound, we used mismatched role-pairings that contrasted social standing. Representative examples include: (i) Peer vs. Upward (*Student → Friend* vs. *Student → Professor*), (ii) Downward vs. Upward (*Employee → Subordinate* vs. *Employee → Supervisor*), and (iii) Higher vs. Lower (*Teacher → Student* vs. *Student → Professor*). All six pairings produced substantially lower scores than true matches, providing a robust lower-bound baseline.

### 5.3. Representative Situation Rankings

To further highlight where models succeed and fail, we examined the best- and worst-performing com-

municative situations. Formal hierarchical contexts, such as *Student → Professor* (thanking a professor for thesis guidance, mean  $\approx 0.91$ ) and *Employee → Supervisor* (informing a supervisor that requested documents are ready,  $\approx 0.92$ ), achieved the highest scores. By contrast, informal peer-level contexts were consistently more difficult, with situations like *Student → Friend* (canceling a study meeting,  $\approx 0.66$ ; asking a friend for homework details,  $\approx 0.70$ ) yielding the lowest scores. This confirms that LLMs align well with structured norms but struggle with register in casual or ambiguous contexts. Peer-level situations remain most difficult, often leading to over-formality. Viewed from the SFL framework, this pattern reflects a partial failure of strata alignment: while models capture Social Relationship and Speech Function in formal contexts, they frequently misrealize the Honorific Style in informal ones.

## 6. Conclusion

We present the **JaSocial Evaluation Framework**, a stratified benchmark for assessing the social intelligence of LLMs in Japanese honorific email generation. Built on Systemic Functional Linguistics, it evaluates outputs across three annotation strata: *Social Context*, *Speech Function*, and *Honorific Style*. Evaluating 16 LLMs on 30 role-based situations, we find that while models handle basic communicative roles well, they often overuse honorifics in casual contexts and struggle with strata requiring social inference. From an SFL perspective, this reflects a disruption in the realization hierarchy: lexicogrammatical forms are reproduced, but the conditioning by social context is weak. The framework thus highlights both the limits of current LLMs and the importance of explicitly modeling context-form relations in future work.

## Acknowledgements

This paper is based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain".

## Ethical Considerations and Limitations

Our dataset consists of constructed business email scenarios and does not involve any personal or sensitive information. Nevertheless, the task of modeling social relationships and honorific usage may reflect cultural norms and implicit biases; these aspects should be interpreted with caution. All annotations were carried out through careful linguistic analysis and team-internal review to ensure consistency. Limitations of this study include the focus on Japanese business email contexts and a relatively small dataset size, which may restrict generalizability across languages and domains. Future work should expand the dataset and explore cross-cultural settings to assess broader applicability.

## Bibliographical References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Xinqi Chen, Koh Mitsuda, Toshiaki Wakatsuki, and Kei Sawada. [rinna/llama-3-youko-8b-instruct](#).
- Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. [Controlling Japanese honorifics in English-to-Japanese neural machine translation](#). In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Daniel Goleman. 1995. *Emotional Intelligence: Why It Can Matter More Than IQ*. Bantam Books.
- Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. [Knowledgenavigator: leveraging large language models for enhanced reasoning over knowledge graph](#). *Complex Intelligent Systems*, 10.
- M. A. K. Halliday and C. M. I. M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*, 4th edition. Routledge, London and New York.
- M.A.K. Halliday. 1985. *An Introduction to Functional Grammar*, 1st edition. Edward Arnold, London.
- Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. 2024. [elyza/llama-3-elyza-jp-8b](#).
- LLM-jp, :, 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, Akim Moustereou, 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](#).
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations (ICLR)*.
- Meta. 2024. [The llama 3 herd of models](#). We used the LLaMA 3.1-8B model available at <https://huggingface.co/meta-llama/llama-3.1-8B> (Accessed: 2025-03-20).

- National Institute of Informatics. 2024a. [Llm-jp-3-1.8b-instruct3](#). Accessed: 2025-03-27.
- National Institute of Informatics. 2024b. [Llm-jp-3-3.7b-instruct3](#). Accessed: 2025-03-27.
- National Institute of Informatics. 2024c. [Llm-jp-3-7.2b-instruct3](#). Accessed: 2025-03-27.
- Jun Ohashi. 2018. [An emerging role-identity and honorifics: A longitudinal study of email exchanges in a Japanese community](#). *Journal of Pragmatics*, 127:36–55.
- Shigeko Okamoto and Janet Shibamoto-Smith. 2016. [Keigo: From official policy to popular pedagogy](#). In *The Social Life of the Japanese Language: Cultural Discourse and Situated Practice*, pages 125–153. Cambridge University Press.
- 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.
- OpenAI. 2024a. [Gpt-4 technical report](#).
- OpenAI. 2024b. [Gpt-4o mini](#). Accessed: 2025-03-27.
- OpenAI. 2024c. [Gpt-4o system card](#).
- Ely Triasih Rahayu. 2018. [Japanese honorific language in various domains](#). In *Proceedings of the Fourth Prasasti International Seminar on Linguistics (Prasasti 2018)*, pages 25–34. Atlantis Press.
- Rakuten Group, Inc., Aaron Levine, Connie Huang, Chenguang Wang, Eduardo Batista, Ewa Szymanska, Hongyi Ding, Hou Wei Chou, Jean-François Pessiot, Johannes Effendi, Justin Chiu, Kai Torben Ohlhus, Karan Chopra, Keiji Shinzato, Koji Murakami, Lee Xiong, Lei Chen, Maki Kubota, Maksim Tkachenko, Miroku Lee, Naoki Takahashi, Prathyusha Jwalapuram, Ryutaro Tatsushima, Saurabh Jain, Sunil Kumar Yadav, Ting Cai, Wei-Te Chen, Yandi Xia, Yuki Nakayama, and Yutaka Higashiyama. 2024. [Rakutenai-7b: Extending large language models for Japanese](#).
- Tokyo Institute of Technology Tohoku University CyberAgent Inc. Tokai National Higher Education RIKEN, Fujitsu and Ltd. Research System, Kotoba Technologies Japan Co. 2023. [Fugaku-llm: A large-scale Japanese language model developed using the fugaku supercomputer](#). Accessed: 2025-03-27.
- 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. <https://arxiv.org/abs/2404.01657>.
- Ryo Sekizawa and Hitomi Yanaka. 2023. [Analyzing syntactic generalization capacity of pre-trained language models on Japanese honorific conversion](#). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (\*SEM 2023)*, pages 40–47, Toronto, Canada. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Edward L Thorndike. 1920. [Intelligence and its uses](#). *Harper's Magazine*, 140:227–235.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671*.

## Language Resource References

- Namgi Han and Nobuhiro Ueda and Masatoshi Otake and Satoru Katsumata and Keisuke Kamata and Hirokazu Kiyomaru and Takashi Kodama and Saku Sugawara and Bowen Chen and Hiroshi Matuda and Yusuke Miyao and Yugo Miyawake and Koki Ryu. 2024. [llm-jp-eval: Automatic evaluation tool for Japanese large language models](#). llm-jp Evaluation Project. Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing (NLP2024), Japan. In Japanese.

Muxuan Liu, Tatsuya Ishigaki, Yusuke Miyao, Hiroya Takamura, and Ichiro Kobayashi. 2023. [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, Hong Kong, China. Association for Computational Linguistics.

Muxuan Liu and Ichiro Kobayashi. 2022. [Construction and validation of a Japanese honorific corpus based on systemic functional linguistics](#). In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

## Appendix

### A. Training Details and Optimization Formulas

#### A.1. Overall Training Configuration

Our pipeline employs a three-stage cascading supervised training procedure, with the following consistent hyperparameters across all strata:

- Learning rate =  $2 \times 10^{-5}$
- Batch size = 16
- Epochs = 5 per stage

All three strata use the AdamW optimizer without additional learning rate schedulers. Each stage independently saves the best model based on validation accuracy.

#### A.2. Stratum 1: Social Context Classification

Each email is preprocessed by concatenating its components (social context, subject, recipient, and body) with special markers (e.g.,  $[CLS]$  and  $[SEP]$ ). The resulting text is fed into the pre-trained Japanese BERT model (*cl-tohoku/bert-base-japanese-v2*), a 12-layer Transformer encoder with hidden dimension 768, 12 attention heads, and a vocabulary size of 32,000.

The final representation from the  $[CLS]$  token is used for predicting two labels: overall social relationship and specific role relationship. The combined loss is defined as:

$$L = \alpha \cdot L_{social} + (1 - \alpha) \cdot L_{role}, \quad (2)$$

with hyperparameter  $\alpha = 0.5$ .

#### A.3. Stratum 2: Speech Function Classification

The representations for email context ( $E_{mail}$ ), target sentence ( $E_{sent}$ ), and the predicted social and role labels from Stratum 1 ( $P_{social}, P_{role}$ ) are concatenated:

$$F = [E_{mail}; E_{sent}; P_{social}; P_{role}], \quad (3)$$

producing a high-dimensional feature vector. A non-linear transformation is then applied:

$$H = \tanh(W_f \cdot F + b_f), \quad (4)$$

followed by multi-label classification layers for functional roles, including Object of Exchange, role in conversation, and sender's action.

#### A.4. Stratum 3: Honorific Style Classification

Stratum 3 similarly uses a dual-encoder architecture. The email and sentence-level representations are combined with predictions from Stratum 2 (item, Role in Conversation, sender's action):

$$F = [E_{mail}; E_{sent}; P_{obj}; P_{role}; P_{action}], \quad (5)$$

and passed through a non-linear transformation:

$$H = \tanh(W_h \cdot F + b_h), \quad (6)$$

before classifying the honorific style of each sentence.

#### A.5. Inference Procedure

During inference, each input email is tokenized and processed through the three annotation strata in sequence. First, the email-level classifiers (Stratum 1) predict the **Social Standing** and **Role Relationship**. Then, sentence-level classifiers (Stratum 2) predict **Object of Exchange**, **Role in Conversation**, and **Sender's Action**, conditioned on Stratum 1 outputs. Finally, **honorific styles** (Stratum 3) are predicted using both sentence-level features and Stratum 2 predictions. The outputs are structured in JSON format for downstream evaluation and analysis.

All configurations and model architectures were selected based on extensive preliminary experiments to ensure stable training and optimal performance.

## B. Annotation Mappings Used in Training

This section provides the annotation mappings used to train all components of the JaSocial Evaluation Framework. Tables 5, 6, and 7 present the correspondence between original Japanese annotation terms and their standardized English labels across the three annotation strata.

The following tables define the internal ID mappings used for classifier training across the annotation strata. Each category—*Object of Exchange*, *Role in Conversation*, *Sender's Action*, and *honorific style*—is assigned a unique integer ID to enable supervised classification. These mappings are consistent with the English annotation terms defined earlier and are used throughout the training and evaluation pipeline.

Category	Japanese	English
Social Standing	目上 目下 同輩	Higher Lower Peer
Role Relationship	学生 → 友人 学生 → 教授 教員 → 学生 従業員 → 同僚 従業員 → 上司 従業員 → 部下	Student → Friend Student → Professor Teacher → Student Employee → Colleague Employee → Supervisor Employee → Subordinate

Table 5: Annotation mappings for Stratum 1: Social Context

Japanese Label	English Label
情報 / ものとサービス 与える / 求める	Information / Goods and Services Give / Receive
提供:提案 要求:依頼 要求:催促 質問:依頼 質問:問い合わせ 陳述:お知らせ 陳述:感謝 陳述:挨拶 陳述:提案 陳述:断り 陳述:説明 陳述:謝罪	Offer: Proposal Request: Requesting Help Request: Reminder Question: Requesting Help Question: Inquiry Statement: Notification Statement: Gratitude Statement: Greeting Statement: Suggestion Statement: Refusal Statement: Explanation Statement: Apology

Table 6: Annotation mappings for Stratum 2: Speech Function

Role in Conversation	ID
Give	0
Receive	1

Table 9: Mapping for Role in Conversation (Stratum 2).

## C. Evaluation Implementation Details

Our evaluation framework quantifies the social appropriateness of LLM outputs using a vector-based similarity approach. Evaluation is performed across six annotation categories, aligned with the three-stratum structure of our annotation scheme.

### C.1. Probabilistic Annotation Vectors

Each model-generated email is parsed to extract annotations in the following six categories:

- **Stratum 1:** *Social Standing, Role Relationship*
- **Stratum 2:** *Object of Exchange, Role in Conversation, Sender's Action*
- **Stratum 3:** *Honorific Style*

For each category, the classifier outputs a probability distribution over all possible labels, represented as a softmax-based vector  $\mathbf{v}_x^{\text{gen}} \in R^d$ . Human-written emails are annotated with discrete tags and represented as one-hot vectors  $\mathbf{v}_x^{\text{ref}_i}$ .

### C.2. Final Score Computation

To compute similarity for a given category  $x$ , we take the average cosine similarity between the model vector and all  $N$  references written under the same situation:

$$\text{Sim}_x = \frac{1}{N} \sum_{i=1}^N \cos(\mathbf{v}_x^{\text{gen}}, \mathbf{v}_x^{\text{ref}_i}) \quad (7)$$

For the comprehensive evaluation, we concatenate the vectors from all six categories:

$$\mathbf{v}_{\text{all}}^{\text{gen}} = [\mathbf{v}_{SS}, \mathbf{v}_{RR}, \mathbf{v}_{OE}, \mathbf{v}_{RC}, \mathbf{v}_{SA}, \mathbf{v}_{HS}] \quad (8)$$

Similarly, for each reference  $i$ , we compute  $\mathbf{v}_{\text{all}}^{\text{ref}_i}$ . The final score is defined as the average cosine similarity between the model's comprehensive vector and each reference's comprehensive vector:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^N \cos(\mathbf{v}_{\text{all}}^{\text{gen}}, \mathbf{v}_{\text{all}}^{\text{ref}_i}) \quad (9)$$

Japanese Honorific Style	English Label
インフォーマル	Informal
丁寧語	Polite
尊敬語 + 丁寧語	Honorific + Polite
謙讓語 + 丁寧語	Humble + Polite
尊敬語 + 謙讓語 + 丁寧語	Honorific + Humble + Polite
謙讓語 + インフォーマル	Humble + Informal

Table 7: Annotation mappings for Stratum 3: honorific style

Object of Exchange	ID
Goods and Services	0
Information	1

Table 8: Mapping for Object of Exchange (Stratum 2).

Honorific Style	ID
Informal	0
Polite	1
Honorific + Polite	2
Honorific + Humble + Polite	3
Humble + Informal	4
Humble + Polite	5

Sender’s Action (English Label)	ID
Offer: Proposal	0
Request: Requesting Help	1
Request: Reminder	2
Question: Requesting Help	3
Question: Inquiry	4
Statement: Notification	5
Statement: Gratitude	6
Statement: Greeting	7
Statement: Suggestion	8
Statement: Refusal	9
Statement: Explanation	10
Statement: Apology	11

Table 10: Mapping for Sender’s Action (Stratum 2), based on Japanese annotation schema.

### C.3. Score Aggregation and Ranking

We evaluate each model across all social situations and report:

- **Model-level scores:** average similarity across all situations.
- **Role-wise scores:** average performance for each Role Relationship (e.g., Student → Professor).
- **Stratum-wise scores:** average performance per annotation category (e.g., highest on *Sender’s Action*).
- **Situation difficulty:** average model scores per situation, used to rank evaluation difficulty.

## D. Detailed Ranking of Social Situations

### D.1. Top and Bottom Performing Situations

To better understand model performance across different communicative contexts, we ranked all

Table 11: Mapping for Honorific Style annotations (Stratum 3).

social situations by their average similarity scores. Tables 12 and 13 show the top five and bottom five situations based on mean performance across all models.

Situations involving formal hierarchical relationships (e.g., *Student* → *Professor*, *Employee* → *Supervisor*) tend to achieve higher scores, suggesting that LLMs are more consistent when handling structured and expectation-driven interactions. In contrast, informal peer situations (e.g., *Student* → *Friend*) show lower and more variable performance, indicating difficulty in generating pragmatically appropriate responses in casual or ambiguous contexts.

### D.2. Per-Situation Final Score Trends

To investigate model behavior in finer detail, we analyzed the final score distributions for each individual social situation across all 30 scenarios spanning six role relationships. Figures 5 through 10 show heatmaps of model scores per situation, revealing nuanced patterns within each relationship type.

Overall, **upward-facing situations** (e.g., *Student* → *Professor*, *Employee* → *Supervisor*) exhibited strong and stable performance across models. Most models achieved scores above 0.90 in these formal contexts (e.g., Situation 7 in *Student* → *Professor*, where 13 out of 16 models scored higher than 0.93). These cases are typically associated with consistent honorific usage and clearer socio-pragmatic constraints, which are easier for LLMs to model reliably.

In contrast, **peer-level or downward-facing situations** (e.g., *Student* → *Friend*, *Employee* → *Subordinate*) showed much wider variation and lower scores. For example, in Situation 7 under *Student*

Role Relationship	Social Situation (JP)	Social Situation (EN)	Mean	SD
Employee → Supervisor	あなたは11月7日付で川崎支店へ移動することになりました。川崎支店のA課長へ挨拶メールを書いてください。	You have been assigned to transfer to the Kawasaki Branch as of November 7. Please write a greeting email to Section Chief A at the Kawasaki Branch.	0.937	0.096
Student → Professor	あなたはダンス系サークルの代表者です。同じ大学の舞踊コースの担当A教授に指導意見をもらいたいので、その前に挨拶のメールを書いてください。	You are the representative of a dance club. You would like to receive some guidance and feedback from Professor A, who is in charge of the dance course at the same university. Please write a greeting email to the professor before making your request.	0.936	0.084
Student → Professor	あなたは研究実験を行っており、研究室を使用する必要がありますが、教授の承認を待っています。できるだけ早く実験室使用要求を承認し、時間通りに研究を完了できるように教授を促すメールを書いてください。	You are conducting a research experiment and need to use the laboratory, but are currently waiting for the professor's approval. Please write an email to politely urge the professor to approve your request as soon as possible, so that you can complete your research on time.	0.925	0.153
Employee → Supervisor	あなたは上司であるAさんから求められていた書類を完成させました。Aさんに確認してもらう必要があります。この件に関して、Aさんに知らせるメールを書いてください。	You have completed the documents that your supervisor, Mr./Ms. A, had requested. You need Mr./Ms. A to review them. Please write an email to inform Mr./Ms. A about this matter.	0.916	0.158
Student → Professor	あなたは論文を書くために教授に相談を持ちかけ、教授は多くの有益なアドバイスを与えてくれました。最終的に、あなたは論文を無事に完成させ、高い評価を得ました。教授に感謝のメールを書いてください。	You asked your professor for advice while writing your thesis, and the professor gave you a lot of helpful guidance. As a result, you successfully completed your thesis and received a high evaluation. Please write an email to express your gratitude to the professor.	0.907	0.199

Table 12: Top 5 Highest Scoring Social Situations. The table shows the average similarity score (**Mean**) and standard deviation (**SD**) across all evaluated LLMs. Each score is computed based on the model-generated email in response to a given social situation prompt.

→ *Friend*, scores dropped to 0.404 (GPT-3.5) and 0.496 (JP-7.2B), highlighting the difficulty of managing informal tone and socially appropriate requests. Such situations demand more nuanced control of register and cultural appropriateness, which many general-purpose models fail to handle, often defaulting to overly formal or awkwardly casual language.

Even within the same role relationship, some situations were consistently more difficult than others. Under *Teacher* → *Student*, Situations 8 and 9 produced markedly lower scores for several models (e.g., Swallow, ELYZA, Youko), suggesting potential confusion between directive and informative styles, or challenges in expressing gratitude in a downward hierarchy.

These per-situation analyses reinforce our earlier findings: while modern LLMs are increasingly

capable of handling structured, formal interactions, they continue to struggle with informality, interpersonal nuance, and socio-cultural variation. The six situation-level heatmaps are shown in the following figures:

## E. Examples of LLM-Generated Emails

The following examples of LLM-generated emails are taken from the highest and lowest scoring social situations identified in Section D.1. These samples demonstrate how model behavior varies across different social contexts, particularly between formal hierarchical exchanges and informal peer interactions. Each example contains the full output generated by the LLM, which may include prompts or

Role Relationship	Social Situation (JP)	Social Situation (EN)	Mean	SD
Student Friend →	明日は文化祭の出し物を決める議決会があるのですが、友達のAさんに、明日みんなが意見を出してくれないようであれば、なんでもいいので何か一つでも提案してくれるよう促すメールを書いてください。	Tomorrow, your class will discuss what to present at the school festival. If no one offers ideas, write an email encouraging your friend A to propose at least one idea.	0.621	0.257
Student Friend →	あなたは友人Aと勉強会をする予定でしたが、部活が長引いて疲れてしまいました。約束を中止にして欲しいというお詫びのメールを書いてください。	You had planned a study session with your friend A, but you got too tired after club activities. Write an email to apologize and cancel the meeting.	0.660	0.363
Student Friend →	あなたは今日学校で出題された宿題の内容を忘れてしまいました。その上、その宿題の提出は明日となっています。そこで、どのようなメールを送れば、友人から早く宿題の内容を返信してくれるか考えなさい。	You forgot the homework assigned at school today, and it is due tomorrow. Write an email that will prompt your friend to quickly reply with the assignment details.	0.701	0.287
Student Friend →	学生のあなたは友人Aさんから預かっていたレポートを今日の講義で提出するのを忘れてしまいました。このことを謝罪するメールを書いてください。	You forgot to submit the report you were keeping for your friend A during today's lecture. Write an email to apologize for this.	0.713	0.289
Employee Subordinate →	あなたは出張へ出ました。出発前に部下へ出張中に利用するデータをまとめて自身へ送るように指示しました。しかし明日データが必要になると日になってもまだデータは送られていません。部下にデータについて問い合わせてください。	You are on a business trip. Before leaving, you asked your subordinate to compile and send you data needed during the trip. The deadline is tomorrow, but the data has not arrived. Write an email inquiring about it.	0.716	0.271

Table 13: Bottom 5 Lowest Scoring Social Situations. The table shows the average similarity score (**Mean**) and standard deviation (**SD**) across all evaluated LLMs. Each score is computed based on the model-generated email in response to a given social situation prompt.

extra formatting. For evaluation, only the actual email body was extracted. When a model output included multiple email candidates, we selected the first one for analysis.

**Note:** All examples are presented in their original Japanese form. Given the nuance of honorifics and social expressions in Japanese, we refrain from translation to preserve fidelity and ensure consistency with the evaluation setting.

**Social Situation:** あなたは11月7日付で川崎支店へ移動することになりました。川崎支店のA課長へ挨拶メールを書いてください。

#### chatgpt-gpt-3.5-turbo-situation-1:

件名: 川崎支店への異動のご挨拶 宛先: 川崎支店 A課長様

A課長殿

お世話になっております。私、〇〇（お名前）は11月7日付で川崎支店へ異動することになりました。A課長を初めとする川崎支店の皆

様には、これからもよろしくご挨拶申し上げます。

業務を通じて貴重なご指導やアドバイスをいただき、大変感謝しております。川崎支店でも精一杯努める所存でございますので、どうぞよろしくお願い申し上げます。

何かご不明点やご要望等ございましたら、お気軽にお知らせいただければ幸いです。

引き続き、ご指導とご鞭撻を賜りますよう、よろしくお願い申し上げます。

敬具

〇〇（お名前）

#### chatgpt-gpt-4.5-preview-situation<sub>1</sub>

: 件名: 異動のご挨拶 (11月7日付)

A課長

お世話になっております。このたび、11月7日付で川崎支店に異動することになりました〇〇と申します。

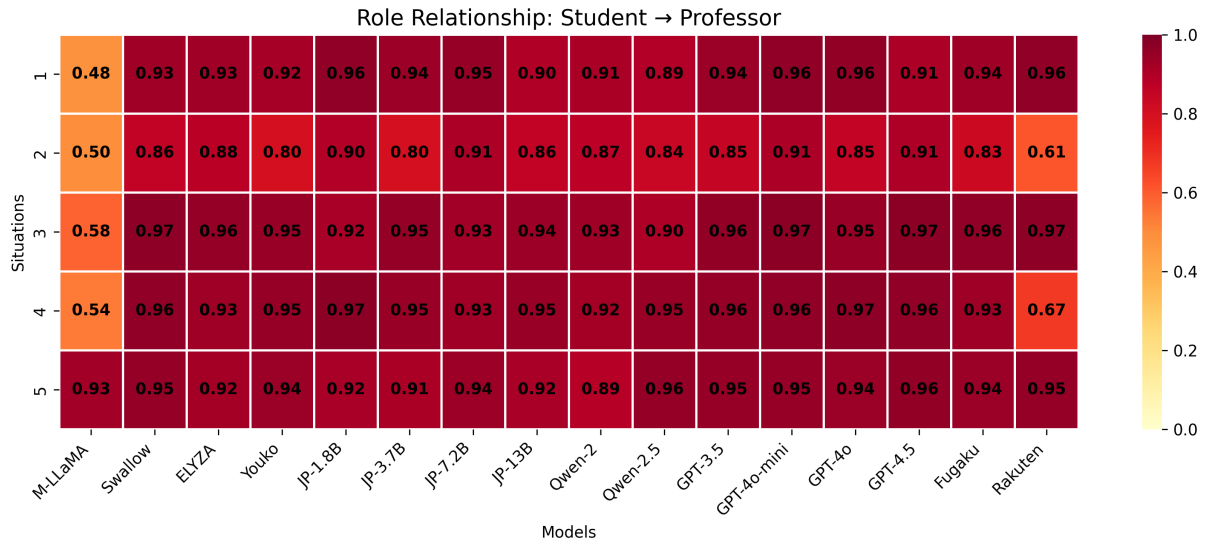


Figure 5: Per-situation model performance for the **Student → Professor** role relationship.

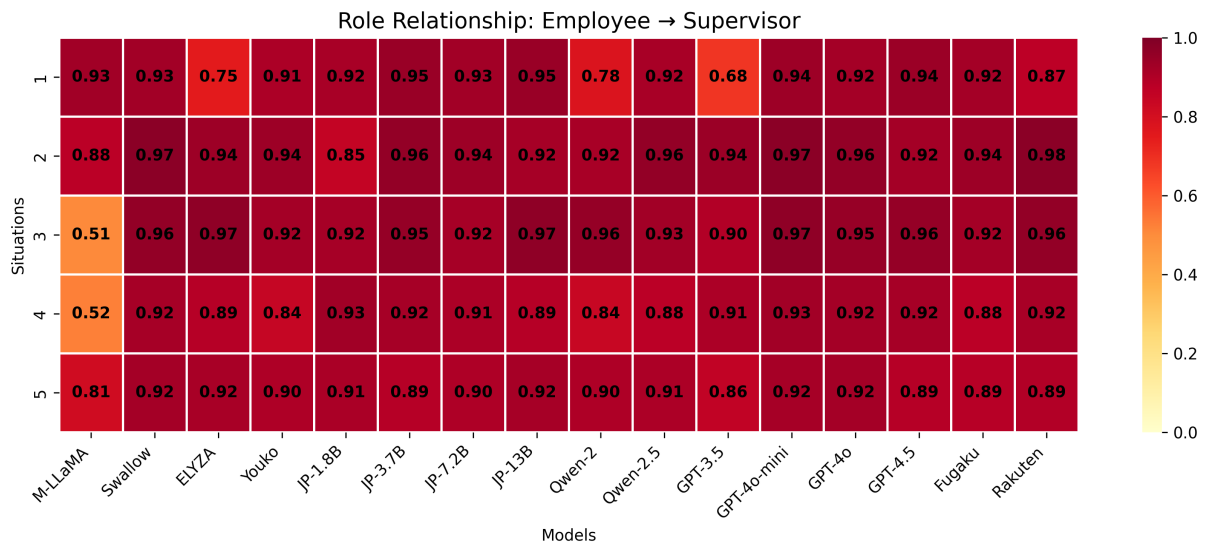


Figure 8: Per-situation model performance for the **Employee → Supervisor** role relationship.

川崎支店への配属は初めてであり、不慣れなことも多くご迷惑をおかけすることもあるかと思いますが、一日でも早く業務を理解し、支店のお役に立てるよう取り組んでまいります。

今後とも何卒ご指導、ご鞭撻のほどよろしくお願い申し上げます。

改めて配属後にご挨拶に伺わせていただきます。まずはメールにてご挨拶申し上げます。

〇〇（あなたの名前）

#### chatgpt-gpt-4o-situation-1:

件名: 異動のご挨拶

A課長様

お世話になっております。〇〇部の〇〇と申します。

この度、11月7日付で川崎支店へ異動すること

となりました。これまで培ってきた経験を活かし、新しい環境でも貢献できるよう努めてまいります。

川崎支店での業務は初めてでございますので、至らぬ点もあるかと思いますが、何卒ご指導ご鞭撻のほどよろしくお願い申し上げます。

お会いできる日を楽しみにしております。どうぞよろしくお願いいたします。

敬具

〇〇部 〇〇（あなたのフルネーム）（あなたの連絡先）

#### chatgpt-gpt-4o-mini-situation-1:

件名: 川崎支店への異動のご挨拶

A課長様

お世話になっております。〇〇（あなたの名

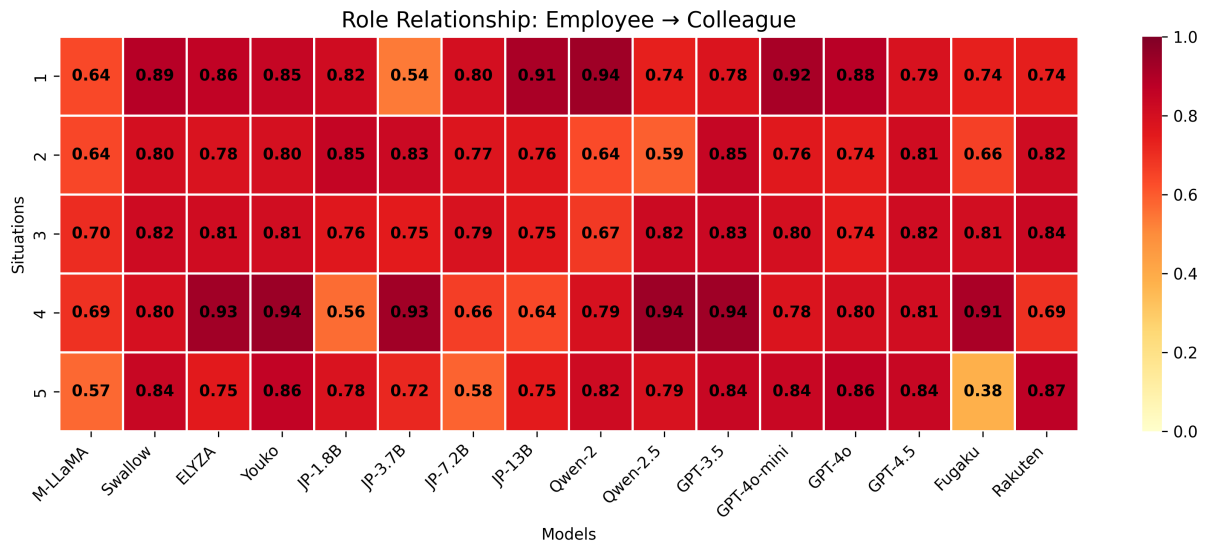


Figure 9: Per-situation model performance for the **Employee → Colleague** role relationship.

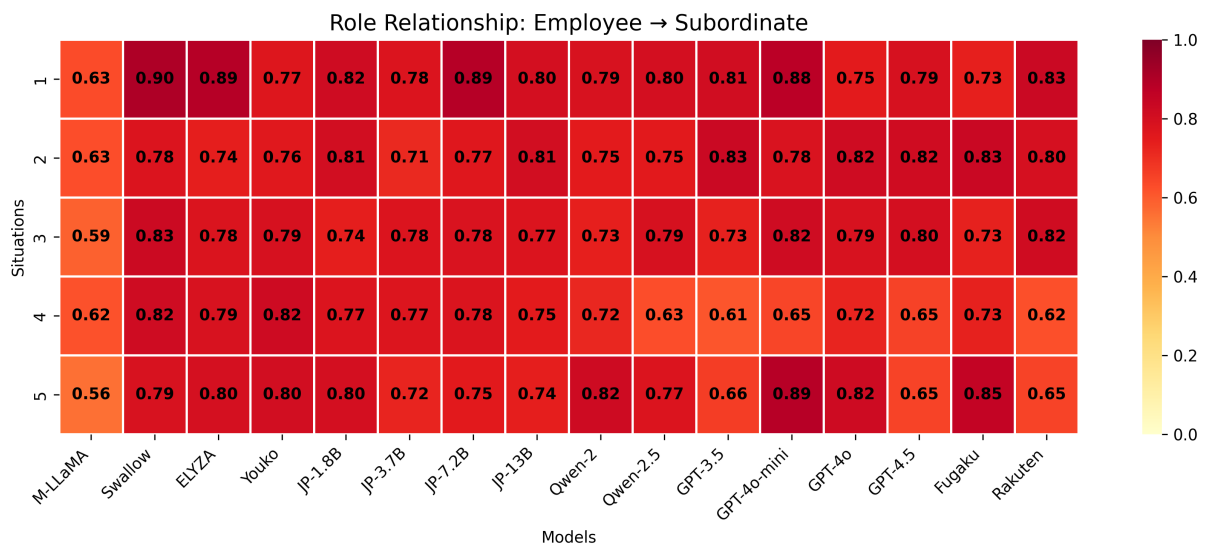


Figure 10: Per-situation model performance for the **Employee → Subordinate** role relationship.

前)です。

このたび、11月7日付で川崎支店へ異動することとなりました。新しい環境での業務に不安もありますが、A課長様をはじめ、川崎支店の皆様と共に成長できることを楽しみにしております。

至らぬ点多いかと思いますが、何卒ご指導のほどよろしくお願い申し上げます。

短いご挨拶となりますが、今後ともどうぞよろしくお願い致します。

敬具

〇〇 (あなたの名前) 〇〇 (あなたの所属部署) 〇〇 (あなたの連絡先)

#### Fugaku-LLM-13B-instruct-situation-1:

以下は、メールの内容を説明する指示です。要

求を適切に満たすメールを書きなさい。件名、宛先 (Aさん、A様など)、本文をそれぞれ書きなさい。

指示: あなたは11月7日付で川崎支店へ移動することになりました。川崎支店のA課長へ挨拶メールを書いてください。

応答: 件名川崎支店への着任挨拶

親愛なるA課長

この度、11月7日より川崎支店に着任いたしました、【お名前】と申します。どうぞよろしくお願い致します。

貴社のチームに加わり、皆様と一緒に働けることを楽しみにしております。川崎支店には素晴らしい人材が揃っていると聞いており、ご迷惑をおかけすることもあるかと思いますが、全力で取り組む所存です。

私の専門知識を活かし、皆様のお役に立てる

よう努力いたします。何か特別な支援や指導が必要な場合は、遠慮なくお尋ねください。

まずは着任のご挨拶まで。

よろしくお願いいたします、[お名前]

### Llama-3-ELYZA-JP-8B-situation-1:

以下は、川崎支店のA課長へ挨拶メールの例です。

件名: 11月7日付で川崎支店へ異動のご挨拶  
A課長様

お世話になっております。

この度、11月7日付で川崎支店へ移動することになりました。突然の移動で、多大なるご迷惑をおかけすることと存じますが、A課長をはじめ支店の皆様にご指導ご鞭撻を賜りながら、業務に邁進したいと考えております。

前任の支店で学ばせていただいたことを活かしつつ、川崎支店の業務に貢献すべく、全力で取り組む所存です。A課長には、ご指導ご鞭撻の程、よろしくお願いいたします。

川崎支店でお会いできる日を楽しみにしております。

この度は、誠にありがとうございました。

敬具

### Llama-3-youko-8b-instruct-situation-1:

件名: ご挨拶と今後の勤務について

A課長

この度、11月7日付けで川崎支店へ異動となりました、[お名前]です。

以前は[現在の部署名]にて勤務しておりましたが、このたび新天地での出発となりました。至らない点も多々あるかと思いますが、精一杯業務に励みますので、何卒よろしくお願いいたします。

また、詳細につきましては、改めてご挨拶に伺う予定であります。その際には、具体的な仕事内容やスケジュールなどをご説明させていただければと存じております。

何卒、ご指導・ご鞭撻を賜りますようお願い申し上げます。

敬具 [お名前]

以上のような文面が、川崎支店のA課長への挨拶メールとして適切かと思えます。

### Llama-3.1-Swallow-8B-Instruct-v0.3-situation-1:

件名: 【挨拶】 11月7日より川崎支店勤務となります

A課長

はじめまして、[あなたの名前]です。

11月7日より、川崎支店に異動となりました。

これまでの経験を活かし、川崎支店の業務に貢献できるよう精一杯努めてまいります。

どうぞよろしくお願いいたします。

### llm-jp-3-1.8b-instruct3-situation-1:

以下は、タスクを説明する指示です。要求を適切に満たす応答を書きなさい。

指示: あなたは11月7日付で川崎支店へ移動することになりました。川崎支店のA課長へ挨拶メールを書いてください。

応答: 件名: 川崎支店への異動に伴うご挨拶

A課長

拝啓 時下ますますご清栄のこととお喜び申し上げます。平素は格別のご高配を賜り、厚く御礼申し上げます。

突然のご連絡となり恐縮ではございますが、私事でこの度、11月7日付にて川崎支店へ異動させていただくこととなりました。これまでの業務において、多くの方々にご指導ご鞭撻をいただき、心より感謝申し上げます。また、皆様からのご支援ご協力のおかげで、無事に職務を遂行することができましたことを、深く感謝しております。

新しい環境での業務は、未知の部分も多いですが、これまでの経験を活かし、微力ながらも新たな挑戦を通じて成長していきたいと考えております。今後とも変わらぬご指導ご鞭撻を賜りますよう、お願い申し上げます。

まずは、略儀ながら書中をもってご挨拶申し上げます。

敬具

[あなたの名前] [あなたの部署名] [連絡先情報]

### llm-jp-3-3.7b-instruct3-situation-1:

以下は、タスクを説明する指示です。要求を適切に満たす応答を書きなさい。

指示: あなたは11月7日付で川崎支店へ移動することになりました。川崎支店のA課長へ挨拶メールを書いてください。

応答: 以下のような挨拶メールを書くことができます。

件名: 異動のご挨拶(川崎支店 A課長様)

本文: ○○部○○課○○係の○○と申します。いつもお世話になっております。この度、私事ではございますが、11月7日付で、○○支店から川崎支店へ異動となりましたので、ご連絡申し上げます。○年間、大変お世話になりました。川崎支店では、これまで以上に頑張りますので、今後ともよろしくお願いいたします。お忙しいところ恐れ入りますが、ご都合がよろしければ、ぜひ川崎支店にお越しいただき、ご挨拶させていただければ幸いです。何かございましたら、どうぞ遠慮なくご連絡くださいませ。何卒よろしくお願いいたします。

○年○月○日 ○○部○○課○○係○○○○

### llm-jp-3-7.2b-instruct3-situation-1:

以下は、タスクを説明する指示です。要求を適切に満たす応答を書きなさい。

