Can LLMs Simulate L2-English Dialogue? An Information-Theoretic Analysis of L1-Dependent Biases

Anonymous ACL submission

Abstract

This study analyzes the ability of Large Language Model (LLM) to simulate non-native second language (L2) English speakers interfered by their prior knowledge of native first language (L1). Specifically, we analyze L1-dependent language interference with L2 (English) dialogues simulated by LLMs with prompting. Our proposed L1-interference evaluation framework focuses on diverse linguistic features such as reference word usage and (in)frequently adopted syntactic constructions biased by L1 (due to, e.g., avoidance behaviors), which are identified through distributional density comparisons using information-theoretic metrics. Our results demonstrate that LLMs can generally emulate L1-dependent linguistic biases reflected in L2 dialogues. Specifically, the impact of native languages varies, for example, with L1s such as Japanese, Korean, and Mandarin significantly affecting tense agreement, Urdu influencing noun-verb collocations, and Thai shaping the use of numerals and modifier, and they agree with real human L2 data. These insights unveil LLMs' potential use for generating diverse L2 dialogues as well as offer a theoretical framework for LLM L2-dialogue evaluation.

1 Introduction

The widespread use of Large Language Models (LLMs) in language communication has opened opportunities to study their ability to simulate human-like language, particularly in second language (L2) communication (Liang et al., 2024; Cherednichenko et al., 2024), as illustrated by the dialogue complexity in Figure 1. Such an L2-speaker simulation will be helpful for, for example, predicting L2 speakers' biases in the pedagogical situation (Settles et al., 2018), emulating diverse agents to simulate the indeed diverse group of people in the world (Ge et al., 2024), and potentially assessing their cognitive plausibility from a language transfer perspective (Aoyama and Schneider, 2024a). However, the ability of LLMs to accurately replicate linguistic patterns of non-native speakers and the systematic influence of L1 knowledge on L2 generation remain underexplored (Chen et al., 2024), especially given the complexities of spoken dialogues in non-native contexts (Fincham and Alvarez, 2024). **This leads us to ask: Can LLMs effectively mimic human-like dialogue performance in L2 contexts?**

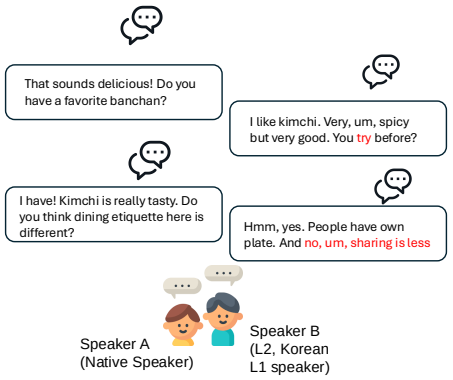


Figure 1: L2 English conversation dialogue from human speakers.

To address this question, it is crucial to understand the role of prior linguistic knowledge in areas such as language education and cross-lingual communication (Brooke and Hirst, 2012). L2 speakers' use of English is often influenced by their L1 traits (Takahashi, 2024), resulting in distinct linguistic patterns, including grammatical constructions and lexical choices in spoken dialogues (Aoyama and Schneider, 2024b; Downey et al., 2023). To investigate whether LLMs simulate similar patterns, we propose an evaluation framework grounded in multiple linguistic perspectives: key features from grammatical/semantic accuracy, fluency, discourse-level cohesion, and pragmatics shape the communicative outcome (Schwandt, 2001; Sun et al., 2021; Gao and Wang, 2024), supported by statistical validation

using an information-theoretic approach (Wu et al., 2020). By analyzing these aspects, we explore how LLMs encode and generate dialogues that align with human linguistic behaviors across different L1 backgrounds.[1]

For benchmark selection, we utilize the ICNALE dataset (Ishikawa, 2023), which includes recordings from 435 human L2 speakers and manual transcripts, comprising approximately 1,600,000 tokens across diverse L1s. An information-theoretic analysis is applied to identify the effectiveness of L1-dependent features by comparing LLM-generated dialogues with human counterparts, providing insights into the influence of L1 information on L2 dialogue generation. To address challenges in mimicking L2 English dialogue generation, as an initial foray, we employ a simple prompting approach (Dong et al., 2022). In this setup, designed dialogue example in prompts is used to help LLMs internalize native knowledge through designed samples, ensuring alignment with the format of the ICNALE human dialogues (interviews with an L1 interviewer and an L2 learner).

Through our exploration, we demonstrate that a fairly simple L1 prompting has significant language-dependent impacts on LLM-generated L2 dialogues. For example, Japanese, Korean, and Mandarin L1 influence tense agreement, Thai and Malay L1 affect speech acts, and Urdu L1 impacts noun-verb collocations, which render intriguing parallels with real human L2 learners. Through our information-theoretic evaluation, we quantitatively show that LLMs generate highly human-like L2 dialogues; this is further supported by manual qualitative analyses as well. These results will serve as a starting point for evaluating LLMs by benchmarking their generation outputs against real human L2 dialogue data. In summary, our contributions are listed as follows:

- We propose a new evaluation framework with eight linguistic features, covering grammatical/semantic accuracy, fluency, cohesion, and pragmatics perspectives, designed to evaluate the impact of L1 information on LLM-generated dialogues. This framework enables systematic analysis of how native language traits (in humans/LLMs) shape linguistic features in cross-lingual dialogue generation.
- We further propose an information-theoretic metric to quantify L1 influence on LLM dialogue generation, revealing L1-dependent differences such as *reference word*, *modifiers* and *numerals* usages.
- We show that, through prompting, LLMs can generate dialogues with varying degrees of non-native-like linguistic features influenced by different L1s, paving a new way for using LLMs to simulate L2 communications.

[1]All codes and dataset can be found https://anonymous.4open.science/r/LLMPirorknowledge-017A/README.md

2 Related Work

2.1 Native and Bilingual Knowledge for LLM

Language Prior Knowledge Native language grammar profoundly influences L2 English communication (Brooke and Hirst, 2013; Chen et al., 2020). This "language interference" effect (Perkins and Zhang, 2024) shapes how L1 speakers interpret, represent, and generate new linguistic information, impacting their dialogue patterns. Native language also plays a growing role in developing/-analyzing LLMs, while prior studies have exclusively focused on sentence-level evaluations (Oba et al., 2023; Yadavalli et al., 2023; Elshin et al., 2024). When it comes to extending to the dialogue level, barriers exist, such as discourse-level contextual dependencies, including cohesion/coherence and nuanced differences in non-native speaker's dialogue strategy (Abe and Roever, 2019; Gao et al., 2024a). On the other hand, LLMs are now generally employed in generating discourse (e.g., chat interactions), opening the opportunity to evaluate their ability to emulate human-like, L1-dependent language intereferences (Jin et al., 2024).

Bilingual Knowledge Bilingual knowledge typically impacts LLM in cross-lingual and multilingual tasks (Miah et al., 2024). For example, leveraging shared grammatical features, bilingual LLM excels with typologically similar language pairs like English-Spanish, improving coherence and fluency through transfer learning (Jeon and Van Roy, 2022). On the other hand, handling distant cross-lingual pairs, such as English-Chinese, poses challenges (i.e., negative language transfer) due to differences in their grammatical features such as word order (Ranaldi and Pucci, 2023), requiring targeted training and alignment of grammatical constructs (Přibáň et al., 2024). In the context of dialogue tasks, limited L2 dialogue data and linguistic inconsistencies sometimes hinder

LLM performance for non-native English speakers (Gan et al., 2024). There are case studies that optimize bilingual knowledge integration, bridge these gaps, and enhance cross-lingual grammatical understanding (Huzaifah et al., 2024), as well as improve LLMs' ability to generate accurate and coherent dialogue, benefiting non-native English users (Han et al., 2024).

2.2 Evaluation of L2 Capabilities of LLM

The evaluation of human-like LLM has become a focal area of research, with recent studies exploring LLM capabilities in interactive tasks, like decision makings (Liu et al., 2024) during collaborative tasks between LLM and human. Recent work extends to knowledge-based capabilities across diverse scenarios, such as text-based dialogues Ou et al. (2024), by evaluating the performance of LLMs by proposing a benchmark on English text dialogues. These works move beyond traditional evaluation tasks which focused solely on factual recall, but offering an understanding of human-like evaluation. However, the gap still exist in interactive dialogues, which has more generalizable context for deploying LLM, such as cross-lingual interactions (Gao et al., 2024a) and language practice (Huzaifah et al., 2024).

Mimicking Human-like L2 Dialogues One of the most critical challenges in developing effective dialogue generation systems for L2 contexts lies in establishing a robust evaluation framework that helps to transfer linguistic knowledge from a speaker's L1 to the target L2 (Li and Qiu, 2023). Such a framework is essential, as it provides a structured way to integrate prior linguistic competence, helping models more intuitively learn meaningful, context-aware utterances (Sung et al., 2024). To bridge these gaps, evaluation protocols should incorporate cross-linguistic benchmarking (King and Flanigan, 2023) and error analysis (Kobayashi et al., 2024) to pinpoint the grammatical errors that frequently occur in specific languages. By systematically analyzing these errors, researchers can gain insights into the underlying issues related to grammatical understanding and representations within LLMs (Cong, 2025). This targeted evaluation process ultimately ensures that LLMs not only understand cross-lingual grammatical constructs but also excel in generating the unique knowledge of each language, leading to more effective and versatile language models in real-world cross-lingual applications (Gao et al., 2024a; Singh et al., 2024; Poole-Dayan et al., 2024).

3 Evaluation Metrics

3.1 Evaluation Framework

Unlike previous studies, we explore LLMs' ability to generate non-native English L2 dialogues and their L capabilities as "L2 speakers" interfered by native L1 knowledge. This shifts the role of LLMs from mere passive evaluators to active participants in dialogue interactions (Fincham and Alvarez, 2024), and has a connection to L2 conversation acquisition studied in applied linguistics (Roever et al., 2023). This leads to an intriguing subsequent question: **How accurately can LLMs simulate L1-dependent language interference in dialogue?**

To this end, we specifically target eight linguistic constructs to evaluate their L2 English generation ability, motivated by L1–L2 interference research (Jackson et al., 2018; Taguchi and Roever, 2020; Millière, 2024). The constructs cover both structural and functional aspects of languages, including *reference word* usage to assess their cohesion, *noun and verb collocations* to capture native-like lexical patterns, and various forms of *agreement* such as *number*, *tense*, and *subject-verb* consistency, which are critical for grammatical accuracy. Additionally, pragmatic constructs like *speech acts* and *modal verbs and expressions* evaluate contextually appropriate language use, reflecting cultural and linguistic nuances often influenced by L1. Together, these metrics provide a comprehensive framework to measure the effectiveness of LLM-generated L2 dialogues, identifying both strengths and areas for improvement in cross-lingual dialogue generation. We summarize these constructs in Table 1.

3.2 Information-Theoretic Metrics

Overview We quantify how similar the specific L2-English usages simulated by LLMs are to those collected by real human L2-English speakers. This is quantified by a particular information-theoretic distance between the dialogues produced by those two groups (LLMs vs. humans); that is, the smaller the score is, the better the LLMs could accurately simulate the real L2-English speakers.

Theoretical introduction We introduce a general information-theoretic framework to analyze the influence of an L1 language on English (L2)

Categories	Features	Definition	Example	Examples in Prompt
Grammatical accuracy	Number Agreement (Watts et al., 2024)	Adjectives and nouns must agree in number and gender (e.g., Spanish, French).	*100 cars*	*"The big cars are red."...*
	Tense Agreement (Oetting et al., 2021)	Ensuring the time of action (past, present, future) aligns across clauses.	*I [did] a task [yesterday].*	*"He has finished his homework."...*
	Subject-Verb Agreement (Jackson et al., 2018)	Matching the verb form to the subject's person and number.	*She [is] amazing.*	*"They are playing football."...*
Semantic accuracy	Modal Verbs and Expressions (Li, 2022)	Verbs that indicate likelihood, ability, permission, or obligation.	*She [might] come to the meeting.*	*"You should complete the project soon."...*
	Quantifiers and Numerals (Zhang and Kang, 2022)	Expressing numbers and amounts uniquely in different languages.	*Some, many, a few*	*"There are ten apples on the table."...*
Fluency	Noun-Verb Collocations (Bahns and Eldaw, 1993)	Common collocations that enhance sentence meaning.	*Drive a car*, *Do a test*	*"He drives a car every day."...*
Cohesion	Reference Word (Chow et al., 2023)	Linguistic devices referring to entities mentioned earlier (anaphora) or later (cataphora).	*She*, *her*, *him*, *he*	*"She went home early."...*
Pragmatics	Speech Acts (Ross and Kasper, 2013)	Actions performed via utterances (assertions, questions, requests, commands).	*"Could you open the window?"* (Indirect request)	*"Can you help me with this task?"...*

Table 1: Linguistic features targeted in our L2-like dialogue generation capability tests for LLMs

learning in dialogues with LLM. Let Y represent a random linguistic variable, as outlined in Table 1, with a distribution $p(Y)$ that captures the properties of the English language shared by the native speakers. Let us additionally assume that learners learn a language with the help of stimuli produced by the respective speakers; thus, in the case of L1 English learners, the learning materials D are generated by Y, following the likelihood $p(D|Y)$. By combining $p(Y)$ and $p(D|Y)$, we model the learning of English dialogue structure through the posterior $p(Y|D)$, which quantifies how effectively an English L1 learner can infer Y from D.

Now, extending this to second language learning, we define another random variable X to represent linguistic properties for L1 **human** native speakers of that L1 language. Here, X acts as a priori information. The relationship between X and Y is described by $p(Y|X) \propto p(Y)p(X|Y)$, where $p(X|Y)$ reflects the likelihood between English and the second language. The data D used to learn their languages (for their L1 and L2) is now defined as $p(D|Y, X)$, and the updated posterior becomes $p(Y|D, X)$, incorporating the prior distribution $p(Y|X)$. The human-like $p(Y|D, X)$ is estimated with the dialogues produced by the real human L2-English speakers of L1 natives (§ 4).

Then, when it comes to LLMs with the respective L1 and L2, their L1 prior knowledge (and their general learning bias) is noted as X'. Our focus is on whether they can have a human-like X (that is, similar to X') when prompted to behave like respective human L2 speakers, which is analyzed through the lens of the L2 behavior similarity between LLMs' $p(Y|D, X')$ and humans' $p(Y|D, X)$. These differences are quantified as a density between these distributions in our experiments. Mathematically, we can characterize this difference with the logarithmic loss function $\ell(Q) = -\log Q$, leading to the following evaluation[2]:

$$
\begin{aligned}
d &= \mathbb{E}_{XX'YD}\left[\ell(p(Y|D, X')) - \ell(p(Y|D, X))\right] \\
&= \mathbb{E}_{XYD}\left[\log \frac{p(Y|D, X)}{p(Y|D)}\right] - \mathbb{E}_{X'YD}\left[\log \frac{p(Y|D, X')}{p(Y|D)}\right] \\
&= I(X; Y|D) - I(X'; Y|D),
\end{aligned}
$$

Where $I(X; Y|D)$ quantifies the mutual information between X and Y given D, it represents the shared information between English (Y) and the

[2]The dependence of D on all X, X', and Y ensures that the posterior $p(Y|D, X)$ differs from $p(Y|D, X')$ due to the different priors $p(Y|X)$ and $p(Y|X')$.

Stats	Dialogues	Tokens	Participants
# ICANLE (Human)	4,250	1,600K	425
# LLM Generated	2,600	1,344K	NA
# Example Dialogue	7 sets (one per each L1)	10K	NA

Table 2: Statistics of L2 Dialogue dataset, including human benchmarks, generated L2 dialogue datasets, and those used in prompting

native language (X) conditioned on the context D. We report d_{bi} in the case that LLMs are instructed to mimic an L2-English speaker with the respective L1. As a baseline, we also compute d_{mono} when no valid L1 information is provided to LLMs as X' (see § 5.2).

4 Annotation Design

As detailed in Section 3, we propose a metric to compare **LLM-generated** L2 dialogues with **human-produced** L2 dialogues. To ensure the reliability of benchmark annotations, we employed a hybrid approach combining automated methods with manual review. This annotation process targeted eight key linguistic constructs that influence dialogue construction from grammatical accuracy to pragmatics, as outlined in Table 1. For this aim, we utilize the International Corpus Network of Asian Learners of English (ICNALE) dataset (Ishikawa, 2018), which includes dialogue response utterances from speakers of 18 diverse native languages: Bahasa Indonesia, Cantonese, English, Mandarin, Japanese, Korean, Filipino, Javanese, Malay, Pakistani, Pashto, Pashtoo, Punjabi, Urdu, Pushto, Tagalog, Thai, and Uyghur.[3] This dataset offers comprehensive information about L2 English speakers with varied L1 backgrounds. Each file in ICNALE contains transcripts of a single L2 speaker's recorded responses on different discussion topics. Examples of these dialogue transcripts can be found in Appendix A.4. For this study, we selected seven linguistically divergent native languages from the dataset (Philippy et al., 2023): Korean (ko), Mandarin (cmn), Japanese (ja), Cantonese (yue), Thai (th), Malay (ms), and Urdu (ur).

4.1 Automated Annotation with GPT-4o

The initial annotation process was performed using GPT-4o (Achiam et al., 2023). Hence we employ few-shot prompting with four examples under each feature, with designed prompts for each feature. For example, for *Reference Word*, we provided four sentences from one dialogue and span the reference word (he, she, her) then provide the full sentence and spanned reference word in few-shot (detailed prompts, provided in Appendix A.3). Each dialogue in the dataset was analyzed using GPT-4o to identify and annotate the specified linguistic entities using a *span-annotation* approach. The resulting annotations were stored in a structured format, such as JSON, to maintain consistency and facilitate efficient manual review.

4.2 Human Validation of LLMs Annotations

To assess the quality of the automated annotations, three volunteer annotators who are proficient bilingual speakers and are all PhD in NLP, and manually reviewed 15% of the annotated dialogues, randomly sampled from the entire set. The annotators are required to judge span-annotation output is correct or not in a brainy way. This process included a cross-validation step to compare the automated span annotations against the human judgment. By combining automation with human oversight, this two-step validation ensured both scalability and reliability. Following the manual review, the GPT-4o annotations achieved an accuracy of 84.1% when compared to the human-validated results. Minor discrepancies were observed, particularly in annotating *Noun-Verb Collocations* within the human L2 dialogues.

4.3 Prompt Refinement and Rechecking

Based on feedback from the manual review, we improved the few-shot examples in the prompts for *Noun-Verb Collocations* to address the identified shortcomings in the automated annotation process. The updated annotations were then subjected to a second round of human validation to ensure they met the required quality standards and passed the manual review. The updated results for *Noun-Verb Collocations* reaches 83.6%. In that case, we use updated prompts and few-shot examples from the second round modification for the formal annotation of *Noun-Verb Collocations*.[4]

[3]For more details, see https://language.sakura.ne.jp/icnale/

[4]Updated prompts refer to https://anonymous.4open.science/r/LLMPirorknowledge-017A/lib/instructions/annotation_instructions/assist_instructions/noun_verb_collocation.txt

5 Evaluation and Discussion

Our generation pipeline consists of two key steps: **(1)** designing and implementing one-shot prompting to simulate L2 English dialogues for seven different L1s, including Cantonese, Malay, Japanese, Korean, Mandarin, Thai, and Urdu by injecting L1 Knowledge grammatical traits; and **(2)** annotating the LLM-generated L2 dialogues and performing a comparative evaluation. Further details are provided below.

5.1 Injecting L1 Knowledge

The proper way to inject the L1 knowledge into LLMs is not obvious; perhaps pre-training LLMs under a bilingual setup from scratch might be plausible, but this requires a huge amount of computing resources. At least in this study, as an initial foray, we inject the L1 knowledge with a simple prompting approach; that is, we employ fine-grained instruction that contains high-level meta-linguistic information of the L1 language and examples of carefully crafted dialogue pairs that capture key dialogue grammatical traits of different native languages (L1s) (Chen, 2023; Hu et al., 2022). Detailed instructions and sample L1 knowledge injection dialogue pairs are provided in Appendix A.1. Each pair consists of at least 20 turns of conversation in L1, with corresponding English (L2) translation. These examples emphasize specific linguistic features, such as Speech Act Politeness in Thai (Srisuruk, 2011), shown in Figure 2. Derived from real human L1 dialogues from xDial-Eval (Zhang et al., 2023), a multilingual open-domain dialogue dataset. The instructions ensure L2 outputs reflect the linguistic characteristics of L1 speakers with prominent grammatical traits, for example, particles in Cantonese, verb conjugations in Japanese and Thai. By standardizing dialogue structures and embedding grammatical traits, this setup offers a controlled framework for generating accurate L2 dialogues. Sample examples for different L1s are provided in Appendix A.1, as space limitations prevent listing all dialogue turns here.

Using the described L1 knowledge injection prompting setup, dialogues are generated to simulate L2 English speakers whose language usage is influenced by their L1 backgrounds under our designed instruction. This is achieved by conditioning the LLM[5] with prompts that capture grammatical features characteristic of L1 speakers. The

[5]For GPT-4o generations, the temperature is set to 0.

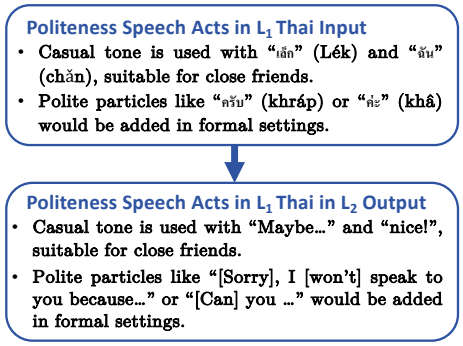


Figure 2: An example for Thai L1 knowledge injection of Speech Acts, we provided full sentences in a complete dialogue context, the utterances were omitted as "..." in this figure

LLM is instructed to "role-play" as an L2 English speaker, emulating realistic behaviors such as tense agreement and politeness strategies. For example, the model is prompted to act as an L2 speaker in an interviewee-interviewer scenario, where the interviewer (a native speaker) follows predefined templates based on ICNALE benchmark datasets. Details of these prompts and scenarios can be found in Appendix A.2. The generated L2 dialogue utterances are then saved in a structured JSON format, ensuring consistency with the ICNALE benchmark datasets as defined in Section 3, in preparation for annotation.

5.2 Annotation and Information-Theoretic Analysis

The generated dialogues' L2 utterances are extracted and annotated (the native speaker utterances get excluded for this stage) for eight linguistic features using the annotation framework outlined in Section 3. The annotated data is then analyzed using an information-theoretic analysis framework to quantify the influence of native L1 knowledge on the distribution density of L2 dialogues. The distribution generally represents the frequency of particular linguistic features annotated in each dialogue. Specifically, we quantify the differences between linguistic feature distribution in human L2 dialogue $p(Y|D, X)$ and that in LLMs-generated dialogues $p(Y|D, X')$ (as prompted in § 5), as shown in § 3.2. We call this distribution distance d_{bi}. This value is compared with the baseline d_{mono} that is computed with the dialogue generated by LLMs *without* the instruction to mimic an L2 English speaker (thus, simply an English monolingual speaker). If the model can, more or less, mimic the L2 English

Lang.	Condition	Distribution distance between humans' and LLMs' generated dialogues (↓)							
		Number Agreement	**Tense Agreement**	**Subject-Verb Agreement**	**Modal Verbs Expressions**	**Quantifiers Numerals**	**Noun-Verb Collocation**	**Reference Word**	**Speech Acts**
yue	d_{bi}	0.099	0.275	0.073	0.052	0.145	0.066	0.109	0.145
	d_{mono}	0.489	0.027	0.318	0.123	0.725	0.029	0.188	0.203
th	d_{bi}	0.130	0.013	0.060	0.120	0.049	0.121	0.097	0.188
	d_{mono}	0.265	0.570	0.913	0.180	9.227	0.190	0.222	0.400
ja	d_{bi}	0.190	0.082	0.087	0.053	0.044	0.073	0.265	0.212
	d_{mono}	0.330	0.514	0.874	0.452	1.954	0.232	0.273	0.520
ko	d_{bi}	0.051	0.019	0.148	0.131	0.033	0.183	0.009	0.109
	d_{mono}	0.259	0.296	0.605	0.069	0.654	0.295	0.108	0.247
ms	d_{bi}	0.092	0.036	0.026	0.065	0.007	0.096	0.027	0.076
	d_{mono}	0.341	0.321	0.477	0.097	1.039	0.080	0.109	0.279
cmn	d_{bi}	0.037	0.038	0.023	0.082	0.027	0.059	0.065	0.161
	d_{mono}	0.375	0.277	0.741	0.212	1.382	0.108	0.099	0.319
ur	d_{bi}	0.050	0.073	0.046	0.126	0.079	0.044	0.062	0.043
	d_{mono}	0.282	0.145	0.386	0.115	0.918	0.046	0.192	0.158

Table 3: The distribution divergences d_{bi} and d_{mono} of LLM-Generated L2 dialogues for different native languages: Cantonese (yue), Thai (th), Japanese (ja), Korean (ko), Malay (ms), Mandarin (cmn) and Urdu (ur)

speaker with the instruction, d_{bi} should be smaller than d_{mono}.

6 Results and Analysis

6.1 Prior Knowledge Impact

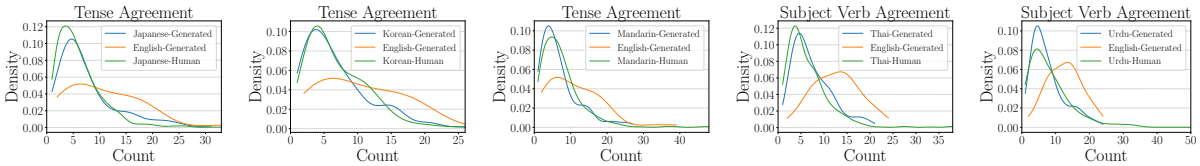
As shown in Table 3, the LLM-generated L2 dialogues exhibit generally consistent and significant improvements across all seven languages after prompting with the L1 information, given the decrease of d_{bi} from d_{mono}. This shows the effectiveness of promoting native linguistic information in L2-like dialogue generation. The eight grammatical constructs listed in Table 1 demonstrate human-like distribution patterns when leveraging native knowledge through L1 knowledge injection learning. This is particularly evident in the categories of *Agreement* and *Pragmatics*, which play an important role in oral communications (Gao et al., 2024b).

Still, in some cases, the L1 injection does not reasonably improve the fit to respective L2 human dialogues. Taking Japanese as an example, the relatively large gap for *Reference Word* demonstrated the difficulties to map pronoun usage in L1-Japanese of L2 speakers with LLM-L2 speakers, which might reflect frequent omissions of pronouns in Japanese spoken expressions, which lacks enough training instances for LLM to infer the language transfer effect from Japanese to English. More generally, the distance in Speech Acts tends to be larger than other linguistic features, suggesting the challenging issue in simulating L2 speakers in discourse, pragmatic level. Nevertheless, for most cases, these divergence measures confirm that LLM tends to, more or less, effectively produce L2-like and context-sensitive dialogues with simple prompting.

6.2 Evaluating LLM L2 Generation via L1 Distance

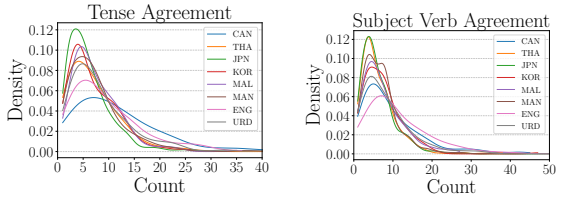
The results of the density comparison shown in Figure 3 indicate subtle but consistent ways in which a speaker's native language influences the LLM's ability to generate English L2 dialogues (see Appendix 3 for full details). For speakers of **Chinese, Japanese, and Korean** that are close in *Tense Agreement* for real L1 (Figure 4a). The generated dialogues exhibit surprisingly L2 human-like patterns in maintaining *Tense Agreement*, which can be further supported by L1 distance density results in Figure 3. The LLMs successfully replicate these patterns for these three languages, indicating a significant transfer from the native language (L1) to the L2 English dialogues.

We also perform further exploratory analyses of inter-L1 differences — which L1 more drastically impacted the L2 English language in human L2 learners (Figure 4). Here, the ENG line (pink) is the baseline (English L1 native). We observe that languages with a feature distant from the English language, such as 'SOV' (Subject-Object-Verb) word order, rather impacted the generated dialogues. Such an L1-dependent degree of impact is, more or less, reflected in the LLMs. For example, in the case of *Subject Verb Agreement*, THA



(a) Japanese-L1 with Tense Agreement (b) Korean-L1 with Tense Agreement (c) Mandarin-L1 with Tense Agreement (d) Thai-L1 with Subject Verb Agreement (e) Urdu-L1 with Subject Verb Agreement

Figure 3: Density results for L2 generation dialogue via different L1s



(a) Tense Agreement of Real L1 (b) Subject-Verb Agreement of Real L1

Figure 4: Density results for human-produced dialogues of different L1s

and JPN are more distant from the English L1's distribution (Figure 4b). This is reflected in the significant decrease from d_{mono} to d_{bi} in Table 3 in the Thai and Japanese languages.

6.3 Qualitative Analysis: LLM L2 Human-like Generations

In addition to the information-theoretic analysis presented in Sections 6.1 and 6.2, we conducted a qualitative analysis of LLM-generated L2 dialogues.[6] As shown in Appendix A.5, we identified three key linguistic features influenced by L1-Urdu: **Word Order, Agreement, and Collocations** (Abbas, 2016). For instance, Urdu's flexible word order leads to L2 English constructions like "Lahore University I study", reflecting Urdu syntax rather than standard English. This flexibility also affects word collocations, resulting in less rigid grammatical structures. Similarly, *Subject-Verb Agreement* errors, such as "It has been, um, about three year now", stem from Urdu grammar, where adjective-noun agreement is prioritized over subject-verb consistency. For Thai speakers, as outlined in Appendix A, generated dialogues reflect *Politeness Levels* typical of informal Thai. Omission of politeness particles, common in informal Thai speech (Yossatorn et al., 2022), leads to responses like "Yes, a lot of plant. Many flower, very beautiful", which replicate casual Thai but may appear abrupt in English due to missing formal markers. These observations demonstrate the influence of L1 linguistic structures on L2 dialogue generation, highlighting how linguistic transfer shapes grammar and syntax in LLM-generated dialogues.

[6]This analysis involved a manual review of 30 generated dialogues per language by native L1 speakers to ensure generation quality.

7 Conclusions

By implementing an automated structured dialogue annotation framework, this study introduces a linguistically informed and information-theoretic evaluation approach to assess LLMs' ability to simulate L2 English dialogues influenced by L1 knowledge. With the benchmark such as the ICNALE dataset, our evaluation ensured consistency by comparing LLM-generated outputs with human-produced data. We demonstrated that LLM-generated L2 dialogues reflect L1-specific influences through a designed L1 knowledge injection mechanism. The results indicate that LLMs effectively replicate native-like L1 linguistic patterns and align closely with human L2 speakers in areas such as dialogue cohesion, grammatical agreement, and pragmatic usage. These insights suggest that the potential to refine LLM evaluation frameworks for better handling linguistic diversity and multilingual contexts, supporting the development of more adaptive and context-aware dialogue systems for speakers of diverse native languages.

8 Limitations

This study has several limitations. First, it relies on the ICNALE dataset as a only benchmark, which may limit the generalizability of the results to languages that are underrepresented in the dataset due to practical limitation in L2 dialogue datasets. Second, the use of predefined templates for one-shot ensures consistency but may constrain the analysis of *spontaneous L2 language behaviors*, such as chit-chat. Furthermore, the study focuses on lin-

guistics features, overlooking the potential impact of socio-cultural bias on each native language use. Future work should address these limitations by incorporating more diverse datasets and examining unscripted interactions to enhance the validity and applicability of the results.

Ethics Statement

This study is conducted under the guidance of the ACL Code of Ethics. The volunteer annotators were all NLP PhDs who are willing to participate in manual checking for this study. We removed all information related to the identification of human volunteer annotators.

References

Qaiser Abbas. 2016. Morphologically rich urdu grammar parsing using earley algorithm. *Natural Language Engineering*, 22(5):775–810.

Makoto Abe and Carsten Roever. 2019. Interactional competence in l2 text-chat interactions: First-idea proffering in task openings. *Journal of Pragmatics*, 144:1–14.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tatsuya Aoyama and Nathan Schneider. 2024a. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Tatsuya Aoyama and Nathan Schneider. 2024b. Modeling nonnative sentence processing with l2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940.

Jens Bahns and Moira Eldaw. 1993. Should we teach efl students collocations? *System*, 21(1):101–114.

Julian Brooke and Graeme Hirst. 2012. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 779–784, Istanbul, Turkey. European Language Resources Association (ELRA).

Julian Brooke and Graeme Hirst. 2013. Native language detection with 'cheap'learner corpora. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*, volume 1, page 37. Presses universitaires de Louvain.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*, pages 3426–3437.

Olga Cherednichenko, Olha Yanholenko, Antonina Badan, Nataliia Onishchenko, and Nunu Akopiants. 2024. Large language models for foreign language acquisition.

Bonnie Wing-Yin Chow, Anna Na-Na Hui, Zhen Li, and Yang Dong. 2023. Dialogic teaching in english-as-a-second-language classroom: Its effects on first graders with different levels of vocabulary knowledge. *Language Teaching Research*, 27(6):1408–1430.

Yan Cong. 2025. Demystifying large language models in second language development research. *Computer Speech & Language*, 89:101700.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

CM Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. 2023. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. *arXiv preprint arXiv:2309.04679*.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, et al. 2024. From general llm to translation: How we dramatically improve translation quality using human evaluation data for llm finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 247–252.

Naiyi Xie Fincham and Aitor Arronte Alvarez. 2024. Using large language models (llms) to facilitate l2 proficiency development through personalized feedback and scaffolding: An empirical study. In *Proceedings of the International CALL Research Conference*, volume 2024, pages 59–64.

Yujian Gan, Changling Li, Jinxia Xie, Luou Wen, Matthew Purver, and Massimo Poesio. 2024. Clarq-llm: A benchmark for models clarifying and requesting information in task-oriented dialog. *arXiv preprint arXiv:2409.06097*.

Rena Gao, Carsten Roever, and Jey Han Lau. 2024a. Interaction matters: An evaluation framework for interactive dialogue assessment on english second language conversations. *arXiv preprint arXiv:2407.06479 (to be appeared at the 31st International Conference on Computational Linguistics 2025)*.

Rena Gao, Jingxuan Wu, Carsten Roever, Xuetong Wu, Jing Wu, Long Lv, and Jey Han Lau. 2024b. Cnima: A universal evaluation framework and automated approach for assessing second language dialogues. *arXiv preprint arXiv:2408.16518*.

Wei Gao and Menghan Wang. 2024. Listenership always matters: active listening ability in l2 business english paired speaking tasks. *International Review of Applied Linguistics in Language Teaching*, (0).

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2024. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Taylor & Francis.

Shin'ichiro Ishikawa. 2018. Icnale: the international corpus network of asian learners of english. *Icnale: the international corpus network of asian learners of english*.

Carrie N Jackson, Elizabeth Mormer, and Laurel Brehm. 2018. The production of subject-verb agreement among swedish and chinese second language speakers of english. *Studies in Second Language Acquisition*, 40(4):907–921.

Hong Jun Jeon and Benjamin Van Roy. 2022. An information-theoretic framework for deep learning. *Advances in Neural Information Processing Systems*, 35:3279–3291.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*, pages 2627–2638.

Brendan King and Jeffrey Flanigan. 2023. Diverse retrieval-augmented in-context learning for dialogue state tracking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5570–5585, Toronto, Canada. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Lexi Xiaoduo Li. 2022. Developmental patterns of english modal verbs in the writings of chinese learners of english: a corpus-based approach. *Cogent Education*, 9(1):2050457.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2024. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Raphaël Millière. 2024. Language models as models of language. *arXiv preprint arXiv:2408.07144*.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Janna B Oetting, Andrew M Rivière, Jessica R Berry, Kyomi D Gregory, Tina M Villa, and Janet McDonald. 2021. Marking of tense and agreement in language samples by children with and without specific language impairment in african american english and southern white english: Evaluation of scoring approaches and cut scores across structures. *Journal of Speech, Language, and Hearing Research*, 64(2):491–509.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170, Mexico City, Mexico. Association for Computational Linguistics.

Kyle Perkins and Lawrence Jun Zhang. 2024. The effect of first language transfer on second language acquisition and learning: From contrastive analysis to contemporary neuroimaging. *RELC Journal*, 55(1):162–178.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. *arXiv preprint arXiv:2305.02151*.

Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. Llm targeted underperformance disproportionately impacts vulnerable users. *arXiv preprint arXiv:2406.17737*.

Pavel Přibáň, Jakub Šmíd, Josef Steinberger, and Adam Mištera. 2024. A comparative study of cross-lingual sentiment analysis. *Expert Systems with Applications*, 247:123247.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the english matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183.

Carsten Roever, Yuki Higuchi, Miyuki Sasaki, Tomoko Yashima, and Makiko Nakamuro. 2023. Validating a test of l2 routine formulae to detect pragmatics learning in stay abroad. *Applied Pragmatics*, 5(1):41–63.

Steven Ross and Gabriele Kasper. 2013. *Assessing second language pragmatics*. Springer.

Thomas A Schwandt. 2001. Understanding dialogue as practice. *Evaluation*, 7(2):228–237.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Vaibhav Singh, Amrith Krishna, Karthika NJ, and Ganesh Ramakrishnan. 2024. A three-pronged approach to cross-lingual adaptation with multilingual llms. *arXiv preprint arXiv:2406.17377*.

Patana Srisuruk. 2011. *Politeness and pragmatic competence in Thai speakers of English*. Ph.D. thesis, Newcastle University.

Weiwei Sun, Chuan Meng, Qi Meng, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Conversations powered by cross-lingual knowledge. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1442–1451.

Mingi Sung, Seungmin Lee, Jiwon Kim, and Sejoon Kim. 2024. Context-aware LLM translation system using conversation summarization and dialogue history. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1011–1015, Miami, Florida, USA. Association for Computational Linguistics.

Naoko Taguchi and Carsten Roever. 2020. *Second language pragmatics*. Oxford University Press.

Chikako Takahashi. 2024. L1 japanese perceptual drift in late learners of l2 english. *Languages*, 9(1):23.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *arXiv preprint arXiv:2406.15053*.

Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. 2020. Information-theoretic analysis for transfer learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2819–2824. IEEE.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Yossiri Yossatorn, Theerapong Binali, Sirisira Chokthawikit, and Cathy Weng. 2022. Thai efl university students' productions of the english past counterfactuals and their influences from interlanguage fossilization. *SAGE Open*, 12(1):21582440221079892.

Chao Zhang and Shumin Kang. 2022. A comparative study on lexical and syntactic features of esl versus efl learners' writing. *Frontiers in Psychology*, 13:1002090.

Chen Zhang, Luis D'Haro, Chengguang Tang, Ke Shi, Guohua Tang, and Haizhou Li. 2023. xDial-eval: A multilingual open-domain dialogue evaluation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5579–5601,

Singapore. Association for Computational Linguistics.

A Appendix

A.1 High-Level Instructions and Input-Output Pairs

A.1.1 General Prompts

Depending on the language, we design explicit L1 knowledge injection learning examples adopted from L2 human data and based on the grammatical traits in expression from each native language

Prompt

Your goal is to generate a realistic conversation in English between one {target language} native speaker and a native English speaker. Read and learn the provided {target language} dialogue and the analysis of grammatical traits. Scene [Optional]: Two friends, {speaker 1} and {speaker 2}, are planning to visit the mall over the weekend and discuss what to do there.

L1 Knowledge Injection Prompt

In this section, we only show a piece of L1 knowledge injection example prompts for different L1s. For more examples from full dialogues, please refer to: `https://anonymous.4open.science/r/LLMPirorknowledge-017A/lib/instructions/context_instructions/`

Scene: Two friends, {speaker A} and {speaker B}, are meeting at a {certain place} for {some discussions}. **Note that this is a template for different example prompt depending on the scene and the contents {...}** are put here as placeholders.
Malay Example
Aiman: Farah, awak ada rancangan hujung minggu ni?
(Farah, awak ada rancangan hujung minggu ni?)
"Farah, do you have any plans this weekend?"
Farah: Tak ada apa-apa pun. Kenapa?
(Tak ada apa-apa pun. Kenapa?)
"No, nothing at all. Why?"
Urdu Example
Ayesha: کیا تم نے نئی لائبریری دیکھی ہے؟
(Kya tum ne nayi library dekhi hai?)
"Have you seen the new library?"
Bilal: ہاں، میں کل لائبریری گیا تھا۔
(Haan, main kal library gaya tha.)
"Yes, I went to the library yesterday."
Japanese Example
Sora: こんにちは 明日何をする予定ですか？
(Konnichiwa, ashita nani o suru yotei desu ka?)
"Hello, what are your plans for tomorrow?"
Aki: 明日は特に予定がありませんが どうしてですか
(Ashita wa toku ni yotei ga arimasen ga, doushite desu ka?)
"I don't have any particular plans for tomorrow. Why do you ask?"
Korean Example
Minji: 지수야, 이번 주말에 시간 있어?
(Jisoo-ya, ibeon jumal-e sigan isseo?)
"Jisoo, do you have time this weekend?"
Jisoo: 응, 있어. 왜?
(Eung, isseo. Wae?)
"Yes, I do. Why?"

Thai Example
Nuch: เล็ก วันเสาร์นี้ว่างไหม?
(Lék wan săo níi wâang mái?)
"Lek, are you free this Saturday?"
Lek: ว่างสิ มีอะไรเหรอ?
(Wâang sì. Mii à-rai rŏe?)
"I free. What's up?"
Mandarin Example
Xiao Ming: 我想去公園玩儿,最近天气很好 。
(Wǒ xiǎng qù gōng yuán wánr, zuì jìn tiān qì hěn hǎo.)
"I want to go to the park; the weather has been great recently."
Xiao Li: 好主意!你想做什么?
(Hǎo zhǔ yì! Nǐ xiǎng zuò shén me?)
"Good idea! What do you want to do?"
Cantonese Example
Mei: 喂,阿Wing,星期六有冇時間呀?
(Wai, a Wing, sing1 kei4 luk6 jau5 mou5 si4 gaan3 aa3?)
"Hey, Wing, do you have time on Saturday?"
Wing: 有呀,你想做咩呀?
(Jau5 aa3, nei5 soeng2 zou6 di1 me1 aa3?)
"Yes, what do you want to do?"

Trait Analysis Prompt

Make sure to follow the following idiomatic expressions and cultural nuances commonly used by {target language} speakers. Keep the tone respectful and in line with traditional {target language} communication styles. **Here we give Malay as an example while we do have specific trait analysis prompts for other languages.**

1. **Particles**
   - "pun": Used for emphasis, e.g., "Tak ada apa-apa pun." (Nothing at all).
   - "ke": Indicates direction, e.g., "pergi ke pusat membeli-belah" (go to the mall).
2. **Aspect Markers**
   - "nak": Informal future marker, e.g., "Saya nak pergi" (I want to go).
   - "dengar": Implied past aspect in "saya dengar food court dia besar" (I heard their food court is big).
3. **Topic-Comment Structure**
   - "Wayang apa yang awak nak tengok?" (What movie do you want to watch?): Topic "Wayang apa" introduces the subject, and "awak nak tengok" comments on it.
4. **Politeness Levels**
   - Formal tone with "saya" (I) and "awak" (you) is polite but casual, suitable for friendly conversations.
   - Politeness can be enhanced with "Encik" or "Cik" for formal contexts.
5. **Verb Serialization**
   - "Makan tengah hari di sana. Lepas tu, nak tengok wayang?" (Have lunch there. After that, shall we watch a movie?): Actions are listed sequentially.

6. **Conjunctions**
    - "dan": Connects clauses, e.g., "banyak kedai baru, dan saya dengar" (many new shops, and I heard).
    - "Lepas tu": Informal for "after that."
7. **Time Expressions**
    - "hujung minggu ni" (this weekend).
    - "pukul 10 pagi" (10 a.m.).
8. **Expressions of Agreement**
    - "Setuju!" (Agreed!).
    - "Boleh!" (Sure!).
9. **Conditional Suggestions**
    - "Kita tengok jadual wayang nanti." (Let's check the movie schedule later): Indicates a planned action.
10. **Adjectives for Excitement**
    - "Bagus tu!" (That's great!) expresses enthusiasm.

A.2 L2 Dialogue Generation Prompts

Prompt

Given the topic: text. Generate a realistic conversation IN ENGLISH with 20 turns between two native Cantonese speakers. Make sure the output is not cut off. Provide the complete English conversation below.

1. **Speaker A (Native Speaker, NS)**
    - Fluent and natural English speaker with clear, concise, and polite phrasing.
    - Provides guidance, asks questions, and may clarify misunderstandings when necessary.
    - Avoids overly complex words or idioms to make the conversation accessible for L2 learners.
2. **Speaker B (Second-Language Speaker)**
    - A non-native English speaker whose proficiency reflects an intermediate-to-upper-intermediate level.
    - Their native language is {language}, please follow the idiomatic expressions and cultural nuances commonly used by {language} speakers.
    - Exhibits typical linguistic influences from their native language, such as:
        - Grammatical mistakes (e.g., "He have" instead of "He has").
        - Limited vocabulary leading to overuse of simple words or circumlocution (e.g., "thing for fixing paper" instead of "stapler").
        - Pronunciation hints if relevant.
        - Uses filler phrases or pauses to reflect real-time language processing (e.g., "Um", "How to say...").
3. **Context**: The conversation is around for some topics or scenes. The L2 speaker is trying to express their thoughts, answer questions, or solve a problem, while the native speaker responds supportively to maintain the flow of the conversation.

4. **Requirements**

   - **Cultural Nuances**: Reflect the L2 speaker's cultural communication style.
   - **Balanced Exchange**: Ensure the dialogue alternates between the two speakers.
   - **Error Patterns**: Highlight realistic mistakes in the L2 speaker's grammar, vocabulary, or syntax. Include occasional self-corrections or clarifications prompted by the native speaker.
   - **Clarity and Empathy**: The native speaker provides clear, friendly responses, avoiding judgment of language mistakes.
   - **Length and Focus**: The conversation should be concise, focusing on the L2 speaker's ability to express their ideas despite language barriers.

L1 Knowledge Injection Prompt

Speaker A (NS): Hi! Thanks for meeting with me today. Can you tell me a little about yourself?
Speaker B (L2): Um, yes. My name is Mei. I am from Hong Kong. I, uh... work in marketing for... four years.
Speaker A (NS): That's great! What kind of marketing work do you do?
Speaker B (L2): I do, um, online... how to say... advertisement? On social media, and also write article.
Speaker A (NS): Oh, social media advertising and content writing?
Speaker B (L2): Yes, yes! Content writing. Sometimes for product launch, or... uh, promotion.
Speaker A (NS): I see. Do you enjoy writing for different audiences?
Speaker B (L2): Yes, very much. But, um... sometime hard because need many idea. Creative, you know?
Speaker A (NS): Absolutely, coming up with fresh ideas can be challenging. How do you find inspiration?
Speaker B (L2): I... ah, read other, um, campaign? And look what people like. Sometimes ask my teammate.
Speaker A (NS): That's a smart approach! Collaboration always helps. What's a campaign you're particularly proud of?
Speaker B (L2): Oh, um, last year I make one for new phone. We use... uh, storytelling to show family connect. Many people like.
Speaker A (NS): Storytelling is very effective. How did you measure its success?
Speaker B (L2): We see, uh, number of share on Facebook and, um... how to say... comment? And we also check sale data.

A.3 L2 Annotation Prompts

Annotation Prompt

- *You are a linguist expert specializing in doing text annotation in the English second language. You will be tasked with making annotations to a given dialogue texts based on some linguistics aspects to compare grammatical features in machine learning models for cross-lingual tasks.*
- The given text are samples in the dialogue passage from second language speakers of English.
- Make sure to keep the annotation format without any change in passage when giving the annotation output.
- A task may ask for one or multiple annotations. Each annotation should be an object with 5 fields:

  - type: the type of annotation
  - annotation sentence: the annotated sentence
  - annotation token: the annotated tokens
  - rationale: the reason why you give the annotation
  - grammar correctness: the annotated grammar feature is aligned with the native English speaker's grammar usage

- Please return a json object which consists of one or multiple modifications.

A.4 Examples from the ICNALE Datasets

Dialogue

- Uh, I think a 100 points is a full points maybe. I think that I have - I maybe have 70 or 75 points.
- No, I - no.
- Um, I think this, uh, starting a new - a new thing I think, this will take a little time, uh, maybe for a month, 2 months, or maybe half an year, but finally you - you will not feel nervous about this.
- Because, uh, when - when you start something, people always - all of the people will feel nervous I think.
- So, just develop your English speaking skills and you will feel confident about.
- I think part-time jobs because I - I am now a student and I - I have no part-time job experience.
- Uh, I think role play. This - uh, because I can have some communications with the teachers and this - this picture - storytelling, I think this is a bit - a little familiar with the TOEFL test, uh, speaking test.

A.5 LLM Generated L2 Dialogue Examples

A.5.1 English Example

Dialogue

Speaker A (NS): Hey Mike, got any plans for the weekend? I was thinking we could do something fun.
Speaker B (NS): Not really. I was just gonna stay home and binge some Netflix. Why, what's up? Do you have something in mind?
Speaker A (NS): Yeah, how about a quick road trip? We haven't done that in ages! I've been itching to get out of the city.
Speaker B (NS): Hmm, that sounds cool. Where were you thinking? Someplace nearby, I hope. Gas prices are crazy.
Speaker A (NS): True! I was thinking we could hit the beach. Maybe that little spot down by Clearwater? It's not too far, and I hear the weather's supposed to be great.
Speaker B (NS): Oh, Clearwater? That could work. What would we do there, though? Just lounge around or do something active?
Speaker A (NS): A bit of both, maybe? We could rent paddleboards or bikes. Then just chill on the sand after. I mean, it's the beach—pretty hard to get bored.
Speaker B (NS): I like the sound of that. Should we leave Saturday morning? Beats sitting around

all weekend.
Speaker A (NS): Yeah, bright and early! Say 8 a.m.? I'll bring snacks if you handle the playlist.
Speaker B (NS): Deal. But no weird indie stuff this time, okay? I'm still scarred from that experimental jazz phase you went through.
Speaker A (NS): Ha! No promises, but I'll try to keep it mainstream. By the way, do you have a cooler? We'll need one for drinks.
Speaker B (NS): Yeah, I've got one. What about parking? Isn't it a nightmare there?
Speaker A (NS): I think there's a lot near the main boardwalk. We'll just have to get there early to snag a spot. Another reason to leave on time!
Speaker B (NS): Gotcha. Alright, sounds like a plan. Let's touch base Friday night to confirm everything.
Speaker A (NS): Sure thing. This is gonna be awesome! See you Friday!
Speaker B (NS): Later, Sarah!

A.5.2 Thai Example

Dialogue

Speaker A (NS): Hello! Is this Nuch?
Speaker B (L2, Thai): Yes, yes, it is Nuch. Who is speaking, please?
Speaker A (NS): It's Lek, your cousin! How have you been doing?
Speaker B (L2, Thai): Oh! Lek! I am good, thank you. How about you?
Speaker A (NS): I'm doing well, thanks. It's been a long time since we talked. How is everything with the family?
Speaker B (L2, Thai): Family is very okay. My mom, she has new hobby... um, plant... planting tree.
Speaker A (NS): Oh, gardening! That sounds nice. Does she have a big garden now?
Speaker B (L2, Thai): Yes, a lot of plant. Many flower, very beautiful. And your family?
Speaker A (NS): They're good too, thanks for asking. My dad is enjoying painting these days.
Speaker B (L2, Thai): Painting? That is nice. Like... picture, um, with... colors?
Speaker A (NS): Exactly! He loves painting landscapes and nature scenes. So, are you still working at the same company?
Speaker B (L2, Thai): Yes, I still work there. It is... um, quite busy. Many project now.

A.5.3 Mandarin Example

Dialogue

Speaker A (NS): Hi, Xiao Li! How are your preparations going for the upcoming exam?
Speaker B (L2, Mandarin): Oh, hello! Uh, I am study very hard, but some part still difficult for me.
Speaker A (NS): I understand. Which parts are you finding difficult?
Speaker B (L2, Mandarin): Um, the... how to say... the part about economy numbers?
Speaker A (NS): Ah, you mean the statistics section?
Speaker B (L2, Mandarin): Yes, statistics. It is, uh, hard to remember all the formulas.
Speaker A (NS): I can help you with that! We can go over some key formulas together if you'd like.
Speaker B (L2, Mandarin): Really? That would be very nice! Thank you.
Speaker A (NS): No problem! Have you tried using flashcards to memorize them?
Speaker B (L2, Mandarin): Flashcard? Oh, I think I know. The small card with... word and answer?
Speaker A (NS): Exactly! They can be very useful for memorizing details quickly.

Speaker B (L2, Mandarin): I will try that. Um, you think study together good idea?
Speaker A (NS): Yes, definitely! Studying together can make it easier to understand tough concepts.

A.5.4 Japanese Example

Dialogue

Speaker A (NS): Hi, everyone. So, we're here to plan our group vacation. Any thoughts on where we should go?
Speaker B (L2, Japanese): Ah, yes. I think... maybe Kyoto is nice? It have many... uh, historical... um, places.
Speaker A (NS): That's a great idea! Kyoto is full of history and culture. What do you think we could do there?
Speaker B (L2, Japanese): We can see... um, temples. And maybe, ah, try... uh, Kyoto food? How you say... kyou-ryouri?
Speaker A (NS): Yes, Kyoto cuisine. That sounds delicious! Are there any other activities you'd like to do?
Speaker B (L2, Japanese): I want to... um, wear kimono. It is very... uh, traditional, right?
Speaker A (NS): Absolutely, wearing a kimono in Kyoto would be quite the experience. How do you feel about transportation? Should we drive or take the train?
Speaker B (L2, Japanese): Um, I think... shinkansen is very fast. But maybe... too expensive?
Speaker A (NS): The bullet train is quick but can be pricey. We could consider an express bus?
Speaker B (L2, Japanese): Oh, yes! Express bus is good idea. Maybe, uh, cheaper?
Speaker A (NS): Definitely could be more budget-friendly. When do you think we should go?
Speaker B (L2, Japanese): Um, maybe next month? I check my... schedule.

A.5.5 Korean Example

Dialogue

Speaker A (NS): Hi Minji, are you prepared for the exam next week?
Speaker B (L2, Korean): Oh, hello! Um, yes, I think so... but not very sure. It's difficult, yes?
Speaker A (NS): It can be challenging. Which part do you find the hardest?
Speaker B (L2, Korean): The, um, history part. Too many dates and name to remember.
Speaker A (NS): I understand. Have you tried making flashcards? They can help with memorization.
Speaker B (L2, Korean): Flashcard? Ah, yes! I make some, but still... um, need more practice.
Speaker A (NS): That sounds like a good start! Maybe we can study together?
Speaker B (L2, Korean): Oh, that would be great! When... um, when can we meet?
Speaker A (NS): How about this weekend? Saturday or Sunday work for you?
Speaker B (L2, Korean): Saturday is good. Um, maybe afternoon?
Speaker A (NS): Perfect! We can meet at the library around 2 p.m.?
Speaker B (L2, Korean): Yes, yes, 2 p.m. good. I will bring flashcards.
Speaker A (NS): Awesome. We can quiz each other and go over the main topics.
Speaker B (L2, Korean): Sounds nice. Thank you for help!

A.5.6 Urdu Example

Dialogue

Speaker A (NS): Oh, you work with computers? That's interesting! What do you do exactly?
Speaker B (L2, Urdu): Yes, um, I am doing software develop... developing. I make, uh, programs and apps.
Speaker A (NS): Software development, that sounds exciting! How long have you been doing that?
Speaker B (L2, Urdu): It has been, um, about three year now. I start after university.
Speaker A (NS): That's quite some time. Which university did you attend?
Speaker B (L2, Urdu): I study at Lahore University. It is good for science, technology... um, these things.
Speaker A (NS): Lahore University is well-known. What was your major?
Speaker B (L2, Urdu): My major was computer science. I always like computers.
Speaker A (NS): That's great! What inspired you to get into computer science?
Speaker B (L2, Urdu): Um, I like solve problems. And, uh, computers are very... um, how to say... powerful for this?

A.5.7 Cantonese Example

Dialogue

Speaker A (NS): Hi there! Can I help you find something today?
Speaker B (L2, Cantonese): Um, yes, please. I looking for... uh, how to say... cleaning thing?
Speaker A (NS): Do you mean cleaning supplies, like a mop or detergent?
Speaker B (L2, Cantonese): Yes, yes! Detergent. I need for washing clothes.
Speaker A (NS): Alright, the laundry detergent is in aisle six. Do you need any help choosing a brand?
Speaker B (L2, Cantonese): Ah, too many brand. Can you recommend? Which is good?
Speaker A (NS): Of course! Tide is quite popular and cleans well. Do you have a preference for liquid or powder?
Speaker B (L2, Cantonese): Uh, I think maybe liquid. Easier to use, I think.
Speaker A (NS): Great choice! Is there anything else you need today?
Speaker B (L2, Cantonese): Um, yes, maybe... how you say... remove spot? On clothes?
Speaker A (NS): Spot remover or stain remover. It's where the laundry detergent is too.
Speaker B (L2, Cantonese): Okay, thank you. I will buy it. Um, question... do you have bags that... um, recycle?
Speaker A (NS): Yes, we have reusable bags at the checkout area. They're a great option for the environment.
Speaker B (L2, Cantonese): Ah, good! I will buy that also. Thank you so much.

A.5.8 Malay Example

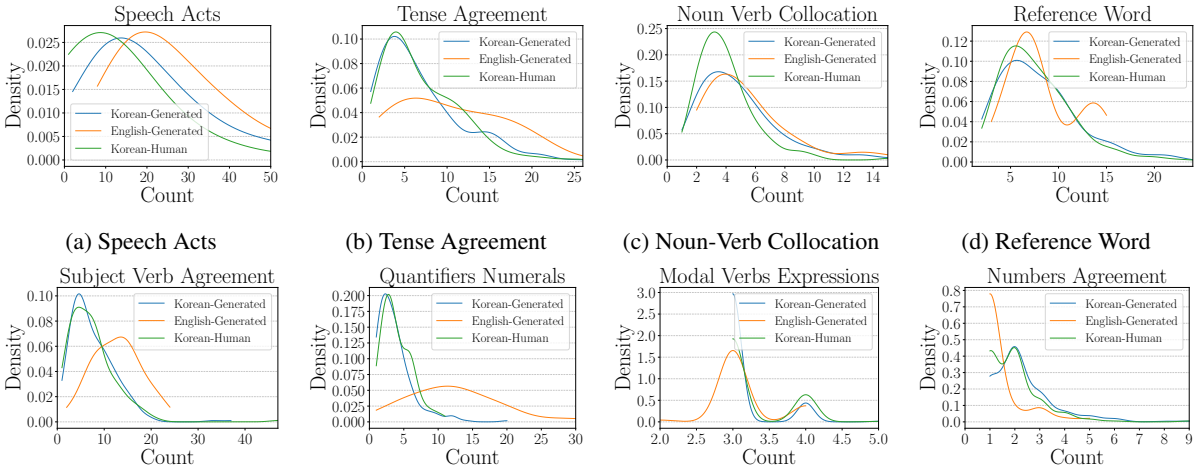
Dialogue

Speaker A (NS): Hi there! I heard Malaysia has a lot of interesting festivals. Can you tell me about one of them?
Speaker B (L2, Malay): Oh, yes! We have many. Um, one famous is Hari Raya Aidilfitri.
Speaker A (NS): Sounds interesting! Can you explain what happens during it?
Speaker B (L2, Malay): Yes, sure. It is, uh... celebration after fasting month, Ramadan.
Speaker A (NS): Oh, right. So, what do people usually do during Hari Raya?
Speaker B (L2, Malay): We, uh, visit family. Have... big meals. Um, special food like rendang,

ketupat.
Speaker A (NS): That sounds delicious! Is there anything else that's part of the celebration?
Speaker B (L2, Malay): Yes, we also... um, give... how to say... small money packets to children.
Speaker A (NS): Ah, like gifts?
Speaker B (L2, Malay): Yes, but... um, we call it "duit raya."

For Other languages generated data, please refer to https://anonymous.4open.science/r/LLMPirorknowledge-017A/README.md for each dialogues.

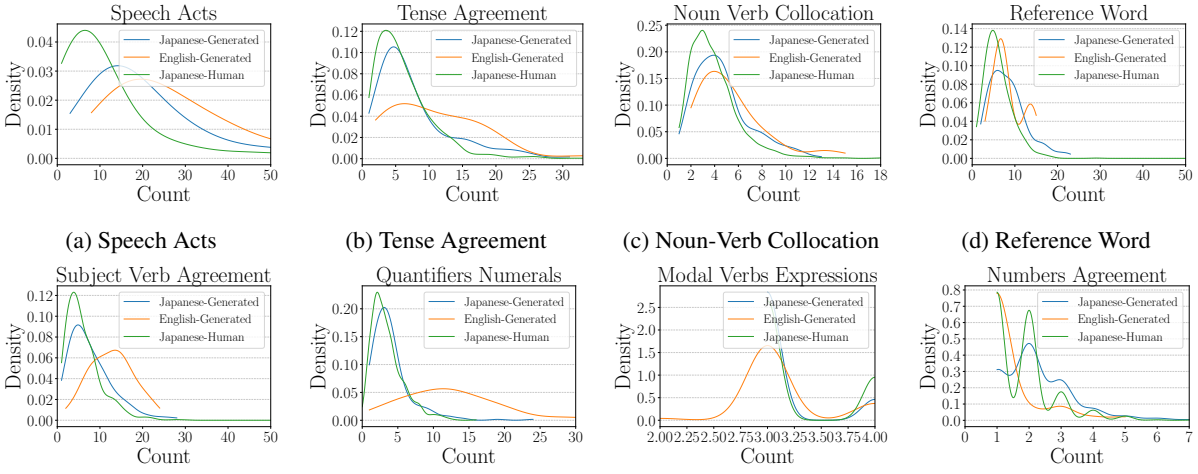
A.6 L2 Density Results



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

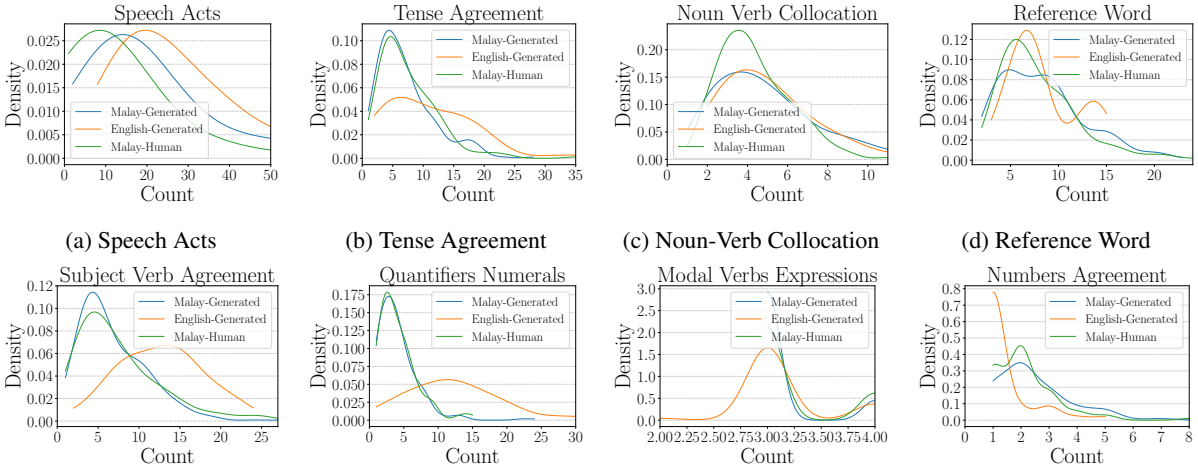
Figure 5: Full density results for L2 generation dialogue via Korean L1s



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

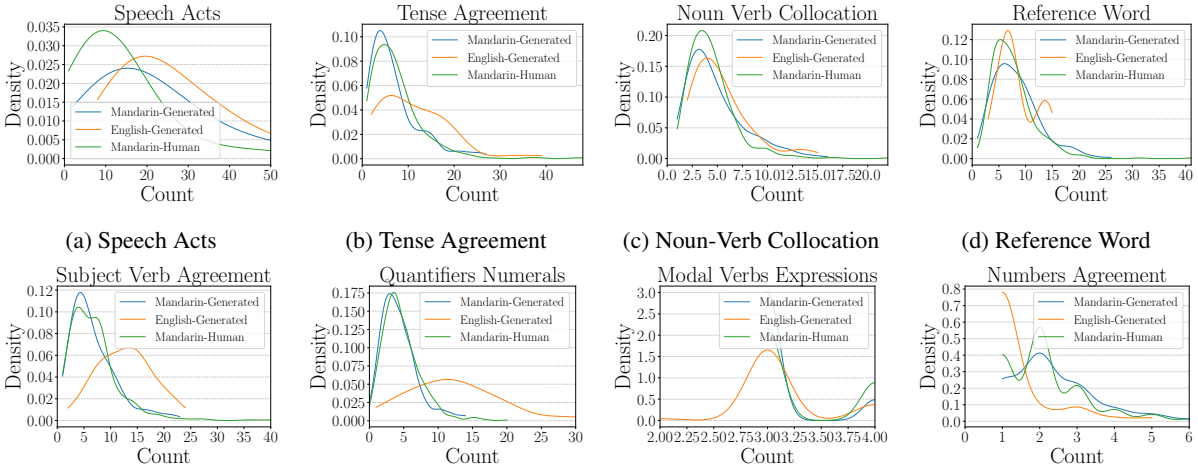
Figure 6: Full density results for L2 generation dialogue via Japanese L1s



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

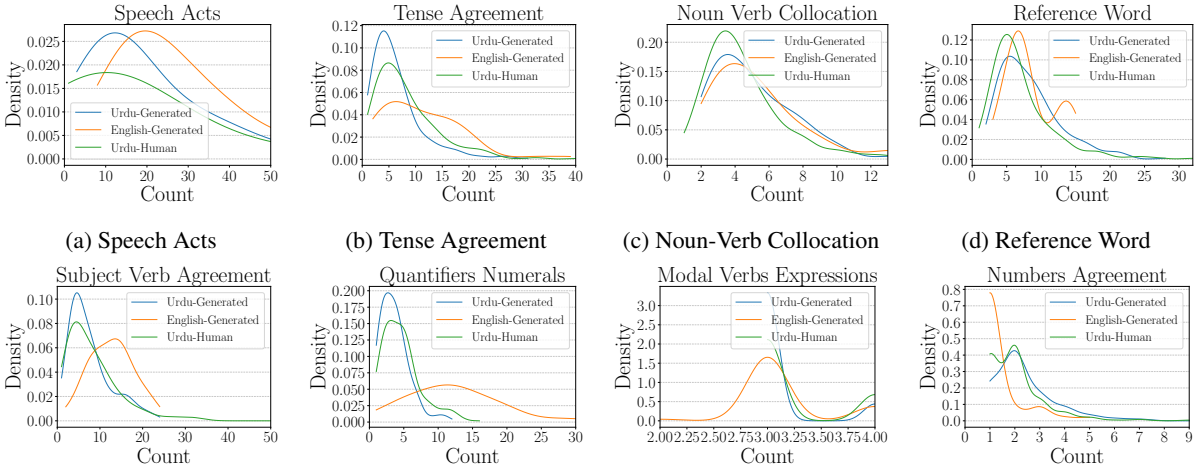
Figure 7: Full density results for L2 generation dialogue via Malay L1s



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

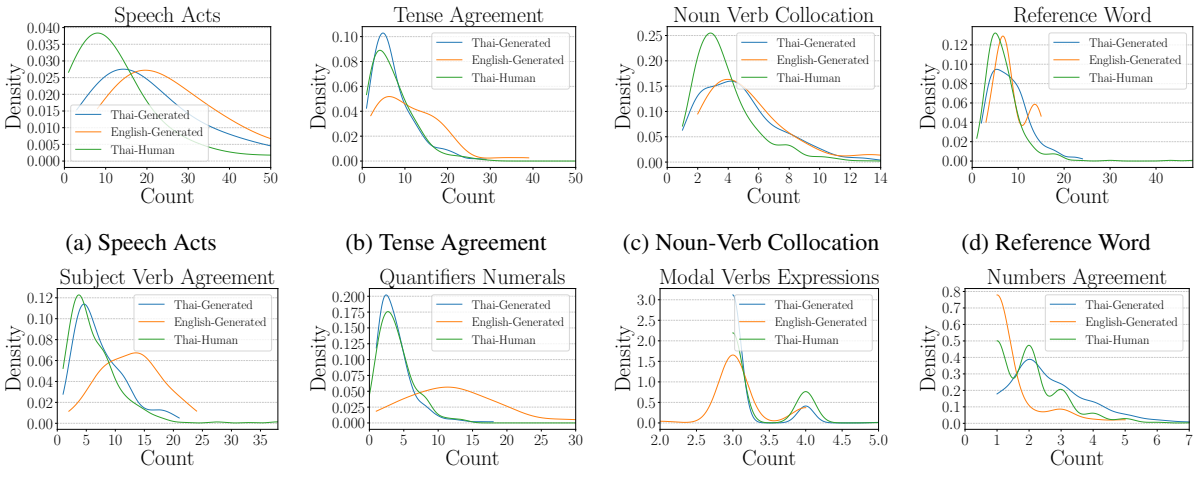
Figure 8: Full density results for L2 generation dialogue via Mandarin L1s



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

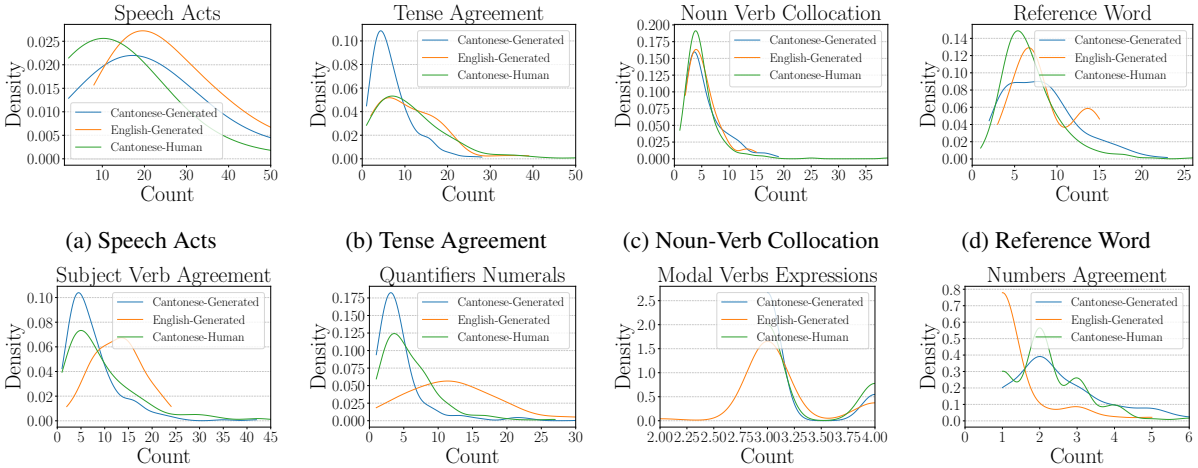
Figure 9: Full density results for L2 generation dialogue via Urdu L1s



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

Figure 10: Full density results for L2 generation dialogue via Thai L1s



(a) Speech Acts (b) Tense Agreement (c) Noun-Verb Collocation (d) Reference Word

(e) Subject Verb Agreement (f) Quantifiers Numerals (g) Modal Verbs Expressions (h) Numbers Agreement

Figure 11: Full density results for L2 generation dialogue via Cantonese L1s