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

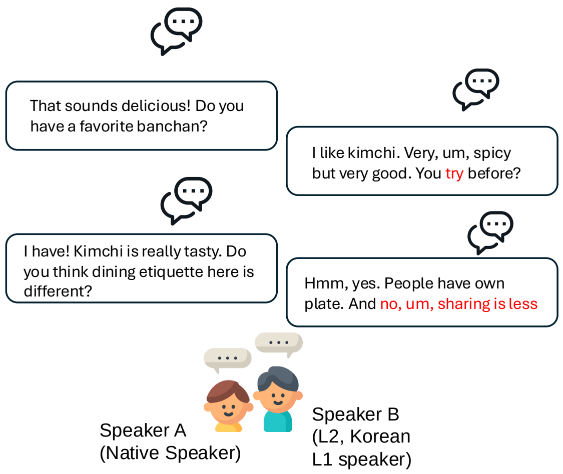
Rena Gao[♡∗], Xuetong Wu[♡∗], Tatsuki Kuribayashi[◇], Mingrui Ye[♣], Siya Qi[♣]
Carsten Roever[♡], Yuanxing Liu[♠], Zheng Yuan[△], Jey Han Lau[♡]
[♡]The University of Melbourne [♣]King's College London
[◇]MBZUAI [♠]Harbin Institute of Technology [△]The University of Sheffield
{rena.gao,carsten}@unimelb.edu.au, {wfyitf,jeyhan.lau}@gmail.com
{tatsuki.kuribayashi}@mbzuai.ac.ae, {yxliu}@ir.hit.edu.cn
{mingrui.ye, siya.qi}@kcl.ac.uk, {zheng.yuan1}@sheffield.ac.uk

Abstract

This study evaluates Large Language Models' (LLMs) ability to simulate non-native English use as observed in human second language (L2) learners interfered with by their native first language (L1). In dialogue-based interviews, we prompt LLMs to mimic L2 English learners with specific L1s (e.g., Japanese, Thai, Urdu) across seven languages, comparing their outputs to real L2 learner data. Our analysis examines L1-driven linguistic biases, such as reference word usage and avoidance behaviors, using information-theoretic and distributional density measures. Results show that modern LLMs (e.g., Qwen2.5, LLAMA3, DeepseekV3, GPT-4o) replicate L1-dependent patterns observed in human L2 data, with distinct influences from various languages (e.g., Japanese, Korean, and Mandarin significantly affect tense agreement, while Urdu influences noun-verb collocations). Our results reveal LLMs' potential for L2 dialogue simulation and evaluation for future educational applications.

1 Introduction

The widespread use of Large Language Models (LLMs) in language communication and education 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 Figure 1. Such an L2-speaker simulation will be helpful for, e.g., predicting L2 speakers' biases in a pedagogical situation (Settles et al., 2018), developing an L2-speaking agent (Timpe-Laughlin et al., 2022), virtual language-learning applications (Bibauw et al., 2022), emulating diverse agents to simulate the diversity of L2 speakers in real world (Ge et al., 2024), and potentially assessing LLMs' cognitive plausibility from a cross-lingual perspective (Aoyama and Schneider, 2024). However, the ability of LLMs to accurately replicate linguistic patterns of non-native speakers and the systematic influence of native language (L1) knowledge on L2 generation remain underexplored (Chen et al., 2024), especially in the dialogue domain (Veivo and Mutta, 2025) and in non-native contexts (Fincham and Alvarez, 2024). This leads us to ask: **Can LLMs effectively mimic human-like dialogue performance in L2 contexts?**



That sounds delicious! Do you have a favorite banchan?

I like kimchi. Very, um, spicy but very good. You try before?

I have! Kimchi is really tasty. Do you think dining etiquette here is different?

Hmm, yes. People have own plate. And no, um, sharing is less

Figure 1: Examples of L2 English dialogue from human speakers, which can generally be biased by their native L1 knowledge, e.g., with particular errors.

To address this question, it is crucial to understand the role of native linguistic knowledge in areas such as language education and cross-lingual communication (Levenston, 1971; Schachter, 1974; Kleinmann, 1977; Brooke and Hirst, 2012). L2 speakers' use of English is often influenced by their L1 traits (Takahashi, 2024), especially for Asian native speakers whose first language (L1) is typologically different from English (Pan, 2024); for example, Mandarin has a looser grammatical tense system and indirect expression of counterfactuals ("if x had... would have y...") (Bloom, 1984), compared to the English

[∗]Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4355–4379
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

grammar. Such differences can result in distinct L1-L2 linguistic patterns (Bailey et al., 2021), including grammatical constructions and lexical choices in spoken dialogues (Levenston, 1971; Schachter, 1974; Kleinmann, 1977; Downey et al., 2023). To investigate whether LLMs simulate similar patterns, we propose an information-theoretic evaluation framework grounded in multiple linguistic perspectives: key features from grammatical/semantic accuracy, fluency, discourse-level cohesion, and pragmatics that shape the communicative outcome (Schwandt, 2001; Sun et al., 2021; Gao and Wang; Santiago-Garabieta et al., 2023). By analyzing these aspects, we explore how accurately various LLMs simulate L2-like dialogues that align with human linguistic behaviors across different L1 backgrounds.[1] For benchmark data, we utilize the ICNALE dataset (Ishikawa, 2023), which includes recordings from 435 human L2 speakers with 18 L1s and manual transcripts comprising approximately 1.6M tokens. An information-theoretic analysis is applied to evaluate LLMs' L1-dependent biases by comparing LLM-generated dialogues (with a prompt to simulate L2 English with a specific L1) with human counterparts. To address challenges in reflecting L1 background in LLMs' L2 generation, as an initial foray, we employ native knowledge injection prompting (e.g., `Simulate L2 English dialogue spoken by Japanese based on provided Japanese native linguistic knowledge`) (Dong et al., 2022; Santiago-Garabieta et al., 2023; Bibauw et al., 2022).

Through our exploration, we demonstrate that simple L1 prompting has a significant impact on LLM-generated L2 dialogues. For example, Japanese, Korean, and Mandarin L1 influence tense agreement, Thai and Malay L1 affect speech acts, while Urdu L1 impacts noun-verb collocations. Our information-theoretic evaluation quantifies their human-like output, which is further supported by qualitative analysis. Ultimately, our study paves the way for using LLMs to simulate human L2 dialogues. Summarizing our contributions:

- 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 LLMs to simulate L2 communications.

2 Related Work

2.1 Bilingual Knowledge for LLMs

L1 interference in humans and LMs Native language profoundly influences L2 language use in humans (Levenston, 1971; Schachter, 1974; Kleinmann, 1977; Brooke and Hirst, 2013). This *language interference* effect biases, for example, the syntactic constructions (Felker et al., 2021) and discourse flows (Bailey et al., 2021) in L2, and the dialogue patterns are not an exception (Veivo and Mutta, 2025). When it comes to neural LMs, the cross-lingual transferability of LMs and their human-likeness has also gained attention, but prior studies have exclusively focused on sentence-level evaluations (Oba et al., 2023; Yadavalli et al., 2023; Elshin et al., 2024). Such perspectives can easily be extended to the dialogue level, involving discourse-level cohesion/coherence and L1-dependent, nuanced differences in dialogue strategy (Abe and Roever, 2019; Gao et al., 2024). Moreover, LLMs are now deployed to generate dialogue (e.g., chat interactions); evaluating their ability in a dialogue scenario generally aligns with their practical usage (Jin et al., 2024; Veivo and Mutta, 2025). That said, our scope is limited to just simulating L2-like language use in a behavioral sense; LM's cognitive plausibility as an L2 learner, while interesting and related, is beyond of the scope of this paper.

Bilingual Knowledge in LLMs 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

[1] All code and dataset can be found `https://github.com/RenaGao/LLMPirorknowledge`

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 to interact (Gan et al., 2024). There are case studies that optimize bilingual knowledge integration 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 Evaluating LLMs in L2-Learning Contexts

Existing studies have explored the use of LLMs in online platforms (Manoharan and Nagar, 2021), personalized language tutoring (Mejeh and Rehm, 2024), and L2 chatbots (Yigci et al., 2024), and their quality is often evaluated by human judgments. Some works proposed automated evaluation tools for L2 interactions (Gao et al., 2025a) and language practice (Huzaifah et al., 2024), yet LLMs' performance of generating non-native-like language in these settings remains underexplored.

Developing effective L2-like dialogue generation systems also requires a robust evaluation framework that can capture and evaluate linguistic knowledge transfer from L1 to L2, particularly for Asian L1 speakers with distinct syntactic structures from English, from their output (Sung et al., 2024; Gao et al., 2025b). The evaluation should incorporate cross-linguistic benchmarking (especially if one explores multiple combinations of L1 and L2) and an accurate error analysis to identify L1-L2-specific grammatical patterns and challenges (Kobayashi et al., 2024). Systematic analysis of their produced errors hopefully provides insights into LLMs' bilingual grammatical understanding and representation, ensuring they not only identify L2-like biases but also can simulate the use of language like L2 speakers, potentially enhancing real-world applications (Cong, 2025; Gao et al., 2024; Singh et al., 2024; Poole-Dayan et al., 2024).

3 Evaluation Metrics

3.1 Evaluation Framework

To assess whether LLMs can accurately simulate L2 dialogues, we target eight linguistic constructs to evaluate their L2 English simulation ability, motivated by L1–L2 interference research (Jackson et al., 2018; Taguchi and Roever, 2020; Millière, 2024; Gao et al., 2025a). 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 in dialogues, reflecting cultural and linguistic nuances often influenced by L1 conversations. 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 exhibited by 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 simulate the real L2-English speakers' patterns.

Theoretical Introduction We propose an information-theoretic framework to explore how a person's first language influences their use of a second language. We use a random variable Y to represent a specific linguistic phenomenon, as shown in Table 1, and the distribution $p(Y)$ describes how frequently this phenomenon occurs.[2] In the case of L1 English acquisition, the English language exposures 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 well

[2] Thus, our focus is on the avoidance behavior of L2 speakers (Levenston, 1971; Schachter, 1974; Kleinmann, 1977), and analyzing the correctness of the phenomenon is left to be our future work.

Categories	Features	Definition	Example	Examples in Prompt
Grammatical Accuracy	Number Agreement (in noun phrase)	Adjectives/determiners and nouns must agree in grammatical number (sometimes involves grammatical gender, e.g., "la/las" and "el/los" in Spanish).	*[100] cars*	*"The big cars are red."...*
	Tense Agreement	The verb tense (i.e., past, present, future) must align with temporal expressions.	*I [did] a task [yesterday].*	*"He has finished his homework."...*
	Subject-Verb Agreement	The verb form must agree with the subject's person and number.	*[She] [is] amazing.*	*"They are playing football."...*
Semantic Accuracy	Modal Verbs and Expressions	The use of modal verbs that indicate likelihood, ability, permission, or obligation.	*She [might] come to the meeting.*	*"You should complete the project soon."...*
	Quantifiers and Numerals	The use of numerical expressions or those related to the amounts, such as quantifiers.	*Some, many, a few*	*"There are ten apples on the table."...*
Fluency	Noun-Verb Collocations	Common collocations that enhance sentence fluency.	*[Drive] a [car], [Do] a [test]*	*"He drives a car every day."...*
Cohesion	Reference Word	The use of linguistic devices referring to entities mentioned earlier (anaphora) or later (cataphora).	*She, her, him, he*	*"She went home early."...*
Pragmatics	Speech Acts	Utterances that serve special functions, such as assertions, questions, requests, or 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

an English L1 learner can infer Y from D. Now, extending this to L2 English acquisition/learning, we define another random variable X to represent linguistic properties for L1 **human** native speakers of that L1 language (non-English). Here, X acts as a **priori knowledge**. According to L2 development theory (Roever and Ikeda 2024; *inter alia.*), L2 learners usually acquire a new language by interacting with native speakers and learning from the linguistic patterns present in the spoken input. With the English (L2) language exposures D and the effect of L1 properties X, the updated posterior for learning Y for L2 becomes $p(Y|D,X) \propto p(D|Y,X)p(Y|X)$ with the assumption that $p(D|Y,X) = p(D|Y)$ as the context D hinges solely on the English linguistic properties Y, which also incorporates 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:[3]

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

where $I(X;Y|D)$ quantifies the mutual information between X and Y given D, and it represents the shared information between English (Y) and the native language (X) conditioned on the context D. Similarly, $I(X';Y|D)$ measures the effectiveness of LLMs in generating L2 English by quantifying the mutual information between the LLM's native language (X') and English (Y) given the context D.[4] We report d_{bi} in the case that LLMs

[3] 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 $p(Y|X')$.

[4] We leave it as future work to align D between humans' and LLMs' L1/L2 learning — an important topic that in-

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'. We will consistently use $\ell(Q) = -\log Q$ in our experiments.

4 Evaluation Framework Annotation Design

We now describe how we annotate the linguistics constructs for dialogues based on the evaluation framework in § 3.1. We used 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. To this end, 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 Asian languages: Bahasa Indonesia, Cantonese, English, Mandarin, Japanese, Korean, Filipino, Javanese, Malay, Pakistani, Pashto, Pashtoo, Punjabi, Urdu, Pushto, Tagalog, Thai, and Uyghur as statistical data in Table 2.[5] 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 (kor), Mandarin (cmn), Japanese (jpn), Cantonese (yue), Thai (tha), Malay (msa), and Urdu (urd).

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

4.1 Automated Annotation with GPT-4o

The initial annotation by GPT-4o (Achiam et al., 2023) with few-shot prompting used four examples per linguistics feature. For *Reference Word*, we selected four sentences from a dialogue, highlighting reference words (e.g., he, she, her) and presenting them in a few-shot format (detailed prompts 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 (JSON).

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 students in NLP, manually reviewed 15% (randomly sampled) of the annotated dialogues. The annotators are required to make a binary judgment whether the span-annotation output is correct. This manual assessment found that the GPT-4o annotations had an accuracy of 84.1%, suggesting that it is a viable approach for automatic annotation.

4.3 Prompt Refinement

Our manual validation revealed consistent errors in constructs like *Noun-Verb Collocations*, where non-collocating tokens (e.g., a little bit *trouble*) were incorrectly annotated by GPT-4o with unnecessary token *trouble*. To address this, we refined the few-shot examples and improved the instructions, and conducted a second human validation. As we see improved accuracy for these constructs, we adopted these updated prompts for all experiments.[6] under the instructions folder.

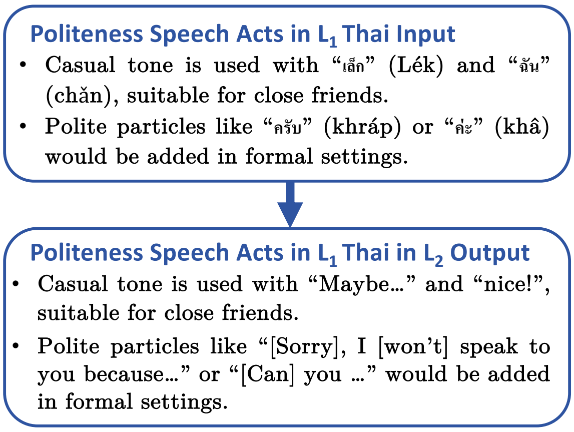
5 L2 Dialogue Simulation

To simulate L2 dialogues using LLMs, we experiment with L1 knowledge injection through prompting: we design an 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 (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 with Thai (Srisuruk, 2011) as shown in Figure 2. In

vestigates the cognitive/developmental plausibility of LM's language learning ability.

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

[6] Full prompts are published in https://github.com/RenaGao/LLMPirorknowledge


Politeness Speech Acts in L_1 Thai Input

- Casual tone is used with "เล็ก" (Lék) and "ฉัน" (chăn), suitable for close friends.
- Polite particles like "ครับ" (khráp) or "ค่ะ" (khâ) would be added in formal settings.

Politeness Speech Acts in L_1 Thai in L_2 Output

- Casual tone is used with "Maybe…" and "nice!", suitable for close friends.
- Polite particles like "[Sorry], I [won't] speak to you because…" or "[Can] you …" would be added in formal settings.

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

addition to the L1 injection prompt, we also provide another set of instructions to generate the L2 dialogues.[7] The 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 interview scenario, where the interviewer (a native English speaker) follows predefined templates based on the ICNALE benchmark datasets. All prompts can be found in Appendix A.2.[8] Table 3 shows an example of LLM-simulated interactions between the English-native interviewer and Japanese-L1 English learner (more examples in Appendix A.5).

Interviewer: Hi, everyone. So, we're here to plan our group vacation. Any thoughts on where we should go?
Japanese-L1: *Ah, yes. I think... maybe Kyoto is nice? It have many... uh, historical... um, places.*
Interviewer: That's a great idea! Kyoto is full of history and culture. What do you think we could do there?
Japanese-L1: *We can see... um, temples. And maybe, ah, try... uh, Kyoto food? How you say... kyou-ryouri?*
Interviewer: Yes, Kyoto cuisine. That sounds delicious! Are there any other activities you'd like to do?
Japanese-L1: *I want to... um, wear kimono. It is very... uh, traditional, right?*
Interviewer: Absolutely, wearing a kimono in Kyoto would be quite the experience. How do you feel about transportation? Should we drive or take the train?
Japanese-L1: *Um, I think... shinkansen is very fast. But maybe... too expensive?*
Interviewer: The bullet train is quick but can be pricey. We could consider an express bus?
Japanese-L1: *Oh, yes! Express bus is good idea. Maybe, uh, cheaper?*
Interviewer: Definitely could be more budget-friendly. When do you think we should go?
Japanese-L1: *Um, maybe next month? I check my... schedule.*

Table 3: Example interactions between LLM-simulated English-native interviewer and Japanese-L1 interviewee (in blue italic). The interviewee parts are evaluated in our experiments.

6 Results and Analysis

We conduct experiments using five large language models (LLMs) for L2 dialogue generation: **LLAMA3-8B, LLAMA3-70B** *(April 2024)*, **Qwen2.5-72B** *(September 2024)*, **DeepSeekV3-685B** *(December 2024)*, and **GPT-4o** *(December 2024)*.

6.1 L1-Specification Impact across LLMs

We present GPT-4o results in Table 4 (full results for all models are in Appendix A.7). GPT-4o generated L2 dialogues exhibit generally consistent and significant improvements across all seven languages after prompting with the L1 information injection prompting, given the decrease (in distance) of d_{bi} (with L1 instruction) from d_{mono} (without L1 instruction), expect for *Quantifiers Numerals*. 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–*Tense Agreement, Number Agreement,* and *Subject-Verb Agreement*, Pragmatics–*Speech Acts*, and *Reference Words*, which play important roles in oral communications (Gao et al., 2025b).

Looking at the results across different LLMs (Appendix A.7), the summary is that performance varies depending on the exact model, but broadly speaking L1 knowledge injection appears to help most models to mimic the L2 dialogue patterns, suggesting our approach is not LLM-dependent. Diving a bit deeper, DeepSeekV3 shows very strong performance similar to GPT-4o's, and LLAMA3-8B performs the worst, although this is perhaps unsurprising since it's the smallest model. This does suggest, however, that model size is an important factor when it comes to the selection of LLM.

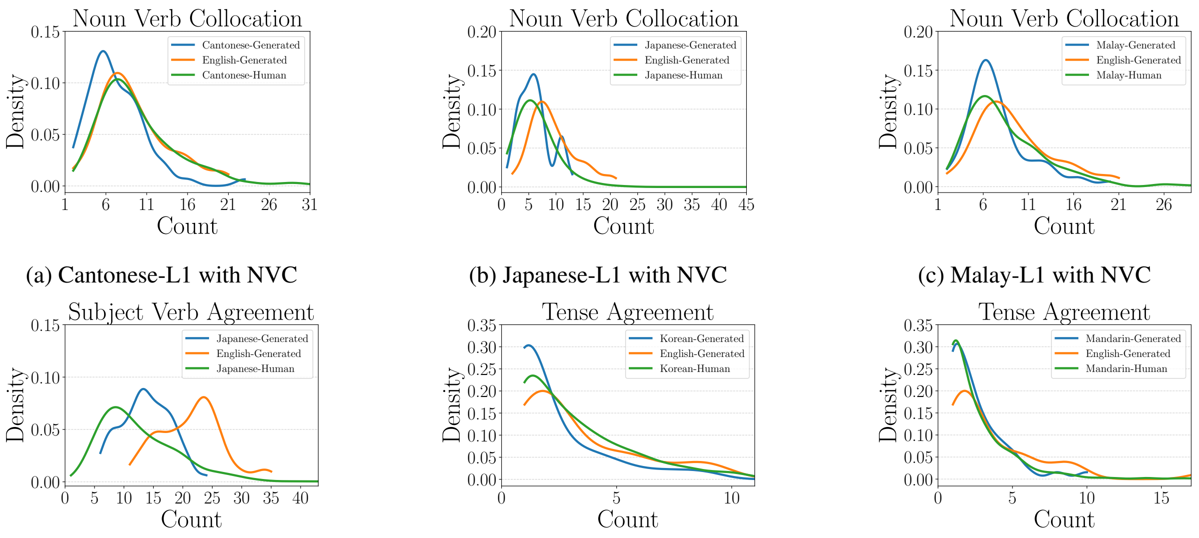
Another consistent result across all LLMs is the

[7] Generation temperature is set to 0 in our experiments.

[8] These example conversations are derived from human L1 dialogues from xDial-Eval (Zhang et al., 2023), a multilingual open-domain dialogue dataset.

		Distribution distance between humans' and LLMs' generated dialogues (↓)							
Lang.	**Condition**	**Number Agreement**	**Tense Agreement**	**Subject-Verb Agreement**	**Modal Verbs Expressions**	**Quantifiers Numerals**	**Noun-Verb Collocation**	**Reference Word**	**Speech Acts**
Cantonese	d_{bi}	0.086	0.021	0.045	0.313	0.367	0.073	0.138	0.373
	d_{mono}	0.231	0.034	0.326	0.158	0.136	0.001	0.449	0.817
Thai	d_{bi}	0.096	0.072	0.311	0.064	0.182	0.050	0.504	0.561
	d_{mono}	0.038	0.257	0.589	0.127	0.024	0.142	0.625	1.099
Japanese	d_{bi}	0.010	0.031	0.160	0.073	0.104	0.014	0.183	1.154
	d_{mono}	0.037	0.305	0.685	0.321	0.090	0.185	0.695	1.894
Korean	d_{bi}	0.043	0.019	0.136	0.049	0.094	0.026	0.321	1.241
	d_{mono}	0.103	0.035	0.394	0.173	0.017	0.144	0.542	2.268
Malay	d_{bi}	0.069	0.156	0.062	0.042	0.113	0.016	0.164	0.771
	d_{mono}	0.109	0.167	0.369	0.082	0.022	0.031	0.438	1.184
Mandarin	d_{bi}	0.030	0.028	0.133	0.023	0.070	0.037	0.261	0.618
	d_{mono}	0.099	0.208	0.455	0.109	0.028	0.091	0.530	1.175
Urdu	d_{bi}	0.041	0.133	0.057	0.091	0.251	0.010	0.205	0.311
	d_{mono}	0.102	0.078	0.291	0.117	0.052	0.035	0.529	0.822

Table 4: The distribution divergences d_{bi} and d_{mono} of GPT-4o generated L2 dialogues for different native languages: Korean (kor), Mandarin (cmn), Japanese (jpn), Cantonese (yue), Thai (tha), Malay (msa), and Urdu (urd) where green indicates d_{bi} is less than d_{mono}, while red indicates the opposite.



(a) Cantonese-L1 with NVC

(b) Japanese-L1 with NVC

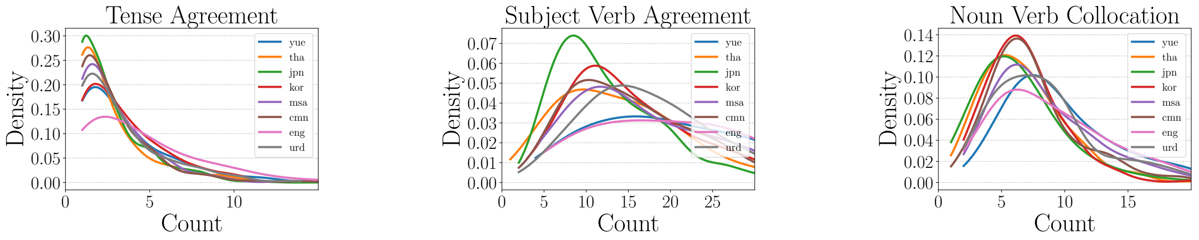
(c) Malay-L1 with NVC

(d) Japanese-L1 with SVA

(e) Korean-L1 with TA

(f) Mandarin-L1 with TA

Figure 3: Density plots for GPT-4o-simulated L2 dialogue with different L1 trairs. NVC represents noun and verb collocations, TA for tense agreement, and NA for number agreement. The blue lines (L2-Generated), orange lines (English-Generated), and green lines (L2-Humans) correspond to LLM-generated dialogue with L1 prompting, that without L1 knowledge injection prompting, and respective human dialogue.



(a) TA of Real L1

(b) SVA of Real L1

(c) NVC of Real L1

Figure 4: Density plots for human-baseline dialogues with different L1s. NVC represents *Noun and Verb Collocations*, TA for *Tense Agreement*, and NA for *Number Agreement*.

		Human–LLM dialogue distribution distance (↓)				
Lang.	**Cond.**	**DeepSeek V3**	**QWEN 72B**	**LLaMA 70B**	**LLaMA 8B**	**GPT-4o**
Cantonese	d_{bi}	0.102	0.137	0.376	0.294	0.367
	d_{mono}	0.051	0.026	0.046	0.086	0.136
Thai	d_{bi}	0.057	0.061	0.093	0.064	0.182
	d_{mono}	0.037	0.098	0.025	0.019	0.024
Japanese	d_{bi}	0.060	0.094	0.060	0.025	0.104
	d_{mono}	0.181	0.286	0.156	0.113	0.090
Korean	d_{bi}	0.021	0.017	0.262	0.085	0.094
	d_{mono}	0.070	0.117	0.031	0.004	0.017
Malay	d_{bi}	0.078	0.060	0.174	0.149	0.113
	d_{mono}	0.034	0.103	0.017	0.015	0.022
Mandarin	d_{bi}	0.108	0.022	0.103	0.085	0.070
	d_{mono}	0.037	0.064	0.008	0.010	0.028
Urdu	d_{bi}	0.072	0.140	0.313	0.054	0.251
	d_{mono}	0.071	0.075	0.025	0.004	0.052

Table 5: The distribution divergences d_{bi} and d_{mono} with different models for *Quantifier Numerals*.

poor performance for *Quantifiers and Numerals*; see Table 5. A likely explanation is that many LLMs prioritize natural, concise, conversational English, where quantifiers and numerals are often omitted when the meaning is clear (e.g., "We bought apples" instead of "We bought *some* apples"). Qwen2.5-72B appears to do better than other LLMs, and it is perhaps due to its stronger exposure to Mandarin (Yang et al., 2024) where quantifiers and numerals are more structurally important. This linguistic influence likely helps the model retain explicit quantifiers and numerals, even in contexts where native English speakers might naturally omit them during conversation.

6.2 L2 Generation Power via L1 Distance

The density comparison results in Figure 3 show a consistent yet subtle influence of a speaker's L1 on GPT-4o's performance in generating L2 English dialogues. We focus on GPT-4o here due to its strong performance in mimicking L1 language patterns (Appendix 3 presents density results for all LLMs).

For speakers of Cantonese, Japanese, and Malay (Figure 3a, 3b, 3c) — languages that share certain structural similarities with English such as *Noun and Verb Collocations* — the generated dialogues generally resemble human-like patterns (except for the Cantonese with Noun Verb Collocation; NVC). This alignment is supported by the L1 distance density results, suggesting that in most cases, LLMs successfully transfer these L1 linguistic features into the L2 English dialogues, particularly when grammatical agreement plays a key role in conveying meaning and maintaining semantic clarity.

To further review the general impact of L1 knowledge within humans, we put together density results for different L1s given a construct in Figure 4. The *eng* line (pink) serves as the baseline, representing native English speakers. We find that L1s with more distant grammatical structures from English — such as those with an 'SOV' (Subject-Object-Verb) word order — tend to induce greater deviations in the generated dialogues. For instance, Japanese and Korean, which follow an SOV structure, exhibit a greater divergence from the English baseline in *Subject-Verb Agreement* (Figure 4b). This is also reflected by the decrease from d_{mono} to d_{bi} in Table 4. In the case of *Tense Agreement* (Figure 4a), all L1s show relatively similar distributions (fewer numbers of agreement than native English speakers), as observed in the density patterns for Korean and Mandarin LLM-generated dialogues, despite their typological differences.

6.3 Qualitative Analysis: LLM L2 Human-like Dialogue Generation

Beyond the statistical analysis, we qualitatively examined 30 LLM-generated L2 dialogues per language across five models, providing an in-depth analysis of L1-specific patterns and LLM-L2 generated traits where the models' generation diverged from target-language norms in ways that go beyond typical L2 transfer patterns. Below, we summarize key features observed: DeepseekV3 reproduces the characteristic omission errors of numerals and quantifiers found among some Asian native L2 learners. An example for Mandarin L1 is: "We make some food. Like... sandwich, fruit...

how to say... drink?" This kind of omission often arises from L1 transfer, where speakers rely on L1 structures or habitual omission of certain function words (Macuch Silva et al., 2024). GPT-4o exhibits word order, agreement, and collocation traits often observed in Urdu-speaking learners following the Subject–object–verb (SOV) structure (Saleem et al., 2021). For instance, “Lahore University I study” and “It has been about three year now.” Additionally, when generating outputs for Thai speakers, GPT-4o captures instances of politeness mismatches by occasionally omitting formal markers, as shown in “Yes, a lot of plant. Many flower, very beautiful.” These intriguing patterns may reflect how learners from such L1 backgrounds may transfer default word ordering or honorific usage from their native language to L2 contexts (Chansamrong et al., 2014). LLAMA3-70B replicates the common challenge with speech acts and modal verbs seen among Korean L1 learners, producing examples like “What time you think is good to go?” This often occurs because of essential differences in how Korean grammar encodes modalities compared to target languages (Mott et al., 2024). Qwen2.5-72B demonstrates pragmatic choices that closely resemble human dialogue, though it occasionally displays unusual modal expressions across languages, such as “Yes, I try. But my picture not very good.” for Mandarin L1. These reflections of L1-driven structures and lexical choices show how underlying knowledge of a first language can shape second language productions. These interesting influences reflect the extent to *which LLMs* internalize L1-specific structures and how to apply them to L2 production, even when the target language’s norms differ.

7 Conclusions

This study introduces an automated dialogue annotation framework and an information-theoretic method to evaluate LLMs’ performance in simulating L2 English dialogues with L1 influence. Using the ICNALE dataset, we compared LLM outputs with human data, showing that LLMs can capture L1-specific patterns through L1 knowledge injection. The results show strong alignment with human speakers in dialogue cohesion, grammar, and pragmatic use, offering insights to improve multilingual dialogue systems for educational applications.

Limitations

This study has several limitations. First, it relies on the ICNALE dataset as only benchmark, which may limit the generalizability of the results to languages beyond Asian languages. Second, the use of predefined templates for few-shot prompting ensures consistency but may constrain the analysis of *spontaneous L2 language behaviors*, such as chit-chat. Furthermore, the study focuses on linguistics 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 PhD students who were willing to participate in manual checking for this study. We removed all information related to the identification of human volunteer annotators. This study was approved by The University of Melbourne ethics board (Human Ethics Committee LNR 1D), Reference Number 2022-24988-32929-3, and data acquisition and analysis have been taken out to according ethical standards., and data acquisition and analysis has been taken out to according ethical standards.

Acknowledgements

This project is supported by the Australian Research Council Linkage Project (ID: LP210200917).

References

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. 2024. 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.

Daniel Bailey, Ashleigh Southam, and Jamie Costley. 2021. Digital storytelling with chatbots: Mapping l2 participation and perception patterns. *Interactive Technology and Smart Education*, 18(1):85–103.

Serge Bibauw, Thomas François, and Piet Desmet. 2022. Dialogue systems for language learning: Chatbots and beyond. In *The Routledge handbook of second language acquisition and technology*, pages 121–135. Routledge.

Alfred H Bloom. 1984. *Caution—the words you use may affect what you say: A response to Au.* Elsevier Science.

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.

Atchara Chansamrong, Chalong Tubsree, and Prateep Kiratibodee. 2014. Effectiveness of cooperative and blended learning to assist thai esl students in learning grammar. *HRD JOURNAL*, 5(2):105–115.

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.

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

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.

Emily Felker, Mirjam Broersma, and Mirjam Ernestus. 2021. The role of corrective feedback and lexical guidance in perceptual learning of a novel l2 accent in dialogue. *Applied Psycholinguistics*, 42(4):1029–1055.

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. 2024. 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, Carsten Roever, and Jey Han Lau. 2025a. Interaction matters: An evaluation framework for interactive dialogue assessment on English second language conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10977–11012, Abu Dhabi, UAE. Association for Computational Linguistics.

Rena Gao and Menghan Wang. Listenership always matters: active listening ability in l2 business english paired speaking tasks. *International Review of Applied Linguistics in Language Teaching*.

Rena Gao, Jingxuan Wu, Carsten Roever, Xuetong Wu, Jing Wu, Long Lv, and Jey Han Lau. 2025b. An interpretable and crosslingual method for evaluating second-language dialogues. *In Proceedings of NAACL 2025, Albuquerque, New Mexico.*

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.

Howard H Kleinmann. 1977. AVOIDANCE BEHAVIOR IN ADULT SECOND LANGUAGE ACQUISITION[1]. *Lang. Learn.*, 27(1):93–107.

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.

Edwards Levenston. 1971. Over-indulgence and under-representation: Aspects of mother-tongue interference.

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*.

Vinicius Macuch Silva, Alexandra Lorson, Michael Franke, Chris Cummins, and Bodo Winter. 2024. Strategic use of english quantifiers in the reporting of quantitative information. *Discourse Processes*, 61(10):498–523.

Ashok Manoharan and Gourav Nagar. 2021. Maximizing learning trajectories: An investigation into ai-driven natural language processing integration in online educational platforms. *International Research Journal of Modernization in Engineering Technology and Science*, 03:01–10.

Mathias Mejeh and Martin Rehm. 2024. Taking adaptive learning in educational settings to the next level: Leveraging natural language processing for improved personalization. *Educational technology research and development*, pages 1–25.

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*.

Terran Mott, Aaron Fanganello, and Tom Williams. 2024. What a thing to say! which linguistic politeness strategies should robots use in noncompliance interactions? In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 501–510.

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.

Zhaoyi Pan. 2024. Impoliteness in polylogal intercultural communication among asian efl learners. *Intercultural Pragmatics*, 21(2):227–254.

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 and Naoki Ikeda. 2024. The relationship between l2 interactional competence and proficiency. *Applied Linguistics*, 45(4):676–698.

Tahir Saleem, Uzma Unjum, Munawar Iqbal Ahmed, and Ayaz Qadeer. 2021. Social distance and speech behavior: A case of pakistani english speakers' apology responses. *Cogent Arts & Humanities*, 8(1):1890410.

Maite Santiago-Garabieta, Rocío García-Carrión, Harkaitz Zubiri-Esnaola, and Garazi López de Aguileta. 2023. Inclusion of l2 (basque) learners in dialogic literary gatherings in a linguistically diverse context. *Language Teaching Research*, 27(6):1532–1551.

Jacquelyn Schachter. 1974. AN ERROR IN ERROR ANALYSIS[1]. *Lang. Learn.*, 24(2):205–214.

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.

Veronika Timpe-Laughlin, Tetyana Sydorenko, and Phoebe Daurio. 2022. Using spoken dialogue technology for l2 speaking practice: What do teachers think? *Computer Assisted Language Learning*, 35(5-6):1194–1217.

Outi Veivo and Maarit Mutta. 2025. Dialogue breakdowns in robot-assisted l2 learning. *Computer Assisted Language Learning*, 38(1-2):30–51.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Defne Yigci, Merve Eryilmaz, Ail K Yetisen, Savas Tasoglu, and Aydogan Ozcan. 2024. Large language model-based chatbots in higher education. *Advanced Intelligent Systems*, page 2400429.

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 L1 Injection Prompts

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 the context instructions folder in: https://github.com/RenaGao/LLMPirorknowledge

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 sǎo níi wâang mái?)
“Lek, are you free this Saturday?”
Lek: ว่างสิ มีอะไรเหรอ?
(Wâang sì. Mii à-rai rǒe?)
“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

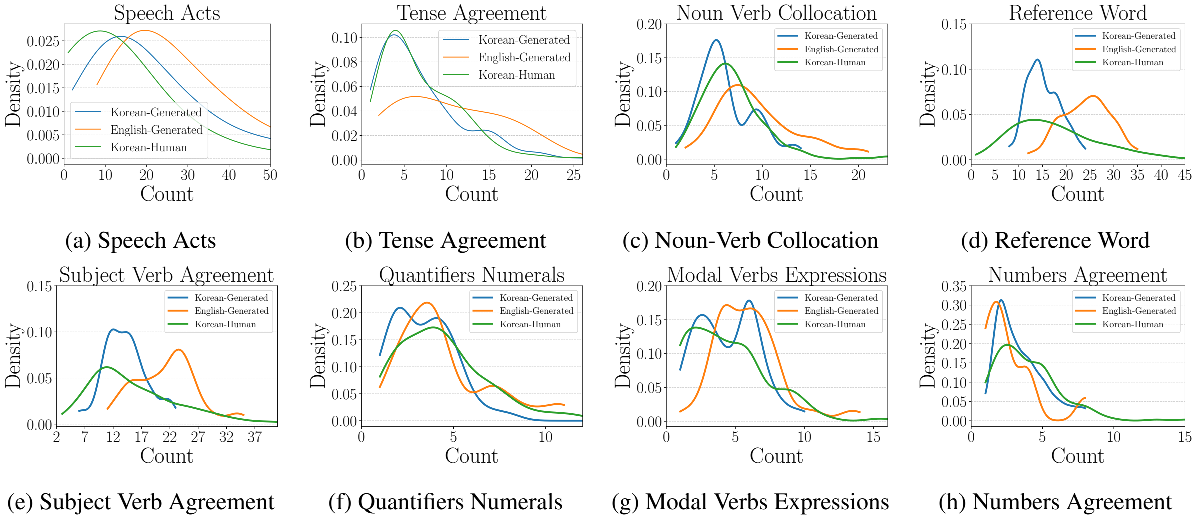
Dialogue

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://github.com/RenaGao/LLMPirorknowledge for each dialogues.

A.6 L2 Density Results for GPT-4o Generations

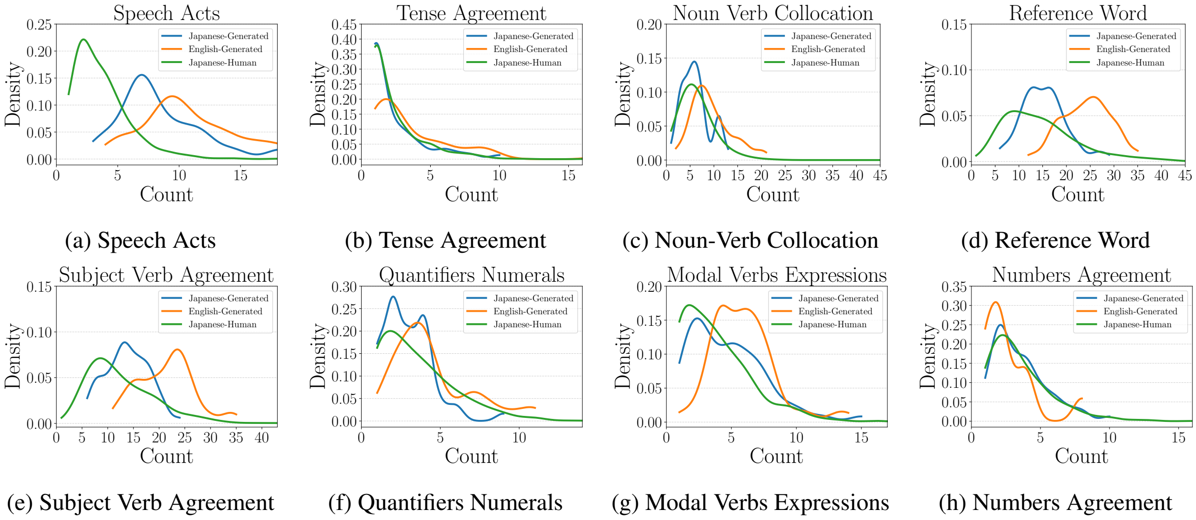
For Other LLMs generated data, please refer to https://github.com/RenaGao/LLMPirorknowledge for each LLM density figure.



(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

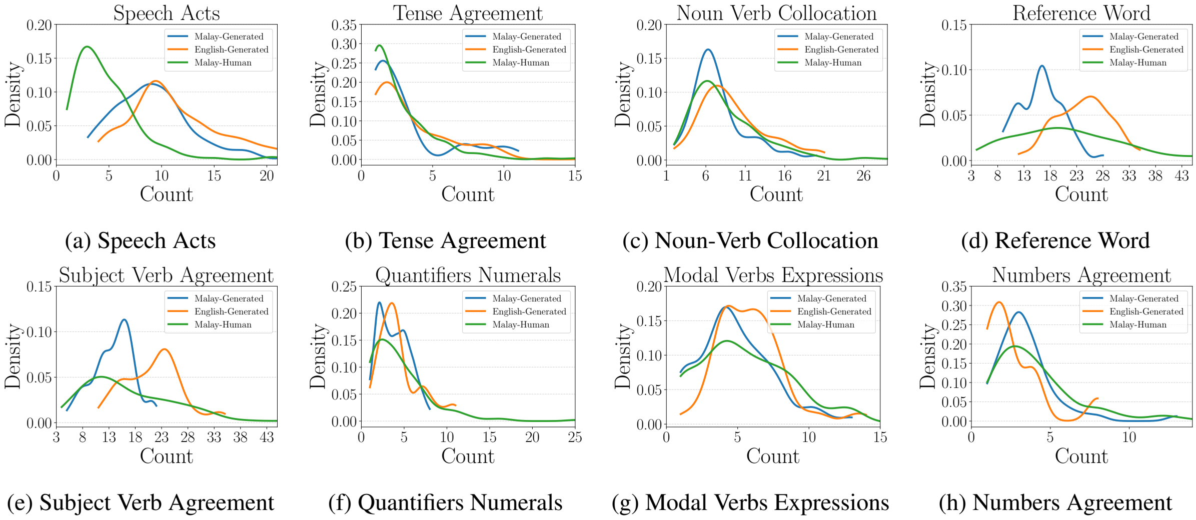
(e) Subject Verb Agreement

(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

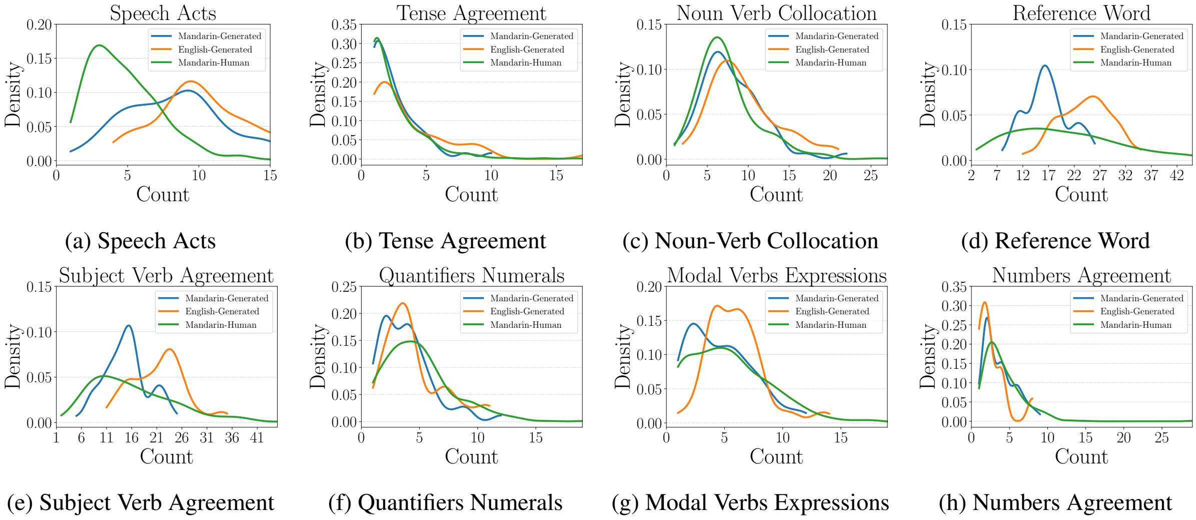
(e) Subject Verb Agreement

(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

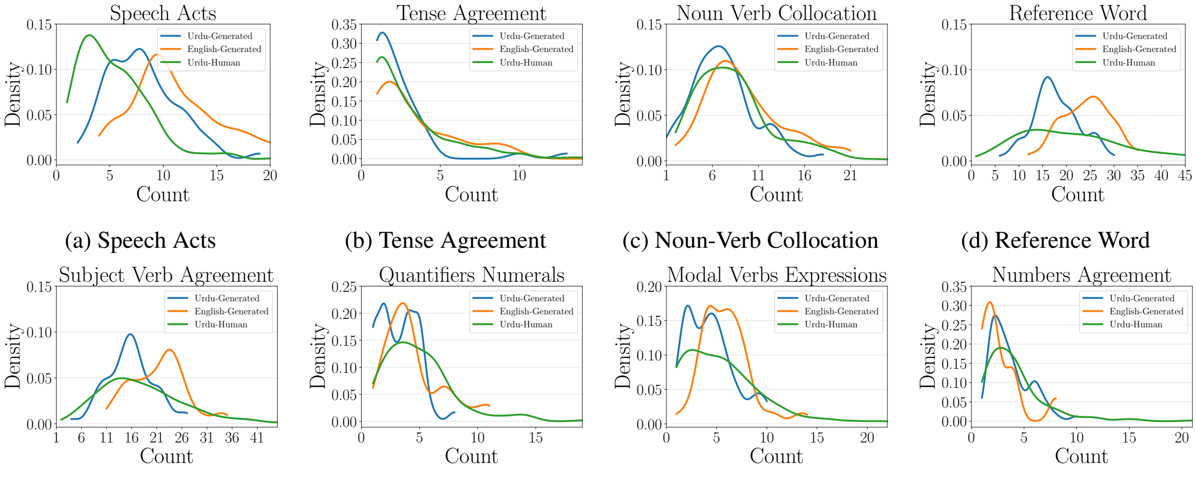
(e) Subject Verb Agreement

(f) Quantifiers Numerals

(g) Modal Verbs Expressions

(h) Numbers Agreement

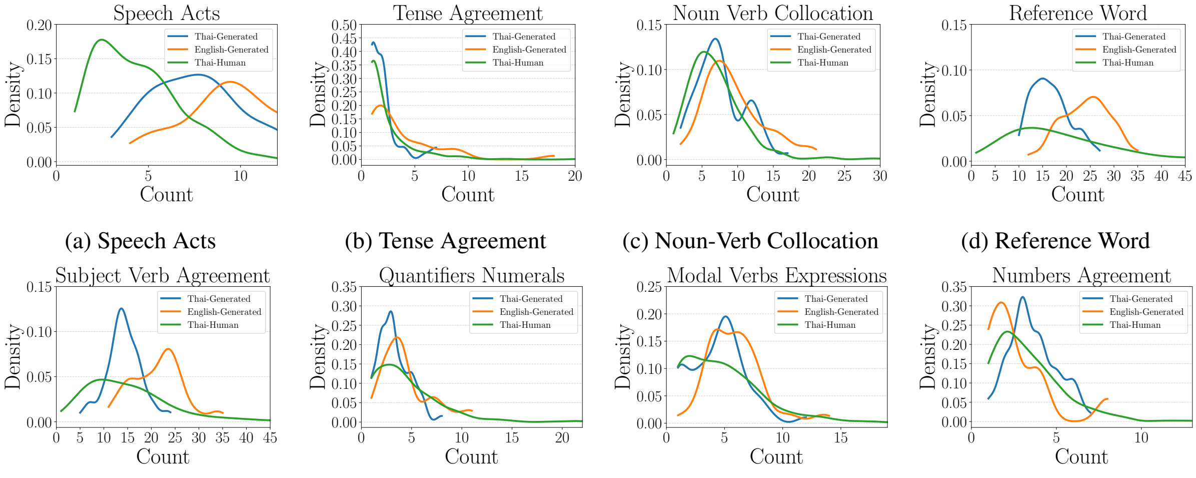
Figure 8: Full density results for L2 generation dialogue via Mandarin L1s



(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

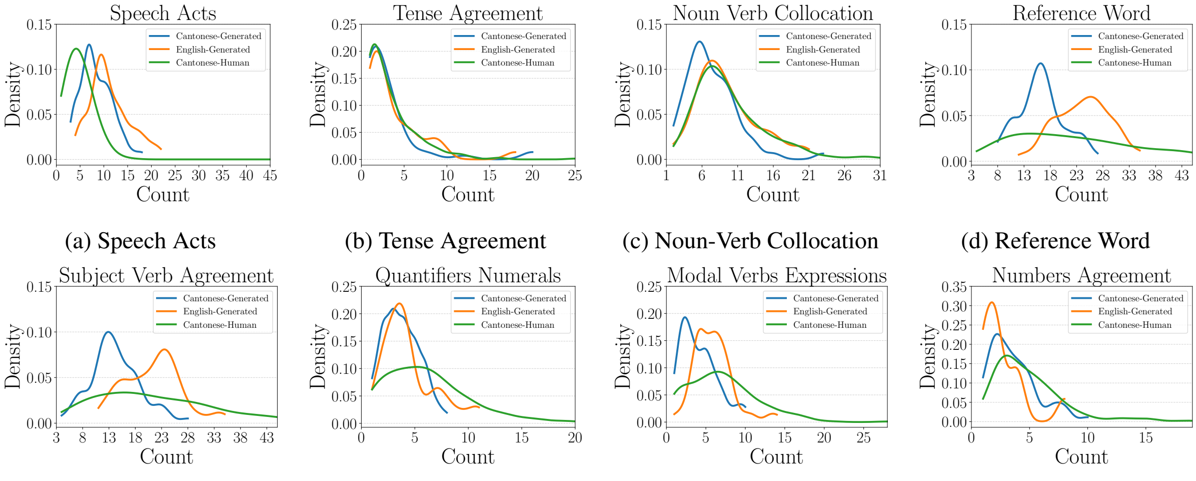
Figure 9: Full density results for L2 generation dialogue via Urdu L1s



(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

A.7 Distance Results under different models

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**
Cantonese	d_{bi}	0.047	0.175	0.034	0.096	0.106	0.030	0.063	0.472
	d_{mono}	0.180	0.159	0.213	0.170	0.057	0.038	0.207	1.109
Thai	d_{bi}	0.053	0.050	0.084	0.117	0.055	0.103	0.038	0.802
	d_{mono}	0.125	0.131	0.439	0.189	0.033	0.217	0.341	1.470
Japanese	d_{bi}	0.055	0.027	0.175	0.142	0.042	0.024	0.266	1.324
	d_{mono}	0.087	0.106	0.510	0.403	0.162	0.252	0.382	2.301
Korean	d_{bi}	0.026	0.027	0.039	0.014	0.023	0.039	0.056	1.611
	d_{mono}	0.141	0.132	0.270	0.188	0.075	0.259	0.234	2.781
Malay	d_{bi}	0.064	0.069	0.024	0.058	0.074	0.020	0.046	0.712
	d_{mono}	0.118	0.084	0.250	0.123	0.031	0.079	0.204	1.523
Mandarin	d_{bi}	0.123	0.024	0.009	0.025	0.107	0.132	0.024	0.666
	d_{mono}	0.099	0.086	0.319	0.135	0.043	0.171	0.267	1.523
Urdu	d_{bi}	0.012	0.038	0.049	0.031	0.080	0.009	0.132	0.399
	d_{mono}	0.116	0.135	0.182	0.155	0.076	0.098	0.266	1.114

Table 6: Distance Results with Deepseek V3 685B

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**
Cantonese	d_{bi}	0.085	0.037	0.040	0.187	0.141	0.005	0.230	0.532
	d_{mono}	0.064	0.029	0.507	0.090	0.026	0.041	0.288	0.879
Thai	d_{bi}	0.294	0.115	0.253	0.053	0.046	0.129	0.247	0.998
	d_{mono}	0.179	0.139	0.801	0.027	0.077	0.264	0.448	1.161
Japanese	d_{bi}	0.211	0.099	0.486	0.061	0.076	0.124	0.559	1.425
	d_{mono}	0.117	0.290	0.928	0.166	0.256	0.306	0.498	1.980
Korean	d_{bi}	0.212	0.023	0.138	0.032	0.020	0.164	0.094	1.784
	d_{mono}	0.158	0.004	0.608	0.054	0.112	0.301	0.324	2.379
Malay	d_{bi}	0.108	0.033	0.050	0.095	0.049	0.086	0.114	0.834
	d_{mono}	0.033	0.138	0.572	0.016	0.087	0.099	0.285	1.253
Mandarin	d_{bi}	0.060	0.040	0.029	0.066	0.018	0.154	0.096	0.813
	d_{mono}	0.074	0.140	0.668	0.016	0.062	0.201	0.361	1.247
Urdu	d_{bi}	0.037	0.026	0.179	0.097	0.131	0.040	0.385	0.792
	d_{mono}	0.045	0.063	0.463	0.022	0.072	0.122	0.364	0.886

Table 7: Distance Results with QWEN2.5 72B

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**
Cantonese	d_{bi}	0.012	0.349	0.018	0.195	0.393	0.009	0.110	0.451
	d_{mono}	0.179	0.007	0.496	0.078	0.050	0.091	0.003	1.160
Thai	d_{bi}	0.071	0.048	0.067	0.022	0.114	0.169	0.173	1.127
	d_{mono}	0.321	0.073	0.788	0.061	0.016	0.398	0.007	1.572
Japanese	d_{bi}	0.020	0.178	0.113	0.032	0.065	0.079	0.256	1.627
	d_{mono}	0.234	0.101	0.912	0.239	0.134	0.443	0.046	2.378
Korean	d_{bi}	0.027	0.035	0.012	0.067	0.289	0.161	0.046	1.612
	d_{mono}	0.332	0.003	0.594	0.090	0.032	0.444	0.005	2.831
Malay	d_{bi}	0.034	0.033	0.053	0.068	0.200	0.025	0.258	0.934
	d_{mono}	0.125	0.043	0.559	0.027	0.010	0.179	0.001	1.585
Mandarin	d_{bi}	0.047	0.068	0.010	0.142	0.123	0.054	0.102	0.862
	d_{mono}	0.190	0.053	0.655	0.031	0.012	0.308	0.005	1.584
Urdu	d_{bi}	0.024	0.093	0.046	0.065	0.337	0.049	0.204	0.571
	d_{mono}	0.148	0.019	0.452	0.040	0.027	0.212	0.007	1.163

Table 8: Distance Results with LLAMA 70B

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**
Cantonese	d_{bi}	0.090	0.469	0.156	0.458	0.294	0.219	0.005	0.088
	d_{mono}	0.046	0.044	0.007	0.097	0.086	0.067	0.021	0.791
Thai	d_{bi}	0.088	0.055	0.023	0.091	0.064	0.012	0.053	0.428
	d_{mono}	0.198	0.060	0.038	0.105	0.019	0.348	0.032	1.041
Japanese	d_{bi}	0.167	0.109	0.046	0.019	0.025	0.010	0.009	0.766
	d_{mono}	0.149	0.168	0.097	0.302	0.113	0.394	0.095	1.848
Korean	d_{bi}	0.044	0.235	0.001	0.058	0.085	0.125	0.000	1.315
	d_{mono}	0.140	0.035	0.025	0.136	0.004	0.380	0.040	2.233
Malay	d_{bi}	0.082	0.237	0.053	0.391	0.149	0.025	0.008	0.409
	d_{mono}	0.045	0.028	0.018	0.046	0.015	0.145	0.015	1.151
Mandarin	d_{bi}	0.011	0.097	0.094	0.176	0.085	0.191	0.028	0.375
	d_{mono}	0.074	0.034	0.031	0.062	0.010	0.264	0.025	1.144
English	d_{bi}	0.008	0.134	0.009	0.062	0.059	0.079	0.015	0.823
	d_{mono}	0.008	0.134	0.009	0.062	0.059	0.079	0.015	0.823
Urdu	d_{bi}	0.099	0.011	0.042	0.188	0.054	0.020	0.010	0.265
	d_{mono}	0.057	0.016	0.016	0.075	0.004	0.174	0.035	0.798

Table 9: Distance Results with LLAMA 8B