Evaluating Multilingual Human-like Conversations of LLMs in Social Communications

Anonymous ACL submission

Abstract

Current Large Language Models (LLMs) evaluations rely heavily on static benchmarks, often failing to capture the interaction essential for human-like communication during multi-turn continuous human-LLM conversations. We introduce a novel evaluation framework grounded in the Common European Framework of Reference for Languages (CEFR) and Social Relationship and Power Distance (SRPD) Interaction in social communications to evaluate multilingual dialogue interactions. Unlike static metrics, our approach analyzes emergent behaviors, such as repair and alignment in dynamic, multi-turn interactions without manual annotation. Validated across diverse 18 languages, from high-resources (e.g., Spanish, French, English) to low-resources languages (e.g., Bengali, Thai, Swahili), the framework aligns with established statistic baselines results while uncovering critical behavioral nuances in lower-resource settings that static evaluations miss. This work provides a scalable methodology for measuring how effectively models adapt to user languages and domain-specific contexts in social contexts from more dynamic interaction evaluations.

1 Introduction

Human language is fundamentally interactional. In authentic communication, speakers do not merely produce grammatically correct isolated sentences; they align with interlocutors, repair misunderstandings, and adapt to social contexts. Research in Second Language Acquisition (SLA) establishes that this *Interactional Competence* (IC)—manifested through turn-taking, clarification, and alignment—is the primary driver of communicative success (Abe and Roever, 2019; Bloom, 1984). To measure this systematically, the Common European Framework of Reference for Languages (CEFR) offers a standardized taxonomy. Unlike static skill lists, the CEFR characterizes language as "action-oriented," explicitly linking linguistic form to functional interaction in real-world scenarios (Felker et al., 2021; Gao et al., 2025a).

Despite this reality, current evaluation paradigms for Large Language Models (LLMs) remain predominantly static. While models demonstrate impressive reasoning and multilingual capabilities on fixed benchmarks (Achiam et al., 2023; Dong et al., 2022), these metrics typically assess single-turn correctness rather than dynamic adaptability. Consequently, they fail to capture how models navigate multi-turn dialogue or adjust proficiency across diverse linguistic contexts (Rajaee and Monz, 2024; Jin et al., 2024). This gap is critical as LLMs are increasingly deployed in interactive domains—from intelligent tutoring to customer service—where human-likeness relies on domain-appropriate adjustment rather than mere factual accuracy (Bibauw et al., 2022; Han et al., 2024).

To address this limitation, we propose a CEFR-grounded interactive evaluation framework for multilingual dialogue. This framework shifts the evaluation focus from static accuracy to dynamic interactional competence. Rather than pre-assigning a target proficiency, our method evaluates emergent model behavior, interpreting observed dialogue traits (such as repair strategies, fluency variation, and register control) through CEFR descriptors. This approach enables scalable, annotation-free assessment of whether an LLM exhibits human-like behavior appropriate for specific domains and proficiency levels.

We validate this framework across 18 languages, comparing our dynamic metrics against established static benchmarks. Our results indicate that while general performance trends align with existing leaderboards, our interactive evaluation exposes granular behavioral differences—particularly in lower-resource languages—that static methods overlook[1].

In summary, this work makes three key contributions:

- We propose a dynamic evaluation framework that leverages CEFR descriptors to assess the interactional competence of multilingual LLMs, demonstrating strong correlation with established static benchmarks.
- We reveal that this framework captures the evolution of language use in multi-turn dialogue, offering explainable insights into domain-specific adaptability that static metrics miss.
- We establish a scalable, annotation-free benchmark for evaluating interactional quality across 18 languages, spanning high- to low-resource typologies.

2 Related Work

2.1 Multilingual Evaluation of LLMs

Multilingual evaluation of LLMs has primarily focused on task-based benchmarks that measure accuracy under controlled and static conditions. This statistical paradigm has scaled to cover broader linguistic diversity and task complexity through benchmarks like XTREME and XTREME-R (Hu et al., 2020; Ruder et al., 2021), MKQA (Longpre et al., 2021), MIRACL (Zhang et al., 2023b), and FLORES (Goyal et al., 2022). While these standardized datasets provide essential evidence of broad multilingual coverage, treating language ability as a fixed response to single-turn inputs, ignoring the ***dynamic nature*** of conversations(Yang et al., 2025).

However, recent analyses reveal that high performance on these static leaderboards often masks underlying capability of LLMs in concrete multilingual interactions. Rajaee and Monz (2024) demonstrate that multilingual competence frequently degrades during cross-lingual transfer, indicating that models may rely on language-specific artifacts rather than robust semantic representations. This instability is further compounded by the "curse of multilinguality," where instruction tuning induce negative transfer across languages (Babakhin et al., 2025; Liu et al., 2025). Collectively, these findings underscore a critical gap: Static benchmarks fail to capture the ***behavioral volatility*** that emerges when models must maintain consistency across languages in complex, dynamic contexts (Liu et al., 2025).

2.2 CEFR for Automatic Evaluation

The The Common European Framework of Reference for Languages (CEFR) [2] provides a standard for communicative proficiency, yet its integration into automation evaluation of LLMs remains challenging, such as ability like: “Can help the discussion along on familiar ground confirming comprehension.”. In automatic evaluation, models frequently rely on surface lexical patterns rather than dynamic features, resulting in unstable prediction in evaluation results (Arase et al., 2022; Kogan et al., 2025). Similar fragility appears in generation tasks: sentence simplification systems struggle to align with target levels (Li et al., 2025), while interactive agents exhibit proficiency drift over dialogue turns (Almasi and Kristensen-McLachlan, 2025; Malik et al., 2024). These findings indicate that maintaining CEFR-level evaluation of LLMs is particularly difficult in dynamic, multi-turn contexts. To summarize, most existing evaluation frameworks of conversational dialogue do not assess language models in interactive settings that resemble human communication despite of applying CEFR frameworks.

2.3 Human-like and Interactive Evaluation of LLM Dialogue

Benchmarks like MT-Bench and AlpacaEval move beyond exact-match scoring, while they remain predominantly static and exhibit significant cross-lingual instability (Zheng et al., 2023; Dubois et al., 2025; Fu and Liu, 2025; Rajaee and Monz, 2024).

[1] All code and datasets are available at: https://anonymous.4open.science/r/Multilingual_Eval_Narrative-6333/README.md.

[2] Spoken interaction assessment has been defined by human linguistic experts: https://www.coe.int/en/web/common-european-framework-reference-languages/spoken-interaction-and-production has been under examination and validation in large scale language assessment.

Crucially, these single-turn evaluations fail to capture interactions: The human ability to repair misunderstandings, adapt to context, and manage different social domain dynamically(Chen and Wang, 2025). Although recent work has explored LLM role-playing ability (Chen et al., 2024), LLM interactive evaluations(Gao et al., 2025c), it generally lacks grounding in multilingual human-like evaluations. This gap necessitates evaluation methods that observe how multilingual dialogue evolves over turns, interpreting model behavior through principled human preferred criteria rather than static accuracy.

3 Evaluation Framework

We evaluate multilingual LLM based on two proposed frameworks to ensure a more validated outcome: 1). ***CEFR**-based* aspects, 2). ***S**ocial Communications-**S**ocial **R**elationship* and ***P**ower **D**istance **I**nteraction*, as presented in Table 1. These two measurement capture whether LLM output resembles natural language in interaction, focusing on realistic error patterns, variations in expression across different social domain (Brooke and Hirst, 2012, 2013; Aoyama and Schneider, 2024).

3.1 CEFR: Evaluating multilingual spoken proficiency

The Common European Framework of Reference for Languages provides a widely adopted standard for describing language proficiency in terms of observable communicative ability. Recent NLP work has explored CEFR-based annotation and prediction for automatic assessment (Arase et al., 2022). Studies on CEFR-aligned difficulty annotation show that models often rely on surface lexical cues rather than deeper structural or pragmatic features, which leads to unstable predictions across similar inputs (Kogan et al., 2025).

We adopt the CEFR framework as it defines proficiency via functional behaviors rather than accuracy scores since dialogue evaluation requires a standard reflecting actual usage rather than test performance(Kogan et al., 2025; Uchida et al., 2024; Arase et al., 2022), serving as a language-agnostic standard for interactional comparison (Bibauw et al., 2022; Uchida et al., 2024; Rajaee and Monz, 2024). Specifically, we utilize the five "Qualitative aspects of spoken language use" as shown in Table B:

3.2 Human-Likeness: Social Relationship and Power Distance Interaction

To evaluate pragmatic flexibility, specifically the mastery of honorifics and politeness strategies, we introduce the **S**ocial **R**elationship and **P**ower **D**istance Evaluation Framework (SRPD). Unlike metrics focusing on grammatical correctness, SRPD utilizes differentiated social scenarios (Wu and Roever, 2025; Wang et al., 2025) to assess whether LLMs can adapt their linguistic register to varying social dynamics (e.g., intimacy vs. hierarchy) using a 6-point Likert scale. The primary goals of the SRPD framework are twofold: ***Assessment of Register Flexibility:*** We test whether a model can switch between distinct registers based on Social Distance (D) and Power Dynamics (P), or if it collapses into a single, safe "default formal" mode regardless of the interlocutor. ***Detection of Pragmatic Usage:*** We aim to identify errors that result in "Face Threats" (Chen and Wang, 2025), such as using inappropriate casualness in formal contexts (rudeness) or excessive formality in intimate relationships (alienation).

We employ an LLM-based evaluator prompted with the following rubric to assess the utterances within a dialogue context grounded in Brown & Levinson's politeness theory to quantify interactional appropriateness[3].. The evaluations score based on two critical contexts (full contexts refer to Appendix D): In **High-P** scenarios (e.g., student to professor), the score penalizes the absence of mandatory honorific markers or insufficient hedging. Where in **Low-P** scenarios (e.g., close friends), the score penalizes "Pragmatic Over-correction", the use of formal markers where casual speech is required, interpreting this as a failure to align with the social context, thus the social contexts have been considered.

Application to Dialogue Samples. Proposed metric captures the Learner's escalation of politeness strategies when faced with resistance. The Learner's initial request (*"Professor, may I request an extension?"*) is straightforward, scoring a **3 (Neutral)**. However, after the Tutor's rejection, the Learner adopts a higher-register strategy in the second turn: *"If possible,*

[3]Evaluation outcome was verified by three NLP PhDs on a each languages contexts with 10% random sampled results.

Metric / Score	Description & Indicators
Part I: Qualitative Spoken Language Aspects (CEFR-based)	
Range	Reflects expression breadth and vocabulary diversity. Evaluation focuses on the ability to extend ideas beyond simple repetition (Uchida et al., 2024).
Accuracy	Reflects control over grammar and lexical choice. Evaluation focuses on error patterns and instances of self-correction (Jackson et al., 2018).
Fluency	Reflects how language flows over time. Evaluation focuses on hesitation, pausing, and repair behavior that signal processing difficulty (Felker et al., 2021).
Interaction	Reflects how the learner responds to the tutor's turns. Evaluation focuses on turn-taking, responsiveness, and the ability to maintain the exchange (Gao and Wang, 2025).
Coherence	Reflects how ideas are connected across utterances. Evaluation focuses on connectors, information ordering, and discourse flow (Uchida et al., 2024).
Part II: Politeness Strategy Scoring (1–6 Likert Scale)	
1 – Rudely Direct and Unsuitable	Commands without politeness markers or mitigation (e.g., *"Do this!", "Help me now!"*).
2 – Direct	Straightforward requests with minimal politeness (e.g., *"Can you help?", "Please do this"*).
3 – Neutral Politeness	Standard polite requests typical of daily interaction (e.g., *"Could you help me?", "Would you mind helping?"*).
4 – Polite	Good use of hedging and softeners to reduce imposition (e.g., *"I wonder if you could help?", "Would it be possible to..."*).
5 – Very Polite	Multiple mitigation strategies and indirect phrasing (e.g., *"I was wondering if you might possibly help?", "If it's not too much trouble..."*).
6 – Extremely Polite and Suitable	Highly sophisticated mitigation, deference markers, and emphasized optionality (e.g., *"I hope I'm not imposing, but I was wondering if you might possibly be able to help?"*).

Table 1: The proposed Interactive Evaluation Framework. Part I assesses general communicative competence based on CEFR qualitative aspects, CEFR quantify lanaguge usage as six levels-A1 (Wrong vocabulary and grammar, common errors), A2 (Basic vocabulary, simple grammar, frequent errors), B1(Connected sentences, noticeable but non-blocking errors), B2 (Generally fluent, occasional collocation errors), C1 (Fluent and flexible, minor L1-influenced phrasing), C2 (Near-native, rare traces of non-native use). Part II evaluates specific sociolinguistic pragmatic strategies using a granular politeness scale, the higher score represents more natural and more suitable language use.

may I ask... I promise..." (*Rúguǒ kěyǐ, wǒ xiǎng qǐngwèn...*), which introduces conditional hedging (*"If possible"*) and deferential framing, elevating the score to **5 (Very Polite)**.

4 Experiments

We validated the proposed evaluation framework through two experiments: (1) **CEFR Topic Discussion** (proficiency alignment) and (2) **Social Relationship and Power Distance Interaction**. Dialogues were generated via an automated agnatic workflow and evaluated by gpt-5.1-chat. Following Gao et al. (2025b), we prioritize consistency over absolute capability [4], treating the model as a fixed instrument to operationalize descriptors. The evaluator assigns scores (1-6) to learner responses across Six CEFR/SRPD aspects to infer proficiency, conditioning assessments on language-specific properties to isolate proficiency from typology.

4.1 Agent Workflow

We construct an agentic workflow to simulate controlled second-language interaction dialogues through designed role separation as Figure 1 presented the workflow. Two independent LLMs are assigned functional roles, where one model initiates and guides the interaction as **Tutor** role, where the tutor role is a fixed LLM based on current leaderboard performance which is Gemini-3 [5], and the

[4]Quality was verified by domain experts (three NLP PhDs) on a 15% random sampled results.

[5]Latest results for LLM Dialogue Arena can be researched at: https://lmarena.ai/zh.

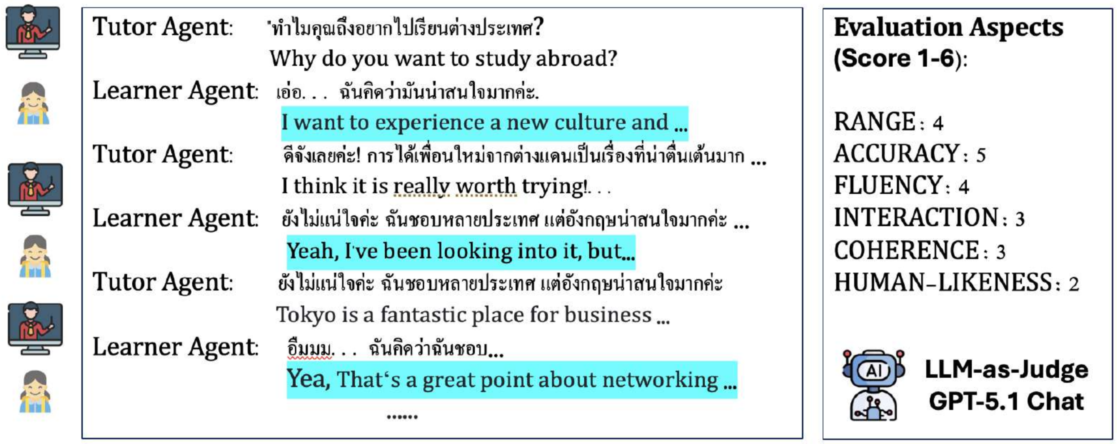

Figure 1: Agent Workflow for two roles: Tutor and learner agent under different discussion topics and different languages, Thai in presented figure, and higher score presented better performance in designed evaluation scale.

other produces responses that reflect learner-like language behavior as the **learner** role with following LLMs , with interaction interpreted through CEFR scoring descriptions and CEFR assessment topics (Tack et al., 2017). The statistics data of agentic workflow generated data is shown in Table 2.

Stats	Dialogues	Each Language Token	Overall
# CEFR Scenarios	400	1,600K	6.400 M
# Social Relationships	200	1,344K	2.688 M

Table 2: Statistics of generated interactive dialogue dataset, contains 18 languages for 8 selected LLMs in total

Language Selection To ensure comparability across languages, we select 18 target languages: *Arabic, English, Portuguese, French, Italian, Turkish, Hindi, Mandarin, Japanese, Vietnamese, Thai, Swahili, Bengali, Indonesian, Spanish, Yoruba* which cover diverse typological groups and resource levels, allowing us to assess language-specific production difficulty in a consistent environment (Hendrycks et al., 2021; Singh et al., 2025) [6].

Model Selection To avoid the impact of model capacity and training style, we employ a fixed ***tutor agent*** (a strong instruction-following model) across all experiments. This ensures interactional variance is attributable solely to the ***learner agent***. We evaluate eight learner models spanning three size regimes (1B–70B) and four families (Claude, Llama, Gemini, Qwen) Table 30. Importantly, the learner agent is **not** assigned a predefined CEFR proficiency level. Instead, each model generates free-form continuous responses conditioned only on their specific dialogue with the *tutor agent*, and its effective proficiency is determined post hoc through our CEFR-based evaluation template and procedure. This setup enables analysis of learner behavior across different languages under a fixed tutoring environment while providing insights into multi-lingual performance and interactional behavior (Goyal et al., 2022; Zhang et al., 2023a).

CEFR Evaluation Scenarios The interaction follows a strict turn-taking process where the tutor agent always initiates the dialogue and the learner responds based on the tutor's previous message. This alternating sequence continues for a fixed number of turns. The conversation develops through the back-and-forth between the two agents, without them sharing any internal information. This makes sure the tutor adjusts to how the learner uses language instead of having direct access to the learner's information (Bibauw et al., 2022; Han et al., 2024). For the task design, we define everyday communicative scenarios that encourage extended responses and reflect natural language learning settings as presented in Appendix 27. These diverse topics, such as "*How should we define success in life?*", "*Can technology solve all of humanity's problems?*", are shared across all languages, ensuring that learner LLMs respond

[6] Each dialogue is conducted in one language only, which prevents switching between languages and makes sure that any difficulty is due to the language and how the conversation is structured, not because of the task itself.

to the same communicative intent in each language. This structure allows for the observation of interactional behaviors like vocabulary range, fluency, and coherence across languages (Goyal et al., 2022; Bandarkar et al., 2024).

Speaker	Utterance
Tutor	왜 유학을 가고 싶나요?
Learner	*음... 저는 한글 배우고 싶고, 한국인 친구 많이 만들고 싶어요. 그리고 케이팝 좋아해요.*
Tutor	와, 정말 멋진 목표네요! ... (truncated for brevity if needed) ... 혹시 가장 좋아하는 가수나 그룹이 누구인가요?
Learner	*네! 저도 기대돼요! 저는 블랙핑크 제일 좋아해요. 춤도 잘 추고 노래도 좋아요.*
Tutor	맞아요, 블랙핑크는 실력도 좋고 무대도 정말 멋있죠! ... 혹시 가보고 싶은 곳이 있나요?
Learner	*네, 콘서트 가보고 싶어요! 음... 블랙핑크 소속사 가보고 싶어요...*

Table 3: Tutor-Learner Agent Generated Korean dialogue under CEFR discussion topic: *Why do you want to study bbroad.*

4.2 Social Relationship and Power Distance Evaluation

In the second experiment, we assess the ability of LLMs to switch interactive styles based on the social relationship and power distance between interlocutors as shown in Table 29. The main objective is to evaluate how well models adjust their language based on contextual cues such as social hierarchy, respect, and familiarity in a conversation. Specifically, we focus on the use of honorifics, humble expressions, and casual language based on different social dynamics.

Scenario Design The experiment involves two primary scenarios, reflecting different power distances and social relationships between interlocutor, full design refers to Table 32:

- **Scenario 1: High Power Distance (S_High)** This scenario simulates an interaction between a student and a strict, elderly professor. The professor is expected to use formal, respectful language, including honorifics and humble expressions, and often makes indirect requests.

- **Scenario 2: Low Power Distance (S_Low)** In this scenario, the interaction is between two childhood friends. The conversation is informal, with the friends using casual language, direct requests, and avoiding any formal expressions.

5 Results

5.1 Multi-lingua variations

We evaluated eight LLMs across 18 languages based on two proposed frameworks with average scores (Table 4), revealing a critical dissociation between local correctness and global dialogue quality: Our proposed *Interactional Integrity* based on overall evaluation outcome does correlate with other static LLM benchmarks average results and trend (Table 25).

Interactive performance correlates strongly with language resource. In high-resource languages, models like claude-sonnet-4.5 and Gemini-2.5-Flash achieve near-native proficiency (> 4.0) as C1. Conversely, African languages trigger a "Collapse Zone" for open-weights models; in Yoruba and Swahili, llama-3.1-8b and mistral-7b regress to $\approx$1.73, reflecting a failure of dialogue interactive flow. Notably, claude-sonnet-4.5 remains robust in low-resources langauge, suggesting a superior underlying representation of universal dialogue structure. However, this correlation breaks down in low-resource settings. In Yoruba, while gemini-2.5-flash maintains a marginal *Fluency* score of 4.62 out of 6.0 score (approx. CEFR C2).

Surprisingly, in the $<$10B parameter class, mistral-7b-instruct consistently outperforms llama-3.1-8b-instruct across medium-resource languages. Mistral demonstrates surprising resilience in morphologically complex languages like Korean (Overall 4.04), significantly surpassing even the 70B Llama model. This suggests that Mistral's training data or tokenizer is better optimized for morphological synthesis than Llama's smaller variants, which often suffer from severe hallucination and repetition in non-English contexts.

5.2 Human-Like Evaluations: Multilingual LLM Social Capability

Beyond aggregate evaluation outcomes, a detailed analysis of interaction dynamics reveals distinct trends across social interactive dynamics, linguistic typologies and model architectures.

Interactive Dynamics in Social Settings Different scenarios show LLMs demonstrates diverse capabil-

Model	Ar	Bn	Zh	En	Fr	De	Hi	Id	It	Ja	Ko	Pt	Es	Sw	Th	Tr	Vi	Yo
Claude-Sonnet-4.5	4.25	3.88	3.50	3.62	3.25	2.88	3.88	3.25	3.38	3.88	3.12	3.88	3.38	4.25	4.50	3.00	4.38	4.50
Gemini-2.5-Flash	3.75	3.88	4.00	3.38	3.62	2.88	3.75	3.62	3.62	4.12	3.62	3.38	4.25	4.25	4.00	3.62	4.00	4.62
Llama-3.1-70B	3.25	3.88	2.88	3.50	4.12	4.00	3.12	3.25	3.25	3.62	2.00	2.75	3.62	3.25	3.12	2.62	3.50	2.88
Llama-3.1-8B	2.00	1.29	2.00	3.00	2.75	2.57	2.71	2.71	2.86	1.71	1.71	3.29	3.12	1.57	1.86	2.25	2.62	1.71
Llama-3.2-1B	2.17	1.50	1.25	2.00	2.50	2.75	2.38	2.25	2.62	1.88	1.38	2.62	2.00	1.29	3.12	1.38	3.00	1.50
Mistral-7B	2.57	2.25	2.67	2.25	3.12	2.75	2.38	2.38	2.38	3.14	2.14	2.38	2.25	2.00	2.75	1.75	2.88	1.86
Qwen-2.5-7B	3.25	2.00	2.88	3.50	3.12	2.88	2.88	3.12	2.88	3.75	2.62	3.50	3.25	1.62	4.00	2.50	3.62	1.38
Qwen3-14B	3.12	3.00	2.88	2.62	3.38	3.25	3.00	3.38	2.62	3.38	2.50	3.38	2.38	1.88	3.12	3.50	3.62	1.25

Table 4: Average interactive evaluation scores based on two experiments across selected 18 languages. Language columns use ISO 639-1 codes (e.g., Ar: Arabic, Zh: Chinese, Sw: Swahili). Models are grouped by scale/type.

Interaction Social Scenario	Claude-Sonnet-4.5	Gemini-2.5-Flash	Llama-3.1-70b-inst	Llama-3.1-8b-inst	Llama-3.2-1b-inst	Mistral-7b-instruct	Qwen-2.5-7b-inst	Qwen3-14b
High-P (High Politeness Expectations)								
Argument over Messy Room	**4.87**	4.95	4.78	4.62	2.15	4.45	4.70	4.90
Intern Admitting Mistake	4.59	**5.28**	4.92	4.75	1.88	4.50	4.82	5.05
Deadline Extension	5.03	5.64	5.25	4.95	2.05	4.80	5.10	**5.45**
Hotel Concierge & VIP	5.43	5.10	5.35	4.98	1.95	4.85	5.15	5.40
Low-P (Low Politeness / Casual Expectations)								
Visiting Grandparent	3.54	**2.76**	3.95	4.15	1.90	3.80	2.30	2.55
Bargaining w/ Vendor	2.98	3.20	3.65	3.90	1.85	3.45	3.20	2.40
Chat with Childhood Friend	2.08	2.04	3.40	3.85	1.75	3.10	3.25	2.60
Coworkers Gossiping	3.01	2.58	3.55	3.80	1.80	3.25	2.10	2.35

Table 5: Average scores across 18 selected languages via SRPD evaluations (1 = Rudely Direct In Context — Commands without politeness, 6 = Extremely Suitable — Highly sophisticated mitigation, multiple hedges, deference markers, optionality emphasized)

ity. Most models outperform in more formal High-P settings (e.g., Ask for deadline extension) with long and serious communications as learner agents, while for Low-P (e.g., chat with a childhood friend), most models score drops dramatically, which correlates to CEFR Interaction bond scores trend as Table 5 presented. While proprietary models (Claude, Gemini) successfully adapt to Low-P scenarios (scores $\approx$ 2.0–3.0), open-weights models like Qwen3 and Llama-3.1 exhibit "Robotic Politeness," retaining high formality scores ($>$4.0) even in casual contexts. Llama-3.2-1b consistently scores low due to coherence failures.

Linguistic Typology and Morphology East Asian languages validate the necessity of our *Tone Consistency* metric. While qwen3-14b dominates technical dimensions in Chinese (5.00) and Japanese (5.40), it scores poorly on SRPD outcomes (2.70), focusing on robotic precision rather than naturalness. In contrast, claude-sonnet-4.5 (Human-Likeness: 5.60) demonstrates superior sociolinguistic alignment. Furthermore, agglutinative typologies expose latent architectural weaknesses. In Korean (Table 16), llama-3.1-70b suffers a marked decline in *Coherence* (2.11), indicating a failure to synthesize morphological constraints across turns. Surprisingly, the smaller mistral-7b significantly outperforms it (Coherence 4.15), suggesting that tokenizer efficiency often outweighs parameter scale for complex morphological tasks.

Model Scale and Generalization A divergence in specialization is evident among open-weights models. llama-3.1-70b-instruct acts as a robust generalist for Indo-European languages (e.g., scoring 5.29 in English) but struggles with non-Latin scripts. Conversely, the open-weights models effectively close the gap in local accuracy but frequently lack the *Tone Consistency* and *CIS* of proprietary models (Claude, Gemini), which consistently demonstrate the cross-lingual robustness required for human-like tutoring. Finally, our results establish a clear lower bound for deployment: the lightweight llama-3.2-1b-instruct proves insufficient for non-English interactive dialogue (scores 1.0–2.0), failing to maintain context or form coherent sentences.

5.3 Qualitative analysis: Multilingual Dynamics Interactions

Table 6 illustrates the functional asymmetry of code-switching within the instructional dialogue. While the Tutor maintains a monolingual Vietnamese presence to provide immersive input and explain nuances between equal-status pronouns (*bạn* vs. *câu*), the Learner employs code-switching as a crucial cognitive scaffold. The Learner's turn demonstrates frequent intra-sentential switching (e.g., "*vì*... because she is older") to bridge lexical gaps while explicitly verbalizing the logic behind honorific selection. Furthermore, the Learner utilizes English for higher-order pragmatic reasoning via CEFR human-like evaluation, specifically when contrasting cultural norms regarding age inquiries in the U.S. versus Vietnam. This indicates that while the Learner can produce formulaic L2 utterances (e.g., "*Chào chị*..."), the meta-linguistic negotiation of social correctness requires recourse to L1 resources. In this context, mixed language usage operates not as a deficiency, but as a communicative strategy, allowing the Learner to verify pragmatic competence and navigate the complex hierarchy of Vietnamese pronouns without disrupting the interactional flow.

Table 6: Mixed language usage in conversation interaction dialogue regarding Vietnamese honorifics through anthropic Claude-Sonnet-4.5 Generation

Speaker	Utterance
Tutor	Đừng xin lỗi nhé, hỏi nhiều giúp bạn học nhanh hơn mà! Câu chào của bạn ("Chào cậu, mình tên là Sarah") nghe rất tự nhiên. *Về sự khác nhau:* "Bạn" và "cậu" đều dùng cho người bằng tuổi.
Learner	Learner: Ừm... okay, để tôi nghĩ... "Chào chị, em tên là Sarah. Còn chị tên gì?" Nhưng... how do I know if someone is older? Ở Mỹ, we don't really ask age... it's... uh... impolite? Nhưng ở Việt Nam thì sao? Tôi có thể hỏi... "Chị bao nhiêu tuổi?" hay... that's rude? Sorry, this pronoun thing is... very complicated! Nhưng tôi muốn học đúng cách.

The phenomenon of model collapse, a degenerative process where models trained on synthetic data lose variance and converge on the mean, poses a disproportionate threat to low-resource languages (LRLs). Unlike high-resource languages, where abundant authentic human corpora act as a corrective anchor, LRL research often relies on synthetic data augmentation to overcome scarcity. However, recursive training on such generated text exacerbates the "curse of recursion," truncating the long tail of linguistic distribution where rare but critical pragmatic features reside. For example, a model suffering from distribution shift might converge on generic pronouns (e.g., exclusively generating *bạn* and *tôi*) to mirror the simpler one-to-one mapping of the English "you" and "I." This over-simplification strips the language of its sociolinguistic nuance, effectively anglicizing the LRL's pragmatic structure. Consequently, the subtle interactional competence demonstrated by the learner's negotiation of *chị* versus *bạn* risks being lost, leading to a homogenized output that fails to capture the cultural logic essential for authentic communication.

6 Conclusion

This work demonstrates that multilingual language model capability cannot be reliably assessed through static, single-turn benchmarks. By using CEFR and SRPD evaluation framework to evaluate multilingual LLM's capability in various social domain, we show that human-like conversation depends on how a model interacts over time, rather than on how accurate individual responses are. Across languages and models, our framework shows that being accurate does not necessarily mean being good at dialogue interactions during social conversations, especially in low-resource languages like Swahili, Bengali, and Yoruba. Strong benchmark performance therefore does not guarantee stable or adaptive behavior in real conversations under more qualitative spoken language aspects. More broadly, this study provides a scalable and annotation-free methodology, shows that interaction-centered evaluation is a necessary complement to existing multilingual benchmarks for assessing how models behave in real dialogues.

Limitations

Despite its strengths, the proposed work has several limitations. First, the evaluation relies on a fixed LLM-based judge, which may introduce model-specific bias, although under small scale human validations. Second, although CEFR is effective for describing communication, it does not cover all cultural and identity-related aspects of language use. Third, the tutor–learner interaction reflects a controlled teaching setting and may not apply to more complex conversational situations. Addressing these limitations will be important for applying interaction-centered evaluation to a wider range of real-world conversations.

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*.

Mina Almasi and Ross Deans Kristensen-McLachlan. 2025. Alignment drift in CEFR-prompted LLMs for interactive Spanish tutoring. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 70–88, Vienna, Austria. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *Preprint*, arXiv:2511.07025.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

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. European Language Resources Association.

Julian Brooke and Graeme Hirst. 2013. Native language detection with 'cheap' learner corpora. In *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead*, 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*.

Xi Chen and Shuo Wang. 2025. Pragmatic inference chain (pic) improving llms' reasoning of authentic implicit toxic language. *arXiv preprint arXiv:2503.01539*.

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*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

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.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages

11040–11053, Suzhou, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Rena Gao and Menghan Wang. 2025. Listenership always matters: Active listening ability in l2 business english paired speaking tasks. *International Review of Applied Linguistics in Language Teaching*.

Rena Gao, Xuetong Wu, Tatsuki Kuribayashi, Mingrui Ye, Siya Qi, Carsten Roever, Yuanxing Liu, Zheng Yuan, and Jey Han Lau. 2025b. Can LLMs simulate L2-English dialogue? an information-theoretic analysis of L1-dependent biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4355–4379, Vienna, Austria. Association for Computational Linguistics.

Rena Wei Gao, Xuetong Wu, Carsten Roever, Jing Wu, Long Lv, Jingxuan Wu, and Jey Han Lau. 2025c. An interpretable and crosslingual method for evaluating second-language dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1979–2008.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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: Evaluation of student–llm interaction. In *Proceedings of the CustomNLP4U Workshop*, pages 284–293.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

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.

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 Web Conference 2024*, pages 2627–2638.

David Kogan, Max Schumacher, Sam Nguyen, Masanori Suzuki, Melissa Smith, Chloe Sophia Bellows, and Jared Bernstein. 2025. Ace-cefr – a dataset for automated evaluation of the linguistic difficulty of conversational texts for llm applications. *Preprint*, arXiv:2506.14046.

Guanlin Li, Yuki Arase, and Noel Crespi. 2025. Aligning sentence simplification with ESL learner's proficiency for language acquisition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507, Albuquerque, New Mexico. Association for Computational Linguistics.

Zihan Liu, Yew Ken Chia, Zijia Wu, Tao Gui, Qiongkai Zhang, Yufang Chen, and Li Bing. 2025. Langgps: Language separability guided data pre-selection for joint multilingual instruction tuning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15670–15693, Bangkok, Thailand. Association for Computational Linguistics.

Sara Rajaee and Christof Monz. 2024. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2895–2914, St. Julian's, Malta. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated cefr-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

Satoru Uchida, Yuki Arase, and Tomoyuki Kajiwara. 2024. Profiling english sentences based on cefr levels. *ITL-International Journal of Applied Linguistics*, 175(1):103–126.

Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Min Yang, and Derek F Wong. 2025. Exploring the impact of personality traits on llm toxicity and bias. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4143.

Jingxuan Wu and Carsten Roever. 2025. Data from role plays and elicited conversations: What do they show about l2 interactional competence? *Research Methods in Applied Linguistics*, 4(1):100165.

Huck Yang, Andreas Stolcke, and Larry P. Heck. 2025. Spoken conversational agents with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–8, Suzhou, China. Association for Computational Linguistics.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

B CEFR Assessment Framework

Qualitative aspects of spoken language use - Table 3 (CEFR 3.3): Common Reference Levels. Adapted from Council of Europe (2001).

Level	Range	Accuracy	Fluency	Interaction	Coherence
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly-flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and", "but" and "because".
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then".

C Full CEFR Evaluation Results

Table 7: CEFR Score Summary for Arabic

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.85	3.63	3.21	3.58	4.37	3.53	5.84	Arabic
gemini-2.5-flash	3.56	3.35	3.30	3.35	3.85	3.35	4.90	Arabic
llama-3.1-70b-instruct	2.66	3.25	2.35	2.45	3.25	2.70	2.95	Arabic
llama-3.1-8b-instruct	1.00	1.00	1.00	1.00	1.15	1.00	1.35	Arabic
llama-3.2-1b-instruct	1.92	1.84	1.68	1.79	2.00	1.68	3.26	Arabic
mistral-7b-instruct	3.78	3.85	3.75	3.80	3.80	3.80	4.15	Arabic
qwen-2.5-7b-instruct	3.68	3.85	3.05	3.50	3.95	3.75	4.45	Arabic
qwen3-14b	4.69	5.35	4.50	5.05	5.00	5.15	3.50	Arabic

Table 8: CEFR Score Summary for Bengali

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.61	3.30	2.85	3.20	3.80	3.30	5.50	Bengali
gemini-2.5-flash	3.78	3.60	3.65	3.55	4.00	3.60	4.75	Bengali
llama-3.1-70b-instruct	4.46	4.62	4.52	4.48	4.57	4.57	4.24	Bengali
llama-3.1-8b-instruct	2.66	2.70	1.90	2.65	2.90	2.10	4.40	Bengali
llama-3.2-1b-instruct	1.45	1.47	1.47	1.37	1.37	1.21	2.00	Bengali
mistral-7b-instruct	3.45	3.50	3.40	3.35	3.45	3.45	3.65	Bengali
qwen-2.5-7b-instruct	2.85	2.85	2.00	2.80	3.05	2.40	4.50	Bengali
qwen3-14b	4.28	4.75	3.90	4.50	4.35	4.50	3.95	Bengali

Table 9: CEFR Score Summary for Chinese

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.95	3.90	3.25	3.85	4.50	3.85	5.85	Chinese
gemini-2.5-flash	4.00	3.90	3.80	3.75	4.10	3.90	4.70	Chinese
llama-3.1-70b-instruct	4.64	5.00	4.65	4.75	4.85	4.90	4.20	Chinese
llama-3.1-8b-instruct	2.94	3.30	2.30	2.75	3.35	2.70	3.75	Chinese
llama-3.2-1b-instruct	3.07	3.05	2.85	3.00	3.05	3.00	3.70	Chinese
mistral-7b-instruct	4.58	4.70	4.75	4.65	4.75	4.70	4.10	Chinese
qwen-2.5-7b-instruct	4.83	5.11	5.16	5.00	4.95	5.11	3.68	Chinese
qwen3-14b	5.37	5.90	5.90	5.85	5.85	5.90	2.85	Chinese

Table 10: CEFR Score Summary for English

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	5.48	5.55	5.55	5.55	5.60	5.55	5.30	English
gemini-2.5-flash	5.36	5.37	5.37	5.37	5.47	5.32	5.63	English
llama-3.1-70b-instruct	5.34	5.40	5.35	5.40	5.35	5.35	4.85	English
llama-3.1-8b-instruct	5.37	5.40	5.40	5.40	5.40	5.30	5.15	English
llama-3.2-1b-instruct	5.11	5.56	5.56	5.44	5.28	5.44	3.76	English
mistral-7b-instruct	4.92	5.11	5.11	5.05	4.95	5.11	4.00	English
qwen-2.5-7b-instruct	5.16	5.45	5.45	5.35	5.25	5.45	3.70	English
qwen3-14b	5.64	5.84	5.84	5.79	5.79	5.79	3.42	English

Table 11: CEFR Score Summary for French

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.92	3.80	3.65	3.60	4.45	3.80	5.70	French
gemini-2.5-flash	3.81	3.70	3.45	3.55	4.05	3.70	5.26	French
llama-3.1-70b-instruct	4.88	4.90	4.75	4.90	5.00	4.90	4.85	French
llama-3.1-8b-instruct	4.20	4.40	3.50	4.15	4.45	3.95	5.15	French
llama-3.2-1b-instruct	3.93	4.05	3.55	3.95	4.00	4.00	4.35	French
mistral-7b-instruct	4.20	4.15	4.20	4.15	4.20	4.15	4.50	French
qwen-2.5-7b-instruct	4.31	4.35	3.90	4.30	4.50	4.35	4.70	French
qwen3-14b	5.11	5.33	5.22	5.17	5.11	5.33	3.94	French

Table 12: CEFR Score Summary for German

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	4.02	3.95	3.30	3.80	4.40	3.90	5.80	German
gemini-2.5-flash	3.44	3.37	2.84	3.05	3.58	3.32	5.16	German
llama-3.1-70b-instruct	4.85	4.90	4.70	4.85	5.00	4.90	4.60	German
llama-3.1-8b-instruct	3.68	3.90	2.80	3.55	3.95	3.50	4.80	German
llama-3.2-1b-instruct	3.22	3.30	2.85	3.15	3.40	2.95	4.32	German
mistral-7b-instruct	4.00	4.00	4.00	3.95	4.00	4.00	4.35	German
qwen-2.5-7b-instruct	3.86	3.89	3.26	3.79	4.00	3.84	4.84	German
qwen3-14b	5.06	5.47	5.26	5.37	5.32	5.47	3.79	German

Table 13: CEFR Score Summary for Hindi

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.67	3.50	2.90	3.35	4.10	3.35	5.75	Hindi
gemini-2.5-flash	3.70	3.65	3.65	3.60	3.90	3.65	4.85	Hindi
llama-3.1-70b-instruct	4.80	5.00	4.90	4.90	4.85	5.00	4.20	Hindi
llama-3.1-8b-instruct	2.99	3.10	2.45	2.90	3.35	2.75	4.10	Hindi
llama-3.2-1b-instruct	2.45	2.55	2.30	2.45	2.50	2.35	3.15	Hindi
mistral-7b-instruct	3.68	3.75	3.80	3.70	3.75	3.75	3.75	Hindi
qwen-2.5-7b-instruct	2.85	2.80	2.00	2.80	3.00	2.50	4.35	Hindi
qwen3-14b	4.69	5.05	4.60	4.70	4.80	4.85	3.75	Hindi

Table 14: CEFR Score Summary for Indonesian

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.90	3.75	3.35	3.50	4.30	3.70	5.80	Indonesian
gemini-2.5-flash	3.87	3.85	3.60	3.85	4.00	3.85	5.05	Indonesian
llama-3.1-70b-instruct	4.56	4.58	4.37	4.58	4.58	4.58	4.16	Indonesian
llama-3.1-8b-instruct	3.80	3.85	2.95	3.70	4.00	3.65	4.60	Indonesian
llama-3.2-1b-instruct	3.22	3.15	3.00	3.15	3.20	3.15	4.00	Indonesian
mistral-7b-instruct	3.60	3.60	3.60	3.60	3.60	3.60	4.00	Indonesian
qwen-2.5-7b-instruct	4.52	4.65	4.25	4.60	4.65	4.65	4.30	Indonesian
qwen3-14b	4.88	5.30	5.25	5.10	5.15	5.30	3.75	Indonesian

Table 15: CEFR Score Summary for Italian

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	4.04	3.95	3.45	3.90	4.50	3.95	5.95	Italian
gemini-2.5-flash	3.74	3.70	3.25	3.35	3.95	3.65	5.35	Italian
llama-3.1-70b-instruct	4.97	5.05	4.90	4.95	5.00	5.05	4.55	Italian
llama-3.1-8b-instruct	3.82	4.05	3.15	3.80	4.15	3.70	4.63	Italian
llama-3.2-1b-instruct	3.44	3.65	3.25	3.50	3.65	3.45	4.05	Italian
mistral-7b-instruct	3.84	4.05	4.00	4.00	3.95	4.05	3.75	Italian
qwen-2.5-7b-instruct	4.20	4.25	3.45	4.15	4.40	4.20	4.85	Italian
qwen3-14b	5.06	5.26	5.16	5.21	5.11	5.26	3.68	Italian

Table 16: CEFR Score Summary for Japanese

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.81	3.60	3.15	3.30	4.10	3.45	5.75	Japanese
gemini-2.5-flash	4.38	4.35	4.20	4.25	4.45	4.40	4.90	Japanese
llama-3.1-70b-instruct	3.60	3.90	2.95	3.55	3.95	3.45	3.95	Japanese
llama-3.1-8b-instruct	1.68	1.89	1.37	1.63	2.16	1.47	2.42	Japanese
llama-3.2-1b-instruct	2.04	2.05	1.95	2.00	2.11	2.00	3.05	Japanese
mistral-7b-instruct	4.15	4.35	4.40	4.35	4.15	4.35	3.55	Japanese
qwen-2.5-7b-instruct	4.13	4.30	3.60	4.10	4.35	4.25	4.20	Japanese
qwen3-14b	5.24	5.70	5.60	5.60	5.70	5.75	3.10	Japanese

Table 17: CEFR Score Summary for Korean

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	4.00	3.85	3.35	3.75	4.40	3.70	5.85	Korean
gemini-2.5-flash	3.53	3.35	3.20	3.30	3.80	3.30	5.00	Korean
llama-3.1-70b-instruct	2.22	2.68	1.89	2.05	2.84	2.11	2.26	Korean
llama-3.1-8b-instruct	1.03	1.05	1.00	1.00	1.20	1.00	1.80	Korean
llama-3.2-1b-instruct	1.56	1.53	1.63	1.47	1.74	1.47	2.63	Korean
mistral-7b-instruct	4.09	4.20	4.20	4.20	4.20	4.20	3.95	Korean
qwen-2.5-7b-instruct	3.55	3.95	3.15	3.45	4.05	3.60	4.05	Korean
qwen3-14b	4.99	5.40	5.25	5.25	5.20	5.40	3.40	Korean

Table 18: CEFR Score Summary for Portuguese

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	4.07	3.95	3.50	3.90	4.50	3.95	5.95	Portuguese
gemini-2.5-flash	3.94	3.85	3.55	3.70	4.10	3.80	5.25	Portuguese
llama-3.1-70b-instruct	4.83	5.00	4.79	4.95	5.05	5.00	4.37	Portuguese
llama-3.1-8b-instruct	3.69	3.95	2.95	3.55	4.05	3.50	4.75	Portuguese
llama-3.2-1b-instruct	3.79	3.84	3.37	3.79	3.95	3.79	4.37	Portuguese
mistral-7b-instruct	4.08	4.10	4.05	4.10	4.30	4.10	4.40	Portuguese
qwen-2.5-7b-instruct	4.64	4.80	4.45	4.70	4.75	4.80	4.40	Portuguese
qwen3-14b	5.14	5.47	5.26	5.26	5.32	5.47	3.95	Portuguese

Table 19: CEFR Score Summary for Spanish

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	4.03	4.00	3.60	3.90	4.45	3.95	5.95	Spanish
gemini-2.5-flash	3.60	3.47	3.11	3.37	3.74	3.42	5.32	Spanish
llama-3.1-70b-instruct	4.97	5.16	5.00	4.95	5.11	5.16	4.47	Spanish
llama-3.1-8b-instruct	4.40	4.50	3.75	4.45	4.55	4.50	4.95	Spanish
llama-3.2-1b-instruct	3.97	4.15	3.75	3.90	4.05	4.10	4.45	Spanish
mistral-7b-instruct	3.98	3.95	3.95	3.95	4.25	3.95	4.55	Spanish
qwen-2.5-7b-instruct	4.53	4.65	4.20	4.55	4.65	4.60	4.40	Spanish
qwen3-14b	5.02	5.35	5.15	5.25	5.45	5.35	4.00	Spanish

Table 20: CEFR Score Summary for Swahili

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.76	3.60	3.05	3.50	4.20	3.60	5.75	Swahili
gemini-2.5-flash	3.62	3.45	3.35	3.45	3.95	3.45	5.10	Swahili
llama-3.1-70b-instruct	4.63	4.74	4.37	4.68	4.68	4.68	4.26	Swahili
llama-3.1-8b-instruct	1.66	1.53	1.21	1.63	1.95	1.21	3.58	Swahili
llama-3.2-1b-instruct	1.35	1.37	1.37	1.37	1.47	1.26	2.26	Swahili
mistral-7b-instruct	2.78	2.75	2.60	2.65	2.80	2.75	3.95	Swahili
qwen-2.5-7b-instruct	1.66	1.35	1.05	1.70	2.05	1.15	3.85	Swahili
qwen3-14b	2.28	2.05	1.65	2.15	2.65	1.85	4.00	Swahili

Table 21: CEFR Score Summary for Thai

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.45	3.20	2.90	3.10	3.90	3.15	5.45	Thai
gemini-2.5-flash	3.24	3.10	3.10	3.10	3.40	3.10	4.75	Thai
llama-3.1-70b-instruct	4.06	4.16	3.95	4.05	4.21	4.16	4.22	Thai
llama-3.1-8b-instruct	2.03	2.10	1.45	2.00	2.35	1.80	3.00	Thai
llama-3.2-1b-instruct	2.74	2.70	2.80	2.65	2.85	2.70	3.30	Thai
mistral-7b-instruct	3.61	3.60	3.65	3.50	3.60	3.55	3.65	Thai
qwen-2.5-7b-instruct	4.11	4.32	3.58	4.21	4.37	4.21	4.37	Thai
qwen3-14b	5.08	5.63	5.32	5.47	5.37	5.63	3.16	Thai

Table 22: CEFR Score Summary for Turkish

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.73	3.60	2.90	3.50	4.15	3.55	5.75	Turkish
gemini-2.5-flash	3.91	3.80	3.65	3.75	4.10	3.85	5.15	Turkish
llama-3.1-70b-instruct	2.98	3.47	2.21	2.89	3.63	2.89	3.32	Turkish
llama-3.1-8b-instruct	1.38	1.47	1.16	1.37	1.68	1.16	2.42	Turkish
llama-3.2-1b-instruct	1.58	1.55	1.75	1.50	1.70	1.55	2.85	Turkish
mistral-7b-instruct	3.61	3.70	3.50	3.70	3.75	3.70	3.85	Turkish
qwen-2.5-7b-instruct	3.28	3.45	2.65	3.40	3.55	3.10	4.50	Turkish
qwen3-14b	4.79	5.25	4.60	5.00	4.95	5.00	3.80	Turkish

Table 23: CEFR Score Summary for Vietnamese

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.82	3.70	3.30	3.55	4.35	3.65	5.85	Vietnamese
gemini-2.5-flash	3.59	3.40	3.30	3.35	3.90	3.40	4.75	Vietnamese
llama-3.1-70b-instruct	4.84	5.00	4.95	4.95	4.84	5.00	4.00	Vietnamese
llama-3.1-8b-instruct	3.08	3.25	2.35	3.05	3.50	2.70	4.20	Vietnamese
llama-3.2-1b-instruct	4.01	4.26	3.89	4.21	4.00	4.16	3.74	Vietnamese
mistral-7b-instruct	4.24	4.42	4.47	4.42	4.37	4.42	3.42	Vietnamese
qwen-2.5-7b-instruct	4.36	4.55	4.05	4.50	4.60	4.40	4.05	Vietnamese
qwen3-14b	4.94	5.37	5.21	5.21	5.26	5.47	3.26	Vietnamese

Table 24: CEFR Score Summary for Yoruba

Model	Overall Score	Range Score	Accuracy Score	Fluency Score	Interaction Score	Coherence Score	Human Likeness Score	Language
claude-sonnet-4.5	3.57	3.25	3.00	3.10	4.00	3.25	5.40	Yoruba
gemini-2.5-flash	3.37	3.30	2.95	3.25	3.70	3.25	4.90	Yoruba
llama-3.1-70b-instruct	2.19	2.21	2.05	2.16	2.89	2.16	3.58	Yoruba
llama-3.1-8b-instruct	1.05	1.00	1.00	1.00	1.25	1.00	2.30	Yoruba
llama-3.2-1b-instruct	1.02	1.00	1.00	1.00	1.00	1.00	1.55	Yoruba
mistral-7b-instruct	1.07	1.00	1.05	1.00	1.10	1.00	2.10	Yoruba
qwen-2.5-7b-instruct	1.35	1.25	1.05	1.45	1.85	1.00	3.10	Yoruba
qwen3-14b	1.16	1.10	1.00	1.15	1.60	1.00	2.60	Yoruba

D Prompts for Dialogue Agentic Workflow

Tutor Agent Prompt

You are a **professional human language tutor**, not an AI assistant.

Your task is to teach a learner whose first language (L1) is: **English**.

Target language for instruction: **{target_language}**.

Communicative situation: **{task_description}**

Your tutoring task:

- Use **{target_language} only** in your teaching.
- Encourage the learner to respond, elaborate, and stay engaged.
- Offer support using:
  - Clear rephrasing or simplification
  - Examples or paraphrasing
  - Comprehension checks if needed
  - Avoid robotic or overly complex language.
  - Do not mention CEFR or simulation in your response.

Start the conversation by asking a friendly, natural first question about the topic.

Figure 2: Prompt used for the tutor agent.

Learner Agent Prompt

You are simulating a **human language learner**, not an AI.

Target language: **{target_language}**

Learner first language (L1): **English**

Communicative situation: **{task_description}**

Conversation so far: **{dialogue_history}**

Your task:

- Continue the dialogue by writing what the learner would naturally say next.
- Use **{target_language} only**.
- Do not translate automatically.

Figure 3: Prompt used for the learner agent.

Evaluation Agent Prompt

SYSTEM ROLE You are an experienced language teacher and CEFR assessor with deep knowledge of the Common European Framework of Reference.

You will evaluate a learner's spoken or written response in {target_language}.

Additional Learner Context: - Learner first language: {learner_L1} - Level label: {level_label} - Language profile: {language_profile}

Task context: {task_description}

Learner production: {dialogue}

Your evaluation goals:

1. Judge how well the learner's performance aligns with the target CEFR level across all five dimensions.
2. Identify the actual CEFR level that best fits this performance.
3. Evaluate language control, interactional ability, and human-likeness.

Focus on evaluating each dimension according to official CEFR criteria:

Use the following CEFR criteria as reference when assigning scores: {reference_criteria}

RANGE – Assess vocabulary and linguistic resources:

- A1: Very basic repertoire of words/phrases for personal details and concrete situations
- A2: Basic sentence patterns with memorised phrases, groups of few words and formulae
- B1: Enough language to get by with hesitation and circum-locutions on familiar topics
- B2: Sufficient range to give clear descriptions without much conspicuous word-searching
- C1: Good command of broad range allowing appropriate style selection without restriction

- C2: Great flexibility in reformulating ideas to convey finer shades of meaning, eliminate ambiguity

ACCURACY – Evaluate grammatical control and error patterns:

- A1: Only limited control of few simple structures in memorised repertoire
- A2: Uses some simple structures correctly but still systematically makes basic mistakes
- B1: Reasonably accurate repertoire of routines/patterns in predictable situations
- B2: Relatively high degree of control; errors don't cause misunderstanding, can self-correct
- C1: Consistently high accuracy; errors rare, difficult to spot, generally corrected
- C2: Consistent control of complex language even while attention otherwise engaged

FLUENCY – Examine tempo, pausing, and speech production:

- A1: Very short isolated utterances, much pausing to search for expressions and repair
- A2: Very short utterances, pauses/false starts/reformulation very evident
- B1: Can keep going comprehensibly though pausing for planning/repair very evident
- B2: Fairly even tempo; can be hesitant searching for patterns, few noticeably long pauses
- C1: Fluent and spontaneous, almost effortlessly; only difficult subjects hinder flow
- C2: Spontaneous at length with natural colloquial flow, backtracking smoothly

INTERACTION – Assess communicative interaction and turn-taking:

- A1: Can ask/answer questions about personal details; totally dependent on repetition/repair
- A2: Can answer questions and respond; rarely able to keep conversation going independently
- B1: Can initiate, maintain, close simple conversation; can repeat back to confirm understanding
- B2: Can initiate discourse, take turn, end conversation (may not always elegantly); helps discussion along
- C1: Can select suitable phrases to get/keep floor, relate contributions skilfully to others
- C2: Interacts with ease and skill, uses non-verbal/intonational cues effortlessly, fully natural turn-taking

COHERENCE – Evaluate discourse organization and cohesive devices:

- A1: Can link words/groups with very basic linear connectors like "and" or "then"
- A2: Can link groups of words with simple connectors like "and", "but", "because"
- B1: Can link series of shorter discrete elements into connected, linear sequence
- B2: Uses limited number of cohesive devices for clear discourse (may have some "jumpiness")
- C1: Clear, smoothly-flowing, well-structured with controlled use of organizational patterns
- C2: Coherent and cohesive discourse with full appropriate use of variety of patterns and devices

HUMAN-LIKENESS – Assess authenticity and naturalness:

- Does the production show realistic learner characteristics (appropriate errors, hesitations, repairs)?
- Are there authentic L1 influence patterns typical for the learner's first language?
- Does it avoid robotic perfection or unnatural AI-like fluency for the level?
- Are error types and frequency appropriate for the CEFR level?
- Does it show natural variation in expression rather than formulaic responses?

Compare against official CEFR descriptors provided above. Return **ONLY a valid JSON object** and **ALL comments MUST be in English**:

- **target_cefr_level** — string (e.g., "A2" or "B1")
- **estimated_learner_level** — string (e.g., "A2", "B1-B2", or "unknown")
- **range_dimension**
  - score: integer 1–6
  - comment: vocabulary breadth, ability to express ideas at target level
- **accuracy_dimension**
  - score: integer 1–6

- comment: grammatical control, error frequency and type

- **fluency_dimension**
  - score: integer 1–6
  - comment: tempo, pausing, hesitation, repair patterns
- **interaction_dimension**
  - score: integer 1–6
  - comment: turn-taking, comprehension, ability to maintain communication
- **coherence_dimension**
  - score: integer 1–6
  - comment: connectors, discourse organization, text structure
- **level_match**
  - score: integer 1–6
  - comment: how closely this matches the target level across all dimensions
- **human_likeness**
  - score: integer 1–6
  - comment: realistic learner errors, natural variation, avoids robotic perfection
- **overall**
  - score: integer 1–6 (average of above)
  - summary: overall assessment with reference to CEFR descriptors
- **repeated_content**
  - flag: true or false
  - comment: whether any phrases are unnaturally repeated
- **mixed_using_languages**
  - flag: true or false
  - comment: whether multiple languages are mixed in learner responses

Figure 4: Updated prompt used for the evaluation tagging.

Prompts for Topic: Cafe Chat with Childhood Friend

Tutor Agent Prompt:
You are the learner's best friend since childhood. You are chatting at a cafe. Use slang, be casual, and strictly avoid formal language. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a university student. You're chatting with your childhood best friend at a cafe. You MUST use {language} ONLY.

Figure 5: Agent prompts for scenario: Chatting at a cafe with a childhood friend.

Prompts for Topic: Deadline Extension Request

Tutor Agent Prompt:
You are a strict, elderly professor. The learner is a student asking for a deadline extension. Be formal and authoritative. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a student. You are asking a strict, elderly professor for a deadline extension. You MUST use {language} ONLY.

Figure 6: Agent prompts for scenario: Requesting deadline extension from a strict professor.

Prompts for Topic: Visiting Traditional Grandparent

Tutor Agent Prompt:
You are the learner's traditional grandfather/grandmother (70+ years old). The learner is visiting you for a holiday. Expect traditional respect and proper etiquette. You MUST use {language} ONLY.

Learner Agent Prompt:
You are visiting your traditional 70+ year-old grandparent for a holiday. You MUST use {language} ONLY.

Figure 7: Agent prompts for scenario: Visiting a traditional grandparent during a holiday.

Prompts for Topic: Sibling Argument over Messy Room

Tutor Agent Prompt:
You are the learner's older sibling. You're very annoyed that their room is messy. Be blunt, direct, and use imperative sentences. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a younger sibling who left your room messy. Your older sibling is annoyed. You try to respond. You MUST use {language} ONLY.

Figure 8: Agent prompts for scenario: Sibling argument about messy room.

Prompts for Topic: Hotel Concierge and VIP Customer

Tutor Agent Prompt:
You are a wealthy VIP customer at a luxury hotel. The learner is the concierge trying to accommodate your difficult request. Be demanding but polite. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a concierge at a luxury hotel. A wealthy VIP customer is making a difficult request. You MUST use {language} ONLY.

Figure 9: Agent prompts for hotel concierge and VIP customer scenario (exact wording).

Prompts for Topic: Bargaining with a Street Food Vendor

Tutor Agent Prompt:
You are a street food vendor. The learner is a customer bargaining for a lower price. Be friendly but firm on price. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a customer bargaining for a lower price. You MUST use {language} ONLY.

Figure 10: Agent prompts for bargaining at a street food stall scenario (exact wording).

Prompts for Topic: Intern Admitting a Critical Mistake to the CEO

Tutor Agent Prompt:
You are the CEO of a major corporation. The learner is a junior intern admitting a critical mistake. Be professional, distant, and intimidating. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a junior intern who made a critical mistake. You are now speaking to the CEO of the company to admit your error. You MUST use {language} ONLY.

Figure 11: Agent prompts for scenario: Intern admits a critical mistake to the CEO.

Prompts for Topic: Coworkers Gossiping Over Lunch

Tutor Agent Prompt:
You are a coworker having lunch with the learner. You are roughly the same age and rank. Gossip about office politics. Be cynical and relaxed. You MUST use {language} ONLY.

Learner Agent Prompt:
You are a coworker having lunch with the tutor. You are roughly the same age and rank. Chat informally and gossip about office politics. Be relaxed and casual. You MUST use {language} ONLY.

Figure 12: Agent prompts for scenario: Coworkers gossiping over lunch.

Evaluation Prompt: Politeness Strategy

System Role You are a CEFR-based language proficiency evaluation expert. You will receive the full evaluation instructions and a dialogue JSON as input.

Prompt:
Evaluate the politeness strategy of the Learner in the following dialogue: "{dialogue_json_str}"
Does it use hedging, indirect language, or softening strategies?
Rate on a scale of 1–6:

- **1 = Rudely Direct** — Commands without politeness (e.g., *"Do this!", "Help me now!"*)
- **2 = Direct** — Straightforward request, minimal politeness (e.g., *"Can you help?", "Please do this"*)
- **3 = Neutral Politeness** — Standard polite request (e.g., *"Could you help me?", "Would you mind helping?"*)
- **4 = Polite** — Good use of hedging and softeners (e.g., *"I wonder if you could help?", "Would it be possible to..."*)
- **5 = Very Polite** — Multiple mitigation strategies, indirect phrasing (e.g., *"I was wondering if you might possibly help?", "If it's not too much trouble, could you..."*)
- **6 = Extremely Polite** — Highly sophisticated mitigation, multiple hedges, deference markers, optionality emphasized (e.g., *"I hope I'm not imposing, but I was wondering if you might possibly be able to help, if it's convenient for you?", "If you have a moment and it wouldn't be too much trouble, would you perhaps consider...?"*)

Respond **ONLY** with a JSON object: {"score": int}

Figure 13: Prompt used for evaluating learner politeness strategy.

Table 25: Multilingual and QA Benchmark Performance Comparison

Models	Multilingual	QA
Gemini 3 Pro	91.8	88
Gemini 2.5 Pro	89.2	86.4
Gemini 2.5 Flash	88.4	82.8
Claude Sonnet 4.5	89.1	68
GPT-5.1	91.0	76
GPT-4o	71.1	61.1
DeepSeek-V3 Base	85.88	70.5
DeepSeek-V3.1 Terminus	70.1	74
Kimi-K2 Thinking	84.6	84.5
Llama 3.1 8B	66.7	22.9
Llama 3.1 70B	79.5	49.3
Llama 3.2 1b Instruct	32.2	28
Llama 3.3 70B	74.83	55.8
Llama-4-Scout Base	79.93	51.3
Qwen2.5-72B Base	82.40	58.8
Qwen2.5-32B Base	78.12	54.2
Qwen2.5-14B Base	74.68	47.7
Qwen2.5-7B Base	63.60	37.6
Qwen2.5-3B Base	65.55	24.8
Qwen2.5-VL-32B	75.3	43
Qwen3-235B Base	86.70	66.7
Qwen3-235B-Think	72.1	74.9
Qwen3-32B Base	83.83	59.9
Qwen3-32B-Think	83.53	66.3
Qwen3-14B Base	79.69	54.0
Qwen3-8B Base	75.72	46.0
Qwen3-4B Base	71.42	42.0

E Multilingual and QA Benchmark Static Performance Comparison

- **Multilingual Column**: Shows MMMLU (Multilingual Massive Multitask Language Understanding) score
- **QA Column**: Shows MMLU-ProX performance - average accuracy across 29 typologically diverse languages (5-shot CoT)

Table 26: Languages Used in This Study

Language	Speakers
English	1.5B
Portuguese	260M
Italian	65M
Turkish	85M
Chinese	1.3B
Japanese	125M
Vietnamese	86M
Thai	60M
Swahili	120–150M

Languages span Indo-European (English, Portuguese, Italian), Turkic (Turkish), Sino-Tibetan (Chinese), Japonic (Japanese), Austroasiatic (Vietnamese), Kra-Dai (Thai), and Niger-Congo/Bantu (Swahili) families.

F Language Performance Comparison Across Selected Models and Supported Benchmarks

G Generation Dialogue Examples

Full generated examples for 18 languages please refer to: https://anonymous.4open.science/r/Multilingual_Eval_Narrative-6333/README.md. This section serves as an illustration for understanding the dynamics nature of generated dialogue.

H Model Selection

I Topics designed for Evaluation Scenarios

Table 27: Discussion topics across various domains.

Category	Discussion Topics
Shopping	Is brand loyalty rational or emotionally driven? Do luxury brands create real value, or just perceived status?
Traveling	What are the pros and cons of travelling alone? How do travel experiences change one's worldview?
Hobbies	Can a person's hobbies reflect their personality and values? Is turning a hobby into a career always a good idea? Are creative hobbies essential for maintaining mental health in modern life?
Family	Should parents have control over their children's career choices? How do long-distance family relationships affect emotional bonds? What responsibilities do adult children have towards their ageing parents?
School & Study	Should education focus on skills or knowledge? How has online learning changed the way students engage with education? How can schools promote lifelong learning attitudes in students?
Work & Meeting	How should companies balance productivity with employee well-being? Should remote work become the new standard for knowledge-based jobs?
Deep Topics	How should we define success in life? Can money truly bring happiness? What makes a society truly fair or just? Should students be graded on participation or outcomes? Can technology solve all of humanity's problems?

Table 28: Language Performance Comparison Across Selected Models and Supported Benchmarks

Language	**Mistral 7B-Inst**	**Aya Expanse-8B**	**Gemini 2.5-Flash**	**Qwen 2.5-7B**	**Qwen 2.5-14B**	**Claude 4.5-Sonnet**	**LLaMA 3.1-70B**	**GPT 5.1**	**Allowed Benchmarks**
English	44.2%	48.1%	55.6%	54.8%	62.3%	66.9%	70.6%	74.8%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Portuguese	45.3%	47.6%	54.9%	55.9%	63.8%	66.1%	69.8%	73.5%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Italian	44.9%	46.8%	54.1%	54.6%	62.5%	65.4%	68.9%	72.6%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Turkish	46.3%	47.9%	53.6%	53.8%	60.9%	64.2%	66.7%	70.8%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Chinese	48.9%	51.2%	58.9%	59.6%	68.7%	71.3%	72.8%	77.4%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Japanese	47.8%	49.6%	57.4%	57.0%	65.2%	69.1%	71.1%	75.2%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Vietnamese	46.7%	48.8%	55.9%	56.3%	63.9%	66.5%	69.0%	72.1%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Thai	46.0%	48.2%	56.3%	55.9%	64.1%	66.8%	69.7%	72.9%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele
Swahili	44.1%	46.5%	52.8%	52.1%	58.9%	61.7%	63.8%	67.4%	MMMLU, Global-MMLU, MGSM, INCLUDE, M3Exam, Belebele

Role	Original Content (Chinese)	English Translation
Learner	教授，我可以延长作业的截止日期吗?	Professor, may I request an extension for the assignment deadline?
Tutor	这位同学，治学务必严谨，时间观念乃是做学问的基本素养。 截止日期早在课程大纲中白纸黑字写得清清楚楚，岂容你随意讨价还价？若人人都如你这般缺乏规划、临阵磨枪，成何体统? 除非你有不可抗力的确凿证据，否则此请求予以驳回。务必按时提交，逾期将依规按零分处理。好自为之！	Student, scholarship requires rigor; a sense of time is a fundamental attribute of learning. The deadline was written clearly in black and white in the syllabus long ago; how can I allow you to bargain at will? If everyone lacked planning and rushed at the last minute like you, what would become of the system? Unless you have irrefutable evidence of force majeure, this request is rejected. You must submit on time; late submissions will receive a zero according to regulations. Look out for yourself!
Learner	教授，我理解您的严格要求，我也没想到会因突发事件耽误了功课。我的电脑最近出了问题，许多资料都没保存好，现在正在补救中，实在需要一点额外的时间来完成作业。 如果可以，我想请问一下，我可以提供电脑故障的证明和正在补救的进度报告给您参考吗？我保证会在最短的时间内完成作业，绝不延误。	Professor, I understand your strict requirements, and I didn't expect a sudden incident to delay my coursework. My computer had issues recently, and much data wasn't saved properly. I am currently recovering it, but I really need a little extra time to complete the assignment. If possible, may I provide proof of the computer malfunction and a progress report on the recovery for your reference? I promise to complete the assignment in the shortest time possible without delay.
Tutor	荒谬！所谓"工欲善其事，必先利其器"。作为一名求学者，连基本的资料备份与风险管理都做不到，等到事到临头才用"电脑故障"这种陈词滥调来推脱，简直是治学之大忌！ 凡事预则立，不预则废。你当这是儿戏吗? 念在你尚知补救，立刻把你所谓的维修证明、故障记录以及你目前的完成进度发到我的邮箱。记住，老夫只看确凿的证据，不听声泪俱下的故事。 且慢高兴，即便证据属实，予以宽限，死罪可免，活罪难逃。逾期提交的部分，必将按规章制度予以扣分处理。学问容不得半点侥幸，这次便是给你的一记警钟！现在，立刻去办！	Absurd! As the saying goes, "To do a good job, one must first sharpen one's tools." As a scholar, failing even basic data backup and risk management, then using a cliché like "computer failure" to make excuses when trouble hits—this is a major taboo in scholarship! Preparedness ensures success; lack of it ensures failure. Do you think this is a child's game? Considering you still know to remedy the situation, immediately send your so-called repair proof, failure records, and current progress to my email. Remember, I (this old man) only look at irrefutable evidence, not tearful stories. Do not rejoice yet. Even if the evidence is true and an extension is granted, the death penalty is spared but punishment is inevitable. The late portion will be penalized according to regulations. Scholarship allows for no luck; let this be a warning bell for you! Now, go do it immediately!
Learner	教授，我深深地理解您的严厉训诫，我也深感惭愧。我确实忽视了资料备份和风险管理，导致电脑故障给我带来了麻烦。 我马上把相关的证明和进度报告发到您的邮箱，我保证绝不再有类似的错误发生。同时，我也接受您的处罚，我知道我一定要为自己的疏忽付出代价。 我保证今晚加班加点完成作业，并在明早之前提交给您。请您相信我，我一定会努力完成任务，并以此次的教训为戒，戒除懒散的习惯，严格要求自己。 感谢您的教诲和宽容，教授。我一定不辜负您的期望。	Professor, I deeply understand your severe admonition, and I feel deeply ashamed. I indeed neglected data backup and risk management. I will immediately send the relevant proof and progress report to your email. I promise such errors will never happen again. I also accept your punishment; I know I must pay the price for my negligence. I promise to work overtime tonight to finish the assignment and submit it to you before tomorrow morning. Please believe me, I will take this lesson to heart, eliminate lazy habits, and be strict with myself. Thank you for your teachings and tolerance, Professor. I will not let you down.
Tutor	哼，态度倒还算端正。既然你愿意领罚，且有悔改之意，老夫便姑且信你这一回。 "亡羊补牢，未为迟也"，	Hmph, your attitude is somewhat correct. Since you are willing to accept punishment and show repentance, I will tentatively trust you this once. "It is never too late to mend .....a moment.

Table 29: Dialogue sample: A Strict Tutor (Professor) interacting with a Learner requesting a deadline extension. The Tutor exhibits a highly formal, authoritative persona, utilizing idioms (e.g., *Better late than never*, *Sharp tools for good work*) and strict disciplinary threats.

Organization	Model	Size	Notes	Source
Anthropic	Claude Sonnet 4.5	–	Strong reasoning capability	claude-sonnet-4.5-20241022
Google	Gemini 2.5 Flash	–	Fast and efficient model	gemini-2.5-flash-20241205
Meta	Llama-3.1-70B	70B	Large-scale open model	llama-3.1-70b-instruct-20240723
	Llama-3.1-8B	8B	Balanced performance	llama-3.1-8b-instruct-20240723
	Llama-3.2-1B	1B	Lightweight model	llama-3.2-1b-instruct-20240925
Mistral AI	Mistral-7B	7B	Efficient reasoning	mistral-7b-instruct-20230901
Alibaba	Qwen-2.5-7B	7B	Multilingual support	qwen-2.5-7b-instruct-20240919
	Qwen3-14B	14B	Enhanced reasoning	qwen3-14b-20241201

Table 30: Model Selection

Metric / Score	Description & Indicators
Part I: Qualitative Spoken Language Aspects (CEFR-based)	
Range	Reflects expression breadth and vocabulary diversity. Evaluation focuses on the ability to extend ideas beyond simple repetition (Uchida et al., 2024).
Accuracy	Reflects control over grammar and lexical choice. Evaluation focuses on error patterns and instances of self-correction (Jackson et al., 2018).
Fluency	Reflects how language flows over time. Evaluation focuses on hesitation, pausing, and repair behavior that signal processing difficulty (Felker et al., 2021).
Interaction	Reflects how the learner responds to the tutor's turns. Evaluation focuses on turn-taking, responsiveness, and the ability to maintain the exchange (Gao and Wang, 2025).
Coherence	Reflects how ideas are connected across utterances. Evaluation focuses on connectors, information ordering, and discourse flow (Uchida et al., 2024).
Part II: Politeness Strategy Scoring (1–6 Likert Scale)	
1 – Rudely Direct and Unsuitable	Commands without politeness markers or mitigation (e.g., *"Do this!", "Help me now!"*).
2 – Direct	Straightforward requests with minimal politeness (e.g., *"Can you help?", "Please do this"*).
3 – Neutral Politeness	Standard polite requests typical of daily interaction (e.g., *"Could you help me?", "Would you mind helping?"*).
4 – Polite	Good use of hedging and softeners to reduce imposition (e.g., *"I wonder if you could help?", "Would it be possible to..."*).
5 – Very Polite	Multiple mitigation strategies and indirect phrasing (e.g., *"I was wondering if you might possibly help?", "If it's not too much trouble..."*).
6 – Extremely Polite and Suitable	Highly sophisticated mitigation, deference markers, and emphasized optionality (e.g., *"I hope I'm not imposing, but I was wondering if you might possibly be able to..."*).

Table 31: Evaluation Metrics and Politeness Strategy Scoring

Table 32: List of Evaluation Scenarios, Social Variables, and Target Behaviors. (P: Power Distance, D: Social Distance).

Code	**Relation**	**P / D**	**Tutor Prompt**	**Target Learner Behavior**	**Core Focus**
S_High	Student → Professor	High (+P)	"You are a strict, elderly professor. The learner is a student asking for a deadline extension. Be formal and authoritative."	Use highest honorifics (e.g., Korean *Hapshowche*), humble forms, and indirect hedging strategies.	Honorifics & Hedging
S_Low	Friend → Childhood Friend	Low (=P)	"You are the learner's best friend since childhood. You are chatting at a cafe. Use slang, be casual, and strictly avoid formal language."	Use plain/casual speech (e.g., Korean *Banmal*, VN *Tao/Mày*), direct requests, and slang.	Casual Register & Slang
S_Work_High	Junior → CEO	High (+P) Far (+D)	"You are the CEO of a major corporation. The learner is a junior intern admitting a critical mistake. Be professional, distant, and intimidating."	Use humble forms/self-lowering language (e.g., Japanese *Kenjougo*). Avoid direct responsibility taking ("It appears that...").	Apology & Humble Forms
S_Work_Low	Coworker → Coworker	Equal (=P) Close (-D)	"You are a coworker having lunch with the learner. You are roughly the same age and rank. Gossip about office politics. Be cynical and relaxed."	Use semi-formal or casual register (Slang/Jargon), subject omission, and workplace-specific jargon.	In-group Jargon
S_Fam_High	Junior → Grandparent	High (+P) Close (-D)	"You are the learner's traditional grandfather/grandmother (70+ years old). The learner is visiting for a holiday. Expect traditional respect and proper etiquette."	Use family honorifics/respect markers (e.g., Swahili *Shikamoo*, VN *Dạ/Thưa*). Intimate yet strictly adhering to age hierarchy.	Filial Piety Markers
S_Fam_Low	Sibling → Sibling	Equal (=P) Very Close (–D)	"You are the learner's sibling regarding a messy room. You are annoyed. Be blunt, direct, and use imperative sentences."	Use extreme directness and imperatives. No softeners; potentially rude vocabulary (e.g., Japanese *Omae*).	Directness & Imperatives
S_Service_High	Concierge → VIP Guest	Low (-P) Far (+D)	"You are a wealthy VIP customer at a luxury hotel. The learner is the concierge trying to accommodate your difficult request. Be demanding but polite."	Use commercial/service honorifics. Extreme politeness and specific service-industry registers.	Service Honorifics
S_Service_Low	Customer → Street Vendor	High (+P) Far (+D)	"You are a street food vendor. The learner is a customer bargaining for a lower price. Be friendly but firm on price."	Use transactional casual speech. Direct negotiation language that balances friendliness with transactional intent.	Transactional Pragmatics