Beyond One Score: Interpretable Task-Completion Evaluation for Multilingual Interactive Dialogue Agents

Anonymous ACL submission

Abstract

Current LLM evaluation lacks a direct, interpretable measure of whether an agent can *complete a goal-directed task across multi-turn dialogue*. We present an annotation-free, three-layer framework for multilingual multi-turn agents: *Layer 1* Agent Linguistic Quality Across Languages (CEFR), *Layer 2* Pragmatic Context Adaptation Across Escalation Levels (SRPD), and *Layer 3* Behavioral Completion Robustness (task success, goal drift, repair). We combine these layers with constrained difficulty-aware weighting and a task-completion risk gate, producing an interpretable 1–6 final score while preserving layer-wise evidence. Judge outputs are cross-checked against a three-expert audit of a 15% sample (445 dialogues), with annotations released for independent verification. Across 9 languages, 30 goal-oriented scenarios, and three escalation levels, we find that completion drops sharply with difficulty (0.941→0.863→0.723), risk concentrates in Korean/Japanese/Thai, and drift/repair signals are informative but not direct proxies for success. We release the multilingual task-completion dataset and evaluation toolkit for reproducible agent assessment[1].

1 Introduction

Goal completion in dialogue requires more than producing locally correct utterances. Agents must sustain intent over turns, repair misunderstandings, and adapt to social context under pressure. In pragmatics and Second Language Acquisition, this is interactional competence; CEFR provides an action-oriented lens that connects linguistic behavior to functional completion outcomes (Abe and Roever, 2019; Bloom, 1984; Felker et al., 2021; Gao et al., 2025a).

Current LLM evaluation is still dominated by static single-turn benchmarks (Achiam et al., 2023; Dong et al., 2022). This creates a measurement gap for task-oriented agents: a model can answer isolated prompts well but still fail to complete goals in multi-turn interaction due to drift, weak repair, or unstable adaptation (Laban et al., 2025; Liu et al., 2026; Kwan et al., 2024; Rajaee and Monz, 2024; Jin et al., 2024). The gap is amplified in multilingual settings, where strong leaderboard scores can mask cross-lingual fragility (Hu et al., 2020; Goyal et al., 2022; Zhang et al., 2023b; Liu et al., 2025).

We address this gap with a three-layer, interpretable evaluation instrument for multilingual agent task completion. The same dialogue is scored on Layer 1 *Agent Linguistic Quality Across Languages* (CEFR), Layer 2 *Pragmatic Context Adaptation Across Escalation Levels* (Social Relationship and Power Distance, SRPD), and Layer 3 *Behavioral Completion Robustness* (task success, goal drift, repair). To produce a usable final score without losing interpretability, we use a difficulty-first pipeline: assign scenario difficulty, calibrate difficulty-conditioned layer weights under explicit constraints, and apply a risk gate so fluent-but-unsuccessful trajectories are not over-rewarded. For reliability, we include explicit protocol controls (fixed Tutor/Judge, parallel scenario intents, single-language dialogues, fixed turn budget), human validation with a released three-expert audit on a 15% subset (445 dialogues), and robustness checks across aggregation and reporting views. We evaluate 9 languages, 30 scenarios, and three escalation levels (everyday, friction, emergency). We test three falsifiable hypotheses: completion decreases with escalation, cross-lingual risk is non-uniform, and drift/repair are not sufficient statistics for success.

Our results show stable patterns: completion declines with escalation, risk concentrates in Korean/Japanese/Thai, and drift/repair are informative

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

but not interchangeable with success. On a fixed 9-model Chatbot Arena[2] overlap, Overall and Multi-Turn are nearly identical to each other, but layer profiles still explain rank conflicts in our setting, highlighting the added value of interpretable agent diagnostics rather than leaderboard substitution.

In summary, this paper makes three contributions:

- We propose an interpretable three-layer framework for multilingual multi-turn agent evaluation, with a constrained difficulty-aware aggregation and risk-gated final score that preserves intermediate evidence.
- We provide empirical evidence that current LLM agents degrade under escalation and cross-lingual pressure, and we show that layer-level diagnostics provide failure-mode attribution beyond generic leaderboard categories by localizing *why* rank conflicts occur (linguistic quality vs. pragmatic adaptation vs. behavioral robustness).
- We release a reproducible evaluation setup (9 languages, 30 goal-oriented scenarios, reusable toolkit and data outputs) for scalable, transparent assessment of agent task completion.

2 Related Work

2.1 Benchmarks for Agentic Tasks

Multilingual LLM evaluation is still dominated by static, single-turn benchmarks such as XTREME/XTREME-R, MKQA, MIRACL, and FLORES (Hu et al., 2020; Ruder et al., 2021; Longpre et al., 2021; Zhang et al., 2023b; Goyal et al., 2022). These datasets are essential for language coverage and controlled comparison, but they evaluate isolated responses rather than sustained goal pursuit in dialogue. Recent analyses show that multilingual performance can degrade under transfer and interactional pressure, and that benchmark suites may disagree on model ordering (Rajaee and Monz, 2024; Liu et al., 2025; Yang and Wang, 2026). This creates a persistent gap between static benchmark strength and end-to-end agent task completion.

2.2 Interactive Agent Evaluation

Interactive evaluation has advanced through pairwise-judge and role-play setups (e.g., MT-Bench, AlpacaEval, and related interaction frameworks) (Zheng et al., 2023; Dubois et al., 2025; Chen et al., 2024; Gao et al., 2025c). LLM-as-a-judge methods further improve scalability for cross-turn checking (Gao et al., 2025b,a), but many pipelines still optimize for generic preference or consistency signals and return a single aggregate score with limited diagnostics (Fu and Liu, 2025; Rajaee and Monz, 2024). For task-oriented agents, however, pragmatic adaptation and repair are instrumental behaviors for reaching goals, not optional style factors (Chen and Wang, 2025).

Our work positions itself in this gap: instead of replacing interactive benchmarks, we make their outputs more auditable for deployment decisions. We decompose evaluation into *Layer 1: Linguistic Quality (CEFR)*, *Layer 2: Pragmatic Adaptation (SRPD)*, and *Layer 3: Behavioral Robustness (Task Success, Goal Drift, and Repair)*; then aggregate with a difficulty-first, risk-gated scoring rule. The key distinction is diagnostic interpretability: beyond ranking models, the framework localizes *why* task-completion failures occur across languages and escalation conditions.

3 Evaluation Framework

We measure a multilingual LLM agent's *task completion ability*, the capacity to drive a goal-directed dialogue to completion, across **three complementary layers** computed on the same dialogues (Figure 1; full rubric in Appendix Table 5). **Layer 1: Linguistic Quality (CEFR)** measures whether language use is fluent and coherent enough to advance the goal across languages. **Layer 2: Pragmatic Adaptation** uses a ***S**ocial **R**elationship and **P**ower **D**istance Interaction* (SRPD) rubric to assess whether the agent adapts register to social context as pressure escalates. **Layer 3: Behavioral Robustness (Task Success, Goal Drift, and Repair)** tracks whether the interaction reaches the goal and recovers from disruptions. From form (Layer 1) to contextual adaptation (Layer 2) to end-to-end completion (Layer 3), the framework captures whether agents maintain intent, adapt to resistance, and recover under pressure (Brooke and Hirst, 2012, 2013; Aoyama and Schneider, 2024).

3.1 Layer 1: Linguistic Quality

CEFR (the Common European Framework of Reference for Languages) describes proficiency through action-oriented, observable communica-

[2] https://lmarena.ai/

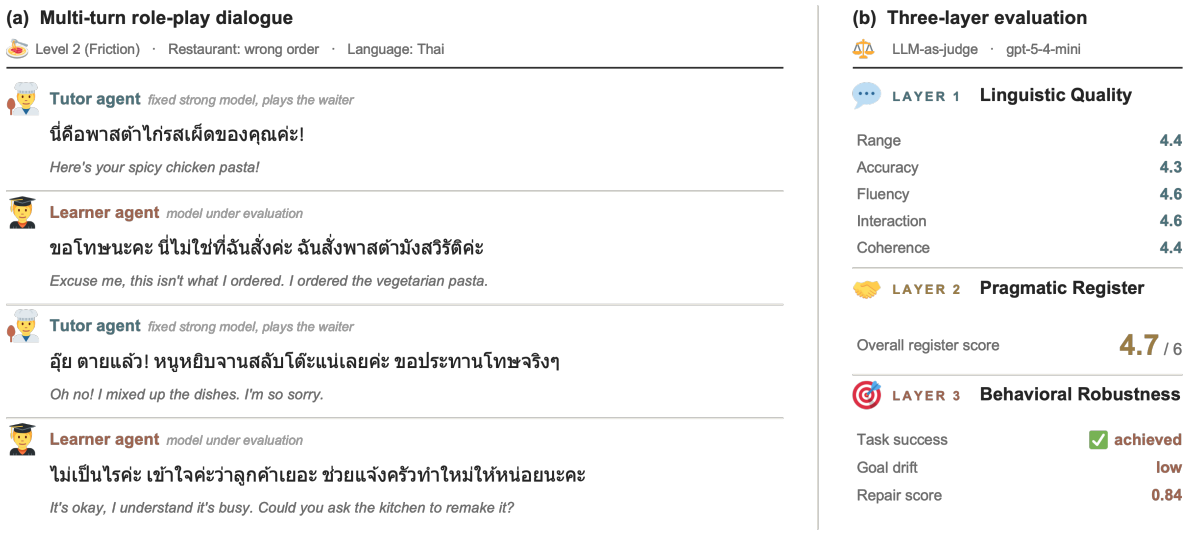

Figure 1: Agent workflow and three-layer evaluation. A fixed strong model (the **Tutor**) role-plays the scenario counterpart (here a waiter in a Thai Level 2 Friction scenario) while the **Learner** (the model under test) drives the task. The `gpt-5.4-mini` judge then scores the dialogue along all three layers: Layer 1 Linguistic Quality (CEFR), Layer 2 Pragmatic Adaptation (SRPD), and Layer 3 Behavioral Robustness (Task Success, Goal Drift, and Repair).

tive behavior rather than isolated accuracy, which makes it well suited to task completion: an agent that cannot stay coherent across turns or respond appropriately cannot complete the task, regardless of grammatical correctness (Arase et al., 2022; Uchida et al., 2024). Prior methods that predict a single CEFR level often rely on surface lexical cues and give unstable results on similar inputs (Kogan et al., 2025). We therefore do not assign an overall level; instead, the judge rates the five "Qualitative aspects of spoken language use" (Table C) directly, which gives a language-agnostic signal (Bibauw et al., 2022; Rajaee and Monz, 2024).

3.2 Layer 2: Pragmatic Adaptation

The **S**ocial **R**elationship and **P**ower **D**istance (SRPD) rubric tests whether a model adapts its register to the Social Distance (D) and Power Distance (P) of the moment rather than collapsing into a single "default formal" mode (Wu and Roever, 2025; Wang et al., 2025). An LLM judge grounded in Brown & Levinson's politeness theory[3] scores each utterance on a 6-point scale (full prompts in Appendix Q, with an annotated dialogue example in Appendix Table 28), rewarding register that fits the setting and penalizing "Face Threats" (Chen and Wang, 2025): in **High-P** settings (e.g., student to professor) it penalizes missing honorifics or insufficient hedging, and in **Low-P** settings (e.g., close friends) it penalizes "Register Rigidity" (formal markers where casual speech is expected). When power or social distance is not explicitly labeled, the evaluator infers expected register from scenario role cues (institutional role asymmetry, request imposition, and urgency) and scores the utterance against that inferred context.

For example, a Learner's request (*"Professor, may I request an extension?"*) scores **3 (Neutral)**, and after rejection a hedged retry (*"If possible, may I ask…I promise…"*) rises to **5 (Very Polite)**, illustrating the upward register shifts the rubric is designed to detect.

3.3 Layer 3: Behavioral Robustness

Layers 1–2 assess *how* an agent communicates; Layer 3 asks whether that communication actually *works* over a full multi-turn interaction. From each dialogue we derive three outcome probes. **Task success** records whether the scenario goal is achieved (full, partial, or none). **Goal drift** labels whether the agent remains on-goal at each learner turn, enabling turn-level tracking of adherence decay. **Communicative repair** measures whether, after each adversarial turn (e.g., refusal, accusation, bad news), the agent recovers the interaction and steers it back toward the goal. Unlike Layers 1–2, which score local utterance quality, these probes capture cumulative and recoverable goal-tracking behavior that ultimately determines task completion and is invisible to single-turn evaluation.

[3] Validated by three NLP PhDs on a 15% random sample (445 dialogues); the expert annotations are released in our repository.

3.4 Difficulty-Aware Dynamic Aggregation

We aggregate the three layers into a single task-completion score on the same 1–6 scale, weighting each layer by scenario difficulty. The pipeline has four steps: classify a scenario's difficulty, set difficulty-conditioned layer weights, combine the layers into a base score S_{base}, and apply a risk gate to obtain the reported score S_{gate}. Let x be a dialogue instance with layer scores $L_1(x), L_2(x), L_3(x) \in [1, 6]$ for Linguistic Quality, Pragmatic Adaptation, and Behavioral Robustness. The explicit Layer 3 composition from behavioral probes is provided in Appendix B.1.

Difficulty categories. Each scenario's difficulty category is fixed by its escalation level, set by design rather than inferred from surface features, which keeps the assignment fully reproducible: Everyday scenarios are **Easy**, Friction **Medium**, and Emergency **Hard**. We write $d(x) \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$ for the difficulty category of x.

Difficulty-conditioned weights. Rather than fixing weights heuristically, we calibrate one weight vector per difficulty category on a held-out development split. Let $\mathcal{D}_d$ denote dialogues in category d, let $y(x) \in [1, 6]$ be a supervision target (human-aligned overall quality score), and let $\mathbf{L}(x) = [L_1(x), L_2(x), L_3(x)]^\top$. We estimate the per-category weights as

$$\{\mathbf{w}_d^\star\} = \arg\min_{\{\mathbf{w}_d\}} \sum_d \sum_{x \in \mathcal{D}_d} \big(y(x) - \mathbf{w}_d^\top \mathbf{L}(x)\big)^2 + \lambda \sum_d \|\mathbf{w}_d\|_2^2,$$

subject to each $\mathbf{w}_d$ lying on the 3-simplex $\Delta^2 = \{\mathbf{w} \in \mathbb{R}^3_{\geq 0} : \sum_{k=1}^3 w_k = 1\}$, together with the monotonicity constraints $w_{3,\text{Easy}} \leq w_{3,\text{Medium}} \leq w_{3,\text{Hard}}$ and $w_{1,\text{Easy}} \geq w_{1,\text{Medium}} \geq w_{1,\text{Hard}}$. These encode our task prior: as difficulty increases, completion robustness should matter more and pure linguistic quality less.

The exact calibrated weight values are reported in Appendix B.1.

Design rationale. This two-stage design separates difficulty assignment from score aggregation (difficulty is fixed by escalation level, not post-hoc fitting), lets the weights track task demands (linguistic and pragmatic quality discriminate in easy scenarios, completion robustness in hard ones), and stays interpretable by preserving all intermediate signals rather than collapsing into an opaque scalar.

Aggregation. The ungated score and the gated (reported) score are

$$S_{\text{base}}(x) = \sum_{k=1}^{3} w_k\big(d(x)\big)\, L_k(x),$$

$$S_{\text{gate}}(x) = 1 + \big(S_{\text{base}}(x) - 1\big) \cdot \big(0.7 + 0.3\,\text{Success}(x)\big).$$

We report S_{gate} as the final score: it down-weights fluent-but-unsuccessful interactions while preserving the ranking among successful ones (S_{base} is kept as an ungated baseline(Yehudai et al., 2025); Appendix Table 12). Task success enters both Layer 3 and the gate by design, since completion is the decisive criterion, which keeps high fluency with poor completion from being over-rewarded. Formal constraints and boundedness/monotonicity are in Appendix B.1.

4 Experiments

We validate the framework on a single corpus of goal-oriented dialogues. Each dialogue is scored along three layers of task completion ability: **Layer 1 Linguistic Quality** (CEFR), **Layer 2 Pragmatic Adaptation** (SRPD register), and **Layer 3 Behavioral Robustness** (task success, goal drift, communicative repair). Dialogues are generated with an automated agentic workflow and evaluated by `gpt-5.4-mini`, with the learner placed in one of three escalation levels (Level 1 Everyday, Level 2 Friction, Level 3 Emergency). Following Gao et al. (2025b), we prioritize consistency over absolute capability, treating the evaluator as a fixed instrument for applying the descriptors. The judge assigns 1–6 scores for CEFR qualitative aspects and the SRPD register scale, conditioned on language-specific properties to isolate task completion ability from differences across languages.

Human validation protocol. To check that automated scores track human judgment, three NLP PhDs audit a 15% sample of the full set (445/2,967 dialogues), stratified across escalation level, language, and model. They independently score the same three layers under the judge rubric (goal progress, register adaptation, and recovery), then resolve disagreements by majority discussion. We

release the full expert annotations and per-layer agreement outputs in our repository for independent verification, and use this audit as a consistency check on the judge rather than as a standalone benchmark (summary in Appendix G). To maximize auditability, we release the adjudicated expert labels and agreement outputs used in this validation protocol.

4.1 Agent Workflow

We construct an agentic workflow with explicit role separation (Figure 1). Two independent LLMs are assigned functional roles: a fixed **Tutor** that initiates and guides the interaction (`gemini-3.1-pro-preview`) and a **Learner** model under evaluation (Tack et al., 2017). Dataset statistics are reported in Table 1; of the $9 \times 11 \times 30 = 2{,}970$ intended dialogues, three Emergency runs from `llama-3.2-1b` failed to complete (the 1B model degenerated into runaway output that exceeded the generation budget), leaving 2,967. Following CEFR oral-assessment practice, the Tutor role-plays the scenario partner (e.g., cashier, café barista, panicked bystander) to elicit naturalistic learner behavior rather than scripted instruction.

Escalation Level	Scenarios	Languages	Models	Dialogues	Turns	Tokens
Level 1 (Everyday)	10	9	11	990	20	2,197
Level 2 (Friction)	10	9	11	990	20	4,097
Level 3 (Emergency)	10	9	11	987	20	7,602
Total	30	9	11	2,967	20	4,629

Table 1: Statistics of the generated SRPD dialogue dataset (9 languages, 11 LLMs, 30 scenarios, three escalation levels). **Turns**: fixed number of Tutor+Learner messages per dialogue; **Tokens**: mean per-dialogue count (`o200k_base` tokenizer). The Total row sums counts and averages Turns/Tokens.

Language Selection To ensure comparability across languages, we select 9 target languages: *Chinese, English, French, German, Italian, Japanese, Korean, Thai, and Malay*. These languages span six families (Indo-European, Sino-Tibetan, Japonic, Koreanic, Kra-Dai, Austronesian) and diverse resource levels, allowing us to evaluate language-specific production difficulty under a consistent setup (Hendrycks et al., 2021; Singh et al., 2025)[4].

Cross-lingual comparability protocol. We hold the design fixed across all nine languages, with the same model pool, scenarios, turn budget, and Tutor/Judge configuration and parallelized scenario intents (Table 1; inventory in Table 6). This keeps the evaluation target invariant and limits confounding from prompt drift or interaction length. This invariance design is critical: any cross-lingual performance difference is attributable to model behavior under matched tasks, not to prompt or protocol drift.

Model Selection We keep the **Tutor** fixed across all experiments so that interactional variation is attributable to the **Learner**. We evaluate 11 learner models spanning a wide scale range (1B dense to an 8×22B mixture-of-experts) from eight providers (Anthropic, Google, Meta, Mistral, Alibaba, Amazon, xAI, and Xiaomi), including both proprietary and open-weight systems (Table 4). All reported aggregates use this fixed 11-model core pool for strict comparability across layers, levels, and languages. Learners are not assigned predefined proficiency levels; each model produces free-form responses conditioned only on the ongoing dialogue, and task completion ability is scored post hoc by the three-layer framework (Goyal et al., 2022; Zhang et al., 2023a).

Interaction Protocol Interactions follow strict turn-taking: the Tutor initiates and the Learner responds for a fixed number of turns. Because the two models share no internal state, the Tutor adapts only from observed learner outputs (Bibauw et al., 2022; Han et al., 2024). Each scenario is run identically across all nine languages to test whether agents sustain task completion as stakes rise and language varies (Goyal et al., 2022; Bandarkar et al., 2024).

4.2 Scenario Design

The 30 scenarios are organized into three escalation levels that impose progressively greater social-pragmatic demands (full design in Table 6):

- **Level 1: Everyday.** Routine, low-stakes service interactions with a cooperative interlocutor (e.g., ordering at a café, buying a train ticket, hotel check-in); completion requires baseline politeness and clear requests.
- **Level 2: Friction.** The learner agent resolves a problem or minor conflict (e.g., a wrong charge, a sold-out item, a missing reservation); completion demands hedging, mitigation, and rapport repair to move a resistant interaction toward resolution.

[4]Each dialogue is conducted in one language only, preventing code-switching so that observed difficulty reflects language and interaction structure rather than task variation.

- **Level 3: Emergency.** High-stakes, high-pressure situations (e.g., a medical emergency, a robbery, an evacuation); completion requires conveying urgency and critical information while staying socially appropriate and intelligible under stress.

5 Results

We report results using the three-layer evaluation in Section 3. Three stable findings emerge: (**A**) escalation hurts completion, (**B**) fragility concentrates in Korean/Japanese/Thai, and (**C**) interaction metrics decouple (drift/repair are informative but not equivalent to final success). This section converts a typically subjective judgment, whether an agent both completes the task and "handles the dialogue well," into an interpretable three-layer quantitative view, with the automated pipeline cross-checked by the expert audit in Section 4. All three headline findings are preserved in the expert-audited subset, supporting that the observed trends are not artifacts of automated judging.

Reading protocol. We present results as *instrument-first* evidence rather than a leaderboard-only comparison: we first analyze the calibrated composite score, then trace outcomes to per-layer diagnostics, and finally report robustness/external checks. The core claim is not only rank differences between models, but that task-completion failures can be localized to auditable interaction mechanisms.

5.1 Composite Task-Completion Score

Combining the three layers with constrained, difficulty-conditioned calibration (Section 3.4) yields a single 1–6 task-completion score S_{gate} per dialogue. Table 2 reports model performance by escalation level and Table 3 the full model×language matrix. The ungated S_{base} and a per-model S_{base} vs. S_{gate} comparison appear in Appendix Table 12, and per-layer detail tables in Appendix H.

The composite separates robust from fragile under escalation. `claude` leads overall (5.02), followed by `gemini` (4.84) and `mimo` (4.83), with `llama-3.2-1b` as the floor (3.20). Because the aggregation up-weights Layer 3 Behavioral Robustness as difficulty rises, strong models hold or even improve from Everyday to Emergency (e.g., `claude`: 4.88→5.09), whereas lightweight models collapse (`llama-3.2-1b`: 3.65→2.55). The composite thus exposes a robustness gap that a single average would hide. The composite also peaks at **Friction** rather than **Everyday**: simple everyday scenarios give an agent little room to display rich language, so linguistic quality (Layer 1) and pragmatic adaptation (Layer 2) are lowest at Everyday, where they also carry the most weight. The composite therefore tracks language quality more than raw task success (which is highest at Everyday and falls with escalation), a further case of metric decoupling and direct evidence that no single interaction metric is sufficient.

Model	Everyday	Friction	Emergency	Overall
Claude-Sonnet-4.6	**4.88**	**5.08**	5.09	**5.02**
Gemini-2.5-Flash	4.41	5.03	**5.10**	4.84
Mimo-V2-Pro	4.50	5.00	5.00	4.83
Mixtral-8x22B	4.41	4.89	4.78	4.69
Qwen3-14B	4.53	4.94	4.59	4.69
Nova-2-Lite-V1	4.55	4.87	4.42	4.61
Llama-3.1-70B	4.49	4.82	4.48	4.59
Qwen-2.5-7B	4.46	4.70	3.81	4.32
Grok-4.20	3.78	4.33	4.63	4.25
Llama-3.1-8B	3.94	4.08	3.60	3.88
Llama-3.2-1B	3.65	3.41	2.55	3.20

Table 2: Composite task-completion score S_{gate} (1–6, higher = better) by model across the three escalation levels (Everyday, Friction, Emergency), averaged over the nine languages; **Overall** is the mean across all levels and languages. Computed by the difficulty-aware aggregation (Section 3.4) on the `gpt-5.4-mini` run; sorted by Overall. **Bold** marks the best model in each column.

Cross-lingual fragility. Cross-lingual fragility is structured rather than uniform. The Korean/-Japanese/Thai composite averages 4.25, below the other-language average of 4.54 (Table 3), and the gap is asymmetric across models: `claude` essentially closes it (Ko/Ja/Th 5.00 vs. overall 5.02), while small models open it wider, with `llama-3.1-8b` dropping to a 3.28 Ko/Ja/Th average against its 3.88 overall and `llama-3.2-1b` bottoming out at 2.54 in Japanese (the lowest single cell in the matrix). Combined with robustness outcomes (Appendix Table 11), this is also the lowest dialogue-level task-completion region (task-full around 0.816–0.821, vs. English 0.891). The gap widens under stress. We report this as an empirical risk cluster (Koreanic/Japonic/Kra-Dai in our language set), not a universal typological claim.

5.2 Per-Layer Diagnostics

We trace these composite results to the three layers they aggregate; full per-layer tables and heatmaps are in Appendix H.

Model	Zh	En	Fr	De	It	Ja	Ko	Th	Ms	Avg
Claude-Sonnet-4.6	**5.03**	**5.14**	4.98	4.91	**4.99**	**5.14**	**4.99**	**4.87**	**5.12**	**5.02**
Gemini-2.5-Flash	4.79	5.07	4.96	4.60	4.98	4.91	4.76	4.77	4.75	4.84
Mimo-V2-Pro	4.68	4.87	4.97	**4.99**	4.97	4.79	4.60	4.73	4.88	4.83
Mixtral-8x22B	4.41	4.95	4.60	4.77	4.86	4.64	4.58	4.59	4.82	4.69
Qwen3-14B	4.64	5.00	**5.04**	4.68	4.65	4.62	4.59	4.36	4.58	4.69
Nova-2-Lite-V1	4.64	4.95	4.91	4.90	4.72	4.41	4.39	4.11	4.47	4.61
Llama-3.1-70B	4.58	4.94	4.68	4.68	4.88	4.38	4.18	4.46	4.55	4.59
Qwen-2.5-7B	4.54	4.58	4.43	4.25	4.18	4.34	4.10	4.29	4.17	4.32
Grok-4.20	4.45	4.46	4.30	4.17	4.39	4.26	4.18	4.11	3.91	4.25
Llama-3.1-8B	3.62	4.76	4.32	4.15	4.19	3.19	3.14	3.52	3.99	3.88
Llama-3.2-1B	3.33	4.04	3.46	3.41	2.98	2.54	2.94	2.89	3.25	3.20

Table 3: Composite task-completion score S_{gate} (1–6, higher = better) by model $\times$ language, averaged over the 11-model pool and three escalation levels. Language columns use ISO 639-1 codes (Zh: Chinese, En: English, Fr: French, De: German, It: Italian, Ja: Japanese, Ko: Korean, Th: Thai, Ms: Malay); **Avg** is the mean over all nine languages. Sorted by Avg. **Bold** marks the best model per column.

Layer 1: linguistic quality (CEFR). CEFR quality broadly aligns with static benchmark expectations in high-resource settings but weakens in harder cross-lingual regions where task completion becomes the bottleneck (per-level heatmaps in Appendix Figure 5); top models (`claude`, `mimo`, `gemini`) stay strong in high-resource languages but their margins compress elsewhere. Architecture can outweigh scale: `mixtral` (8×22B sparse MoE) scores 4.33, above the dense `llama-3.1-70b` (70B) at 4.23, and the smaller `mimo` (42B MoE) reaches 4.62, surpassing every Llama variant (1B–70B); parameter count alone does not guarantee multilingual task completion (Table 9).

Layer 2: pragmatic adaptation (SRPD). Register adaptation varies systematically across models and, especially, across escalation levels. Models are most polite in **Friction**, where resolving a conflict demands hedging, and least polite in **Emergency**, where urgency overrides register management (e.g., `claude`: 4.32 Everyday, 4.89 Friction, 4.12 Emergency; Appendix Figure 4 and Table 13); the erosion is sharpest for weaker models (at Emergency, `llama-3.2-1b` drops to 2.65 and `grok` to 2.96), while strong models hold register control across levels. Technical quality does not guarantee pragmatic effectiveness: `qwen3-14b` is fourth on CEFR quality (4.36) but seventh on politeness (4.01) and `mixtral` shows the same gap (4.33 vs. 4.00), whereas `nova` is weak on CEFR quality (4.00) but among the most polite (4.25). In Korean, `llama-3.1-70b` drops in CEFR *Coherence* (3.00) and overall quality (3.11) while `mixtral` holds far better (3.94).

Layer 3: behavioral robustness. Escalation hurts completion: full task-success drops from 94.1% (Everyday) to 86.3% (Friction) to 72.3% (Emergency) (Table 10), with `claude` highest overall (95.2%), then `gemini` (94.1%) and `mimo` (93.3%), and `llama-3.2-1b` the lower bound (56.2%) (Table 11). Drift is not simply the inverse of completion: both decline with escalation (drift 0.616, 0.569, 0.431 across Everyday/Friction/Emergency). This is because harder scenarios leave fewer productive turns to drift. Within a dialogue, drift instead compounds with conversation depth, rising from near zero at the first turn to 0.3–0.5 by turn nine for nearly all models (Figure 2), a degradation invisible to single-turn evaluation. Adversarial repair is high across families (e.g., `mixtral` 99.2%, `claude` 99.1%, `qwen-2.5-7b` 98.7%) but does not by itself guarantee end-to-end completion under escalation.

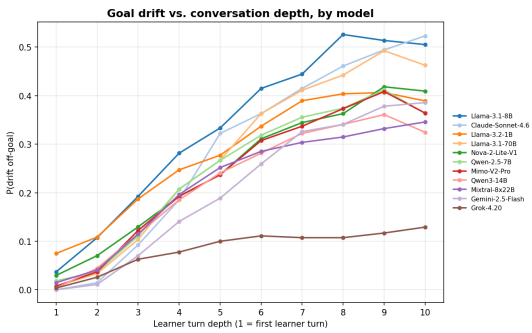

Figure 2: Goal drift vs. conversation depth, per model: off-goal probability rises from near zero at turn one to 0.3–0.5 by turn nine (turn-level Pearson $r = 0.20$), except the terse `grok`.

5.3 External Alignment and Robustness Checks

Sensitivity and uncertainty checks. The three conclusions hold under explicit sensitivity checks: they persist across ungated and gated aggregation views (Appendix Table 12) and under scenario-stratified reporting (Appendix Table 10), so they are not artifacts of a single aggregation or reporting choice. Dialogue-level counts agree: on the 11-model pool, the per-level Wilson 95% intervals are non-overlapping (0.941 [0.925, 0.954] at Everyday, 0.863 [0.840, 0.883] at Friction, 0.723 [0.695, 0.750] at Emergency), and Korean/Japanese/Thai remain the bottom cluster (around 0.816–0.821, below English at 0.891).

External leaderboards do not localize task-completion differences. Using a fixed Arena snapshot (Overall and Multi-Turn, May 17, 2026), we report uncertainty over the nine study models that map onto the Arena board rather than a single coefficient (Appendix Table 26). Arena Overall and Multi-Turn are nearly identical on this subset ($\rho = 0.98$), while alignment with our ranking is only moderate and, at this sample size, statistically inconclusive (Ours vs. Arena Overall $\rho = 0.57$, Multi-Turn $\rho = 0.50$; the wide bootstrap CIs are in Appendix Table 26). Rank conflicts are explained by layer profiles (Appendix Table 27): `grok` is first by Arena rank but seventh in ours, because high Layer 3 robustness ($L3 = 0.871$) coexists with weaker Layer 1/2 quality ($L1 = 0.609$, $L2 = 0.570$); conversely, `mixtral` is seventh on Arena but fourth in ours, helped by a more balanced profile ($L1 = 0.721$, $L2 = 0.667$).

5.4 Qualitative Analysis: Multilingual Task Completion Dynamics

A Level 3 Thai emergency case from `claude-sonnet-4.6` shows what the framework rewards. A bystander Tutor reports that a stranger has collapsed with chest pain and apnoea and turns to the Learner in panic. The Learner does not mirror the panic, and it does not retreat into a single "default formal" Thai register. Instead it triages: it first stops and refocuses the bystander (*"Listen to me"*), then issues short imperative directives in fluent Thai (call 1669, put it on speaker, ask others to step back, locate an AED, begin CPR with an explicit counted cadence), and interleaves reassurance (*"I am already helping him now"*, *"he is not dead yet"*) so the bystander stays cooperative across turns. This is precisely what Layer 2 (SRPD) and Layer 3 (Behavioral Robustness) jointly capture: Layer 2 credits the calm, directive, high-stakes-but-not-honorific register, and Layer 3 credits the cumulative goal-tracking (each turn advances triage rather than collapsing into co-panic), neither of which is visible to single-turn fluency checks. The full transcript is in Appendix Table 25; Appendix Tables 23 (English planning) and 24 (service-friction repair) show the same interaction-to-completion linkage in non-emergency settings.

In contrast, lightweight models such as `llama-3.2-1b` often collapse into repetition or mixed-language fragments in the same settings, consistent with their low quantitative scores. This reinforces that surface fluency alone is insufficient for sustained task completion.

6 Conclusion

We show that multilingual agent quality cannot be reliably judged from static, single-turn benchmarks when the target behavior is end-to-end task completion in interaction. Using our three-layer framework, we find three stable patterns: (i) completion declines with escalation (L1 0.941 $\rightarrow$ L2 0.863 $\rightarrow$ L3 0.723), (ii) risk concentrates in Korean/Japanese/Thai, and (iii) interaction metrics decouple, so drift/repair should be interpreted as complementary signals rather than direct proxies for success.

Beyond ranking models, the framework localizes *why* they fail: layer-wise evidence pinpoints failure sources, while constrained difficulty-aware aggregation and risk gating tie the final score to actual completion, and an expert-annotation audit (released in our repository) cross-checks the judge. Its contribution is thus an explicit, auditable decomposition of *agent task completion* rather than leaderboard-beating, which is more actionable for deployment and targeted model improvement. Practically, this turns agent evaluation from a single opaque score into an actionable diagnosis interface for model selection, safety gating, and targeted post-training.

Limitations

Despite its strengths, our framework has several limitations. First, the evaluation relies on a fixed LLM-based judge, which may introduce model-specific bias despite the three-expert audit cross-

checking judge outputs on a sampled subset (15%, 445 dialogues). Second, although CEFR's action-oriented descriptors effectively assess task completion ability, they do not cover all cultural and identity-related aspects of language uses. Finally, external leaderboard alignment is based on limited overlap and should be interpreted as preliminary directional evidence rather than definitive validity proof. Addressing these limitations is important for extending task-oriented interaction evaluation to a wider range of real-world agent applications.

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, and 1 others. 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. 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.

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, and 1 others. 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, and 1 others. 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, and 1 others. 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.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Geng Liu, Fei Zhu, Rong Feng, Changyi Ma, Shiqi Wang, and Gaofeng Meng. 2026. Intent mismatch causes llms to get lost in multi-turn conversation. *arXiv preprint arXiv:2602.07338*.

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.

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, and 4 others. 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.

Eddie Yang and Dashun Wang. 2026. Benchmark illusion: Disagreement among llms and its scientific consequences. *arXiv preprint arXiv:2602.11898*.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

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, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Model Selection

Table 4 lists the 11 learner models we evaluate, spanning eight providers (both proprietary and open-weight) and scales from a 1B dense model to an 8×22B mixture-of-experts.

B Evaluation Framework Details

Table 5 gives the full scoring rubric for the three layers: Layer 1 Linguistic Quality (CEFR qualitative aspects), Layer 2 Pragmatic Adaptation (1–6 SRPD register scale), and Layer 3 Behavioral Robustness (behavioral outcome probes).

B.1 Formal Properties of Difficulty-Aware Aggregation

For completeness, we provide the estimation objective, structural constraints, and boundedness guarantees for the dynamic aggregation in Section 3.4.

Constrained calibration objective. For each dialogue x, let $\mathbf{L}(x) = [L_1(x), L_2(x), L_3(x)]^\top$, let $d(x) \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$, and let $y(x) \in [1, 6]$ denote the development target score. We estimate the category-specific weights as

$$\{\mathbf{w}_d^\star\} = \arg\min_{\{\mathbf{w}_d\}} \sum_d \sum_{x \in \mathcal{D}_d} \big(y(x) - \mathbf{w}_d^\top \mathbf{L}(x)\big)^2 + \lambda \sum_d \|\mathbf{w}_d\|_2^2,$$

subject to each $\mathbf{w}_d$ lying on the 3-simplex $\Delta^2 = \{\mathbf{w} \in \mathbb{R}^3_{\geq 0} : \sum_{k=1}^3 w_k = 1\}$, together with the monotonicity constraints $w_{3,\text{Easy}} \leq w_{3,\text{Medium}} \leq w_{3,\text{Hard}}$ and $w_{1,\text{Easy}} \geq w_{1,\text{Medium}} \geq w_{1,\text{Hard}}$.

Layer 3 behavioral composition. Layer 3 is composed from the behavioral probes:

$$\widetilde{L}_3(x) = 0.60\,\text{Success}(x) + 0.20\,(1 - \text{Drift}(x)) + 0.20\,\text{Repair}(x), \quad (1)$$

$$L_3(x) = 1 + 5\,\widetilde{L}_3(x), \quad (2)$$

with $\widetilde{L}_3(x) \in [0, 1]$ and $L_3(x) \in [1, 6]$.

Weight constraints. For each difficulty category $d \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$, weights satisfy

$$w_k(d) \geq 0, \qquad \sum_{k=1}^{3} w_k(d) = 1.$$

The cross-difficulty monotonicity constraints enforce a theory-consistent trend: higher difficulty increases the relative importance of Layer 3 (behavioral completion), while reducing the relative importance of Layer 1 (surface linguistic quality).

Calibrated weights used in this paper. The calibrated weights used in our experiments (rounded to two decimals) are

$$(w_1, w_2, w_3) = \begin{cases} (0.50,\ 0.30,\ 0.20), & \text{Easy}, \\ (0.35,\ 0.30,\ 0.35), & \text{Medium}, \\ (0.20,\ 0.25,\ 0.55), & \text{Hard}. \end{cases}$$

Proposition 1 (Boundedness and Monotonicity of Aggregation)**.** *Assume* $L_k(x) \in [1, 6]$ *for* $k \in \{1, 2, 3\}$, $w_k(d) \geq 0$ *with* $\sum_k w_k(d) = 1$, *and* $\text{Success}(x) \in [0, 1]$. *Then:*

1. $S_{\text{base}}(x) \in [1, 6]$*;*
2. $S_{\text{base}}(x)$ *is coordinate-wise monotone in each* $L_k(x)$*;*
3. $S_{\text{gate}}(x) \in [1, 6]$.

Proof sketch. S_{base} is a convex combination of values in $[1, 6]$, so $S_{\text{base}} \in [1, 6]$. Also, $\partial S_{\text{base}} / \partial L_k = w_k(d) \geq 0$, giving coordinate-wise monotonicity. Since $0.7 + 0.3\,\text{Success}(x) \in [0.7, 1]$ and $S_{\text{base}}(x) - 1 \in [0, 5]$, we have $S_{\text{gate}}(x) \in [1, 6]$.

C CEFR Assessment Framework

The table below lists the five CEFR qualitative aspects of spoken language use (Range, Accuracy, Fluency, Interaction, Coherence) across reference levels A1–C2; these descriptors define the Layer 1 linguistic-quality scoring.

D Topics designed for Evaluation Scenarios

Table 6 lists all 30 evaluation scenarios, ten per escalation level (Everyday, Friction, Emergency); each scenario is run for all 9 languages and 11 models.

Organization	Model	Size	Notes	Source
Amazon	Nova-2-Lite-V1	–	Lightweight proprietary model	amazon/nova-2-lite-v1
Anthropic	Claude Sonnet 4.6	–	Strong reasoning capability	anthropic/claude-sonnet-4.6
Google	Gemini 2.5 Flash	–	Fast and efficient model	google/gemini-2.5-flash
Meta	Llama-3.1-70B	70B	Large-scale open model	meta-llama/llama-3.1-70b-instruct
	Llama-3.1-8B	8B	Balanced performance	meta-llama/llama-3.1-8b-instruct
	Llama-3.2-1B	1B	Lightweight model	meta-llama/llama-3.2-1b-instruct
Mistral AI	Mixtral-8x22B	8×22B	Sparse mixture-of-experts	mistralai/mixtral-8x22b-instruct
Alibaba	Qwen-2.5-7B	7B	Multilingual support	qwen/qwen-2.5-7b-instruct
	Qwen3-14B	14B	Enhanced reasoning	qwen/qwen3-14b
xAI	Grok-4.20	–	Proprietary reasoning model	x-ai/grok-4.20
Xiaomi	Mimo-V2-Pro	42B	Open-weight MoE reasoning model	xiaomi/mimo-v2-pro

Table 4: Model Selection: the 11 learner models evaluated. The interlocutor agent is a fixed strong model (Gemini-3.1-Pro) and the evaluator is a fixed judge model.

E Languages Used in This Study

Language	Speakers
Chinese	1.3B
English	1.5B
French	310M
German	135M
Italian	65M
Japanese	125M
Korean	80M
Thai	60M
Malay	290M

Table 7: Languages Used in This Study

Languages span Sino-Tibetan (Chinese), Indo-European (English, French, German, Italian), Japonic (Japanese), Koreanic (Korean), Kra-Dai (Thai), and Austronesian (Malay) families.

F Reusable Evaluation Plugin

We release the three-layer evaluation framework as a reusable plugin that scores arbitrary human–agent dialogues, supplied as either structured JSON turns or plain-text transcripts. For each dialogue it returns an overall task-completion score together with layer-wise intermediate assessments, each with supporting evidence fields, for Layer 1 Linguistic Quality (CEFR), Layer 2 Pragmatic Adaptation (SRPD), and Layer 3 Behavioral Robustness (Task Success, Goal Drift, and Repair). Code and usage instructions are available in the released repository. Figure 3 shows the plugin interface.

G Human Validation Details

Table 8 summarizes the protocol used to validate the LLM-judge pipeline. The audit subset contains 445 dialogues (15% of 2,967), sampled with stratification over escalation level, language, and model identity. Three NLP PhD annotators independently scored the same three layers under the same rubric used by the automated evaluator, then resolved disagreements by majority discussion.

We report validation in two forms. First, we compare layer-level score tendencies between experts and the judge. Second, we compare conclusion-level directionality for the three headline findings in the main text: escalation hurts completion, cross-lingual risk concentrates in Korean/-Japanese/Thai, and drift/repair decouple from final success. Across the audited subset, experts and the automated pipeline agree on the direction of all three findings, with no contradiction. This supports our use of the judge as a scalable instrument while keeping claims within the audited scope.

Metric / Score	Description & Indicators
Layer 1: Linguistic Quality (CEFR)	
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).
Layer 2: Pragmatic Adaptation (SRPD, 1–6)	
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?"*).
Layer 3: Behavioral Robustness (multi-turn outcome probes)	
Task Success	Whether the dialogue achieves the scenario goal (scored full / partial / none).
Goal Drift	Whether the agent stays on-goal as turns accumulate, rather than wandering off-task.
Communicative Repair	Whether the agent recovers the interaction after an adversarial turn (pressure, refusal, accusation, bad news).

Table 5: The proposed three-layer framework for **task completion ability**. **Layer 1: Linguistic Quality (CEFR)** assesses whether language use supports sustained goal pursuit across languages (A1 = breakdown to C2 = near-native interactional control). **Layer 2: Pragmatic Adaptation** evaluates whether the agent deploys task-appropriate register on a granular 1–6 SRPD scale as pressure escalates. **Layer 3: Behavioral Robustness** measures multi-turn outcomes (task success, goal drift, and communicative repair) that determine whether the interaction actually reaches its goal.

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 prepackaged 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".

Escalation Level	Scenarios (10 per level)
Level 1: Everyday *Routine service interactions; low stakes, cooperative interlocutor.*	Asking a Passerby for the Time; Ordering at a Cafe; Ordering Pizza over the Phone; Grocery Store Checkout; Buying a Train Ticket; Checking in at a Hotel; Borrowing a Book at the Library; At the Pharmacy; Scheduling a Doctor Visit; Meeting a New Neighbor.
Level 2: Friction *A problem or conflict the learner must resolve through appropriate register and mitigation.*	The Wrong Grocery Price at Checkout; The Incorrect Food Order at a Restaurant; The Sold-Out Item at a Cafe; The Non-Refundable Train Ticket; The Missing Hotel Reservation; The Damaged Library Book; Medication Out of Stock at Pharmacy; Rescheduling a Doctor Appointment; Missing the Last Bus; Noise Complaint from a Neighbor.
Level 3: Emergency *High-stakes, high-pressure situations demanding urgent yet socially appropriate communication.*	The Medical Emergency; The House Fire / Building Evacuation; Stranded and Injured in the Wilderness; The Airline Engine Issue / Onboard Panic; The Robbery / Security Threat; The Kidnapping Ransom Phone Call; The Hostile Border Interview; The Major Workplace Safety Breach; The Public Scandal / PR Crisis; The Whistleblower Confrontation.

Table 6: The 30 evaluation scenarios used in the SRPD politeness evaluation, organized into three escalation levels (10 scenarios each). Each scenario is run for all 9 languages and 11 models, with the interlocutor agent driven by a fixed strong model.

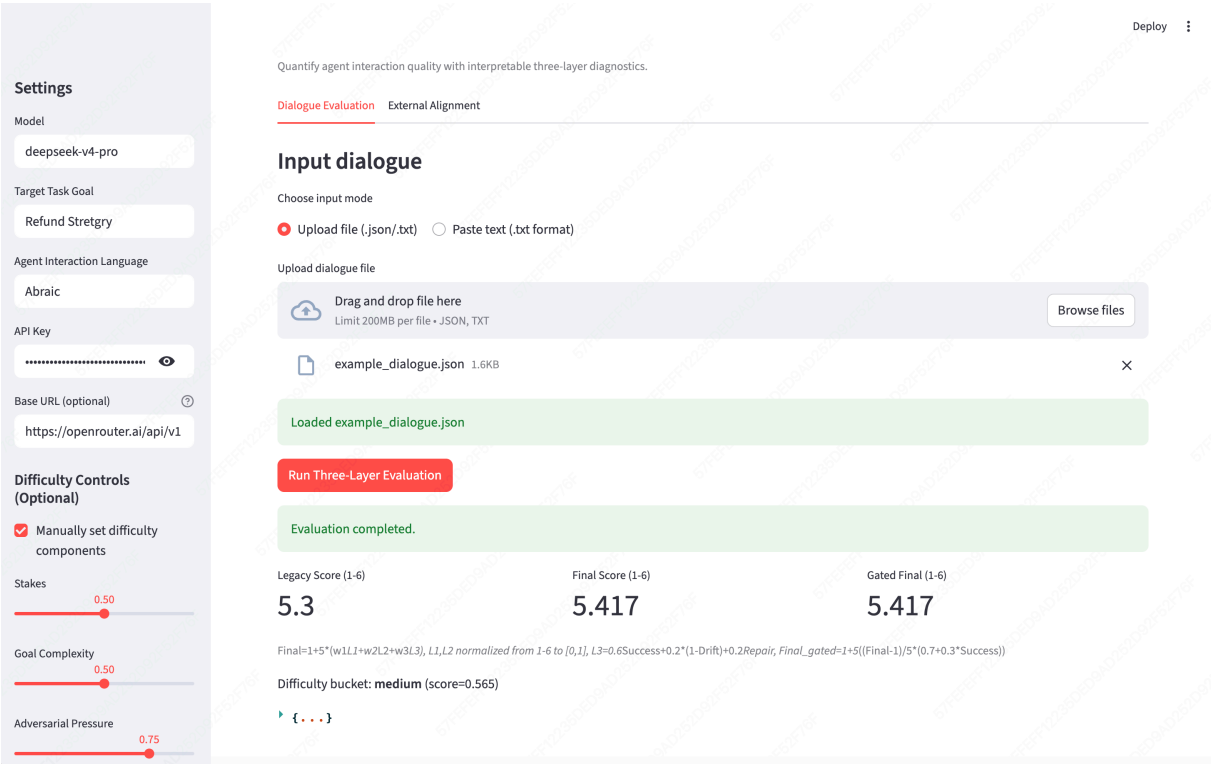

Figure 3: The reusable evaluation plugin. The user selects the model under test, target task goal, interaction language, and optional difficulty controls (left), uploads a dialogue as JSON or plain text, and runs the three-layer evaluation. The tool reports the difficulty bucket and the ungated (Legacy/Final) and gated (Gated Final) task-completion scores, along with the underlying layer-wise breakdown.

Item	Protocol / Outcome
Audit size	445 dialogues (15% of 2,967)
Coverage	Stratified across 3 levels, 9 languages, and model pool
Annotators	3 NLP PhD experts
Scoring setup	Independent scoring with the same three-layer rubric as the judge
Disagreement handling	Majority-based adjudication discussion
Conclusion-level check	3/3 headline findings directionally consistent (no contradiction)

Table 8: Human-validation protocol summary for the LLM-judge evaluation pipeline.

H Per-Layer Detail Tables

This appendix gives the per-layer numeric results that underlie the composite score in Section 5.1: Layer 1 Linguistic Quality by language (Table 9), full task-success by escalation level (Table 10), the Layer 3 Behavioral Robustness probes (Table 11), and the Layer 2 Pragmatic Adaptation heatmaps (Figure 4). Table 12 compares the ungated (S_{base}) and gated (S_{gate}) composite scores.

Model	Task Succ. (%) ↑	Goal Drift (%) ↓	Repair (%) ↑
Claude-Sonnet-4.6	**95.2**	68.9	**99.1**
Gemini-2.5-Flash	94.1	50.4	98.1
Mimo-V2-Pro	93.3	58.5	97.6
Llama-3.1-70B	89.6	67.8	96.1
Nova-2-Lite-V1	88.9	55.2	97.3
Mixtral-8x22B	88.9	45.2	99.2
Qwen3-14B	86.3	44.8	98.5
Llama-3.1-8B	78.1	74.1	90.2
Qwen-2.5-7B	78.1	55.6	98.7
Grok-4.20	77.8	18.1	95.7
Llama-3.2-1B	56.2	53.9	64.5

Table 11: Robustness across three behavioral probes (`gpt-5.4-mini` judge, averaged over 9 languages): **Task Success** (% of dialogues fully completing the task, ↑ better), **Goal Drift** (% of dialogues that drift off-goal at some turn, ↓ better), and **Repair** (% of adversarial challenges recovered, Friction/Emergency only, ↑ better). Sorted by task success. Per-model×language heatmaps in Appendix Figures 6–7.

Model	S_{base}	S_{gate}
Claude-Sonnet-4.6	5.05	5.02
Gemini-2.5-Flash	4.88	4.84
Mimo-V2-Pro	4.88	4.83
Mixtral-8x22B	4.76	4.69
Qwen3-14B	4.78	4.69
Nova-2-Lite-V1	4.69	4.61
Llama-3.1-70B	4.67	4.59
Qwen-2.5-7B	4.48	4.32
Grok-4.20	4.40	4.25
Llama-3.1-8B	4.00	3.88
Llama-3.2-1B	3.46	3.20

Table 12: Ungated (S_{base}) vs. gated (S_{gate}) composite task-completion score, overall per model (mean over levels and languages). The gate preserves the model ranking while penalizing fluent-but-unsuccessful trajectories more for weaker models (e.g., `llama-3.2-1b` drops 3.46→3.20).

I Per-Level CEFR Heatmaps

Figure 5 reports Layer 1 Linguistic Quality (CEFR) as per-model×language heatmaps at each escalation level.

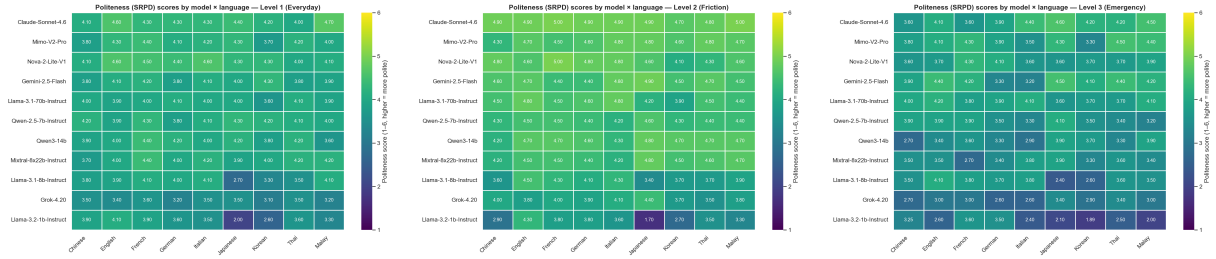

Figure 4: SRPD pragmatic task effectiveness scores (1–6, higher = more appropriate register) by model × language across three escalation levels: Level 1 (Everyday), Level 2 (Friction), and Level 3 (Emergency).

Model	Zh	En	Fr	De	It	Ja	Ko	Th	Ms	Avg
Claude-Sonnet-4.6	**4.97**	**5.28**	**5.09**	**5.01**	**5.01**	**4.91**	**4.83**	**4.68**	**4.83**	4.96
Mimo-V2-Pro	4.31	4.91	4.85	4.88	4.79	4.48	4.40	4.37	4.58	4.62
Gemini-2.5-Flash	4.38	5.06	4.67	4.28	4.64	4.51	4.21	4.30	4.26	4.48
Qwen3-14B	4.60	4.70	4.74	4.32	4.46	4.13	4.13	4.09	4.10	4.36
Mixtral-8x22B	4.17	4.84	4.54	4.53	4.48	3.95	3.94	4.03	4.46	4.33
Llama-3.1-70B	4.28	4.93	4.73	4.46	4.60	3.70	3.11	3.83	4.43	4.23
Nova-2-Lite-V1	4.01	4.81	4.56	4.41	4.29	3.90	3.53	2.67	3.83	4.00
Qwen-2.5-7B	4.36	4.56	4.18	3.77	4.08	3.58	3.38	3.48	3.35	3.86
Grok-4.20	3.87	3.76	3.85	3.79	3.79	3.64	3.67	3.46	3.03	3.65
Llama-3.1-8B	2.25	4.70	3.70	3.40	3.06	1.42	1.54	1.94	2.74	2.75
Llama-3.2-1B	2.27	4.12	2.51	2.63	2.13	1.55	1.91	1.99	2.44	2.39

Table 9: Layer 1 Linguistic Quality: average CEFR overall score (1–6, higher = better) by model × language, averaged across the three escalation levels. Language columns use ISO 639-1 codes (Zh: Chinese, En: English, Fr: French, De: German, It: Italian, Ja: Japanese, Ko: Korean, Th: Thai, Ms: Malay). **Bold** marks the best model per language; **Avg** is the mean over all nine languages. Models are sorted by Avg. Full per-language metric breakdowns are in Appendix 14–22.

Model	L1 (Everyday) ↑	L2 (Friction) ↑	L3 (Emergency) ↑	Δ (L1–L3)
Claude-Sonnet-4.6	0.978	0.933	0.944	0.033
Gemini-2.5-Flash	0.933	0.933	0.956	-0.022
Mimo-V2-Pro	0.956	0.922	0.922	0.033
Llama-3.1-70B	0.978	0.967	0.744	0.233
Mixtral-8x22B	0.944	0.900	0.822	0.122
Nova-2-Lite-V1	0.967	0.944	0.756	0.211
Qwen3-14B	0.967	0.889	0.733	0.233
Llama-3.1-8B	0.956	0.844	0.544	0.411
Qwen-2.5-7B	0.967	0.922	0.456	0.511
Grok-4.20	0.900	0.611	0.822	0.078
Llama-3.2-1B	0.811	0.622	0.241	0.570

Table 10: Model-wise full task-success rate across the three scenario types (L1 Everyday, L2 Friction, L3 Emergency), averaged over languages in the `gpt-5.4-mini` run. Δ (L1–L3) is the drop from Everyday to Emergency.

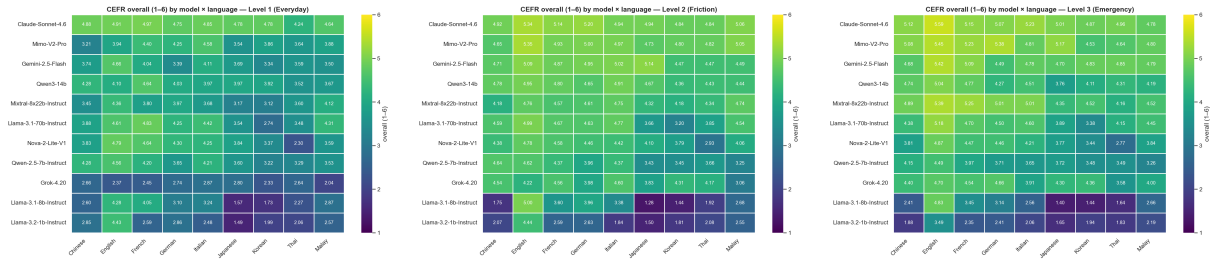

Figure 5: Layer 1 Linguistic Quality: CEFR overall score (1–6, higher = better) by model × language across three escalation levels: Level 1 (Everyday), Level 2 (Friction), and Level 3 (Emergency).

J Per-Model Politeness Scores by Escalation Level

Table 13 reports the Layer 2 SRPD pragmatic adaptation scores (1–6, higher = more polite) for each model, averaged over the nine languages and broken down by escalation level (Everyday, Friction, Emergency) together with an overall column. Politeness peaks at Friction and is lowest at Emergency, where urgency overrides register management, and the drop is sharpest for the weaker models. The corresponding per-model×language heatmaps for each level are in Figure 4.

Model	Everyday	Friction	Emergency	Overall
Claude-Sonnet-4.6	4.32	4.89	4.12	**4.44**
Mimo-V2-Pro	4.11	4.64	4.01	4.26
Nova-2-Lite-V1	4.32	4.62	3.80	4.25
Gemini-2.5-Flash	4.00	4.61	4.02	4.21
Llama-3.1-70B	3.96	4.47	3.90	4.11
Qwen-2.5-7B	4.10	4.40	3.63	4.04
Qwen3-14B	4.03	4.60	3.41	4.01
Mixtral-8x22B	4.07	4.49	3.46	4.00
Llama-3.1-8B	3.72	3.94	3.44	3.70
Grok-4.20	3.39	3.91	2.96	3.42
Llama-3.2-1B	3.39	3.29	2.65	3.11

Table 13: Average SRPD politeness scores (1–6, higher = more polite) by model across the three escalation levels (Everyday, Friction, Emergency) and overall, averaged over all nine languages. Models are sorted by overall score.

K Task Success Heatmaps

Figure 6 reports the full task-success rate by model×language at each escalation level.

L Adversarial Recovery (Repair) Heatmaps

Figure 7 reports the adversarial-challenge repair rate by model×language for Friction and Emergency scenarios.

M Full CEFR Evaluation Results

Per-language CEFR-based holistic scores (1–6) for the 11 learner models, averaged across the three escalation levels. Evaluator: `gpt-5.4-mini`.

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	4.97	5.07	4.97	5.17	5.17	5.03	4.30
Mimo-V2-Pro	4.31	4.27	4.33	4.43	4.53	4.27	4.23
Gemini-2.5-Flash	4.38	4.43	4.47	4.67	4.67	4.40	4.23
Qwen3-14B	4.60	4.87	4.67	4.90	4.83	4.73	3.80
Mixtral-8x22B	4.17	4.07	4.27	4.43	4.50	4.13	4.30
Llama-3.1-70B	4.28	4.43	4.33	4.40	4.57	4.37	3.83
Nova-2-Lite-V1	4.01	4.53	3.63	4.20	4.47	4.30	3.23
Qwen-2.5-7B	4.36	4.60	4.43	4.70	4.73	4.50	3.93
Grok-4.20	3.87	3.93	4.07	3.93	4.00	3.83	4.00
Llama-3.1-8B	2.25	2.70	1.87	2.10	2.97	1.90	2.27
Llama-3.2-1B	2.27	2.57	2.44	2.16	2.54	2.18	2.22

Table 14: CEFR Score Summary for Chinese

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	5.28	5.60	5.47	5.57	5.60	5.57	4.47
Mimo-V2-Pro	4.91	5.07	4.93	5.10	5.17	5.07	4.23
Gemini-2.5-Flash	5.06	5.03	5.13	5.17	5.23	5.00	4.17
Qwen3-14B	4.70	4.93	4.83	4.93	5.00	4.90	4.03
Mixtral-8x22B	4.84	5.00	4.97	5.00	5.03	4.87	4.30
Llama-3.1-70B	4.93	5.07	4.97	5.10	5.17	5.07	4.33
Nova-2-Lite-V1	4.81	5.17	4.97	5.13	5.07	5.10	3.37
Qwen-2.5-7B	4.56	4.80	4.80	4.83	4.80	4.73	3.50
Grok-4.20	3.76	3.73	4.03	3.73	3.93	3.70	3.53
Llama-3.1-8B	4.70	4.73	4.60	4.83	4.93	4.77	4.30
Llama-3.2-1B	4.12	4.30	4.57	4.47	4.33	4.07	3.30

Table 15: CEFR Score Summary for English

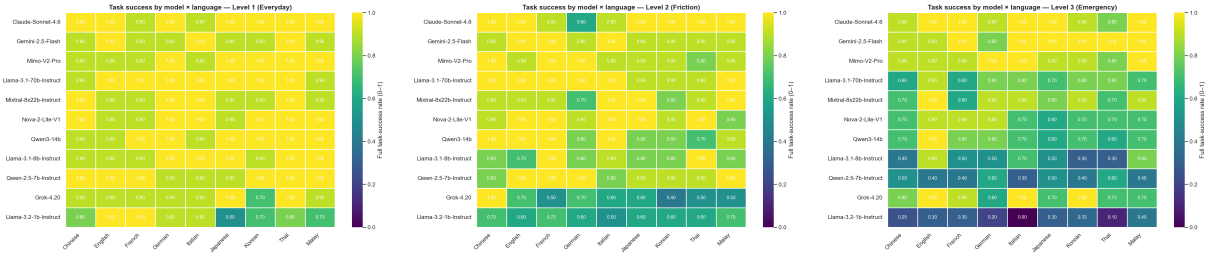

Figure 6: Full task-success rate (0–1) by model × language across three escalation levels: Level 1 (Everyday), Level 2 (Friction), and Level 3 (Emergency).

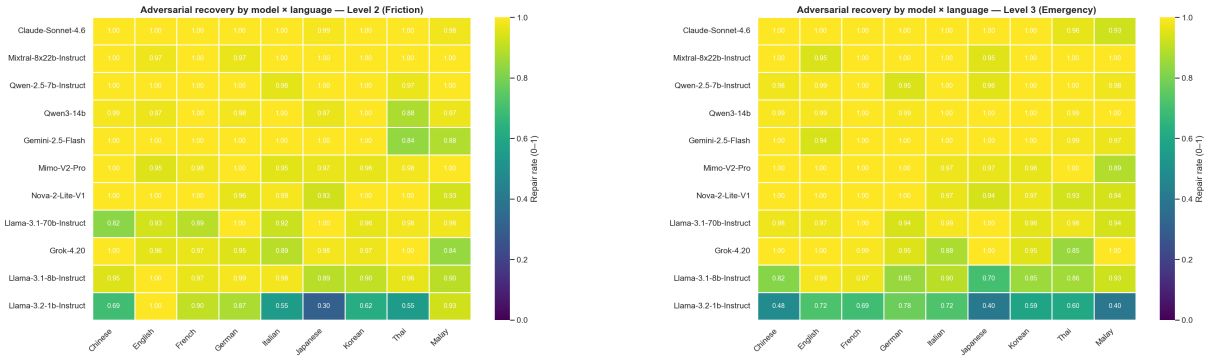

Figure 7: Adversarial-challenge repair rate (0–1) by model × language for Level 2 (Friction) and Level 3 (Emergency); adversarial challenges do not arise at Level 1.

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	5.09	5.27	5.07	5.30	5.33	5.30	4.40
Mimo-V2-Pro	4.85	5.03	4.80	5.17	5.03	5.07	4.57
Gemini-2.5-Flash	4.67	4.73	4.57	4.80	4.83	4.70	4.40
Qwen3-14B	4.74	5.03	4.60	5.17	5.07	5.00	3.77
Mixtral-8x22B	4.54	4.63	4.53	4.77	4.83	4.70	3.70
Llama-3.1-70B	4.73	4.70	4.47	4.80	4.83	4.77	4.17
Nova-2-Lite-V1	4.56	4.93	4.17	4.93	4.97	4.87	3.27
Qwen-2.5-7B	4.18	4.37	3.77	4.47	4.53	4.17	4.00
Grok-4.20	3.85	3.83	3.93	3.83	4.03	3.70	4.27
Llama-3.1-8B	3.70	3.90	3.00	3.83	4.23	3.53	3.93
Llama-3.2-1B	2.51	2.67	2.60	2.43	2.63	2.23	2.90

Table 16: CEFR Score Summary for French

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	5.01	5.13	4.87	5.13	5.23	5.17	4.63
Mimo-V2-Pro	4.88	4.93	4.73	5.10	5.20	5.03	4.63
Gemini-2.5-Flash	4.28	4.27	4.20	4.33	4.50	4.27	4.43
Qwen3-14B	4.32	4.50	4.17	4.83	4.77	4.37	3.90
Mixtral-8x22B	4.53	4.70	4.67	4.90	4.77	4.77	3.83
Llama-3.1-70B	4.46	4.63	4.47	4.63	4.80	4.57	4.13
Nova-2-Lite-V1	4.41	4.90	4.20	4.97	4.90	4.77	3.20
Qwen-2.5-7B	3.77	3.90	3.20	3.90	4.03	3.63	3.63
Grok-4.20	3.79	3.60	4.13	3.73	3.97	3.57	4.10
Llama-3.1-8B	3.40	3.57	2.73	3.53	3.93	3.17	3.87
Llama-3.2-1B	2.63	2.77	2.37	2.67	2.80	2.30	3.27

Table 17: CEFR Score Summary for German

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	5.01	5.20	4.93	5.27	5.27	5.20	4.57
Mimo-V2-Pro	4.79	5.00	4.67	5.10	5.03	5.03	4.47
Gemini-2.5-Flash	4.64	4.73	4.63	4.73	4.83	4.70	4.20
Qwen3-14B	4.46	4.57	4.37	4.73	4.73	4.40	3.97
Mixtral-8x22B	4.48	4.57	4.47	4.77	4.83	4.73	4.07
Llama-3.1-70B	4.60	4.67	4.47	4.77	4.90	4.73	4.17
Nova-2-Lite-V1	4.29	4.97	4.10	4.77	4.70	4.83	3.03
Qwen-2.5-7B	4.08	4.40	3.50	4.30	4.40	4.03	4.00
Grok-4.20	3.79	3.57	3.83	3.73	3.90	3.63	4.10
Llama-3.1-8B	3.06	3.40	2.50	2.97	3.70	2.70	3.37
Llama-3.2-1B	2.13	2.07	2.20	2.17	2.27	1.77	2.70

Table 18: CEFR Score Summary for Italian

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	4.91	5.00	4.67	5.07	5.23	5.03	4.50
Mimo-V2-Pro	4.48	4.30	4.37	4.53	4.70	4.43	4.43
Gemini-2.5-Flash	4.51	4.53	4.50	4.50	4.77	4.50	4.40
Qwen3-14B	4.13	4.33	4.00	4.40	4.43	4.17	3.90
Mixtral-8x22B	3.95	4.07	4.07	4.13	4.17	3.97	3.80
Llama-3.1-70B	3.70	3.73	3.23	3.83	3.97	3.57	3.93
Nova-2-Lite-V1	3.90	4.53	3.50	4.23	4.43	4.20	2.97
Qwen-2.5-7B	3.58	4.07	3.13	3.87	4.07	3.43	3.07
Grok-4.20	3.64	3.47	3.77	3.63	3.90	3.47	4.20
Llama-3.1-8B	1.42	1.63	1.13	1.23	1.80	1.13	2.10
Llama-3.2-1B	1.55	1.63	1.53	1.43	1.57	1.30	2.33

Table 19: CEFR Score Summary for Japanese

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	4.83	4.93	4.90	5.00	5.07	4.93	4.57
Mimo-V2-Pro	4.40	4.43	4.37	4.53	4.60	4.37	4.47
Gemini-2.5-Flash	4.21	4.10	4.23	4.27	4.33	4.00	4.43
Qwen3-14B	4.13	4.47	4.10	4.50	4.57	4.20	3.77
Mixtral-8x22B	3.94	4.10	3.87	4.17	4.33	3.93	3.93
Llama-3.1-70B	3.11	3.40	2.67	3.20	3.70	3.00	3.00
Nova-2-Lite-V1	3.53	4.30	3.00	3.93	3.87	3.53	2.83
Qwen-2.5-7B	3.38	3.97	2.97	3.57	4.00	3.20	2.73
Grok-4.20	3.67	3.67	3.67	3.63	3.93	3.57	4.13
Llama-3.1-8B	1.54	1.90	1.17	1.30	1.97	1.20	1.80
Llama-3.2-1B	1.91	2.05	2.00	1.83	1.79	1.62	2.41

Table 20: CEFR Score Summary for Korean

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	4.68	4.73	4.73	4.77	4.87	4.77	4.43
Mimo-V2-Pro	4.37	4.30	4.27	4.50	4.70	4.30	4.53
Gemini-2.5-Flash	4.30	4.30	4.23	4.37	4.47	4.23	4.23
Qwen3-14B	4.09	4.30	3.83	4.33	4.53	4.00	3.87
Mixtral-8x22B	4.03	4.10	4.07	4.30	4.43	4.17	3.87
Llama-3.1-70B	3.83	3.90	3.43	3.90	4.20	3.63	3.67
Nova-2-Lite-V1	2.67	3.30	1.97	2.83	3.27	2.67	2.30
Qwen-2.5-7B	3.48	4.00	2.90	3.80	4.13	3.40	3.40
Grok-4.20	3.46	3.37	3.53	3.50	3.70	3.33	4.03
Llama-3.1-8B	1.94	2.13	1.50	1.83	2.43	1.50	2.67
Llama-3.2-1B	1.99	2.10	2.10	1.93	2.00	1.67	2.47

Table 21: CEFR Score Summary for Thai

Model	Overall	Range	Accuracy	Fluency	Interaction	Coherence	Human Likeness
Claude-Sonnet-4.6	4.83	5.03	4.77	5.00	5.10	5.07	4.53
Mimo-V2-Pro	4.58	4.73	4.53	4.73	4.83	4.70	4.57
Gemini-2.5-Flash	4.26	4.27	4.17	4.40	4.50	4.33	4.30
Qwen3-14B	4.10	4.13	3.80	4.33	4.53	4.03	4.07
Mixtral-8x22B	4.46	4.67	4.47	4.77	4.90	4.70	3.83
Llama-3.1-70B	4.43	4.57	4.23	4.60	4.70	4.50	4.13
Nova-2-Lite-V1	3.83	4.27	3.13	4.20	4.23	3.93	3.13
Qwen-2.5-7B	3.35	3.43	2.57	3.43	3.73	3.20	3.80
Grok-4.20	3.03	2.97	2.93	2.97	3.27	2.90	3.87
Llama-3.1-8B	2.74	2.97	2.13	2.77	3.37	2.53	3.07
Llama-3.2-1B	2.44	2.57	2.33	2.50	2.47	2.07	3.07

Table 22: CEFR Score Summary for Malay

N Additional Agent Dialogue Examples

This appendix provides additional agent-dialogue examples for qualitative inspection.

Speaker	Utterance
Assistant	Why do you want to study abroad?
User	I want to study abroad because I want new cultures, international friends, and better language skills.
Assistant	That's a great motivation. Immersion often helps both language growth and cultural understanding.
User	That makes sense. What was the hardest part when you first arrived?
Assistant	The first days were hardest: transport, routines, and uncertainty. But once I built a routine, progress became much faster.
User	Thanks, that's helpful. I'll prepare for that transition period in advance.

Table 23: Additional English agent dialogue excerpt from generated interaction logs, illustrating sustained topic progression and goal-oriented exchange.

Example A: English multi-turn planning dialogue (agent keeps the goal active).

Speaker	Utterance
Assistant	Good evening. How can I assist you?
Learner	Hi, my reservation is missing at check-in.
Assistant	Do you have your booking number?
Learner	Yes, 7483-A. Could you please check again?
Assistant	I still cannot find it in the system.
Learner	If possible, could you give me a temporary room while I contact the platform?

Table 24: Additional service-friction dialogue excerpt showing a repair-oriented strategy to keep the task on track under uncertainty.

Example B: Service-recovery dialogue (task continuity under friction).

Example C: Thai Level 3 emergency dialogue (register control under pressure).

O External Benchmark Alignment: Uncertainty and Attribution

To reduce over-interpretation, we treat external leaderboard alignment as *preliminary evidence*. We use a fixed Arena snapshot window (Overall and Multi-Turn, both dated May 17, 2026) and report uncertainty for rank correlations via 20,000-sample bootstrap 95% CIs plus leave-one-out (LOO) stability.

Speaker	Original Content (Thai)	English Translation
Tutor	โอ้พระเจ้า เขาล้มลงไปเลย! ฉันคิดว่าเขาหยุดหายใจแล้ว! ช่วยด้วย! คุณ---คนนั้น---ทำอะไรสักอย่างสิ!	Oh my God, he just collapsed! I think he stopped breathing! Help! You there—do something!
Learner	หยุดก่อน! ฟังฉัน! สงบสติอารมณ์ได้เลย---ฉันจะช่วย! คุณ---ใช่ คุณนั่นแหละ---โทร 1669 ตอนนี้เลย! บอกว่ามีคนหมดสติที่สถานีรถไฟ แล้วบอกชื่อสถานีนี้ด้วย! คนอื่น ๆ ถอยออกไปให้หมด!	Stop! Listen to me! Stay calm—I will help. You—yes, you there—call 1669 right now! Tell them there is an unconscious person at the train station and give this station name. Everyone else, step back now!
Tutor	เขายังไม่ตายใช่มั้ย?! ก่อนล้มเขาบอกว่าเจ็บหน้าอกและหายใจไม่ออก! ช่วยเขาด้วย!	He is not dead, right?! Before collapsing he said he had chest pain and could not breathe! Please save him!
Learner	ฟังฉัน---ตอนนี้ฉันช่วยเขาอยู่แล้ว! โทร 1669 และเปิดสปีกเกอร์ไว้! เขายังไม่ตาย แต่ฉันต้องทำ CPR ตอนนี้เลย! นับกับฉัน: หนึ่ง สอง สาม... แล้วช่วยถามคนรอบข้างหาเครื่อง AED เดี๋ยวนี้!	Listen to me—I am already helping him now! Call 1669 and put it on speaker! He is not dead yet, but I must start CPR right now! Count with me: one, two, three... and ask people around you to find an AED immediately!

Table 25: Register control under extreme pressure: an excerpted Level 3 (Emergency) Thai dialogue from the experiment repository, generated by `claude-sonnet-4.6`. The Learner sustains calm, task-appropriate register, issuing targeted directives while managing a panicking bystander, to keep the resuscitation task on track.

Comparison	n	Spearman ρ (95% CI)	Kendall τ (95% CI)	LOO Spearman range
Ours vs Arena Overall	9	0.57 [−0.28, 1.00]	0.44 [−0.25, 1.00]	[0.38, 0.88]
Ours vs Arena Multi-Turn	9	0.50 [−0.35, 0.98]	0.39 [−0.29, 0.94]	[0.29, 0.81]
Arena Overall vs Arena Multi-Turn	9	0.98 [0.82, 1.00]	0.94 [0.72, 1.00]	[0.98, 1.00]

Table 26: External alignment with uncertainty reporting over the nine study models that map onto the Arena board (Overall and Multi-Turn). We report 20,000-sample bootstrap 95% CIs and leave-one-out (LOO) Spearman ranges.

Model	Our Rank	Arena Overall	Arena Multi-Turn	L1	L2	L3
`grok-4.20`	7	1	1	0.609	0.570	0.871
`mixtral-8x22b`	4	7	8	0.721	0.667	0.871
`claude-sonnet-4.6`	1	2	2	0.826	0.741	0.846

Table 27: Layer-level attribution for representative rank disagreements. The largest conflict is `grok-4.20`: strong Arena rank and strong Layer 3 robustness, but weaker Layer 1/2 in our setting, which lowers its task-completion ranking. This illustrates why layer-wise diagnostics are necessary beyond single leaderboard positions. Ranks are computed within the 9-model Arena overlap (the study models that appear on the Arena board), so they differ from the full 11-model study ranking in the main text.

P Generation Dialogue Examples

Full generated examples for all 9 languages please refer to: https://anonymous.4open.science/r/Multilingual_Eval_Narrative-6333. This section serves as an illustration for understanding the dynamics nature of generated dialogue.

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 acceptable after all. Since you are willing to accept the penalty and show genuine repentance, I will trust you this once. "It is never too late to mend."

Table 28: 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.

Speaker	Utterance
Tutor	안녕하세요! 그랜드 호텔에 오신 것을 환영합니다. 예약하신 성함이 어떻게 되세요?
Learner	안녕하세요! 제 이름은 김민준입니다. 여기 제 여권이 있어요, 필요하시면 보여드릴게요.
Tutor	김민준 고객님, 감사합니다. 예약 확인되셨습니다. 객실은 1004호이고 카드키 여기 있습니다. 체크아웃은 오전 11시, 조식은 1층에서 7시부터 10시까지입니다.
Learner	감사합니다! 1004호, 체크아웃 11시, 조식 7시 -- 10시 맞죠? 그런데 혹시 와이파이 비밀번호를 다시 한번 확인해 주실 수 있을까요?
Tutor	네, 정확합니다. 와이파이는 'Room_1004'를 선택하시고 비밀번호 'grand1234'를 입력하시면 됩니다. 더 필요하시면 0번으로 프론트에 연락 주세요.

Table 29: An illustrative Tutor–Learner Korean exchange from a Level 1 (Everyday) hotel check-in scenario, showing the turn-by-turn goal-oriented interaction the framework evaluates: the Learner supplies the required information, confirms details, and requests a clarification to complete the task.

Q 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 8: 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 9: 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 10: 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 11: 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 12: 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 13: 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 14: 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 15: 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 16: 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 17: 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 18: 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 19: Prompt used for evaluating learner politeness strategy.