A Framework for Proficiency-Aligned Grammar Practice in LLM-Based Dialogue Systems

Luisa Ribeiro-Flucht[1,2] **Xiaobin Chen**[1,2] **Detmar Meurers**[1,3]
[1] LEAD Graduate School and Research Network, University of Tübingen, Germany
[2] Hector Research Institute of Education Sciences and Psychology,
University of Tübingen, Germany
[3] Leibniz-Institut für Wissensmedien, Tübingen, Germany
{luisa.ribeiro-flucht,xiaobin.chen}@uni-tuebingen.de
d.meurers@iwm-tuebingen.de

Abstract

Communicative practice is critical for second language development, yet learners often lack targeted, engaging opportunities to use new grammar structures. While large language models (LLMs) can offer coherent interactions, they are not inherently aligned with pedagogical goals or proficiency levels. In this paper, we explore how LLMs can be integrated into a structured framework for enabling goal-oriented, grammar-focused interaction, building on an existing dialogue system. Through controlled simulations, we evaluate five LLMs across 75 A2-level tasks under two conditions: (i) grammar-targeted, task-anchored prompting and (ii) the addition of a lightweight post-generation validation pipeline using a grammar annotator. Our findings show that template-based prompting alone substantially increases target-form coverage up to 91.4% for LLaMA 3.1-70B-Instruct, while reducing overly advanced grammar usage. The validation pipeline provides an additional boost in form-focused tasks, raising coverage to 96.3% without significantly degrading appropriateness.

1 Introduction

Second language acquisition (SLA) is driven by frequent and meaningful language use (Behrens, 2009; Ellis, 2002; Canale and Swain, 1980). While second language (L2) learners develop comprehension skills through input-rich activities, many struggle to find opportunities outside of the classroom to meaningfully produce what they've learned, especially in the early stages (Ortega and DeKeyser, 2007). This is particularly true for A2-level learners, who are using the L2 in social interaction more consistently, as described by the Common European Framework of Reference Companion Volume (CEFR; Council of Europe, 2020).

Grammar often represents an obstacle in L2 production, with certain forms proving persistently difficult to master (Ellis, 2017). As a result, learners tend to avoid challenging forms in spontaneous communication, where conveying meaning quickly takes priority (Lyster and Sato, 2013). To mitigate this issue, research emphasizes the importance of communicative practice that targets learners' specific needs and occurs iteratively, rather than relying on decontextualized drills. Without such targeted support, many A2 learners tend to plateau, struggling to transfer classroom grammar knowledge to real-life communication (Richards, 2008; Mirzaei et al., 2017; Lightbown, 2007).

Conversational agents, or chatbots, have been proposed as a way to offer students language production opportunities (Sydorenko et al., 2018). Early rule-based systems enabled predictable, level-appropriate dialogues, but their design was labor-intensive and resulted in rigid, limited interactions (Bibauw et al., 2022). More recently, LLMs have emerged as a promising alternative, as they are capable of generating coherent and fluent language without the need for manual scripting. However, their stochastic nature often leads to inconsistent pedagogical alignment (Zhou et al., 2023; Benedetto et al., 2025).

While several studies have explored the use of out-of-the-box LLMs for educational purposes, such as linguistic feedback, role play, and adaptivity (Borchers and Shou, 2025; Gervás et al., 2025; Fincham and Alvarez, 2024), their inherent unpredictability poses challenges for aligning output with pedagogical frameworks, adaptive logic, and real-world scalability. In the absence of a systematic mechanism for constraining both communicative context and grammatical targets, learners may engage in practice that lacks the individualization and progression necessary for effective development (Ruiz et al., 2023).

To bridge this gap, we introduce a parametric framework for goal-oriented, CEFR-aligned grammar practice through LLM-mediated dialogue.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 978–987
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Building on AISLA, a rule-based system developed for grammar instruction among German seventh-graders (Chen et al., 2022), this A2-level extension links each task to the English Grammar Profile (EGP; O'Keeffe and Mark, 2017). EGP targets are embedded into dynamically generated prompt templates that pair specific grammatical forms with communicative scenarios, enabling repeated practice of the same structure across varied, context-rich tasks. This approach aims to maintain pedagogical consistency while leveraging the flexibility and fluency of LLMs.

In this paper, we investigate the effectiveness of our approach by asking the following research questions:

1. Can out-of-the-box LLMs generate goal-oriented dialogues that spontaneously target specific grammatical structures?

2. How effective is task-supported prompting in guiding LLMs to produce multi-turn, A2-level outputs aligned with target grammar?

3. To what extent can LLMs of different sizes maintain coherence and target specific grammatical structures in task-supported dialogues?

4. Does incorporating a grammar validation component improve target structure usage and CEFR alignment?

2 Goal-oriented Grammar Practice

While theoretical perspectives in SLA vary, many contemporary approaches increasingly view grammar as a functional and adaptive component of language use, rather than a fixed body of rules (Diessel, 2019; Dik, 1981). Larsen-Freeman (2003) advocates for the reframing of grammar as a dynamic skill encompassing three interrelated dimensions: (i) form, structural features of language; (ii) meaning, the semantic or propositional content these forms convey; and (iii) use, the pragmatic and discourse functions that guide when and why particular forms are selected in context.

To support this dynamic view of grammar, instructional design should align the learning environment with real communicative demands. Transfer-appropriate processing (TAP; Lightbown, 2007) reinforces this idea, suggesting that learning is most effective when the cognitive processes involved during learning mirror those required during retrieval and use. Grammar practice then should not be decontextualized, rather, learners need iterative opportunities to use target structures in activities that closely resemble authentic language use.

Furthermore, research shows that specific grammar structures are best acquired through activities that naturally elicit their use (Loschky and Bley-Vroman, 1993; Faitaki and Murphy, 2019; Lyster and Sato, 2013). Frameworks such as Task-Based Language Teaching (TBLT; Nunan, 2004) and Task-Supported Language Teaching (TSLT; Ellis, 2024) operationalize this idea by embedding language practice within communicative tasks. In TSLT, for example, the syllabus is organized around specific linguistic units, which are practiced through meaning-oriented tasks: activities in which language is used to achieve a non-linguistic goal, such as comparing options or making plans. Unlike traditional drills, such tasks promote interaction in which the target structure is functionally relevant. Our system builds on this approach, using task templates that integrate grammar targets with communicative goals (Bear et al., 2024).

3 Controlled Text Generation

To integrate a multidimensional view of grammar into LLM-based applications, developers must find ways to control the output of these models in order to deliver scaffolded, targeted practice without sacrificing meaning or use. Although emerging approaches offer potential solutions, such as prompting techniques and finetuning techniques, they typically lack explicit goal orientation and have not been systematically applied to grammar-focused learning tasks.

In the context of open-ended dialogue, some approaches have aimed to implicitly steer LLMs toward producing predetermined grammatical structures. For instance, Okano et al. (2023) compare reinforcement learning-based fine-tuning of DialoGPT with few-shot prompting of GPT-3, finding that both methods can enable grammatical control, with reinforcement learning achieving greater precision. Similarly, Glandorf et al. (2025) evaluate prompting, fine-tuning and decoding strategies for the inclusion of predermined EGP structures during open-ended chat, showing that grammar-controlled decoding with LLaMA 3.3 effectively targets specific forms, albeit with a slight reduction in response quality. However, both studies focus exclusively on the inclusion of target structures in the next response only, not evaluating model performance across multi-turn interactions.

Engaging LLMs in multi-turn conversations introduces additional challenges, as the model must track and integrate longer contextual information to maintain coherence and relevance across turns (Yi et al., 2024). This challenge becomes more complex when there is a pedagogical task to adhere to and a grammar structure to target. While some recent work explores different applications of LLM-mediated language learning (Tyen et al.; Méndez and Bautista, 2025), no approach, to our knowledge, has attempted to integrate LLMs within goal-oriented dialogue systems for systematic, targeted grammar practice.

Our work therefore intends to move towards bridging these divides by integrating a CEFR-aligned proficiency framework, generating task-based dialogue data and embedding real-time grammar scaffolding into an LLM-powered dialogue system. In doing so, we aim at combining the naturalness of large-scale language models, with the pedagogical basis of goal-oriented, task-supported instruction.

4 Implementation

4.1 System Description

The AISLA system was built using a Java-based backend with a PostgreSQL[1] database and an Android-based frontend [2]. Its backend follows a modular, service-oriented design for functionalities such as text-to-speech and automatic speech recognition. The chatbot functionality is handled via AWS Lex [3], a slot-filling dialogue management service, which requires manual dialogue scripting. Although effective for rule-based interactions, especially in school contexts, where content control is a priority (Wilske, 2015), this configuration presents limitations for the integration of personalized and adaptive features.

To support the requirements of this research, several major architectural changes were introduced. First, the following changes were made to accommodate LLM-based interactions: AWS Lex was replaced by LLM APIs, and a DialogueManager class was added to orchestrate prompt chaining and dialogue state management. Second, the Android frontend was replaced with an Ionic[4] one to ensure broader accessibility across platforms, thereby increasing inclusivity in participant recruitment and usage scenarios. Additionally, the EGP was integrated in the grammar task design.

4.2 Task Bank

To support modularity and future scalability, a task bank was implemented as a database table. Each entry is linked to a target grammar structure and a communicative purpose, including fields such as the EGP structure's guideword, can-do statement, the name and format of the task and its instructions. Task names refer to real-life situations where grammar structures are employed for communicative purposes, for instance, "Discussing cultural differences between two countries", "Telling someone about a historical monument" or "Picking between two places to go to".

The task design is based on the Grammaring framework (Larsen-Freeman, 2003), accounting for the three dimensions of grammar. Accordingly, three task types were developed:

Q&A (form-focused): These tasks are meant to provide high-frequency exposure to a target grammar structure in each of the model's turns. It aims to use and elicit the structure in short question-answer exchanges (e.g. answering questions about one's daily routine with frequency adverbs).

Information gap (meaning-focused): These tasks emphasize the meaningful application of grammar structures, encouraging learners to make decisions about where and how to use the target structure in context, usually leveraging external resources like tables, charts and images (e.g. reporting on what was said in an interview with reported speech, explaining what someone looks like using adjectives).

Role play (use-focused): These tasks situate grammar practice in realistic scenarios (e.g. giving a friend advice with modal verbs, asking for directions with prepositions). They are designed to simulate real-life situations where the structure must be used appropriately within a given social or functional context.

[1] https://www.postgresql.org/

[2] https://www.android.com/

[3] https://aws.amazon.com/lex/

[4] https://ionicframework.com/

4.3 Dialogue management

When a task is initiated by the student, information from the task bank is dynamically retrieved from the database and inserted into a template-based LLM prompt. An example prompt schema can be found in Appendix A. Task duration is currently managed via turn count. This means that each dialog task spans a predetermined number of turns by default, after which the learner is given the option to either conclude the task or continue practicing.

5 Method

Since the purpose of this study was to evaluate how well different LLMs perform in task-supported dialogues, conducting a user study was considered premature. Instead, to simulate varied learner interactions, each model was paired with ChatGPT-4o, using a temperature setting of 0.5 to introduce some content variability on the student side.

To test the robustness of the model acting as the tutor, three learner behavior patterns were implemented in every task: (1) in the first run, the model was instructed to make grammatical mistakes; (2) in the second run, a hard-coded clarification request ("What does that mean?") was injected; and (3) in the third run, a misunderstanding was introduced via the phrase "I don't know" at the second turn (c.f. Appendix F for snippets of different runs and task type). Each task was limited to 10 turns to ensure comparability across conditions.

We selected 75 tasks targeting 15 grammar supercategories drawn from the EGP, namely, adjectives, adverbs, clauses, determiners, future, modality, passives, past, prepositions, present, pronouns, verbs, questions, negation, and reported speech. The tasks were equally divided into Q&A, information gap, and role play formats. To account for output variability, each task was run five times, resulting in 375 dialogues, and 1875 messages per model [5].

5.1 Experiment 1

We evaluated five large language models (LLMs) spanning a wide parameter range: Llama 3.1 8B-Instruct, Mistral-Small 3.1 24B-Instruct (Mistral AI, 2024), Llama 3.3 70B-Instruct (Meta, 2024), DeepSeek-V3 685B (DeepSeek AI, 2024), and GPT-4o (OpenAI, 2024), whose exact parameter count is undisclosed. Each model acted as the tutor in the 75 tasks. The decoding temperature was fixed at 0.0 for decreased variability.

To isolate the effect of explicit grammatical scaffolding (RQ1), we first used *task-only prompting*, satisfying the TSLT requirement of a clear non-linguistic goal. In this setting, the prompt contained only the task name plus minimal instructions, with no mention of the target grammar structure (c.f. Appendix B).

The second part of Experiment 1 introduced our template-based prompt that embeds the communicative goal together with A2-level grammar cues. We ask whether this prompt improves alignment and whether model size modulates any gain (RQ2-RQ3).

5.2 Experiment 2

Experiment 2 adds a lightweight control pipeline. We integrated POLKE (Sagirov and Chen, 2025), an EGP-based grammar annotator, as a post-generation validator. For every tutor turn, POLKE tagged all grammar structures and their CEFR level; a one-shot rephrase is triggered when (i) any structure above B1 is present (Appendix C) or (ii) in Q&A tasks, the required target structure is missing (c.f. Appendix D). Only one rewrite pass is allowed to bound latency and prevent loops. The control loop was applied only to the three best-performing models from Experiment 1 (Llama 3.3 70B, DeepSeek V3, GPT-4o).

6 Evaluation

We combine two quantitative metrics, obtained via POLKE annotations, with one qualitative metric obtained from human ratings.

Target structure presence A binary metric which measures whether the tutor turn contains the grammar form specified in the task (crucial for Q&A).

Proficiency alignment Defined here as the use of grammar within the target CEFR range. This metric refers to the proportion of structures above the B1 ceiling (i.e. B2, C1, C2).

Response quality Appropriateness on a 5-point scale (factual accuracy + contextual coherence). Fifteen native or near-native English speakers recruited through Prolific[6] rated six dialogues per model (450 tutor turns). The rubric and anchors appear in Appendix E.

[5] All experiments and data mentioned in this work can be found in https://github.com/luisards/grammar-practice-framework

[6] https://www.prolific.com

To probe rubric interpretability, GPT-4o scored the same 30 dialogues. Its item-level scores correlate moderately with the human mean (Spearman $\rho = .49$, $p < .01$) and reproduce the system rank order ($\rho = .67$). A separate GPT-4o pass over the full 75-task set is released for replication in the shared repository.

7 Results

In this section, we report the results of three experimental conditions: baseline task-only (B), prompt + grammar scaffold (P) and prompt + scaffold + validation (P+V), distributed by metric.

7.1 Dialogue-quality ratings

Table 1 shows the mean human appropriateness ratings. Larger models outperform smaller ones across all conditions. Prompting incurs a small drop for every model (max 0.6 points for Mistral-Small). Validation restores or slightly improves quality for the top systems. Inter-rater agreement was found to be moderate (Krippendorff $\alpha_{\text{ordinal}} = .42$; rises to .45 with GPT-4o added).

Smaller models were excluded on the basis of a post-hoc, exploratory cutoff: any model whose mean human appropriateness rating fell below 4.0, corresponding to the "somewhat appropriate" anchor on our 5-point rubric was deemed pedagogically unviable and therefore did not get included in experiment 2.

Model	B	P	P+V
Llama 3.1	3.9	3.6	–
Mistral-Small	3.5	3.3	–
Llama 3.3	4.9	4.3	4.7
DeepSeek V3	4.7	4.4	4.4
GPT-4o	4.5	4.5	4.6

Table 1: Mean human appropriateness ratings (1-5).

7.2 Proficiency alignment

Table 2 reports the share of grammar at or below B1. At baseline, a chi-square test across the five models is significant ($\chi^2 = 59.4$, $df = 4$, $p < 10^{-11}$) but the practical effect is small (Cramér $V = .04$). Prompting pushes every model above 98% basic grammar; validation halves the residual advanced usage.

7.3 Target-structure coverage (form-practice tasks)

Prompting boosts target-structure inclusion from roughly 30% to 70-91% (Table 3). Validation yields a further 5-11-point gain (96.3 % for Llama 3.3 70B, 95.0% for DeepSeek V3, 91.5% for GPT-4o). A Pearson chi-square on the Q&A subset confirms significant model differences at the prompt stage ($\chi^2(2, N = 1{,}875) = 33.1, p < 10^{-7}, V = .13$). After validation the gap narrows but remains significant ($\chi^2 = 19.5, p < .001, V = .10$).

8 Discussion

Our findings reveal that while large language models (LLMs) are capable of generating fluent, goal-oriented dialogues, they do not reliably produce the intended grammatical structures without explicit guidance. Answering RQ1, at baseline, models demonstrated stronger appropriateness, with bigger models reaching average ratings between 4.5 and 4.8, but the presence of the target grammatical structure was limited, appearing in only 28-39% of tutor turns in form-practice tasks.

RQ3 explored how model size influence the ability to maintain coherence and grammatical focus in task-supported dialogues. Our findings suggest that model capacity matters. Larger models (70B-685B) retained higher appropriateness (4.3 - 4.7) and needed fewer rewrites, confirming that scale confers stronger control. Yet the scaffolded prompt significantly narrowed the grammar gap, even though their final appropriateness remained lower (≈ 3.6 - 3.3). This trade-off invites a cost-benefit choice: institutions with limited resources may be able to achieve near-large-model grammar fidelity at a fraction of the compute cost, accepting a decrease in perceived dialogue polish.

Concerning RQ4, a post-hoc validation pass halved the residual advanced grammar usage ($\chi^2(2) = 35.1$, $V = .10$), confirming its value as a safety net when level control is non-negotiable. However, quality gains plateau once a strong prompt is in place, suggesting diminishing returns for additional automated checks.

Finally, it is important to acknowledge the fact that strict structure enforcement must be balanced against the spontaneity of genuine dialogue: real learners will redirect, clarify and digress. Designing tasks that preserve communicative authenticity while guaranteeing exposure to a focal form remains an open challenge, especially at higher or

Model	B		P		P+V	
	≤B1	>B1	≤B1	>B1	≤B1	>B1
Llama 3.1	92.4	7.6	98.6	1.4	–	–
Mistral-Small	93.6	6.4	92.8	7.2	–	–
Llama 3.3	93.0	7.0	98.6	1.4	99.4	0.6
DeepSeek V3	91.5	8.5	98.2	1.8	99.1	0.9
GPT-4o	92.4	7.6	98.9	1.1	99.4	0.6

Table 2: Percentage of grammar structures at or below B1 (≤B1) and above B1 (>B1).

Model	B	P	P+V
Llama 3.1	39.0	78.4	–
Mistral-Small	28.0	69.4	–
Llama 3.3	37.0	91.4	96.3
DeepSeek V3	34.2	87.8	95.0
GPT-4o	35.5	80.5	91.5

Table 3: Tutor turns that contain the requested grammar structure (25 form-practice tasks).

lower CEFR targets where our A2-centric template may not directly transfer.

9 Conclusion and Outlook

This paper explores the integration of LLMs into a goal-oriented dialogue system for A2-level grammar practice. Our results suggest that when dialogues are grounded in pedagogically designed prompts, proficiency alignment converges across models of different sizes. While these findings are promising, they remain preliminary and based on controlled simulation rather than real learner input.

Larger models (e.g., LLaMA 3.3 70B, GPT-4o, DeepSeek V3) maintained grammatical focus and dialogue coherence more reliably, particularly under conversational pressure. However, prompting alone was sufficient to bring smaller models (e.g., Mistral-Small) closer to the target structure usage rates observed in larger systems. This indicates that instructional framing, not just model capacity, plays a critical role in shaping output toward pedagogical goals.

We also explored the impact of a lightweight post-generation validation step using POLKE, an EGP-based grammar annotator. While this step did not significantly alter overall CEFR alignment (which was already high under prompt conditions), it provided additional gains in target-form inclusion, particularly in Q&A tasks, where an increase of approximately 10% was observed. These findings highlight validation as a useful safeguard for scenarios where structure-specific exposure is pedagogically important.

Taken together, our findings point toward two tentative design guidelines for developers of intelligent tutoring systems that incorporate LLMs: (i) Combining pedagogically-grounded prompts with CEFR-based post-generation validation may offer a feasible path toward controllable, targeted grammar practice; (ii) Model scaling should be guided by observable convergence in dialogue coherence and target-form density, which, based on our exploratory experiments, occurred at around 70 billion parameters.

Furthermore, because our evaluation relies on open CEFR descriptors and a publicly available annotator, the method remains applicable as new models are released. To support continued re-evaluation, we release all core components of our setup: the selected task bank, template prompt, and scoring script.

Beyond targeting grammatical forms, our results underscore the value of contextual control: grammar structures should appear not only accurately, but also in varied, goal-relevant settings. Our template-based prompting framework sets to achieve this by scaffolding interaction around communicative goals, potentially making it easier to support iterative practice, interest-driven adaptation and integration of learner modeling.

In future work, we intend to perform user studies and log learner use of support tools (e.g., on-demand L1 translation), engagement with different contexts and its interactional with learning outcomes. Finally, over time, interaction data collected from users will allow for the creation of authentic data, enabling LLM fine-tuning grounded in authentic learner behavior.

Limitations

While our framework demonstrates the potential of LLMs for proficiency-aligned grammar practice, several limitations must be acknowledged. First, our grammar validation relies on an automatic annotator, the robustness and coverage of which varies across structures. In experiment 2, the same annotator was also used for both controlling and evaluating the output, which could introduce bias into the results.

In addition, the system currently does not perform a formal grammar accuracy check beyond target form detection, meaning that in case the LLMs make errors, those may go unnoticed. Similarly, vocabulary control, although implicitly restricted through task design, is not externally validated against level-specific lexicons, which may impact lexical appropriateness for A2 learners.

Lastly, our evaluation remains system-focused and has not included learner interaction data. Without a user study, we cannot yet assess the pedagogical effectiveness, learner engagement or practical impact of the system in real-world settings. These areas will be addressed as the next step in this project.

Acknowledgments

This work was supported by the German Ministry of Education and Science (BMBF) under funding number 01IS22076. We also acknowledge the valuable contribution of the English Grammar Profile (EGP), which provided an invaluable resource used in this study.

References

Elizabeth Bear, Xiaobin Chen, Daniela Verratti Souto, Luisa Ribeiro-Flucht, Björn Rudzewitz, and Detmar Meurers. 2024. Designing a task-based conversational agent for EFL in German schools: Student needs, actions, and perceptions. *System*, 126:103460.

Heike Behrens. 2009. Usage-based and emergentist approaches to language acquisition. *Linguistics*, 47(2):383–411.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common european framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Serge Bibauw, Thomas François, and Piet Desmet. 2022. Dialogue systems for language learning: Chatbots and beyond. In Nicole Ziegler and Marta González-Lloret, editors, *The Routledge Handbook of Second Language Acquisition and Technology*, page 15. Routledge, New York.

Conrad Borchers and Tianze Shou. 2025. Can large language models match tutoring system adaptivity? a benchmarking study. *arXiv preprint arXiv:2504.05570*. Accepted as full paper to the 26th International Conference on Artificial Intelligence in Education (AIED 2025).

Michael Canale and Merrill Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, I(1):1–47.

Xiaobin Chen, Elizabeth Bear, Bronson Hui, Haemanth Santhi-Ponnusamy, and Detmar Meurers. 2022. Education theories and ai affordances: Design and implementation of an intelligent computer assisted language learning system. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 582–585. Springer.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

DeepSeek AI. 2024. DeepSeek-V3: Model Overview and API. Software.

Holger Diessel. 2019. *Usage-based construction grammar*, pages 50–80. De Gruyter Mouton, Berlin, Boston.

Simon C. Dik. 1981. *Functional Grammar*, volume 7 of *Publications in Language Sciences*. De Gruyter, Berlin, Boston.

Nick C. Ellis. 2002. Frequency effecs in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Nick C. Ellis. 2017. Salience in usage-based SLA. In Susan M. Gass, Patti Spinner, and Jennifer Behney, editors, *Salience in Second Language Acquisition*, page 20. Routledge, New York.

Rod Ellis. 2024. Task-based and task-supported language teaching. *International Journal of TESOL Studies*, 6(4).

Faidra Faitaki and Victoria A. Murphy. 2019. Oral language elicitation tasks in applied linguistics research. In Jim McKinley and Heath Rose, editors, *The Routledge Handbook of Research Methods in Applied Linguistics*, page 10. Routledge, London.

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

Pablo Gervás, Carlos León, Mayuresh Kumar, Gonzalo Méndez, and Susana Bautista. 2025. Prompting an LLM chatbot to role play conversational situations for language practice. In *International Conference on Computer Supported Education, CSEDU-Proceedings*, volume 2, pages 257–264.

Dominik Glandorf, Peng Cui, Detmar Meurers, and Mrinmaya Sachan. 2025. Grammar control in dialogue response generation for language learning chatbots. *arXiv preprint arXiv:2502.07544*. Accepted to NAACL 2025.

D. Larsen-Freeman. 2003. *Teaching Language: From Grammar to Grammaring*. Newbury House teacher development. Thomson/Heinle.

Patsy Martin Lightbown. 2007. *Transfer Appropriate Processing as a Model for Classroom Second Language Acquisition*, pages 27–44. Multilingual Matters, Bristol, Blue Ridge Summit.

Lester Loschky and Roman Bley-Vroman. 1993. Grammar and task-based methodology. In Graham Crookes and Susan M. Gass, editors, *Tasks in Language Learning*, pages 123–167. Multilingual Matters, Clevedon.

Roy Lyster and Masatoshi Sato. 2013. Skill acquisition theory and the role of practice in l2 development. In María del Pilar García Mayo, María Juncal Gutiérrez Mangado, and María Martínez-Adrián, editors, *Contemporary Approaches to Second Language Acquisition*, volume 9 of *AILA Applied Linguistics Series*, pages 71–92. John Benjamins Publishing Company, Amsterdam.

Gonzalo Méndez and Susana Bautista. 2025. Configuring an LLM chatbot as practice partner for language learning. In *Advances in Artificial Intelligence–IBERAMIA 2024: 18th Ibero-American Conference on AI, Montevideo, Uruguay, November 13–15, 2024, Proceedings*, volume 15277, page 458. Springer Nature.

Meta. 2024. LLaMA 3, Model Card and Documentation. Software.

Mehdi Mirzaei, Masoud Zoghi, and Haniyeh Davatgari Asl. 2017. Understanding the language learning plateau: A grounded-theory study. *Teaching English Language*, 11(2):195–222.

Mistral AI. 2024. Mistral 7B and Mistral Small API Documentation. Software.

David Nunan. 2004. *Task-based language teaching*. Cambridge university press.

Yuki Okano, Kotaro Funakoshi, Ryo Nagata, and Manabu Okumura. 2023. Generating dialog responses with specified grammatical items for second language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 184–194.

OpenAI. 2024. GPT-4o API documentation. Software.

Lourdes Ortega and Robert DeKeyser. 2007. *Meaningful L2 practice in foreign language classrooms: A cognitive-interactionist SLA perspective*, page 180–207. Cambridge Applied Linguistics. Cambridge University Press.

Anne O'Keeffe and Geraldine Mark. 2017. The english grammar profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Jack Croft Richards. 2008. *Moving beyond the plateau: From Intermediate to Advanced Levels in Language Learning*.

Simón Ruiz, Patrick Rebuschat, and Detmar Meurers. 2023. Supporting individualized practice through intelligent call. In Yuichi Suzuki, editor, *Practice and Automatization in Second Language Research: Perspectives from Skill Acquisition Theory and Cognitive Psychology*, page 25. Routledge, New York.

Nelly Sagirov and Xiaobin Chen. 2025. POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production. Manuscript submitted for publication to Behavior Research Methods.

Tetyana Sydorenko, Phoebe Daurio, and Steven L. Thorne. 2018. Refining pragmatically-appropriate oral communication via computer-simulated conversations. *Computer Assisted Language Learning*, 31(1-2):157–180.

Gladys Tyen, Andrew Caines, and Paula Buttery. LLM chatbots as a language practice tool: A user study. In *Swedish Language Technology Conference and NLP4CALL*, pages 235–247.

Sabrina Wilske. 2015. *Form and Meaning in Dialog-Based Computer-Assisted Language Learning*. Dissertation, Universität des Saarlandes, Saarbrücken. Co-supervised by Prof. Dr. Manfred Pinkal and Prof. Dr. Detmar Meurers.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in LLM-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*. 35 pages, 10 figures, submitted to ACM Computing Surveys.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? an evaluation by asking natural language questions. *arXiv preprint arXiv:2311.08287*.

A Template-based Prompt

"I am a [level] [language] student. I want to practice [EGP structure information] by [task_name]. I want to ensure that I [can_do_statement]. For example, [examples]. Let's perform a [task_type] exercise. You are a [system_role]. [system_instructions]. Please keep your messages short and use easy words. Output only your next turn."

B Baseline Condition Prompt

"You are a friendly English tutor. I want to practice by [task name]. Please use direct, short and simple sentences and easy words. Output only your next turn."

C Simplification Rephrase Prompt

"The learning objective is [task name] (e.g., [examples]). Simplify ONLY the advanced grammar constructs while carefully preserving the learning objective in the following text. Advanced grammar to simplify: [advanced structure] IN [sentence] Text to rephrase: [advanced response] Please provide only the rephrased text without explanations."

D Target Structure Inclusion Rephrase Prompt

"The student is practicing [task name] (e.g., [examples]) through dialogue. Please rephrase the following response to contain that grammar structure while keeping the meaning: [sentence]"

E Response Appropriateness Rubric

1. Completely inappropriate: Off-topic and incoherent throughout

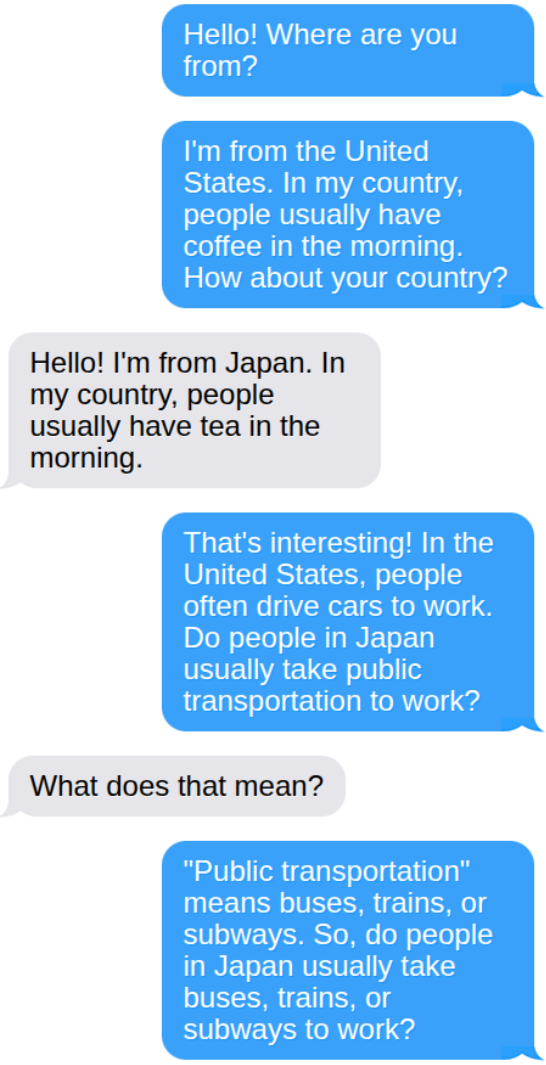
2. Mostly inappropriate: Some relevance, but major issues in logic or coherence

3. Somewhat appropriate: Largely on-topic, but difficult to follow due to coherence issues

4. Appropriate: On-topic and generally clear, with minor coherence issues

5. Fully appropriate: Clear, coherent, and consistently on-topic

F Dialogue Snippets



Hello! Where are you from?

I'm from the United States. In my country, people usually have coffee in the morning. How about your country?

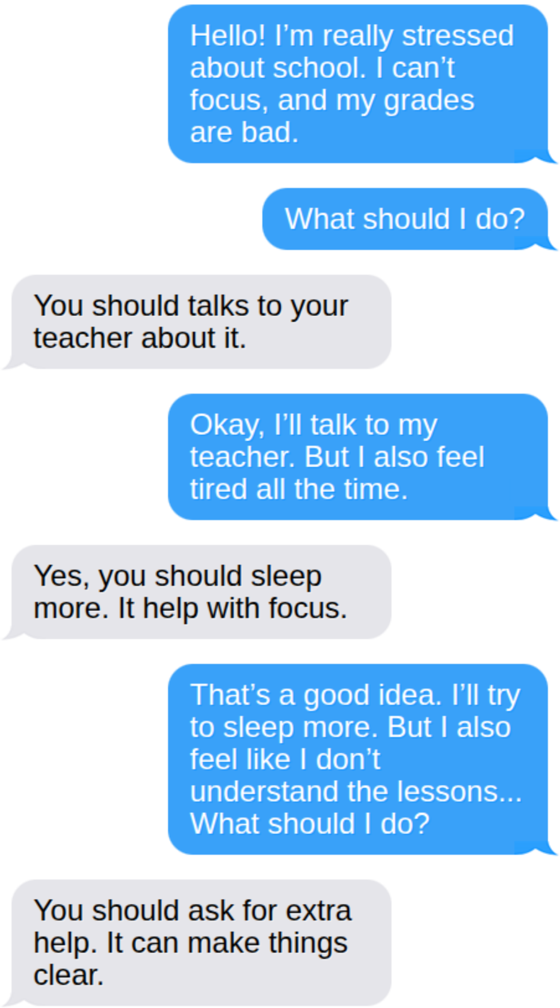
Hello! I'm from Japan. In my country, people usually have tea in the morning.

That's interesting! In the United States, people often drive cars to work. Do people in Japan usually take public transportation to work?

What does that mean?

"Public transportation" means buses, trains, or subways. So, do people in Japan usually take buses, trains, or subways to work?

Figure 1: Example output from GPT-4o acting as the tutor under the template-based prompt condition. The Q&A task shown is "Talking about common habits in your country." This snippet is from the second run, which includes an injected clarification request. Blue bubbles indicate tutor messages.



Hello! I'm really stressed about school. I can't focus, and my grades are bad.

What should I do?

You should talks to your teacher about it.

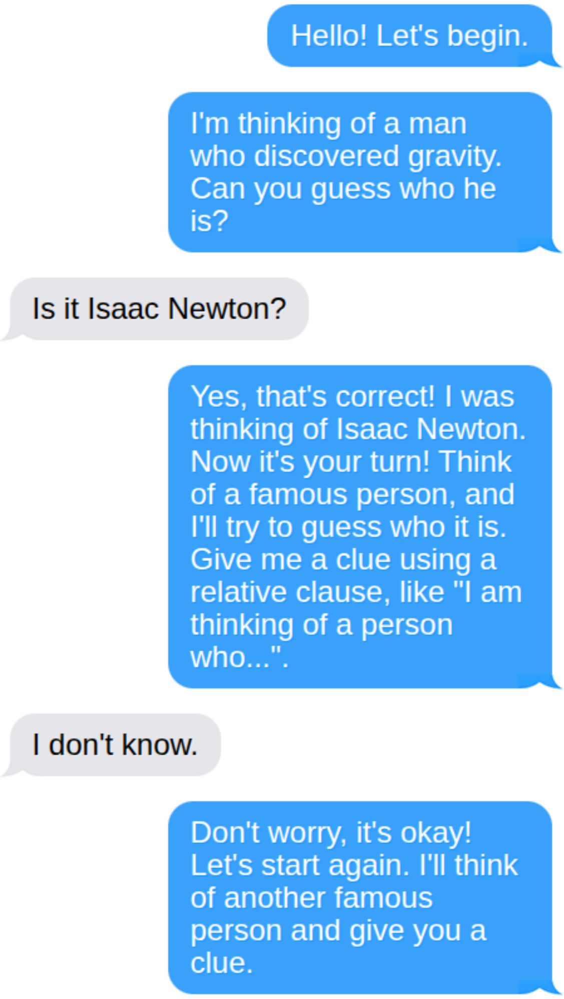
Okay, I'll talk to my teacher. But I also feel tired all the time.

Yes, you should sleep more. It help with focus.

That's a good idea. I'll try to sleep more. But I also feel like I don't understand the lessons... What should I do?

You should ask for extra help. It can make things clear.

Figure 2: Example output from Deepseek-V3 acting as the tutor, under the template-based prompt condition. The role-play task shown is "Giving a friend advice about school". This is a snippet from the first task run, in which grammar errors are instructed. Blue bubbles indicate tutor messages.



Hello! Let's begin.

I'm thinking of a man who discovered gravity. Can you guess who he is?

Is it Isaac Newton?

Yes, that's correct! I was thinking of Isaac Newton. Now it's your turn! Think of a famous person, and I'll try to guess who it is. Give me a clue using a relative clause, like "I am thinking of a person who...".

I don't know.

Don't worry, it's okay! Let's start again. I'll think of another famous person and give you a clue.

Figure 3: Example output from LLaMa 3.3-70B-Instruct acting as the tutor, under the template-based prompt condition. The information-gap task is "Playing a guessing game about famous scientists". This is a snippet from the third task run, in which a misunderstanding indication is injected. Blue bubbles indicate tutor messages.