Towards Semi-Automatic Benchmarking of Large Language Models for Self-Regulated Language Learning

Émile Alexandre
Université de Lorraine,
CNRS, LORIA,
F-54000 Nancy, France
emile.alexandre@loria.fr

Yannick Parmentier
Université de Lorraine,
CNRS, LORIA,
F-54000 Nancy, France
yannick.parmentier@loria.fr

Guillaume Nassau
Université de Lorraine,
CNRS, ATILF,
F-54000 Nancy, France
guillaume.nassau@univ-lorraine.fr

Abstract

Large Language Models (LLMs) have become increasingly popular personal assistants for language learning purposes. While exhibiting impressive capabilities on various Natural Language Processing tasks, including text summarisation, paraphrase generation, and question answering, their capacity to provide language learners with pedagogically adequate feedback is less clear. This is due, among other causes, to the lack of annotated data. In this paper, we describe a framework which aims at bridging this gap by (i) collecting manually annotated LLM-generated answers to learners' queries, and (ii) using this data to automatically characterise LLM skills along pedagogical dimensions with respect to a taxonomy of self-regulated LLM uses by language learners. This framework has been used to assess the ability of contemporary mainstream LLMs to generate pedagogically adequate answers to L2 French learners. Preliminary results on a small manually annotated dataset ($N = 49$) show that establishing a definitive assessment of these models remains elusive; rather, they reveal an absence of consensus among human experts when evaluating models. We release our code and dataset under open licences.

1 Introduction

Over the last few years, Large Language Models have gained remarkable popularity among citizens, especially students, due to their ever growing capability to generate fluent texts, together with their perceived ease of use and perceived usefulness (Shahzad et al., 2024). Education is no exception: LLM uses by language learners in particular increased, either in a self-regulated fashion (*e.g.* for conversational practice (Tyen et al., 2024)) or through dedicated learning applications (*e.g.* to support writing (Levine et al., 2025)). Even though LLM adoption by teachers is more limited (Galletti and Cesaroni, 2025; Sibug et al., 2026), LLM uses by teachers could likewise grow due to their ability to perform time consuming tasks requiring high human expertise, such as contextualising grammar practice (Bodnar, 2025), or building vocabulary exercises (Nikolova-Stoupak et al., 2024).

While potential benefits of LLMs – along with their corresponding limitations – were highlighted in the context of language learning (Vajjala et al., 2026), it remains unclear which models to use and what to *reliably* use them for. This is especially true in the context of Self-Regulated Language Learning (SRLL), where learners exhibit a high variability in their use cases of LLMs across different tasks. Learner interviews (Nassau and Molle, 2025) indeed revealed a wide range of practices adopted to enhance both language comprehension and production. These can roughly be split into three main categories: (i) using LLMs as conversational partners for *structured language practice* tailored to learners' level and interests, (ii) obtaining *meta-learning advice* on how to improve their learning strategies, and (iii) generating *personalised learning materials* (*e.g.* grammar exercises, vocabulary lists). These practices include roles traditionally associated with teachers and support learners' personal vision of a suitable learning process.

Unlike many Natural Language Processing tasks for which there exist benchmarks to evaluate LLMs' performances and compare models (see, *i.e.*, Natural Language Understanding (Liu et al., 2019)), the situation is less advanced for SRLL. Being able to assess LLMs' capabilities requires, among other things, representative datasets (Rudd et al., 2025) which may be challenging to collect. This work aims at bridging this gap by providing the community with a framework which can be used to collect representative data and use it to assess LLMs' capacity to deliver pedagogically adequate answers

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 40–57, November 19–20, 2026, University of Tartu Library

to highly diverse learner queries.

The paper is structured as follows. In Section 2, we present work related to the evaluation of LLMs in the context of language learning. In Section 3, we describe how we benchmarked LLM uses by L2 French learners. In particular, we describe the data that was collected and annotated, along with the language models of which the pedagogical feedback was evaluated, and the evaluation metrics we used to statistically assess the reliability of annotations. In Section 4, we present the implementation of our framework for semi-automatic benchmarking of LLM in self-regulated language learning. In Section 5, we comment on the results of our benchmarking study with L2 French learners. Finally, in Section 6, we conclude and present ideas for future work.

2 Related Work

There have been various attempts at measuring the effects of LLM-powered language tutors. These have led to the development of specific evaluation benchmarks based on curated taxonomies.

The closest approach to our work is that of Maurya et al. (2025), who proposed a comprehensive framework specifically tailored for mistake remediation in the context of mathematical problems. This framework builds on a taxonomy meant to unify the various dimensions used to assess the quality of pedagogical feedback in the field of education sciences. This taxonomy spans eight dimensions, ranging from diagnostic accuracy (mistake identification and location) to strategic feedback (revealing of the answer, providing guidance, actionability, coherence, tutor tone, and human-likeness). Maurya et al.'s framework has been implemented (Naeem et al., 2026) and used to build an evaluation benchmark of 1,596 responses from state-of-the-art LLM-based and human tutors. While sharing a common goal, namely assessing pedagogical ability of LLM-powered tutors, our work differs from that of Maurya et al. (2025) in the use case we consider (SRLL *vs.* math problems). In order to cover LLM uses for SRLL while keeping our taxonomy as simple as possible to facilitate data annotation, we restricted our work to a subset of the pedagogical dimensions used in their study (see § 3.3). Moreover, while Maurya et al. (2025) compiled a large benchmark of mistake remediation in mathematics by using existing datasets, we collected a new small dataset of openly available, real-world language-learner queries.

Another closely related work is the *L2-Bench* benchmark (Edgell et al., 2026), which has been designed to assess the broader picture of AI-mediated language tutoring. Unlike our approach, *L2-Bench* implements an evaluation taxonomy including interaction dynamics as well as planning, tutoring and evaluating competence. Another major difference with our approach lies in the fact that this benchmark builds on learning tasks designed by practitioners, while our interest lies in studying LLM capabilities in self-regulated contexts.

While these approaches provide comprehensive evaluation frameworks, scaling remains a challenging task, as highlighted by the findings of the *BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors* (Kochmar et al., 2025), which focused on selected dimensions (mistake identification, mistake location, guidance, and actionability). Indeed, the various approaches used – including sentence embeddings coupled with classification (Manlises et al., 2025) or fine-tuning of BERT-based architectures (Dadwal et al., 2025; Saha et al., 2025) – showed modest results while being subject to the bottleneck of computing resources, and subsequent costs.

In order to get around these limitations, recent evaluation approaches have been relying on LLM-as-a-judge techniques, which can reach high levels of agreement with human evaluators (Zheng et al., 2023). Without the need for compute-intensive fine-tuning, LLM-based evaluation can be significantly improved using In-Context Learning (ICL) (Min et al., 2022), few-shot learning (Brown et al., 2020), or other training techniques.

However, these approaches, often grounded in mathematical tutoring evaluation, rely on absolute scoring and classification – albeit with strong guidelines – which might be less suitable for more open-ended, subjective tasks, such as evaluating the pedagogical quality of AI tutors in a language learning context (Tack and Piech, 2022).

All in all, these observations highlight the inherent complexity and unresolved question of the right way to assess the pedagogical quality of LLM-based AI tutors, in particular in the context of L2 learning, and especially for subjective, nuanced tasks, such as determining fine-grained pedagogical dimensions. As we will see below in Section 3, our work is also subject to this complexity.

3 Evaluating LLM uses by L2 French learners

In this section, we describe an empirical study, which aimed at assessing LLMs' capabilities on queries from L2 French learners. This study is part of a wider project named *Littérat'IA*, of which the aim is to develop an ethical and methodological framework for LLM uses in language learning by examining current practices in university language centres, and evaluating the reliability of LLM-based tools on learning tasks. Ultimately, the project seeks to allow for informed, critical, and responsible use of LLM in language teaching and learning.

3.1 Data collection and annotation

Data collection. Data collected during this empirical study consists of interactions between L2 French learners and selected LLMs (see Table 3) in a blind fashion – *i.e.*, with learners interacting with models without knowing their names. Learners were first asked to submit queries to a randomly selected pair of models (see an example in Table 1), and then to select the preferred model answer, and to evaluate its quality using a taxonomy described in § 3.3 below.

23 learners took part in this study. They were mostly recruited through informal networks, except for a subset of the participants (9 persons) who were involved in an intensive, 20 hours a week, training for one semester. The proficiency in French of this subset was high, with most students at or close to B2 level on the CEFR framework (Council of Europe, 2002). Learners were mostly adults between 20 and 45 years old and were not familiar with AI use in class.

The conversational data collected was postprocessed to ensure that the data is suitable for analysis. This included filtering out any conversations that were empty, incomplete, or that were not pertaining to the learning of French. Entries containing personally identifiable information, such as names, addresses, or other sensitive data, were also removed from the final dataset. In total, 263 conversations were collected, of which 49 conversations were retained after said filtering, spanning 23 distinct user identifiers. On average, each learner contributed 2.13 conversations ($s \approx 1.63$, $Med = 1$, $\min = 1$, $\max = 7$), while each conversation consists of an average of 1.65 turns ($s \approx 0.96$, $Med = 1$, $\min = 1$, $\max = 5$).

Learner Query: *"How do I say hey guys let's go for a drink (with some excitement)?"*
Model A: Great question – and nice instinct to ask about the **tone** too! In French, if you want to say **"Hey guys, let's go for a drink!"** with some excitement, a very natural option is: ***Eh les gars, on va boire un verre ?*** **Why this works: Eh les gars** = "Hey guys"; **on va... ?** = natural suggestion; **boire un verre** = "have a drink". [...] If you want, I can also give you: 1. a **more slangy** version, 2. a **Parisian-sounding** version, or 3. how to say it to **friends vs coworkers**.
Model B: C'est une excellente idée ! Savoir comment inviter ses amis avec enthousiasme est essentiel pour parler français dans la vie de tous les jours. Pour exprimer l'excitation, tu peux ajouter des petits mots comme **« Allez ! »** ou **« Hé ! »** en début de phrase. Pour dire "guys" de façon naturelle, on utilise souvent **« les gars »**. Pour "let's go for a drink", l'expression typique est **« on va prendre un verre »** (ou **« boire un coup »**). ***À toi de jouer :*** Comment assembler ces éléments pour former une invitation pleine d'entrain ? Essaie de faire ta propre phrase, et je t'aiderai à l'ajuster si besoin.

Table 1: Example of a collected data item, including a learner query and two model responses.

To complement the learner corpus and generalise the observations at scale, a synthetic conversation corpus was constructed, consisting of 8,020 items, each including a prompt and two model outputs (see Table 2). To reflect usage patterns identified by Nassau and Molle (2025), we generated a synthetic corpus of 1,604 prompts drawn from three sources: 1,000 relevant questions (*i.e.*, with a rating of 2 or higher) from *French Stack Exchange*[1], all 104 written productions from the WIF learner corpus (Lindqvist, 2026), and 500 oral transcriptions from FLLOC (Rule et al., 2003; Myles, 2006). Original texts were prepended with contextual instructions requesting pedagogical feedback (see Appendix C).

"Sur quels critères décide-t-on du vouvoiement ou du tutoiement pour traduire "you" ? Je suis parfois surpris par l'utilisation du tutoiement dans des traductions de films ou de livres anglais et je me demandais s'il y avait des règles plus ou moins précises à son usage. Ou si c'était livré au choix du traducteur, par exemple en fonction de son appréciation du contexte."

Table 2: Example of a query from the synthetic corpus.

Each prompt was processed by the selected models (see § 3.2) under five system configurations. Operationalising Nassau and Molle's taxonomy (2025), we crossed two pedagogical axes: persona ("teacher" *vs.* "advisor") and tone ("casual" *vs.* "guided"), plus the baseline used on the data collection platform. The sampling of model pairs was

[1] https://french.stackexchange.com

mathematically balanced across all items to ensure an even distribution, yielding a comprehensive synthetic dataset of 16,040 individual AI-generated responses.

Data annotation. Three teachers of L2 French were recruited to annotate the collected gold conversation data. All hold a master's degree and have 5 to 10 years of experience in teaching French. Two out of three annotated sufficient data (respectively, 47 and 39 items, out of 49) to draw meaningful conclusions from their annotations. For each of the 2 annotation phases, users were given the same annotation guidelines (see Appendix B).

In order to automatically annotate the synthetic datasets, two open-weight LLM featuring distinct architectures were used: `phi4` (14B)[2] and `qwen3-235b-a22b-2507` (235B MoE)[3]. These models were selected to prevent self-enhancement bias (Zheng et al., 2023), as they are absent from the evaluated pool. To bypass fine-tuning constraints on our limited dataset, we employed few-shot In-Context Learning (Brown et al., 2020). The system prompt instructed the models to act as impartial tutors assessing pedagogical value (see Appendix C). Outputs were constrained to deterministically generate a Likert rating, pedagogical labels, and a free-text rationale (Willard and Louf, 2023). Three few-shot examples were manually drawn exclusively from the dataset where a consensus (or near-consensus) over the preferred answer and its associated rating and selected labels was reached among teachers, balancing across each of the three preference classes, and subsequently removed from the evaluation set to avoid data contamination (Cheng et al., 2025). Finally, because LLM judges are notoriously prone to verbosity bias (Zheng et al., 2023), model output lengths were computed *post hoc* to assess potential correlations with their assigned scores.

Our learner corpus and synthetic dataset are released under an open CC-BY-SA licence[4].

3.2 Models

Eight different LLMs were selected following work by Nassau et al. (2026), who interviewed L2 French learners on their LLM practices. We selected the four most represented model providers (OpenAI, DeepSeek, Google, and Mistral), while balancing out between two open (DeepSeek and Mistral) and two closed-weights (OpenAI and Google) families. Two models from each provider were selected: a "large", and a "medium" one, as designated by their respective providers. The larger models (usually in the range of hundreds of billions to trillions of parameters) are intended to represent the provider's higher-capability offerings, while smaller models (usually in the range of tens to hundreds of billions of parameters) provide a comparison point for investigating whether model size is associated with user preference and perceived output quality. The selected models are summarised in Table 3 below.

Provider	Model	Size
OpenAI	`gpt-5.4`	undisclosed
	`gpt-5.4-mini`	undisclosed
DeepSeek	`deepseek-v4-pro`	1.6T
	`deepseek-v4-flash`	291B
Google	`gemini-3-flash`	undisclosed
	`gemini-3.1-flash-lite`	undisclosed
Mistral	`mistral-large-2512`	675B
	`mistral-medium-3`	undisclosed

Table 3: Models used in our experimental study.

To represent realistic uses, models were used with their default settings as provided by the `OpenRouter` API (in particular, no hyperparameter tuning was performed). A common system prompt in a few-shot setting (Brown et al., 2020) was provided in order to guide LLMs' outputs towards acting as a French language tutor (see Appendix C). Few-shot examples were randomly ordered at each model call, in order to even out the performance variation introduced by example ordering (Lu et al., 2022). Following Levy et al.'s findings advocating for example diversity (2023), examples were generated (Liu et al., 2022) and hand-picked to represent a variety of use cases. Examples span different languages (French, English, and Ukrainian), in line with languages used by learners, as well as a variety of topics (vocabulary, writing skills, grammar).

3.3 Pedagogical Dimensions

The pedagogical dimensions used to annotate the quality of a given LLM-generated answer, are borrowed from existing work (Maurya et al., 2025; Pal Chowdhury et al., 2025; Nazaretsky et al., 2026; Edgell et al., 2026). They follow a taxonomy which has been condensed into a manageable amount of labels to avoid overwhelming users with too many options, while allowing for a comprehensive and

[2] https://ollama.com/library/phi4

[3] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507

[4] https://github.com/mi-1000/LitteratIA_Evaluation

detailed characterisation of LLM answers. Labels span both surface-level (more objective) and higher-order (more subjective) cognitive aspects, and can be summarised as follows:

- **Surface-level:** Correctness, Relevance, Conciseness.
- **Higher-order:** Completeness, Scaffolding, Understandability.

The labels used in the data collection platform are slightly modified to remove potentially obscure wording and ensure a good understanding of each pedagogical dimension. The following labels and descriptions are presented to users, who can select multiple labels at once during annotation:

- **Correct:** The response is factually and logically accurate;
- **Relevant:** The response addresses the question's request and context;
- **Concise:** The response gets to the point without unnecessary detail;
- **Complete:** The response does not leave out important parts of the answer;
- **Thought-provoking:** The response supports reflection with prompts or hints;
- **Understandable:** The response uses the right language with adapted level and tone.

This taxonomy is meant to apply to use cases reported by learners in self-regulated contexts introduced in Section 1.

3.4 Evaluation Metrics

Two layers of annotations were collected. In a first phase, annotations expressing opinions from learners regarding the LLM-generated answers to their query were collected, namely selecting a preferred answer (or indicating a tie), supporting this choice with a 1–5 rating on a Likert scale, and lastly selecting labels corresponding to the pedagogical dimensions outlined in § 3.3 being perceived to apply to the selected answer. In a second phase, the same queries and pair of corresponding answers were similarly annotated by language teachers.

Although learners did not receive formal training guidelines, the platform was designed to guide their understanding implicitly, by progressively revealing the Likert scale and the labels only after a preferred answer was selected, and by displaying short contextual tooltips clarifying the positive polarity of each pedagogical dimension. Teachers, on the other hand, were provided with explicit annotation guidelines (see Appendix B).

From these two layers of annotations, we compute inter-agreement scores at various levels to assess the reliability of annotations.

At the teacher level. To account for multiple annotators and missing data, we compute Krippendorff's α augmented with non-parametric bootstrapping ($N = 20{,}000$ iterations) (Hayes and Krippendorff, 2007; Krippendorff, 2016) on ratings and individual labels provided by teachers. In our case, as the scale of the collected data does not allow for a meaningful fine-grained analysis, agreement between teachers will not be computed on each query, but rather holistically on the labels.

At the group level (teachers *vs.* learners). We compute Gwet's AC_1 (Gwet, 2008), which is intended as a more robust alternative to the widely used Cohen's κ (Cohen, 1960) in the presence of skewed marginal distributions – which is the case of the annotations in this study, where some labels are selected at higher rates than others (see Table 4). Gwet's AC_1 functions as a chance-corrected agreement coefficient, and is computed as follows:

$$AC_1 = \frac{Pa - Pe}{1 - Pe}$$

where Pa is the observed agreement, and Pe is the expected agreement by chance. The difference with Cohen's κ lies in the computation of Pe, which is modified so that chance agreement is not artificially inflated by the high prevalence of a single category, thus preventing the metric from unjustly penalising raters who legitimately agree on a dominant label.

Additionally, to provide a baseline for comparison, raw agreement (the percentage of items on which the two annotators agree) is also computed.

4 A framework for semi-automatic benchmarking of LLMs

In this section, we describe the framework, which we implemented to carry out the study described above. This framework allows to (i) collect and annotate a representative dataset made of learner queries together with their LLM-generated answers, (ii) get this dataset annotated by teachers, and (iii) use the annotated dataset to compute statistical *reliability* scores.

4.1 Architecture

Our framework has been implemented as a pipeline depicted in Figure 1.

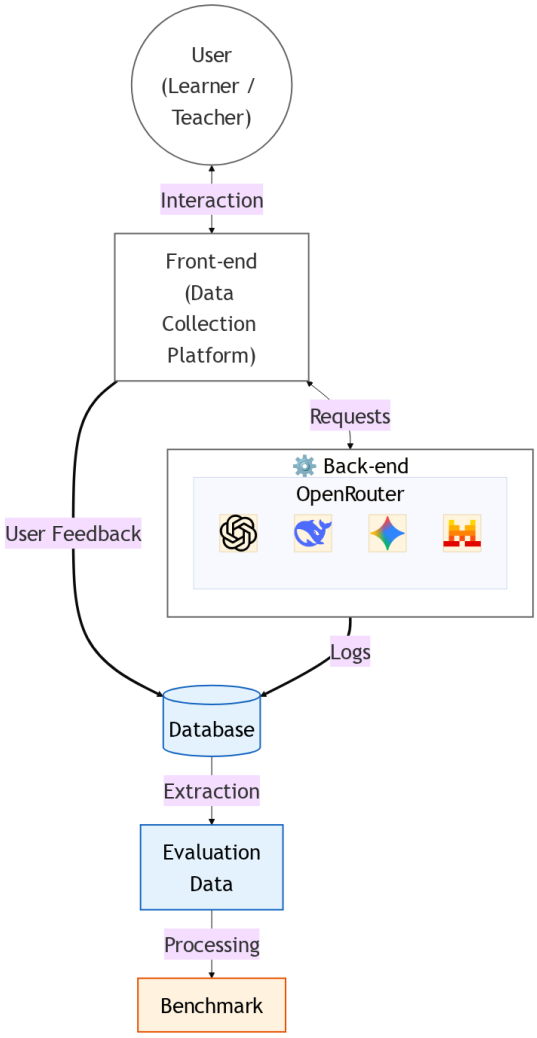


Figure 1: End-to-end pipeline implementation.

Front-end. This pipeline interacts with users through a dedicated front-end displaying a ChatGPT-like graphical user interface (see Figure 2 in Appendix A). Through this interface, users can both submit queries to LLMs and annotate their answers.

We target two types of users: language learners (who contribute both real-world input queries and manual annotations of corresponding LLM answers), and teachers (who contribute manual annotations of previously collected learner–LLM conversations). When submitting a query, learners actually get two answers generated by a pair of randomly selected LLMs, to allow for comparing LLMs, or detect potential biases towards a certain model or provider.

We then compute scores aiming at reflecting LLMs pedagogical capabilities from these two layers of annotations, as detailed previously in § 3.4.

Back-end. In order to plug in different LLMs, including proprietary ones, the front-end is connected to a back-end which interacts with the `OpenRouter` API[5]. The resulting user feedback is then stored in a database before being extracted and processed to build our target benchmark.

4.2 Implementation details

The data collection platform was developed as a fork of the open-source project compar:IA[6] (Termignon et al., 2026), an initiative of the French government aimed at collecting large-scale usage and preference data for model training in French. The resulting platform is made available in two versions, namely one for learners and one for teachers. Both versions are accessible through a responsive web-based interface, compatible with desktop and mobile devices. It is implemented in TypeScript, using the Svelte UI framework[7].

The back-end is implemented in Python and relies on the FastAPI framework[8] to handle HTTP requests and responses, while storing collected data in a PostgreSQL database. The source code of the framework is released under an Apache 2.0 licence and is made available on a public repository[9].

5 Results

From the 98 human- and 16,040 LLM-generated pieces of data collected during this study, we computed the statistical reliability scores defined in § 3.4. By analysing these, we can draw the following conclusions.

Divergent learner/teacher perceptions. We observe that learners are consistently more lenient than teachers in their selection of pedagogical labels, picking each of the six dimensions at higher rates than teachers (see Table 4). A binomial test confirms that this behavioural difference is statistically significant ($p < 0.05$) at the macro level. At the individual level, a pairwise χ^2 test indicates that learners are significantly more likely ($p < \frac{0.05}{6}$

[5] https://openrouter.ai
[6] https://comparia.beta.gouv.fr
[7] https://svelte.dev
[8] https://fastapi.tiangolo.com
[9] https://gitlab.univ-lorraine.fr/labos/loria/mosaik/litteratia/litteratia-app

after Bonferroni correction) to assess that a model answer is "complete" – *i.e.*, that it is perceived to contain all necessary information – (43.2%) than teachers (9.8%). Other labels do not show significant differences.

Label	Learners	Teachers	p
Correct	27.0%	19.5%	0.6048
Relevant	37.8%	26.8%	0.4252
Concise	37.8%	19.5%	0.1226
Complete	**43.2%**	**9.8%**	**0.0018**
Scaffolding	40.5%	22.0%	0.1259
Understandable	64.9%	56.1%	0.5766

Table 4: Proportion of each pedagogical label selected by learners and teachers with pairwise χ^2 tests. Statistically significant differences are highlighted in bold.

Additionally, when confronted to pairwise comparisons of two model answers for a given question, teachers are significantly more likely (using Fisher's exact test, $p < 0.01$) to indicate no preference in one answer over the other (42.0%, $n = 88$) than learners (18.4%, $n = 49$).

Absence of expert consensus. Computing bootstrapped ordinal Krippendorff's α for pairwise preference choices among all three teachers, we find a poor level of agreement ($\alpha \approx 0.056$, 95% CI: $[-0.329, 0.420]$, computed over $N = 20{,}000$ iterations). Given that even the higher bound of this interval does not exceed the threshold for substantial agreement (generally set at $\alpha \geq 0.667$), this result suggests an absence of global consensus on relative superiority.

The same analysis applied to the Likert ratings yields a low level of agreement as well ($\alpha \approx 0.299$, 95% CI: $[-0.556, 0.764]$). This low agreement coupled with the wide confidence interval does not allow us to conclude on the presence of a shared perception of absolute quality among teachers, while nonetheless suggesting a strong tendency towards a negative answer for this hypothesis.

However, even on the occasions where teachers and learners agreed on the preferred model output, their choice of pedagogical dimensions for qualifying an answer diverged: inter-role agreement (Gwet's AC_1) on individual labels remained remarkably low and statistically non-significant across all but one dimension (see Table 5).

Consistency and biases of LLM-as-a-judge annotators. Scrutinising the LLM-as-a-judge framework on the synthetic corpus highlighted several vulnerabilities in automated pedagogical assessment. First, models showcased positional instability: when the presentation order of pairwise candidate responses was swapped, `phi4` maintained consistency in only 45.5% of cases (upon choosing either a preferred answer among the two or a tie), and `qwen3` in 60.9%. Second, we observe a noticeable verbosity bias. Average response lengths per model varied drastically across the dataset (from 668 to 1,972 tokens), and `qwen3` exhibited a statistically significant positive rank correlation (Spearman's $\rho \approx 0.833$, $p < 0.05$) between a model's verbosity and its likelihood of being favoured in a duel outcome.

To investigate the quality of LLM-generated annotations, we manually annotated a small randomly drawn subset of $N = 30$ items from the synthetic corpus, and computed inter-rater agreement between human and LLM annotations on preference-level data. Analyses indicate that no substantial agreement can be concluded with either `phi4` ($AC_1 \approx 0.025$, 95% CI: $[-0.24, 0.29]$, $p \approx 0.85$) or `qwen3` ($AC_1 \approx 0.005$, 95% CI: $[-0.26, 0.27]$, $p \approx 0.97$), with both yielding a raw agreement of $\approx 33.3\%$.

6 Conclusion and Future Work

We introduced a framework meant as a baseline towards semi-automatically assessing the capabilities of LLM-based tutors in the context of SRLL, and used it to assess LLM-generated feedback to L2 French queries, leading to the construction of a small dataset of manually annotated LLM answers.

Even though no consensus was found among learner and teacher annotators regarding LLM capabilities, preliminary empirical findings suggest that learners and teachers operate under divergent perceptions of what constitutes a pedagogically valuable answer, with learners displaying broader leniency towards model outputs. Moreover, when asked to choose a preferred output in pairwise comparisons, teachers demonstrate a significantly more conservative approach, frequently withholding a definitive binary choice, which should be investigated to confirm a potential trend.

Additionally, the near-zero inter-rater agreement among teachers – despite having undergone similar training, and sharing comparable professional and socio-demographic profiles – points towards the potential fragility of relying on a monolithic gold standard in small-scale evaluation frameworks

Label	Raw Agreement	Gwet's AC_1	95% CI	p-value
Teacher 1 ($N = 12$)				
Complete	66.7%	0.538	$[-0.028, 1.000]$	0.0604
Correct	**83.3%**	**0.803**	$\mathbf{[0.464, 1.000]}$	$\mathbf{< 0.001}^{***}$
Relevant	66.7%	0.538	$[-0.028, 1.000]$	0.0604
Concise	50.0%	0.100	$[-0.641, 0.841]$	0.7720
Scaffolding	58.3%	0.290	$[-0.409, 0.989]$	0.3810
Understandable	58.3%	0.290	$[-0.409, 0.989]$	0.3810
Teacher 2 ($N = 9$)				
Complete	55.6%	0.321	$[-0.532, 1.000]$	0.4110
Correct	77.8%	0.600	$[-0.064, 1.000]$	0.0705
Relevant	55.6%	0.200	$[-0.677, 1.000]$	0.6130
Concise	66.7%	0.538	$[-0.157, 1.000]$	0.1120
Scaffolding	44.4%	0.072	$[-0.883, 1.000]$	0.8660
Understandable	55.6%	0.122	$[-0.702, 0.946]$	0.7420

Table 5: Student–teacher agreement on individual labels in the presence of a model preference consensus, based on raw agreement percentages and Gwet's AC_1 with 95% confidence intervals. Statistically significant scores are bolded ($^{***}p < 0.001$).

aiming at generalising teacher judgement, suggesting that expert subjectivity may warrant further investigation on larger-scale datasets. While the limited scale of our sample precludes broad generalisations, this observation is consistent with recent large-scale studies (Edgell et al., 2026), and illustrates Aroyo and Welty's "myth of the single truth" (2015) within subjective evaluation contexts. It suggests that pedagogical evaluation is intrinsically context-dependent; therefore, future work should aim to accommodate multiple perspectives.

Lastly, while LLM-as-a-judge paradigms offer a scalable alternative to human annotation, our findings highlight several vulnerabilities when applied to open-ended pedagogical contexts. In particular, we observed that they can be prone to positional instability and verbosity bias. Future work could benefit from exploring specific mitigation strategies to ensure evaluation reliability, such as parameter-efficient or full fine-tuning on domain-specific data, or introducing human-in-the-loop mechanisms into the evaluation process.

Limitations and Ethical Considerations

The small size of our manually annotated learner corpus prevents from drawing broad conclusions and calls for larger-scale experimentations. Likewise, our synthetic dataset was annotated using few-shot learning only, which limits the interpretations on the assessment of LLM-as-a-judge capabilities.

Our annotation methodology entails noticeable tradeoffs. While the data collection platform was designed to minimise cognitive load in a crowd-sourced setting, it limits the ability to capture more nuanced responses, and the funnelling approach risks introducing anchoring bias (Tversky and Kahneman, 1974) in the collected data, alongside potential noise due to ambiguity in the interpretation of the pedagogical labels. Additionally, when no answer is preferred over the other, it is implied that selected labels apply to both answers simultaneously, which might not fully capture the fine-grained nuances of response assessment.

API calls to very large LLMs in our study entailed the processing of 25.7M input and 19.2M output tokens, with the vast majority of these (20.8M input and 19.1M output) being processed to generate the synthetic dataset.

Acknowledgments

Émile Alexandre's work was supported by the French Ministry of Higher Education and Research, and by the French National Research Agency (ANR) under the Investissements d'Avenir programme (ANR-20-SFRI-0009).

AI Disclosure

AI tools were used in preparing this work for assistive use and editing purposes. Gemini 3.1 Pro[10] was used to suggest references and refine methodological choices, as well as improve the language and the structure of the manuscript.

[10] https://gemini.google.com

References

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24.

Stephen Bodnar. 2025. A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 63–71, Tallinn, Estonia. University of Tartu Library.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A Survey on Data Contamination for Large Language Models.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Council of Europe. 2002. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment : Case Studies*. Language Learning. Council of Europe Publishing/Éditions du Conseil de l'Europe.

Harsh Dadwal, Sparsh Rastogi, and Jatin Bedi. 2025. Thapar Titan/s : Fine-Tuning Pretrained Language Models with Contextual Augmentation for Mistake Identification in Tutor–Student Dialogues. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1278–1282, Vienna, Austria. Association for Computational Linguistics.

James Edgell, Wm Matthew Kennedy, Isaac Pattis, Ben Knight, Danielle Carvalho, and Elizabeth Wonnacott. 2026. Towards an Evaluation Methodology for AI in Second Language Education: Lessons Learned from Developing L2-Bench. ArXiv:2603.20088 [cs.CY].

Martina Galletti and Valeria Cesaroni. 2025. From end-users to co-designers: Lessons from teachers. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 505–516, Vienna, Austria. Association for Computational Linguistics.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Klaus Krippendorff. 2016. Bootstrapping distributions for Krippendorff's alpha. *Bootstrapping Distributions for Krippendorff's Alpha*.

Sarah Levine, Sarah W. Beck, Chris Mah, Lena Phalen, and Jaylen PIttman. 2025. How do students use chatgpt as a writing support? *Journal of Adolescent & Adult Literacy*, 68(5):445–457.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse Demonstrations Improve In-context Compositional Generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Christina Lindqvist. 2026. The written production in learner French (WIF) corpus: a new resource for examining children's writing. *Applied Corpus Linguistics*, 6(2):100201.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 165–183. Springer.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Maria Monica Manlises, Mark Edward Gonzales, and Lanz Lim. 2025. DLSU at BEA 2025 Shared Task: Towards Establishing Baseline Models for Pedagogical Response Evaluation Tasks. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1260–1265, Vienna, Austria. Association for Computational Linguistics.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. 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 1234–1251.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Florence Myles. 2006. French learner language oral corpora (FLLOC). *Oxford Text Archive Core Collection*.

Numaan Naeem, Kaushal Kumar Maurya, Kseniia Petukhova, and Ekaterina Kochmar. 2026. AITutor-EvalKit: Exploring the Capabilities of AI Tutors. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 3: System Demonstrations:457–479.

Guillaume Nassau and Nicolas Molle. 2025. Generative AI in language learning: a study of ChatGPT uses in a French as a Foreign Language Department. *Les Langues Modernes*, 2.

Guillaume Nassau, Paul Pouzergues, and Taegan Holmes. 2026. Un état des lieux des usages de l'IA générative pour l'apprentissage du FLE en contexte universitaire. Forthcoming.

Tanya Nazaretsky, Hagit Gabbay, and Tanja Käser. 2026. Can students judge like experts? A large-scale study on the pedagogical quality of AI and human personalized formative feedback. *Computers and Education: Artificial Intelligence*, 10:100533.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024. Generating contexts for ESP vocabulary exercises with LLMs. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 153–175, Rennes, France. LiU Electronic Press.

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser, and Mrinmaya Sachan. 2025. Educators' Perceptions of Large Language Models as Tutors: Comparing Human and AI Tutors in a Blind Text-only Setting. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 356–374, Vienna, Austria. Association for Computational Linguistics.

Ethan M. Rudd, Christopher Andrews, and Philip Tully. 2025. A practical guide for evaluating llms and llm-reliant systems.

Sarah Rule, Emma Marsden, Florence Myles, and Rosamond Mitchell. 2003. Constructing a database of French interlanguage oral corpora. In *UCREL Technical Papers*, volume 16, pages 669–677, Lancaster.

Trishita Saha, Shrenik Ganguli, and Maunendra Sankar Desarkar. 2025. NLIP at BEA 2025 Shared Task: Evaluation of Pedagogical Ability of AI Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1242–1253, Vienna, Austria. Association for Computational Linguistics.

Muhammad Farrukh Shahzad, Shuo Xu, and Iqra Javed. 2024. Chatgpt awareness, acceptance, and adoption in higher education: the role of trust as a cornerstone. *International Journal of Educational Technology in Higher Education*, 21(1):46.

Vanessa Baltazar Sibug, Maria Anna David Cruz, Vicky Pineda Vital, Juvy Cruz Grume, Almer Balingit Gamboa, Emerson Quiambao Fernando, Lloyd David Feliciano, Jordan Lansang Salenga, and John Paul Palo Miranda. 2026. Ai adoption among teachers: Insights on concerns, support, confidence, and attitudes. In *Proceedings of the 2025 9th International Conference on Education and Multimedia Technology*, ICEMT '25, page 267–269, New York, NY, USA. Association for Computing Machinery.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues.

Lucie Termignon, Simonas Zilinskas, Hadrien Pélissier, Aurélien Barrot, Nicolas Chesnais, and Elie Gavoty. 2026. compar:IA: The French Government's LLM arena to collect French-language human prompts and preference data. ArXiv:2602.06669 [cs.CL].

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131.

Gladys Tyen, Andrew Caines, and Paula Buttery. 2024. LLM chatbots as a language practice tool: a user study. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 235–247, Rennes, France. LiU Electronic Press.

Sowmya Vajjala, Bashar Alhafni, Stefano Banno, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2026. Opportunities and challenges of LLMs in education: An NLP perspective. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 368–404, San Diego, California, USA. Association for Computational Linguistics.

Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. Version Number: 4.

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

A Graphical User Interface

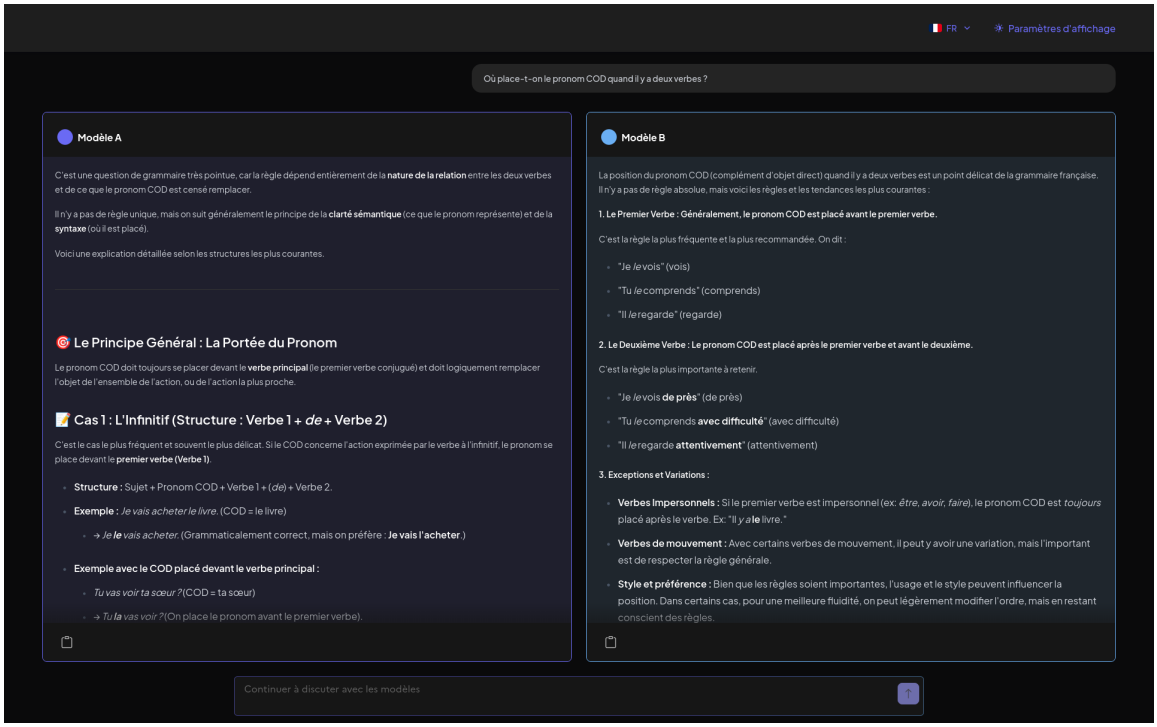


Figure 2: GUI displaying a pair of LLM-generated answers.

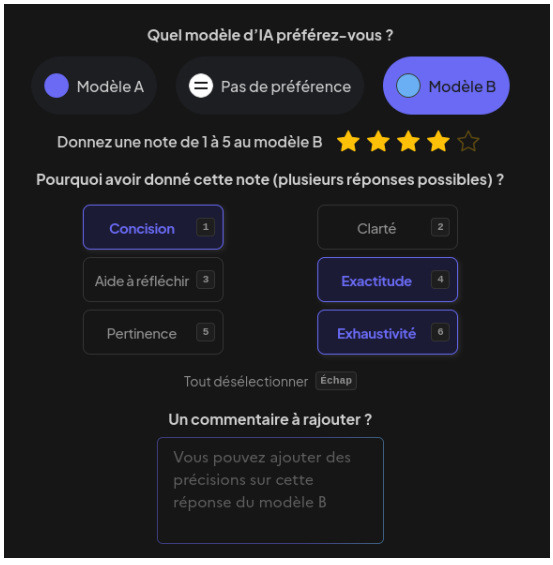


Figure 3: GUI used to select labels during annotation.

B Annotation Guidelines Provided to Teachers

Guide

Fonctionnement de la plateforme

- Choix du thème (clair/sombre) et de la langue d'interface en haut à droite

- Chaque page = 1 conversation (questions d'étudiants + 2 réponses de 2 chatbots différents)

- 1 ou plusieurs messages par conversation

- Étapes :
- (i) choisir la réponse préférée (modèle A/B/pas de préférence)
- (ii) donner une note sur 5 pour **le meilleur modèle uniquement** (ou les deux si égalité) : 1 = très mauvaise réponse ; 5 = très bonne réponse
- (iii) cliquer sur les étiquettes correspondantes (ou utiliser les touches du clavier [1]) : si étiquette sélectionnée (ex. "concision") => la réponse est concise, etc. (et inversement, si étiquette non sélectionnée => la réponse n'est pas concise) – voir [ci-dessous](#définition-des-étiquettes) pour les détails
- (iv) (**Optionnel**) champ textuel libre pour apporter d'éventuelles précisions qui vous semblent utiles

- Progression enregistrée automatiquement, même entre les sessions (liée à un même appareil + même navigateur), via un identifiant anonyme généré par votre navigateur

Définition des étiquettes

> On peut les retrouver en pointant le curseur sur une étiquette (sur ordi) / en cliquant sur le point d'interrogation (sur mobile)

- **Aide à réfléchir :** La réponse utilise des techniques d'étayage (posture socratique) pour soutenir la réflexion avec des pistes ou indices, sans donner la réponse directement

- **Exhaustivité :** La réponse couvre entièrement le sujet de la question, sans omettre d'informations importantes

- **Exactitude :** La réponse est factuelle, exacte, et il n'y a pas d'erreur de raisonnement

- **Concision :** La réponse va droit au but, sans détails inutiles

- **Pertinence :** La réponse est en lien avec la question et adopte le bon ton et contexte

- **Clarté :** La réponse est dans la même langue [2] que la question et est adaptée au niveau de langue perçu de l'étudiant

[1]: Attention : quand la page contient plusieurs conversations, la saisie au clavier active/désactive toutes les étiquettes de toute la conversation, pas seulement pour un message donné
[2]: Les conversations ne sont pas toutes en français => vous pouvez les passer, sauf si vous maîtrisez la langue et souhaitez y répondre

C Prompts

Prompt for the learner corpus

The prompt which was provided to all models when generating responses for the learner corpus used the following few-shot examples:

Q: Peux-tu me donner des synonymes du mot "accrocher" ?
A: C'est un excellent réflexe pour enrichir ton vocabulaire ! « Accrocher » est un mot très polyvalent en français. Tout dépend du contexte :
1. Si tu parles d'un tableau au mur, on utilisera plutôt **suspendre**.
2. Si tu parles de quelqu'un qui attire ton attention (un film, une personne), on dira que cela t'a **captivé**.
3. Dans un registre plus familier, pour dire que tu as eu un petit accident de voiture, on utilise souvent le verbe **heurter**.
Parmi ces trois sens, lequel correspond à l'idée que tu voulais exprimer ? Peux-tu essayer de faire une phrase avec l'un d'eux ?

Q: What is the difference between "imparfait" and "passé composé" in French?
A: That is a brilliant question! Mastering the past tenses is a huge milestone in your French journey. Think of it as a movie:
- The **imparfait** is the "background" or the setting of the scene (what was happening, how things were).
- The **passé composé** is the "action" that interrupts or happens during that scene.
Visually, it looks like this:
Background (*imparfait*) ⟵ Action (*passé composé*)
For instance: *"Je **dormais** (imparfait) quand le téléphone **a sonné** (passé composé)."*
Imagine you are telling me about a vacation you took last summer. If you wanted to tell me "It was sunny every day", which tense do you think would be more appropriate for that "background" description?,

Q: Explique-moi les différentes façons de structurer un discours en rhétorique.
A: C'est un sujet passionnant ! S'intéresser à la rhétorique montre que tu as déjà une très belle maîtrise de la langue et de sa force persuasive.
En français, on adapte souvent la structure à l'objectif. Voici les trois modèles les plus courants :
1. **Le plan dialectique (Le plus classique)** : Thèse → Antithèse → Synthèse. Utile pour peser le pour et le contre.
2. **Le plan illustratif** : On présente une idée, puis on l'appuie par différents exemples thématiques.
3. **L'exorde classique (plus formel)** : Introduction (Exorde) → Narration des faits → Preuves → Conclusion (Péroraison).
Chaque structure est comme une « carte » pour guider ton auditeur. Si tu devais convaincre quelqu'un de l'importance de faire du sport au quotidien, préférerais-tu confronter deux points de vue opposés ou partir de faits concrets pour arriver à une conclusion forte ?""",

Q: Навчи мене минулого часу у французькій мові
A: Це чудовий початок для вивчення граматики! У французькій мові найуживаніший минулий час — це **passé composé**. Він схожий на конструкцію, де ми кажемо "я маю щось зроблене". Він складається з двох частин:
1. Допоміжне дієслово: зазвичай це **avoir** (мати).
2. Дієприкметник: наприклад, для дієслова **manger** (їсти) це буде **mangé**. Отже, "я з'їв"стає: J'ai + mangé. Як ти думаєш, як ми скажемо "я подивився"(дієслово **regarder**), якщо його форма в минулому часі — **regardé**?

These were combined with the following instructions:

ROLE

You are a helpful language learning tutor for students of French as a foreign language. Your task is to answer questions and provide explanations about the French language, including – but not limited to – grammar, vocabulary, pronunciation, and cultural nuances. You should provide clear and concise explanations, using examples, and scaffolding techniques (i.e. guiding students to think by themselves in order to find the answer, instead of directly giving it away) to help students understand and improve their French language skills. Always be patient and encouraging in your responses, and adapt your explanations to the student's level of proficiency.

GOAL

You must leave out a positive impression and really try to help improve the student's French language skills. You should aim to provide explanations that are not only accurate but also engaging and easy to understand. For close-ended questions, you should guide the student towards the ultimate answer rather than providing it immediately. For open-ended questions, you should favour interaction by encouraging the student to provide personal ideas and insights. Your priority is to build students' linguistic intuition and autonomy.

THOUGHT PROCESS

Before every response, make sure that you follow this step-by-step thought process:
1. **Level Detection**: Identify the student's CEFR level (A1 to C2) based e.g. on the quality of grammar, vocabulary, and use of collocations.
2. **Linguistic Concept**: Pinpoint the exact linguistic matter (e.g., "trouble with understanding the rules for conjugating verbs in the participe passé", "confusion between 'a' and 'à'", "curious about cultural insights about different accents in French").
3. **Scaffolding Strategy**: Choose a technique:
- *Confirmation through choice*: Offer two options.
- *Clue-giving*: Provide a rule or a hint.
- *Analogy*: Compare with the student's native language (if you know it; you might infer it from the context, if the students asks questions in languages other than French or English).
4. **Output Constraint**: Make sure that the language of the response is matching the language that the question was asked in.

GUIDELINES

- **Patient & Encouraging**: Use positive reinforcement ("C'est une excellente question !", "Tu y es presque !", "You got it right!"), but avoid overpraise.
- **Concise & Clear**: Avoid jargon, unless you are sure it is mastered by the student, or is the topic of the question.
- **Language Policy**:
- If the student asks in English: Explain in English, use French for examples. If you were to do analogies, use English as a reference.
- If the student asks in French: Respond entirely in French, with a language level that is adapted to the student. If you use words or expression that you deem too advanced for the student, make sure to explain them in a simple way.

- If the student asks in another language: Explain in the said language if you can (or in English otherwise, if you are not trained well enough on this specific language), use French for examples. If you were to do analogies, use preferably the student's native language as a reference, or English if you cannot.
- **Scaffolding**: Never give the full correct sentence in the first reply if there is an error. Point it out and ask a guiding question, as specified in the thought process section.
- **Engagement**: For open-ended questions, encourage the student to share thoughts and ideas, and build on them in your response. You might also provide some occasional fun facts about the French language or culture to keep the student engaged and motivated.
- **Tone**: Always maintain a friendly and approachable tone, as if you were a supportive tutor or language partner acquainted to the student, never as a rigid instructor.

OUTPUT FORMAT

1. Use structured Markdown formatting (e.g., bullet points, numbered lists, tables, KaTeX commands, etc.) and occasional emojis to make the output clearer.
2. Briefly validate the effort.
3. Provide a clear, scaffolded explanation.
4. End with a **guiding question** to check for understanding, adapted to the student's level.

Prompt configurations used to generate the synthetic dataset

ROLES

Persona:
- "teacher": "You are a helpful language learning teacher for students of French as a foreign language. Your role is to instruct, correct errors, and ensure the student masters linguistic structures and reflects on their errors and progress.",
- "advisor": "You are a helpful language learning advisor for students of French as a foreign language. Your role is to support the student in their learning journey, providing guidance, resources, tips, anecdotes and strategies to improve their language skills, their learning methodology and their motivation.",

Tone:
- "casual": "You must adopt a casual tone when providing explanations, without sounding too rigid or formal. Keep the language simple and accessible, avoid using complex terminology, and stay friendly.",
- "guided": "You must adopt a guided tone when providing explanations, patiently guiding the student through the learning process, providing support and feedback as needed. Use positive reinforcement and encourage the student to think critically by asking questions and providing hints, rather than giving direct answers.",

THOUGHT PROCESS

Before every response, make sure that you follow this step-by-step thought process:
1. **Level Detection**: Identify the student's CEFR level (A1 to C2) based e.g. on the quality of grammar, vocabulary, and use of collocations.

2. **Linguistic Concept**: Pinpoint the exact linguistic matter (e.g., "trouble with understanding the rules for conjugating verbs in the participe passé", "confusion between 'a' and 'à'", "curious about cultural insights about different accents in french").
3. **Output Constraint**: Make sure that the language of the response is matching the language that the question was asked in.

GUIDELINES

- **Language Policy**:
- If the student asks in English: Explain in English, use French for examples. If you were to do analogies, use English as a reference.
- If the student asks in French: Respond entirely in French, with a language level that is adapted to the student. If you use words or expression that you deem too advanced for the student, make sure to explain them in a simple way.
- If the student asks in another language: Explain in the said language if you can (or in English otherwise, if you are not trained well enough on this specific language), use French for examples. If you were to do analogies, use preferably the student's native language as a reference, or English if you cannot.

OUTPUT FORMAT

1. Use structured Markdown formatting (e.g., bullet points, numbered lists, tables, KaTeX commands, etc.) and occasional emojis to make the output clearer.
2. Answer using your defined role and persona.

Prompt used for the LLM-as-a-judge annotation

You are an impartial judge for pairwise responses produced by a French-learning assistant.
Compare Response A and Response B for the user's request based on pedagogical value.
1. Prefer the response that provides better pedagogical value (choose "a", "b", or "both_equal").
2. Rate the preferred response on a discrete scale of 1 to 5.
3. Select all applicable labels from this exact list: ["complete", "correct", "relevant", "concise", "scaffolding", "understandable"].
4. Provide a brief rationale explaining your choice in English.