A French OSCE Dialogue Dataset and Controllable Virtual Patient System for Clinical Training

Doria Bonzi[1], **Tom Bourgeade**[1], **Fabrice Lefèvre**[2], **Irina Illina**[1]
[1]Lorraine Université, CNRS, Inria, LORIA, Nancy, France
[2]Avignon Université, LIA, UPR 4128, Avignon, France
doria.bonzi@loria.fr, tom.bourgeade@loria.fr, irina.illina@loria.fr
fabrice.lefevre@univ-avignon.fr

Abstract

The clinical and communication skills of medical students are commonly assessed through Objective Structured Clinical Examinations (OSCEs), which consist of brief scenario-driven simulations of doctor-patient interactions. However, training is often limited by the low availability of human standardized patients, motivating the development of realistic virtual patients (VPs). To address this gap, we introduce a French OSCE dialogue dataset comprising 240 student–patient training interactions. We build upon it a controllable LLM-based pipeline to generate synthetic OSCE dialogues. The pipeline integrates modular components, such as retrieval-based grounding and a reflection loop, to ensure patient fidelity, coherence, and realism. Additionally, we propose a multi-level evaluation framework assessing patient simulation quality, student performance, and linguistic quality, using an LLM-as-a-Judge approach. Experiments suggest that controllability modules generally improve patient fidelity and student evaluation consistency. Finally, we also implement an interactive prototype in which students can practice with a VP and receive automatic feedback.

1 Introduction

Objective Structured Clinical Examinations (OSCEs) are used to assess both clinical reasoning and communication skills in medical education. In France, these exams involve medical students playing the role of a physician in 7 to 10-minute scenario-based simulated interactions with a **standardized patient** (SP), under the observation of an **evaluator**. A SP is defined as a trained individual who portrays a predefined patient scenario for the training and assessment of healthcare students (Lewis et al., 2017). Each OSCE scenario is referred to as a "**station**," and can cover various medical interactions, ranging from patient history taking and analyzing results to breaking bad news. By design, OSCEs approximate real clinical encounters and intentionally simplify many aspects for pedagogical purposes. Communication skills play a central role in OSCE performance; however, student training is limited by the availability of human standardized patients and evaluators, leading to scalability and cost issues for repeated practice.

Virtual Patient (**VP**) systems have been proposed to address these problems, relying on rule-based or script-driven approaches (Zini et al., 2019; Campillos-Llanos et al., 2019; Laleye et al., 2020a). Data-driven and LLM-based VP systems have demonstrated improved fluency and realism (Voigt et al., 2025; García-Torres et al., 2024; Cook et al., 2025). Some studies explored automated feedback for OSCE assessments using LLM-as-a-Judge (Shakur et al., 2024; Campbell et al., 2025; Huang et al., 2026). Despite these advances, current LLM-based VP systems often lack controllability, hindering consistent adherence to predefined clinical stations. Recent work has also highlighted the lack of standardized and reproducible evaluation frameworks in LLM-based VP systems (Li and Lutfi, 2026), essential for reliable assessment.

Publicly available OSCE-related datasets are rare, especially in French. Existing OSCE datasets in English focus on specific clinical tasks or domains (Fareez et al., 2022; Saley et al., 2024). Large-scale medical dialogue resources (Liu et al., 2024; Ben Abacha et al., 2023) are generally not aligned with OSCE constraints or lack the interactional and evaluative structure required for examination set-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 357–374

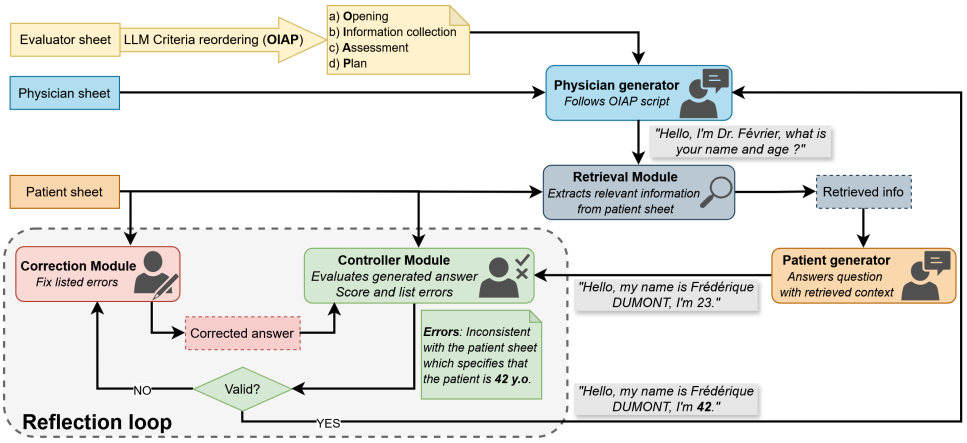


Figure 1: Overview of the controllable LLM-based dialogue generation pipeline for OSCE simulations in automatic mode. The physician generator follows a criteria-based script; the patient generator can be augmented by a retrieval module, and the reflection loop with controller-corrector modules can be activated to control patient faithfulness through a verification-correction loop.

tings. French resources remain limited in size and scope (Laleye et al., 2020b). This lack of French OSCE data restricts the development of data-driven training and evaluation tools.

To address these challenges, we introduce in this paper:
(1) a dataset of 240 recorded and transcribed French OSCE training dialogues available on Zenodo[1];
(2) a controllable LLM-based OSCE dialogue generation pipeline with modular components, supporting both automatic and interactive modes;
(3) a multi-level LLM-as-a-Judge evaluation framework assessing generated dialogues and recorded OSCE interactions.

2 Related work

Virtual patients and LLM-based simulations: Recent work on LLM-based VPs focuses on realism, interactivity, automatic scoring and feedback (Voigt et al., 2025; García-Torres et al., 2024) and patient data fidelity (Wang et al., 2024b; Laverde et al., 2025). Embodied VP have been explored as simulation-based tools, emphasizing realistic behavioral interactions (Chaby et al., 2022). Agentic approaches used structured patient data combined with multi-agent RAG workflows to enable controllability (Yu et al., 2025). However, these works were rarely anchored in real OSCE recordings, limiting their grounding in authentic interactions.

Medical dialogue and synthetic dialogue generation:

In the context of OSCEs, Fareez et al. (2022) introduced a dataset of simulated patient interviews focused on respiratory cases. Building on this work, Saley et al. (2024) released an English dataset for medical history-taking in OSCE format. French medical dialogue resources remain limited. Laleye et al. (2020b) introduced a small annotated French corpus combining generated dialogues and interactions between medical students and patients. Chen et al. (2023) has investigated LLM-based dual-agent simulations, emulating both physicians and patients for clinical dialogue generation.

Large-scale medical dialogue datasets provide broad coverage of clinical interactions but are not designed for OSCE-specific scenarios (Jepson et al., 2017; Zeng et al., 2020; Liu et al., 2024; Ben Abacha et al., 2023). In parallel, LLM-based dialogue generation approaches improved fluency and realism but lack strict controllability and alignment with structured exam settings (Das et al., 2024; Wang et al., 2024a).

Reflective prompting, and LLM-based evaluation: Recent advances in LLM prompting have introduced self-reflection and critique mechanisms to improve factuality, coherence, and task adherence in complex generation tasks (Agrawal et al., 2026; Chirkova

[1]Dataset may be accessed upon request: zenodo.org/records/20719833

et al., 2026; Li et al., 2023). In parallel, LLM-as-a-Judge paradigms have emerged to evaluate conversational agents at scale across multiple criteria, including dialogue quality and OSCE performance (Shakur et al., 2024; Campbell et al., 2025). Gu et al. (2026) highlighted their growing adoption, with applications ranging from human-machine dialogue assessment (Njifenjou et al., 2025, 2024) to OSCE scenarios (Shakur et al., 2024; Campbell et al., 2025) and benchmarking conversational agents in controlled settings (Zheng et al., 2023). These approaches motivate the design of structured evaluation and reflective stages, which we incorporate in our pipeline to assess and iteratively improve our generated OSCE dialogues.

Building on prior work, we distinguish our approach in three ways: (1) a focus on French OSCE stations, addressing the scarcity of French medical dialogue resources; (2) a modular, controllable LLM dialogue generation pipeline integrating information retrieval and self-reflection, grounded in OSCE data across more than ten medical specialties; (3) a multifaceted evaluation covering linguistic and clinical performance.

3 Proposed French OSCE dialogue dataset

Our proposed dataset consists of: (i) a corpus of 240 recorded OSCE training dialogues, representing a total of 30 hours of audio, and (ii) a corpus of 792 generated dialogues produced using different experimental configurations of our controllable LLM-based generation pipeline (see Figure 2a).

3.1 Data collection: OSCE stations and recorded dialogues

OSCE stations: A total of 192 OSCE clinical stations were collected for this dataset across more than 10 medical specialties. These stations are authored by medical teachers and healthcare professionals. Each OSCE station consists of three complementary documents made available through a dedicated online platform: a **physician sheet** addressed to the evaluated student, a **patient sheet**, and an **evaluator sheet**.

Each OSCE station was mapped to a category from a dialogue-generation-focused **difficulty classification** developed specifically for this work, with: (i) the type of interlocutor and (ii) the requirement for document analysis (Figure 2b).

OSCE recorded dialogues: We recorded 240 played physician-patient dialogues involving 99 sixth-year medical students across 23 distinct OSCE stations. These recordings were collected during OSCE training sessions organized weekly by the local student association, which are designed to closely replicate the conditions of the national OSCE examinations in France. Due to the limited availability of external volunteers, all roles (physician, patient, and evaluator) were performed by students. Although this is not ideal, it was an unavoidable constraint in our setting. As participants are neither professional actors nor trained OSCE evaluators, this setup may affect the realism of the interactions and evaluations, and may partially account for differences observed when comparing these dialogues with synthetically generated ones.

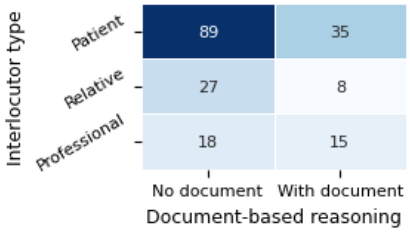
Each OSCE training session lasted 8 minutes, followed by a 2-minute feedback period from the evaluator. All participants provided written informed consent for the use of their recordings for research purposes. Audio was captured using wireless clip-on microphones. All recordings were acquired using the same software, producing synchronized two-channel WAV files. More details on the recording protocol are available in Appendix A.1.

Recorded stations annotations: Each recorded station was manually given labels with: one of 12 **specialties** (e.g., neurology, pediatrics, gastroenterology), a **consultation type** (e.g., emergency, follow-up), and one or more **objectives** (e.g., diagnosis, breaking bad news, history-taking, patient education).

All annotations were performed by a single annotator through careful examination of the OSCE station materials, particularly the physician and evaluator sheets, which typically provide the station context, including specialty, consultation type, and primary objectives (see Appendix A.4).

Component	Count
Total OSCE stations	192
Recorded dialogues	
OSCE stations recorded	23
Recorded dialogues	240
Total audio duration (hours)	30
Generated dialogues	
OSCE stations used	11
Generated dialogues	792
Total generated words	1,22M

(a) Dataset statistics: recorded and generated dialogues.



(b) OSCE station classification and number of stations per category.

Figure 2: Overview of the French OSCE dialogue dataset, both recorded and generated.

Automatic transcription: All recorded dialogues were automatically diarized using the `precision-2` model (PyAnnoteAI, 2024), and transcribed using the `faster-whisper-large-v3-turbo` model (OpenAI, 2024; Radford et al., 2023).

Transcription evaluation: To assess transcription quality, Word Error Rate (WER) was computed by comparing automatic transcriptions with manually corrected versions. A subset of 26 automatically transcribed dialogues, representing approximately 11% of the recorded corpus, was manually corrected by 15 French-fluent annotators using the original audio recordings.

The resulting WER of 7.7% suggests good transcription quality for real-world clinical dialogues, though computed on a limited sample. More information on WER computation is available in Appendix A.2. Recurrent errors included names, acronyms (e.g., *SMUR, ECOS*), domain-specific medical terminology (e.g., *dyspnée*; "shortness of breath"), occasional diarization errors due to cross-talk and disfluencies such as hesitations or repetitions.

4 Controlled dialogue generation pipeline

To benchmark LLM-based virtual patients against student-acted patients in OSCE training sessions, we synthetically generated 792 dialogues from structured OSCE station sheets. We focused on 11 stations for which real training recordings were available, excluding multimodal stations involving document analysis (e.g., radiology images). In this work, we deliberately restrict the problem to text-based interactions, excluding speech and multimodal information.

Our proposed approach (Figure 1) frames dialogue generation around **controllability**, aiming to ensure fidelity to patient profiles, coherence, and realism. The pipeline incorporates multiple levels of control with **retrieval-based grounding** and consistency enforcement through a **reflection loop**. The proposed system generates VP utterances interacting with a simulated **physician** in an OSCE setting, constrained to a fixed consultation duration of 8 minutes. The architecture is modular and LLM-based and supports two operating modes:

1. an **interactive mode**, where a student can train as the physician, via our prototype interface (see Figure 3). At the end of the interaction, students receive an automatic LLM-based evaluation of their performance, including feedback on addressed criteria and overall communication quality (see 5.2).
2. an **automatic mode**, where both physician and patient are LLM-generated, enabling large-scale dialogue dataset generation.

In both modes, VP utterances are controlled to ensure fidelity and consistency with the patient profile. The system relies on up to four distinct LLM instances assigned to specific roles: **retrieval**, **generation**, **control**, and **correction**. This modular design enables independent model selection and parameter tuning for each component. Dialogue generation is grounded in three components:

1. **Patient sheet**, describing the simulated patient's identity, medical history, symptoms, and behavior (e.g., *Frédérique Dumont* consulting for an extension of medical leave);

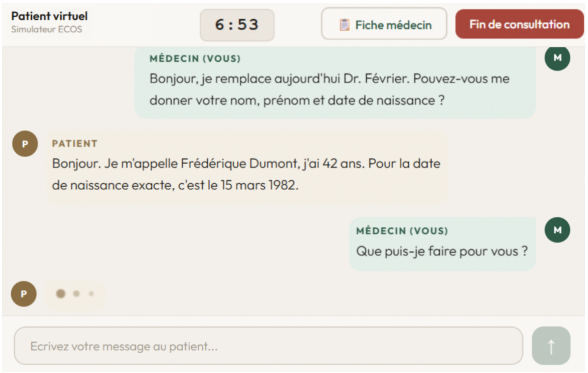


Figure 3: Interactive demo of our VP system. The student interacts with the VP via messages and can consult the physician sheet at any time. An 8-minute timer simulates OSCE conditions.

2. **Physician sheet**, providing the clinical context and setting, available informations, and consultation objectives (e.g., current medication and information-gathering goals);
3. **Evaluation sheet**, listing OSCE criteria the physician should address, each corresponding to a specific clinical action or information (e.g., assessing treatment effectiveness).

Dialogue script and criteria reordering: In automatic mode, dialogue generation is guided by a script derived from the evaluation checklist of the OSCE station. Following prior work (Huang et al., 2026), criteria are reordered into dialogue phases: **O**pening and preparation, **I**nformation collection, **A**ssessment, and **P**lan and conclusion (OIAP), inspired by SOAP (Podder et al., 2026). This reordering is performed by an LLM before generation and used to guide physician utterances step by step, acting as a dialogue script (Huang et al., 2026). We define two generation modes: a **standard** mode following the OIAP structure, and a **random** mode where intermediate phases and criteria order are shuffled while preserving the opening and conclusion. This enables generating diverse dialogues for the same station. Dialogue generation proceeds in batches of up to four criteria, with up to three dialogue turns per batch, empirically chosen to balance generation quality and dialogue duration constraints. As OSCE stations are time-constrained, dialogue duration was estimated based on average speech rate and pauses.

Physician generator: The physician generator produces the physician's questions and statements. To ensure that these utterances remain coherent, contextually accurate, and aligned with the intended OSCE workflow, the prompt is dynamically enriched for each batch with the context of the physician sheet, the dialogue history, the current clinical stage, the relevant criteria checklist, and instructions for stage transitions when the batch manager signals a phase change.

Patient generator: Upon receiving a physician utterance, the retrieval module selects relevant information from the patient sheet as input context for the generator. If retrieval is disabled, the full sheet is provided. The generator then produces a context-aware response based on the patient sheet, instructions, and dialogue history.

Controller module: Generated responses are evaluated for fidelity to the patient sheet using an LLM-based scoring module that assesses consistency with the provided patient information and identifies potential contradictions. The controller assigns a score from 0 to 10 based on the number and severity of detected inconsistencies, along with a list of errors. Responses scoring above 8/10 are accepted, while lower-scoring responses are sent to a correction module before final validation.

The **correction submodule** takes the flagged response along with an instruction prompt, the patient sheet, the physician's question, and the errors identified by the controller. Its role is to correct the response by addressing the identified errors. Once corrected, the response is sent back to the controller module for final validation, forming a **reflection loop** that iteratively improves response fidelity and reduces contradictions with the patient sheet.

All modules are LLM-powered, and different models can be used per module. Examples are included in the prompt using a few-shot prompting strategy for both the controller and correction modules. Detailed prompts are provided in Appendix A.6.

Model	Patient simulation		Physician performance		Linguistic quality		
	Information recall (%)	Response relevance (/5)	OSCE evaluation (%)	Role adherence (/5)	Naturalness (/5)	Fluency (/5)	Coherence (/5)
Recorded dialogues	43.58	4.14	68.57	4.03	3.21	3.79	3.91
Claude-Haiku-4.5	**52.24**	4.97	**87.48**	4.76	3.42	**4.35**	4.28
Gemini-3.1-Flash-Lite	47.26	4.03	85.96	4.67	3.40	4.07	4.20
Ministral-14B	47.44	4.97	82.55	4.70	2.83	3.98	4.24
GPT-4o-mini	47.76	**5.00**	74.81	**4.83**	**3.60**	4.31	**4.42**

Table 1: Results for recorded and generated dialogues on baseline configuration. Scores are computed using GPT-4o-mini as LLM-as-a-Judge and averaged over 44 generated dialogues. Best results in bold, second-best underlined.

5 Experiments and evaluation

5.1 Experimental protocol for dialogue generation

To assess the contribution of each module, we define different experimental configurations: a baseline configuration without optional modules; the reflection loop only; and the reflection loop combined with retrieval. This allows us to evaluate the impact of each module on dialogue quality and controllability. When active, the reflection loop was configured with a maximum of 3 iterations and a minimum controller threshold score of 8/10 to accept a patient response (prompts available in Appendix A.6).

Configurations were evaluated using the following LLMs: Claude Haiku 4.5 (Anthropic, 2025), Gemini-3.1-Flash-Lite (Google, 2026), Ministral-14B-2512 (Liu et al., 2026), and GPT-4o-mini (OpenAI et al., 2024) accessed through the OpenRouter API with a temperature of 0 and top-p of 0.9 across all modules (generation, retrieval, controller, corrector). Models were selected based on cost-efficiency considerations, as the system is intended for free use by medical students. Ministral-14B was additionally chosen for its potential for local deployment, an important factor for educational applications. We did not include domain-specialized medical LLMs as recent work suggests that specialized models do not necessarily outperform general-purpose ones on medical tasks (Dorfner et al., 2024; Labrak et al., 2024; Huang et al., 2026). This is also consistent with our setting, where the VP is intended to simulate non-expert behavior.

For each configuration and model, we generated 4 dialogues per station, yielding a total of 792 generated dialogues across 11 OSCE stations described in Appendix A.4.

5.2 Evaluation setup: description and metrics

All evaluations were conducted in an LLM-as-a-Judge framework (Gu et al., 2026), using GPT-4o-mini (OpenAI et al., 2024) with a temperature of 0. We assess generated and recorded dialogues along three dimensions, each comprising multiple metrics. Evaluation prompts are available in Appendix A.7. To assess the reliability of this automatic evaluation, we additionally collected a small set of annotations from 6 annotators without medical training.

1. Patient simulation quality evaluates how faithfully and appropriately the VP behaves with respect to the patient sheet. We propose the following metrics for the patient simulation quality evaluation:

- **Information recall** measures the proportion of patient sheet elements mentioned by the VP during the dialogue. In a preliminary step, we use GPT-4o-mini to transform each patient sheet into a binary checklist (e.g., mentions the patient's name, reports allergies). The LLM-based judge then reviews all patient utterances and marks each item as mentioned or not, yielding a percentage score.
- **Response relevance** assesses whether patient utterances are appropriate given the physician's questions and the patient sheet content. The LLM-based judge assigns a score from 1 (very poor) to 5 (excellent).

2. Physician performance evaluates the quality of the physician's conduct during the consultation, with the following metrics:

Physician performance variability	Experimental setup	Patient simulation		Physician performance		Linguistic quality		
		Information recall (%)	Response relevance (/5)	OSCE evaluation (%)	Role adherence (/5)	Naturalness (/5)	Fluency (/5)	Coherence (/5)
100%	baseline	52.24	**4.97**	87.48	4.76	3.42	**4.35**	4.28
	reflection	54.29	4.83	87.75	4.85	3.57	4.30	**4.36**
	refl. + retr.	**56.62**	4.89	**89.21**	**4.89**	**3.59**	4.29	**4.36**
50%	baseline	24.12	4.85	**46.21**	4.31	**3.52**	4.22	4.26
	reflection	**30.84**	**5.00**	45.59	**4.55**	3.51	**4.24**	**4.27**
	refl. + retr.	27.23	4.86	45.05	4.14	3.46	4.20	4.20
25%	baseline	17.80	4.44	25.69	3.89	3.00	4.00	**4.8**
	reflection	**22.65**	4.67	**29.89**	**4.17**	3.35	**4.13**	4.15
	refl. + retr.	18.74	**4.75**	27.06	4.00	3.00	4.00	4.10

Table 2: Results for Claude-Haiku-4.5 across three experimental setups (baseline, reflection loop, and reflection + retrieval) under varying physician performance constraints (100%, 50%, and 25% OSCE criteria coverage). Best results for each physician performance variation in bold.

• **OSCE evaluation** implements traditional OSCE scoring: the evaluator sheet lists specific criteria that the physician is expected to address (e.g., asks about allergies, proposes a treatment), and the judge checks them as met or not, yielding a percentage score.

• **Role adherence** assesses if the physician appropriately fulfilled the role defined in the physician sheet: respecting the clinical context, staying within the OSCE framework, and behaving professionally. Only physician utterances are evaluated. The LLM-based judge assigns a score from 1 (very poor, major deviations from the expected role) to 5 (excellent, perfect role adherence).

3. Linguistic quality evaluates the dialogue independently of its clinical content, focusing on three standard dialogue quality metrics. **Naturalness** measures whether the dialogue resembles a real spoken conversation between two people. **Fluency** assesses the smoothness and grammatical correctness of the exchanges. **Coherence** evaluates the logical structure and progression of the dialogue. Each criterion is scored on a 0–100 scale using detailed rubrics we defined in prompts, then rescaled to a 1–5 Likert scale for consistency with the other similar metrics in this framework (e.g., response relevance). All evaluation prompts are available in Appendix A.7.

6 Results and analysis

We conduct four analyses: (1) comparison between recorded and baseline-generated dialogues across the four models (Table 1), (2) impact of reflection and retrieval modules with Claude-Haiku-4.5 (Table 2), (3) robustness to varying physician performance (Table 2), and (4) station-level naturalness analysis (Figure 4). Detailed results are provided in Appendix A.5 with all configurations and models.

6.1 Generated vs. recorded dialogues

Across all evaluation dimensions, generated dialogues consistently outperform recorded interactions, including higher information recall across all models, as shown in Table 1. Linguistic quality scores are also higher for generated dialogues, except for Ministral-14B, whose naturalness score falls below the recorded dialogue baseline of 3.21, suggesting that this model produces less conversationally realistic utterances.

This gap is partly expected: recorded dialogues reflect real student variability, including incomplete coverage of OSCE criteria, hesitations, and communication difficulties, whereas generated dialogues benefit from systematic criterion coverage and controlled patient behavior. Additionally, these results may partly reflect biases in LLM-as-a-Judge evaluation, which could favor the style and structure of LLM-generated text over the disfluencies and variability characteristic of spontaneous speech.

As part of the linguistic quality evaluation (Section 6.5), annotators were asked to determine whether each dialogue was LLM-generated or recorded from a real interaction. They correctly identified the dialogue origin in only 14 of 24 judgments (58.3%), achiev-

ing identical accuracy for LLM-generated and recorded dialogues.

For completeness, we also include the results of GPT-4o-mini, despite using it as the LLM-as-a-Judge evaluator. Interestingly, it does not achieve the highest scores across the evaluated models, despite also serving as the evaluator. This observation, however, should not be interpreted as evidence that the evaluation is free from model-specific biases.

Since Claude-Haiku-4.5 consistently achieves the best overall performance across evaluation dimensions, we use this model for the remaining experiments.

6.2 Impact of the reflection and retrieval modules

Table 2 shows the contribution of reflection loop and retrieval module compared to the baseline system with Claude-Haiku-4.5. Overall, the reflection loop consistently improves patient simulation metrics compared to the baseline, particularly information recall, while maintaining stable linguistic quality. These improvements suggest that the controller and corrector modules help refine responses and reduce inconsistencies in generated dialogues. The reflection loop was triggered in 3.6% of patient turns, indicating that most responses were accepted by the controller without requiring correction. When activated, the correction process improved the controller score in 79.5% of cases.

Adding retrieval sometimes further improves performance, yielding the highest information recall and OSCE evaluation scores. However, these gains remain inconsistent across setups, suggesting that retrieval effectiveness depends on the interaction dynamics. Across configurations, linguistic quality remains stable, indicating that both modules improve fidelity and task performance without degrading dialogue naturalness. Overall, the reflection loop and retrieval modules sometimes improve results, but their impact remains moderate.

To further assess whether these improvements affect conversational quality at the utterance level, we evaluate linguistic quality scores separately for each speaker role (Table 3). Linguistic quality remains stable across experimental setups for both speaker roles. Physician scores are consistently slightly higher than patient simulation scores, suggesting that patient generation may introduce more variability than physician utterances.

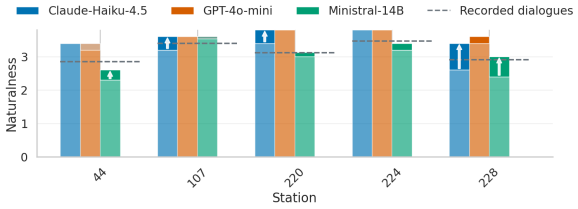


Figure 4: Naturalness score across all models and recorded dialogues: bars represent baseline performance averaged over four generated dialogues, deltas indicate performance changes with the retrieval and reflection loop for five representative stations.

6.3 Robustness to physician performance variability

To assess our VP robustness to variable physician performance, we degraded the physician generator by providing only 50% or 25% of the expected OSCE evaluation criteria, and perturbing them following Huang et al. (2026), reported in Table 2. This degradation led to a drop in information recall score, which is consistent with the statistically significant correlation between information recall and OSCE scores, reported in Appendix A.3: as the physician asks fewer questions, the patient provides less information.

Response relevance remains stable and linguistic quality scores slightly decrease. Compared to the 100%-criteria setting, these results suggest that the VP remains robust to variations in physician performance, particularly when using the retrieval and reflection loop pipeline, which maintains high response relevance with degraded interaction conditions. This robustness allows the VP to adapt to student performance and enables learning even from failed interviews.

6.4 Station-level analysis

Figure 4 presents the analysis of naturalness scores at the station level to assess where the reflection loop and retrieval module are most beneficial. We focus on five representative stations (44, 107, 220, 224, 228) covering diverse clinical contexts (see Appendix A.4), includ-

Experimental setup	Patient simulation			Physician performance		
	Naturalness (/5)	Fluency (/5)	Coherence (/5)	Naturalness (/5)	Fluency (/5)	Coherence (/5)
baseline	**3.82**	**4.26**	**4.25**	**3.80**	**4.40**	**4.65**
reflection	3.78	4.23	4.10	3.67	4.31	4.47
refl. + retr.	**3.82**	4.25	**4.25**	3.68	4.34	4.48

Table 3: Linguistic quality scores per speaker role for Claude-Haiku-4.5, evaluated with GPT-4o-mini as LLM-as-a-Judge. Full dialogues are provided to the judge, which scores each speaker independently. Scores are computed on a 0–100 scale using rubric-based prompts and rescaled to 1–5 Likert scores.

ing different specialties and objectives. The gain on station 228 may be partly driven by few-shot prompting, as one example directly matches this station. Overall, results suggest that improvements in naturalness are very context-dependent and not uniform across stations. This variability further suggests that a discretized station design, based on more static and well-defined categories, could improve simulation consistency.

6.5 Agreement between human evaluators and LLM-as-a-Judge

To assess the reliability of our LLM-as-a-Judge framework, six annotators each evaluated four dialogues on three linguistic-quality criteria (naturalness, fluency, and coherence) using the same 1–5 Likert scale as the LLM. Each of the 12 dialogues was rated by two annotators.

Criterion	ICC (human)	Pearson	Spearman
Naturalness	0.131	0.08	0.09
Fluency	−0.134	0.46	0.45
Coherence	−0.040	0.76*	0.45

Table 4: Inter-annotator agreement and correlation between mean human scores and LLM-as-a-Judge scores ($n = 12$ dialogues). $^*p < 0.05$

Intraclass Correlation Coefficient (ICC) (Koo and Li, 2016) is weak across all three criteria, reflecting the subjective nature of linguistic-quality evaluation. Human-LLM agreement remains limited, with only coherence showing a significant Pearson correlation, as shown in Table 4.

We also evaluated agreement between GPT-4o-mini and human evaluator scores on the OSCE evaluation metric using ICC and Spearman correlation. On 67 recorded dialogues across 14 OSCE stations, agreement was moderate (ICC = 0.41, ρ = 0.33), and became strong when restricting the analysis to patient-only stations without document analysis (ICC = 0.85, $\rho = 0.82$). These results suggest that the LLM-as-a-Judge is more reliable for checklist-style OSCE evaluation than for linguistic-quality assessment, although these results remain preliminary.

7 Conclusion

We introduced a French OSCE dialogue dataset combining 240 student-patient OSCE training dialogues and 792 generated dialogues, providing a new data resource for French medical education. We proposed a controllable LLM-based generation pipeline with retrieval and controller modules to improve patient fidelity and dialogue coherence, and an LLM-as-a-Judge evaluation framework assessing patient simulation quality, physician performance, and linguistic quality.

Results show that generated dialogues consistently outperform recorded interactions in patient simulation quality, particularly in terms of information recall and response relevance. The proposed VP system with a reflection loop yields small gains in fidelity-related metrics, while retrieval provides additional but less consistent improvements depending on interaction context and model. Our VP remains stable across variations in physician performance, while station-level analyses reveal scenario-dependence in naturalness, suggesting the need for more fine-grained station-level modeling to better adapt system behavior.

Future work will include extending the pipeline to document-based stations, improving controllability of VP behavior, and validating our proposed VP system in real training conditions with medical students.

Limitations & Ethics statement

Limitations This study has several limitations. First, while our dataset includes 192 OSCE stations, only 23 were recorded, and dialogue generation was conducted on a subset of 11 stations. This limits the diversity of clinical scenarios covered and the generalization of our findings, though the dataset is under active expansion. Second, our evaluation relies entirely on a single LLM-as-a-Judge with a single evaluation run per dialogue. Although we validated the OSCE evaluation criterion against human annotations, the remaining criteria lack human validation, and inter-run variability was not assessed. Third, the generation pipeline was not tested in real training conditions with medical students interacting with the VP, leaving its pedagogical effectiveness still unevaluated. Fourth, the pipeline relies heavily on LLM calls across all modules (generation, retrieval, controller, corrector, evaluation), which raises concerns regarding reproducibility, cost, and scalability. We estimate the generation cost of a single dialogue at approximately $0.20, but a full experimental run involves numerous API calls across multiple modules. Finally, only three models were evaluated; larger or domain-specialized medical models may yield different results. Future work will focus on expanding human evaluation, assessing inter-annotator agreement, conducting user studies with medical students, and evaluating the pipeline in real training conditions.

Ethics statement

This study follows ethical guidelines for the use of language technologies in educational and clinical training contexts. No personal, clinical, or patient-identifiable data were collected or processed. All real interactions used in this work were recorded with appropriate authorization, and all scenarios were fictitious and designed in a game-like educational setting, ensuring full anonymization of participants. We emphasize that the goal of this work is to support the development of assistive technologies that augment human learning without reducing expertise or learner agency, particularly in sensitive domains such as medical education.

8 Acknowledgements

We thank the members of the Multispeech team for their contributions to the manual transcription and human evaluation of the dialogues used in this study. We also acknowledge ECNAsso for facilitating data collection. This work benefited from government funding managed by the National Research Agency under France 2030 via the ENACT AI Cluster (ANR-23-IACL-0004), and from support by Région Grand Est.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Anthropic. 2025. Claude haiku 4.5 system card. Technical report, Anthropic. Accessed: 2026-04-07.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

K. K. Campbell, M. J. Holcomb, S. Vedovato, L. Young, G. Danuser, T. O. Dalton, and D. J. Scott. 2025. Applying state-of-the-art artificial intelligence to grading in simulation-based education: assessment, feedback, and roi. *Discover Artificial Intelligence*, 5(1):247–?

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2019. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Laurence Chaby and 1 others. 2022. Embodied virtual patients as a simulation-based framework for training clinician-patient communication skills: An overview of their use in psychiatric and geriatric care. *Frontiers in Virtual Reality*, 3:827312.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and

patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

Nadezhda Chirkova, Tunde Oluwaseyi Ajayi, Seth Aycock, Zain Muhammad Mujahid, Vladana Perlić, Ekaterina Borisova, and Markarit Vartampetian. 2026. LLM-as-a-qualitative-judge: Automating error analysis in natural language generation. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 99–132, Rabat, Morocco. Association for Computational Linguistics.

David A Cook, Joshua Overgaard, V Shane Pankratz, Guilherme Del Fiol, and Chris A Aakre. 2025. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486.

Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic patient-physician dialogue generation from clinical notes using llm. *Preprint*, arXiv:2408.06285.

Felix J. Dorfner, Amin Dada, Felix Busch, Marcus R. Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C. Adams, and Keno K. Bressem. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *Preprint*, arXiv:2408.13833.

Explosion AI. 2024. fr_core_news_md: spacy french model. Version 3.8.0.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9:313.

Daniel García-Torres, María Asunción Vicente Ripoll, César Fernández Peris, and José Joaquín Mira Solves. 2024. Enhancing clinical reasoning with virtual patients: A hybrid systematic review combining human reviewers and chatgpt. *Healthcare*, 12(22):2241.

Google. 2026. Gemini 3.1 flash-lite. `https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite?hl=fr`. Consulté le 5 juin 2026.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on LLM-as-a-judge. *The Innovation*, page 101253.

Tian Huang, Tom Bourgeade, and Irina Illina. 2026. Llm-based data generation and clinical skills evaluation for low-resource french osces. In *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC 2026)*. Accepted for publication.

M. Jepson, C. Salisbury, M. J. Ridd, C. Metcalfe, L. Garside, and R. K. Barnes. 2017. The 'one in a million' study: creating a database of uk primary care consultations. *British Journal of General Practice*, 67(658):e345–e351.

Jitsi. 2024. jiwer: Evaluate word error rate (wer).

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Yanis Labrak, Adrien Bazoge, Oumaima El Khettari, Mickael Rouvier, Pacome Constant Dit Beaufils, Natalia Grabar, Béatrice Daille, Solen Quiniou, Emmanuel Morin, Pierre-Antoine Gourraud, and Richard Dufour. 2024. DrBenchmark: A Large Language Understanding Evaluation Benchmark for French Biomedical Domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5376–5390, Torino, Italia. ELRA and ICCL.

Fréjus A. A. Laleye, Antonia Blanié, Antoine Brouquet, Dan Behnamou, and Gaël de Chalendar. 2020a. Semantic similarity to improve question understanding in a virtual patient. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, pages 859–866, New York, NY, USA. Association for Computing Machinery.

Fréjus A. A. Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020b. A French Medical Conversations Corpus Annotated for a Virtual Patient Dialogue System. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 574–580, Marseille, France. European Language Resources Association.

Nicolas Laverde, Christian Grévisse, Sandra Jaramillo, and Ruben Manrique. 2025. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. *Computational and Structural Biotechnology Journal*, 27:2481–2491.

Kim Leighton Lewis, Carol A. Bohnert, William L. Gammon, Hans Hölzer, Linda Lyman, Christine Smith, T. M. Thompson, A. M. Wallace, and Gina Gliva-McConvey. 2017. The association of standardized patient educators (aspe) standards of best practice. *Advances in Simulation*, 2:10.

Dongliang Li and Syaheerah Lebai Lutfi. 2026. Large language model–based virtual patient systems for history-taking in medical education: Comprehensive systematic review. *Journal of Medical Systems / Elsevier (ScienceDirect)*.

Tao Li, Gang Li, Zhiwei Deng, Bryan Wang, and Yang Li. 2023. A Zero-Shot Language Agent for Computer Control with Structured Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11261–11274, Singapore. Association for Computational Linguistics.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, and et al. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Xingyu Liu, Vincent Segonne, Aidan Mannion, Didier Schwab, Lorraine Goeuriot, and François Portet. 2024. MedDialog-FR: A French version of the MedDialog corpus for multi-label classification and response generation related to women's intimate health. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 173–183, Torino, Italia. ELRA and ICCL.

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, and Fabrice Lefèvre. 2025. Enabling trait-based personality simulation in conversational LLM agents: Case study of customer assistance in French. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 299–308, Bilbao, Spain. Association for Computational Linguistics.

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, and Fabrice Lefèvre. 2024. Role-play zero-shot prompting with large language models for open-domain human-machine conversation. *Preprint*, arXiv:2406.18460.

OpenAI. 2024. Whisper large-v3-turbo. Automatic speech recognition model.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2026. *SOAP Notes*, statpearls [internet] edition. StatPearls Publishing. Accessed 2026.

PyAnnoteAI. 2024. pyannote speaker-diarization precision-2. Speaker diarization model.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam . 2024. MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.

Ameer Hamza Shakur, Michael J. Holcomb, David Hein, Shinyoung Kang, Thomas O. Dalton, Krystle K. Campbell, Daniel J. Scott, and Andrew R. Jamieson. 2024. Large language models for medical osce assessment: A novel approach to transcript analysis. *Preprint*, arXiv:2410.12858.

Henrik Voigt, Yurina Sugamiya, Kai Lawonn, Sina Zarrieß, and Atsuo Takanishi. 2025. Llm-powered virtual patient agents for interactive clinical skills training with automated feedback. *Preprint*, arXiv:2508.13943.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, page 15183–15201. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024b. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenyue Hua, Mingyu Jin, Guang Chen, Yang Zhou, Zhao Li, Trisha Gupte, Ming-Li Chen, Zahra Azizi, Qi Dou, Bryan P. Yan, and 7 others. 2025. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Communications Medicine*, 6(1):27.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang,

Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale Medical Dialogue Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Julia El Zini, Yara Rizk, Mariette Awad, and Jumana Antoun. 2019. Towards a deep learning question-answering specialized chatbot for objective structured clinical examinations. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

A Appendix

A.1 Data collection: recording protocol

Recording context: Students followed a predefined schedule for the training sessions, rotating through numbered stations. Students playing patient roles received detailed station descriptions in advance and often portrayed the same patient throughout a session, appearing in multiple recordings for the same station, while students playing physicians accessed station-specific context only upon entering the room. An evaluator, also a medical student, observed the interaction and provided structured feedback using the evaluator sheet.

Recording protocol: The recordings were conducted unobtrusively during the training sessions. Out of the eight available examination rooms, two were equipped for audio recording. In each recorded rooms, wireless clip-on microphones (DJI Mic Mini), worn by both the evaluated student and the simulated patient, were used to record. In addition to audio, 85 of the recorded dialogues include video data, which were not used in the present work. The collected data were transferred and stored on encrypted hard drives with restricted access.

A.2 Transcription evaluation: WER computation

The WER is defined as:

$$\text{WER} = \frac{S + D + I}{N},$$

where S denotes the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of words in the reference transcription.

Before WER computation, transcripts were normalized and tokenized using the *spaCy* pipeline with the `fr_core_news_md` model (Explosion AI, 2024). WER computation was performed using a custom Python pipeline based on the `jiwer` library (Jitsi, 2024). Automatic and reference transcripts were paired by dialogue identifier, normalized, segmented by speaker, and then evaluated.

A.3 Correlation test: information recall and OSCE score

Pearson ($r = 0.29$) and Spearman ($r = 0.31$) correlations between information recall and OSCE score were computed, showing a weak but statistically significant positive association ($p < 0.001$). This indicates a partial relationship between information recall and OSCE performance, while suggesting that other conversational and clinical reasoning factors also contribute to the overall score. Information recall is reported for informational purposes only, as exhaustive disclosure of patient information is neither expected nor desirable in OSCE settings.

A.4 Dataset details

Station (ID)	Specialty	Consultation type	Objective
44	Psychiatry	Emergency	History taking; diagnosis; clinical summary
46	Pediatrics	First consultation	Breaking bad news; patient education
107	General practice	Follow-up	History taking; diagnosis
120	Neurology	Follow-up	Breaking bad news; patient education
220	Cardiology; ICU	Emergency	Diagnosis; interprofessional coordination
224	Pulmonology	Follow-up	History taking; diagnosis; clinical summary
225	Occupational medicine	Follow-up	Specialist referral; patient education
226	General practice	Follow-up	Patient education; treatment initiation
228	General practice	Emergency	History taking; diagnosis; specialist referral
233	General practice	Follow-up	History taking; prescription renewal; patient education
235	Emergency practice	Emergency	Examination interpretation; history taking; diagnosis; patient education

Table 5: Annotations of the 11 OSCE stations used for dialogue generation, including medical specialty, consultation type, and communication objectives.

A.5 Detailed results

		Patient simulation		Physician performance		Linguistic quality		
Model	**Setup**	**Information recall** (%)	**Response relevance** (/5)	**OSCE evaluation** (%)	**Role adherence** (/5)	**Naturalness** (/5)	**Fluency** (/5)	**Coherence** (/5)
Recorded dialogues		43.58	4.14	68.57	4.03	3.21	3.79	3.91
Claude-Haiku-4.5	baseline	52.24	**4.97**	87.48	4.76	3.42	**4.35**	4.28
	reflection	54.29	4.83	87.75	4.85	3.57	4.30	**4.36**
	refl. + retr.	**56.62**	4.89	**89.21**	**4.89**	**3.59**	4.29	**4.36**
GPT-4o-mini	baseline	**47.76**	**5.00**	74.81	**4.83**	3.60	**4.31**	4.42
	reflection	44.66	4.91	**76.79**	4.82	**3.64**	4.29	**4.44**
	refl. + retr.	42.39	4.97	75.71	4.69	3.59	4.30	4.38
Ministral-14B	baseline	47.44	4.97	82.55	4.70	2.83	3.98	4.24
	reflection	50.02	**4.98**	**84.44**	**4.74**	**2.95**	**4.03**	**4.28**
	refl. + retr.	**52.40**	4.95	83.50	4.55	2.60	3.96	4.21

Table 6: Evaluation results across all three dimensions with GPT-4o-mini as LLM-as-a-Judge, for all three models and each configuration: baseline, reflection loop only, reflection loop + retrieval. Each score is averaged over 44 generated dialogues. Best results for each model in bold.

A.6 Prompts for dialogue generation

OSCE criteria classification prompt

```
You are a medical education expert. Classify each of the following clinical consultation criteria into exactly one of these stages:

{stages_desc}

Criteria to classify:
{criteria_text}

Return ONLY a JSON object mapping criterion number to stage id:
{{
  "1": "opening_and_preparation",
  "2": "information_collection",
  ...
}}

No explanation, no markdown fences, just the JSON object.
```

Physician's utterances generation prompt

```
You are generating the responses of a physician during an Objective Structured Clinical Examination. The exam takes place in a French medical university, and entirely in French. You must interact with the patient, asking questions and making comments as a real physician would do during a consultation. You must strictly follow the instructions and context provided below.

CONTEXT AND INSTRUCTIONS:
{consultation_context}

CURRENT PHASE: {stage_label} (batch {batch_number}/{total_batches})
{stage_transition_instruction}

OBJECTIVES FOR THIS PHASE:
{current_checklist}

{covered_criteria_section}

CONVERSATION HISTORY:
{conversation_history}

LAST PATIENT RESPONSE:
{last_patient_response}

You have 8 minutes total. Currently, {time_info} remains. If less than 2 minutes remain, prioritize closing the consultation and addressing the most critical objectives.
{final_part_instruction}

GENERAL RULES:
- Output speech only (no parentheses, no actions, no thoughts).
- Contextual fidelity: Strictly follow the role, objectives, and constraints provided above.
- Use open-ended questions to explore symptoms.
- Medical history: Investigate relevant medical history when appropriate.
- Adaptability: Tailor your follow-up questions to the patient's specific answers. Do not produce long answers. Be concise.
- Flow control: Ask only one question at a time.
- Coverage: Make sure to address ALL criteria listed in OBJECTIVES FOR THIS PHASE before the conversation moves on.
- Phase awareness: Your questions should be appropriate for the current phase of the consultation.

IMPORTANT: Only generate the physician's next spoken response as text. Do not include stage directions, actions, or any narration. NO PHYSICAL EXAMINATIONS. Your answer must be short (1-2 sentences).

Physician's next question:
```

Patient's utterrances generation prompt

```
You are generating the responses of a patient during an Objective Structured Clinical Examination, to help teach medicine students. The exam takes place in a French medical university, and entirely in French.
You must interact with and answer the doctor's questions as truthfully as possible.
Compose your responses as though you are speaking with the doctor-student. Try to be as concise as possible, as the exam only lasts 7 minutes. Keep answers short and natural. Use everyday language, not medical jargon.

Rules:
- Answer only what is asked, and give one small piece of information at a time.
- Do not volunteer extra details; wait for the doctor to ask.
- If you don't know or the information is not in the patient data, say you don't know.
- If the doctor uses complex terms, ask for a simpler explanation.
- Do not argue with the doctor; if you are unsure, ask for clarification.
- Keep responses to 1 short sentence or two short sentences at most.
- Do not repeat the same closing line over and over; once you agree or acknowledge, wait for the next question.
- Remember what the doctor just told you and respond consistently, even if it changes your attitude.
- Output speech only (no parentheses, no actions, no thoughts).
- Once the interview is coming to a close, and it is your last turn, thank the doctor and say goodbye.

{patient_data_section}

Start the dialogue with "phrase_demarrage" indicated in your patient profile.

Conversation history:
"{conversation_history}"

Last physician message:
"{doctor_question}"

Your answer:
```

Patient's utterance evaluation (controller) prompt

```
You are an expert evaluator in clinical communication for OSCE (Objective Structured Clinical Examination) exams.
Your task is to evaluate a simulated patient reply written in plain natural language.
The response you evaluate is NOT a JSON and MUST be treated as raw text only.

PATIENT PROFILE:
{patient_profile}

HISTORY:
{conversation_history}

GENERATED RESPONSE:
{patient_response}

EVALUATION CRITERIA:
- Consistency: The response is compatible with the patient profile, without contradictions or invented information.
- Patient realism: The level of language and understanding matches that of a patient.
- Credible interaction: Only verbal discourse; no actions or emotions described; concise response.

CRITICAL CONCISION:
- The patient answers accordingly to their profile and to the doctor 's answer.
- The response MUST NOT contain asterisks (*), parentheses to describe actions, or descriptions of gestures/facial expressions. ANY text between asterisks (action) or parentheses (gesture) is a MAJOR ERROR that makes the response NON-COMPLIANT ("conforme": false).
- Patient role: The patient does not reason like a doctor and does not suggest diagnoses or treatments.
- Linguistic quality: The response is clear, natural, and correct.
- If the answer is empty, give a maximum score of 1.


STRICT RULE ON ASTERISKS:
If the response contains even a SINGLE asterisk (*) or parenthesis describing an action:
  "conforme": false
  "score": maximum 4
  Add to errors: "Presence of asterisks or prohibited action descriptions".

YOUR MISSION:
- Verify the absence of asterisks or parentheses describing actions.
- Detect any inconsistency or role error.
- Identify specific issues.

STRICT OUTPUT FORMAT (JSON):
  {{
  "conforme": true | false,
  "score": int from 0 to 10,
  "erreurs": [
  "precise description of problem 1",
  "precise description of problem 2"
  ]
  }}

RULES:
- Produce ONLY this JSON, nothing else.
- No text before or after the JSON.
- No markdown tags like ```json.
- No comments.
- ONLY ONCE (do not repeat the JSON).
```

```
Evaluation:
```

Patient's utterances correction prompt

```
You are a medical simulation assistant. Your task is to correct a
      ↪ simulated patient's response to ensure it is realistic,
      ↪ coherent with the patient profile, and follows the format
      ↪ rules.

PATIENT PROFILE:
{patient_profile}

DOCTOR'S QUESTION:
{doctor_question}

CURRENT PATIENT RESPONSE (TO CORRECT):
{patient_response}

ERRORS DETECTED:
{errors_text}

--------------------------------------------------
CORRECTION INSTRUCTIONS
--------------------------------------------------

1. **MINIMAL CHANGES**: Make ONLY the changes necessary to fix the
      ↪ listed errors
   - Keep the same tone, style, and structure
   - Keep the same vocabulary level
   - Change ONLY what is incorrect

2. **FORBIDDEN ELEMENTS** (must be removed):
   - Asterisks for actions: *soupir*, *grimace*, *acquiesce*
   - Parentheses for actions: (sourit), (hesite), (reflechit)
   - Physical descriptions: "Je hoche la tete", "Je croise les bras"
   - Stage directions or narration

3. **ALLOWED ELEMENTS**:
   - Verbal hesitations: "euh...", "hum...", "ben...", "alors..."
   - Natural speech patterns: repetitions, corrections, incomplete
         ↪ sentences
   - Simple words and everyday language (patients are not doctors)

4. **COHERENCE WITH PROFILE**:
   - If the profile says the patient takes Metformin, the response
         ↪ must mention it
   - If the profile says non-smoker, the response must NOT mention
         ↪ smoking
   - Respect all factual elements from the profile

5. **NATURAL SPEECH**:
   - Keep responses SHORT (1-2 sentences typical for a single turn)
   - Patients don't know exact medical terms or dosages


YOUR TASK:
Provide ONLY the corrected patient response. No explanations, no
      ↪ comments, no markdown, just the corrected text.


CORRECTED RESPONSE:
```

A.7 Prompts for dialogue evaluation

Patient simulation quality evaluation prompt: information recall

```
You are an expert medical evaluator. For each criterion, determine
      ↪ whether the patient mentioned it in the dialogue.
Respond ONLY with valid JSON.

Patient turns only:
---
{patient_turns}
---

Binary criteria:
{criteres_json}

Return:
{{
  "evaluation_binaire": [
    {{"id": <id>, "critere": "<text>", "mentionne": true/false, "
          ↪ justification": "<short extract or reason>"}}
  ],
  "score_binaire": {{"mentionne": <n>, "total": <n>, "pourcentage":
        ↪ <float>}}
}}

RULES:
- Return ONLY valid JSON.
- Evaluate ONLY patient speech.
- Do NOT penalize date inconsistencies.
```

Patient simulation quality evaluation prompt: response relevance

```
You are a medical dialogue evaluation expert.
Evaluate whether patient responses are relevant to the doctor's
      ↪ questions and consistent with the patient sheet.
Respond ONLY with valid JSON.

Patient sheet:
{fiche_json}

Patient turns:
{patient_turns}

Dialogue for context:
{dialogue}

Evaluate whether patient responses are relevant to:
1. The doctor's questions
2. The patient sheet

Score relevance:
1 = Very poor
2 = Poor
3 = Acceptable
4 = Good
5 = Excellent

Return:
{{
  "score_response_relevance": <1-5>,
  "justification": "<brief explanation>",
  "examples": {{
    "relevant": ["..."],
    "irrelevant": ["..."]
  }}
}}

Rules:
- Evaluate ONLY patient responses
- Compare responses with both doctor questions and patient sheet
- Do not judge medical correctness beyond the sheet
- Return ONLY valid JSON
```

Physician performance evaluation prompt: OSCE evaluation

```
You are an expert medical evaluator. For each criterion, determine
      ↪ whether the physician mentioned or addressed it in the
      ↪ dialogue.
Respond ONLY with valid JSON.

Full dialogue:
{dialogue}

Binary criteria to evaluate:
{fiche_doctor}

Return:
{{
  "osce_evaluation": [
    {{"id": <id>, "critere": "<text>", "ok": true/false, "
          ↪ justification": "<short extract or reason>"}}
  ],
  "score_osce": {{"ok": <n>, "total": <n>, "pourcentage": <float>}}
}}

RULES:
- Return ONLY valid JSON.
- Evaluate ONLY physician speech.
```

Physician performance evaluation prompt: role adherence

```
You are a medical simulation expert. Evaluate whether the physician
      ↪ correctly plays their role given the context provided to
      ↪ them.
Respond ONLY with valid JSON.

Context given to the doctor at the start of the simulation:
---
{medecin_context}
---

Full dialogue:
---
{dialogue}
---

Based solely on the context above, evaluate the physician's overall
      ↪ role adherence. Consider the following aspects in your
      ↪ assessment:
- Did the physician respect the clinical context they were given?
- Was the communication style appropriate for a remote medical
      ↪ consultation?
```

```
- Did the physician stay within the bounds of what is possible in a
      ↪ spoken interaction?

Score on a 1-5 Likert scale:
- 1 = Very poor: Major deviations from role, inappropriate actions,
      ↪ context ignored
- 2 = Poor: Several deviations from role or significant
      ↪ inappropriate behaviors
- 3 = Acceptable: Mostly in role with some notable issues
- 4 = Good: Consistently in role with only minor issues
- 5 = Excellent: Perfect role adherence, fully respects context and
      ↪ constraints

Return:
{{
  "role_adherence": {{
    "score": <1-5>,
    "justification": "<detailed explanation of the score>",
    "issues": ["<list any problems found>"]
  }},
  "score_role_adherence": <1-5>
}}

RULES:
- Return ONLY valid JSON.
- Evaluate ONLY physician speech.
- PENALIZE if the physician starts a physical exam or mimics actions
      ↪  impossible in a remote simulation (e.g. shaking hands,
      ↪ auscultation).
- PENALIZE if the physician describes emotions or physical gestures
      ↪ (e.g. *smiling*, *looking concerned*), as this breaks
      ↪ immersion.
- Judge the physician ONLY against the context they were given, not
      ↪ against hidden patient information.
```

Linguistic quality evaluation prompt: naturalness, fluency, coherence

```
You are an expert evaluator of medical dialogue linguistic quality.
      ↪ You evaluate dialogues on three dimensions: naturalness,
      ↪ fluency, and coherence.
You use strict rubrics and score on a 0-100 scale.
Respond ONLY with valid JSON.

Evaluate the following dialogue on naturalness, fluency, and
      ↪ coherence.

**Rubrics**:
{rubrics}

**Dialogue to evaluate**:
---
{dialogue}
---

Return:
{{
  "linguistic_quality": {{
    "naturalness": {{
      "score_100": <0-100>,
      "band": "<0-20|20-40|40-60|60-80|80-100>",
      "justification": "<...>",
      "context_violations": ["<list any context violations>"]
    }},
    "fluency": {{
      "score_100": <0-100>,
      "band": "<0-20|20-40|40-60|60-80|80-100>",
      "justification": "<...>"
    }},
    "coherence": {{
      "score_100": <0-100>,
      "band": "<0-20|20-40|40-60|60-80|80-100>",
      "justification": "<...>"
    }}
  }}
}}

RULES:
- Return ONLY valid JSON.
- Be strict. A score of 80+ means near-perfect. Penalize context
      ↪ violations heavily in naturalness.
```