LLM-Based Data Generation and Clinical Skills Evaluation for Low-Resource French OSCEs

Tian Huang, Tom Bourgeade, Irina Illina
Université de Lorraine, CNRS, Inria, LORIA,
F-54000 Nancy, France
tian.huang@loria.fr†, tom.bourgeade@loria.fr, irina.illina@loria.fr

Abstract

Objective Structured Clinical Examinations (OSCEs) are the standard method for assessing medical students' clinical and communication skills through structured patient interviews. In France, however, the organization of training sessions is limited by human and logistical constraints, restricting students' access to repeated practice and structured feedback. Recent advances in Natural Language Processing (NLP) and Large Language Models (LLMs) now offer the opportunity to automatically evaluate such medical interviews, thereby alleviating the need for human examiners during training. Yet, real French OSCE annotated transcripts remain extremely scarce, limiting reproducible research and reliable benchmarking. To address these challenges, we investigate the use of LLMs for both generating and evaluating French OSCE dialogues in a low-resource context. We introduce a controlled pipeline that produces synthetic doctor–patient interview transcripts guided by scenario-specific evaluation criteria, combining ideal and perturbed performances to simulate varying student skill levels. The resulting dialogues are automatically silver-labeled through an LLM-assisted framework supporting adjustable evaluation strictness. Benchmarking multiple open-source and proprietary LLMs shows that mid-size models ($\leq$32B parameters) achieve accuracies comparable to GPT-4o ($\sim$90%) on synthetic data, highlighting the feasibility of locally deployable, privacy-preserving evaluation systems for medical education.

Keywords: Corpus Creation; Tools, Systems, Applications; Dialogue

1. Introduction

Effective clinical and communication skills are essential in healthcare practice, where doctor–patient interviews form the foundation of diagnosis, treatment, and patient trust. However, training opportunities for these skills remain limited, primarily due to the dependence on human participants, which increases costs and reduces accessibility. Recent advances in Natural Language Processing now make it feasible to automate key aspects of this training process, including the generation of realistic medical dialogues and the automated evaluation of student performance.

Objective Structured Clinical Examinations (OSCEs) have become the de facto international standard for assessing medical students through simulated encounters. In the French implementation (*Examens Cliniques Objectifs Structurés, ECOS*), a student plays the role of a doctor interacting with a trained actor serving as a standardized patient (SP), under the observation of an examiner who assesses the student's clinical and communication skills, during 7-10 min pre-defined scenarios, referred to as "**stations**". Only a few French medical schools can organize up to two weekly OSCE training sessions for final-year students preparing for the national exams. However, not all students can attend every session, as limited logistical and human resources restrict the number of available slots.

The ability to perform repeated training with structured feedback would thus significantly benefit students' preparation, and as such, automated evaluation systems leveraging advances in Natural Language Processing and Large Language Models (LLMs) may offer a potential solution. In this work, we focus in particular on the **automated evaluation of clinical skills in doctor-patient dialogue transcripts** (as could be obtained from a speech-to-text model, though we focus on text only here), which could alleviate the workload of, or even completely replace, human examiners in training sessions.

Recent studies have demonstrated the feasibility of LLM-based evaluation of English OSCE transcripts, showing high levels of agreement with human examiners across specific criteria (Shakur et al., 2024; Geathers et al., 2025). However, these works are grounded in the English OSCE tradition, where evaluation relies on standardized, station-agnostic checklists. In contrast, French OSCEs employ highly station-specific and heterogeneous criteria, making direct adaptation of such methods more challenging. Additionally, data for OSCEs, annotated or otherwise, remains extremely scarce, further limiting research and reproducibility in this area. In particular, no public French corpus of OSCE text transcripts has been made available to date. Other research has explored synthetic corpora of medi-

†Work conducted while the author was an intern at LORIA. The institutional email is no longer active; please use thuang44@gmail.com for correspondence.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10592–10602
11–16 May 2026.

cal dialogues as a way to overcome data scarcity (Wang et al., 2024; Das et al., 2024). While these efforts have yielded large-scale resources, they primarily cover English and Chinese, or rely on standardized country-specific structures for clinical interview notes. In contrast, the French OSCE context remains largely low-resource, with scarce annotated data, and no country-wide standardized structure for clinical interviews and notes.

To address these challenges of data scarcity and high variety in evaluation criteria, we propose leveraging LLMs to generate synthetic French OSCE doctor–patient transcripts from existing training scenarios and to automatically evaluate them against their associated clinical-skills checklists. Specifically, our main contributions are as follows: (1) A controlled pipeline is presented for generating French OSCE doctor-patient dialogues, capturing both ideal and suboptimal student performances. We also propose an LLM-assisted labeling framework offering adjustable levels of strictness; (2) several prompting strategies are examined, along with two auxiliary tools designed to support smaller, locally hostable LLMs ($\leq$ 32B parameters) in evaluating transcripts against binary criteria checklists; (3) a benchmarking study across a range of open-source and proprietary LLMs provides an initial assessment of the feasibility of this task in this low-resource educational context.

2. Related Work

AI for Pedagogical Assessment has been increasingly applied in education, with demonstrated benefits for both learning outcomes and scalability. Studies have reported measurable improvements in student outcomes and efficiency, including reductions in examiner workload (Alizadeh and Sameri, 2025). Simulation-based education has also seen the deployment of AI-driven scoring systems that provide large-scale feedback and better return on investment compared to traditional human-based evaluations (Campbell et al., 2025). Automated short-answer grading tools have shown strong correlations with human examiners (Seneviratne and Manathunga, 2025). These advances establish the feasibility of AI for non-interactive, text-based evaluation tasks.

LLMs for OSCE Evaluation have started to be explored as a means of automatically assessing medical students clinical and communication skills. Jamieson et al. (2024) reframed OSCEs criteria as prompts for evaluating post-dialogue clinical notes, while Shakur et al. (2024) demonstrated high agreement ($\kappa = 0.88$) between GPT-4 and human examiners on single criterion, in recorded video transcripts. Geathers et al. (2025) extended this line of work to multi-criteria scoring across 28 generic items(MIRS evaluation rubrics), testing several prompting strategies. Collectively, these studies confirm the feasibility of transcript-based evaluation, but they remain limited to English datasets and standardized criteria. The adaptation to French OSCEs is particularly challenging due to heterogeneous, station-specific criteria and the lack of annotated corpora.

Calibration and Bias in LLM-based Assessment remain key challenges: while LLMs offer scalability and reduced examiner fatigue (Haider et al., 2018), they also introduce variability, with studies documenting systematic differences in grading tendencies (Wei et al., 2025) and biases inherited from training corpora (Santurkar et al., 2023; Gallegos et al., 2024). Surveys of "LLM-as-a-judge" approaches emphasize both their promise for large-scale evaluation and their sensitivity to prompt design and contextual shifts (Gu et al., 2025). These findings motivate the need for explicit calibration mechanisms, and robust validation before deployment in high-stakes educational settings.

Synthetic Medical Dialogue Corpora have been investigated as a means to overcome data scarcity. Wang et al. (2024) introduced NoteChat, a multi-agent framework generating physician–patient interactions conditioned on clinical notes. SynDial (Das et al., 2024) proposed iterative feedback loops to align generated dialogues with case-specific constraints. Large-scale corpora such as *MTS-Dialog* (Ben Abacha et al., 2023) and *MedDialog* (Zeng et al., 2020) provide valuable resources, mainly in English and Chinese, though a French-translated variant, *MedDialog-FR*, has been made available by Liu et al. (2024). Nun et al. (2025) propose a large-scale dataset of transcriptions of simulated real-life medical dispatch calls by French junior emergency responders, alongside an expert-validated emergency dispatch dialog scheme. While these initiatives demonstrate feasibility, they do not capture the specificity of French OSCEs: heterogeneous, station-dependent evaluation criteria, which can sometimes be compositional, consisting of multiple sub-criteria linked by logical operators. This gap underlines the need for controlled synthetic corpora tailored to Francophone OSCE scenarios.

3. Data Generation

3.1. LLM-Based Dialogue Generation

To assess the feasibility of automating OSCE evaluation with LLMs, transcribed doctor–patient interactions are required, as they provide the textual input from which the model can infer, for each evaluation criterion, a binary outcome indicating whether it was met or not (see Fig. 2). However, no pub-

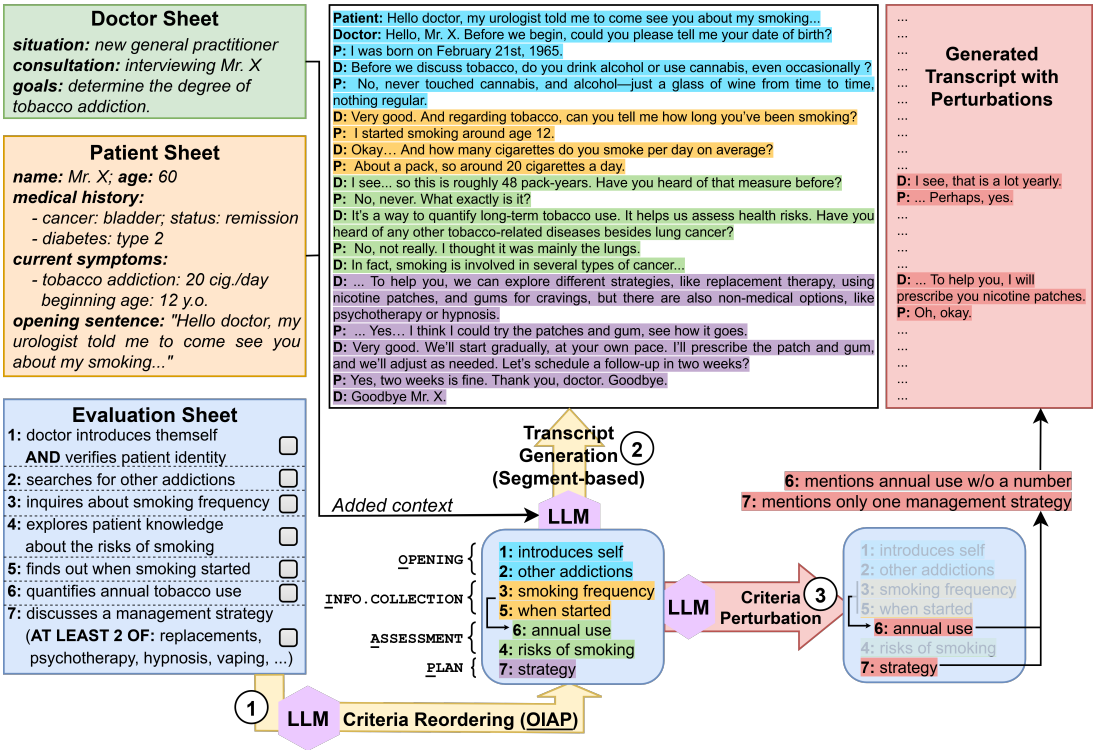

Figure 1: Dialogue Generation pipeline. The Doctor/Patient/Evaluation Sheets are structured documents (JSON). **(1)** The list of binary clinical skills criteria is reordered via LLM, respecting the OIAP structure and presence of dependencies (#6 here depends on #3 and #5). **(2)** A synthetic ideal-performance transcript is generated via LLM in segments using N criteria slices ($N = 2$ in this illustration). **(3)** To generate a less ideal transcript for data variety, some criteria are perturbed, leading to their failure in the generation.

lic corpus of French OSCE transcripts exists, as recordings from both training and examination sessions are confidential and protected by data privacy regulations.

A robust dataset for this task should include examples covering a variety of medical student performance, from ideal to less well executed, matching the range of grades and comments that can be attributed by examiners. Yet obtaining examples of ideal performance is challenging: students seldom reach expert-level execution during training, and transcriptions of tutor- or faculty-led demonstrations, which could serve as references, are comparatively rare. LLM-based dialogue transcript generation thus offers a controlled alternative, allowing experimentation and reproducibility while avoiding the limitations of scarce or sensitive real-world data.

Guiding Criteria: To make such generation feasible and coherent, we rely on the structured materials provided with each OSCE station, which describe the scenario from the perspectives of the doctor, patient, and examiner (see Figure 1). Each French OSCE station comes with three documents that specify its scenario: a **doctor sheet**, which summarizes the situation, patient identity, and goals of the student during the interview; a **patient sheet**, which provides the background information, symptoms, and intended behavior of the patient; and an **evaluation sheet**, which contains a list of observable criteria (e.g., *Inquires about the start of the symptoms*) that the doctor can pass or fail, and which are thus graded in a binary (*done*/*not done*) manner. These criteria are initially presented in an arbitrary order.

The corpus generation process used in this work follows the multi-step pipeline detailed in Figure 1. Inspired by Wang et al. (2024), we propose using evaluation sheets, which typically contain around ~ 20 binary criteria, as a sort of "script" to guide the generation of the doctor's dialogue turns during the simulated interview. First, the list of binary criteria is reordered by an LLM to improve dialogue coherence and naturalness, preserving potential dependencies between criteria. Following, slices of this list are successively fed to a dialogue generation LLM, together with the doctor and patient sheets, to generate segments of the dialogue, which are finally concatenated to obtain a generated transcript of an OSCE station. Because data generation is a one-time process that directly influences corpus quality and entails no data privacy issues, we used an industry-leading large language model (GPT-4o) to ensure the overall quality and realism of the generated dialogues.

Criteria Reordering: The order of these guiding criteria is critical, as they directly impact the order of generated speech turns and thus the naturalness and coherence of the synthetic dialogues. We thus manually picked one of two different ordering strategies, based on the specific context of each OSCE station:

(1) the **OIAP structure** (*Opening and Preparation, Information Collection, Assessment, Plan*), adapted from the SOAP framework (Podder et al., 2023), as operationalized by Wang et al. (2024), with modifications partly inspired by the Calgary-Cambridge model (Kurtz and Silverman, 1996) of medical interview structuring. In Wang et al. (2024), dialogues were generated in the SOAP order—*Subjective* (patient-reported information), *Objective* (clinically observable data), *Assessment*, and *Plan*—with each section produced separately and then concatenated. To better align with French OSCE training practices, we added an *Opening and Preparation* phase covering mandatory preliminary actions such as verifying the patient's identity, as mandated by the 2021 French regulation on patient identity vigilance (Ministère des Solidarités et de la Santé, 2021). We also merged the *Subjective* and *Objective* parts into a single *Information Collection* block to avoid unnatural dialogue breaks. For instance, when one evaluation criterion requires the student to inquire about smoking frequency (*Subjective*) and another to calculate the pack-year value (*Objective*), separating them would cause unnatural distance during generation. The OIAP definition is included in the prompt to guide the LLM in ordering the station's evaluation criteria.

(2) a **context-driven ordering**, used for atypical cases where the OIAP sequence would be unnatural, such as when a patient insists from the beginning on being referred for bariatric surgery. In such situations, the doctor may need to explain the preconditions for surgery (which would normally fall under the *Assessment* phase) rather than starting with the usual *Opening* or *Information Collection* phases. For these stations, the LLM is instructed through a different prompt to order the evaluation criteria directly based on the given doctor–patient context.

Criteria Perturbation: For our objective of assessing whether LLMs can reliably evaluate medical student performance in OSCEs, the generated corpus needs to include simulated doctor–patient interactions that reflect different levels of student performance. To obtain such diversity while maintaining dialogue coherence, we introduce a **perturbation** step in which a subset of the criteria guiding dialogue generation are replaced with **perturbed variants**, which deliberately distort their original intent, prompting the virtual doctor to perform incorrect or incomplete actions. The generated dialogue thus tends to fail the original unperturbed criteria, producing a suboptimal yet coherent doctor performance. However, some criteria can have dependencies: for example, a doctor cannot quantify the annual consumption of tobacco without asking the patient their smoking frequency (see Figure 1). We thus task `GPT-4o` with identifying what we refer to as *leaf criteria*, that is, those whose correct execution does not depend on other criteria. Only these leaf criteria are then considered for perturbation. Perturbations enable the generation of more diverse data while preserving overall coherence in generated dialogue transcripts. Examples are illustrated in Figure 1, e.g., an original criterion requiring the doctor to quantify the patient's annual use of tobacco is perturbed such that providing a number is no longer required, resulting in failing the original criterion.

Once reordered, dialogue generation proceeds in sequential slices of N criteria (we found $N = 4$ provides good quality generated transcripts after preliminary experiments): for each slice, prompts combine station context (doctor and patient sheets), previous segments of the generated transcript, and the target criteria. The LLM then generates a new dialogue segment, which is concatenated with the previous ones. The criteria within a segment may not follow the slice's order, allowing for variations that enhance variety without affecting coherence (e.g., criteria #3 and #5 are swapped in Figure 1). Minor post-processing prompting is performed, namely on the first and final segments, to ensure dialogue openings and closings are present, improving naturalness. From available training scenario material, we selected **10 representative OSCE stations** focused exclusively on doctor–patient dialogue, excluding those requiring external artifacts (use of medical devices, diagnostic maneuvers, or documents) or atypical criteria, resulting in a total of **179 binary criteria**. For each of these 10 stations, two synthetic dialogues were generated and grouped into: a **unperturbed** (optimal execution) corpus and a **perturbed** (with a chosen 50% perturbation rate, determined through preliminary experiments) corpus.

Since the LLM treats the evaluation criteria as guidance for dialogue generation, the passing or failing of any specific criterion cannot be guaranteed. To confirm the influence of this guidance, after automatic silver-labeling (see Section 3.2 below), we measured the proportion of failed criteria in both the **unperturbed** and **perturbed** corpora (Table 1). As expected, the **perturbed** dialogues showed a markedly higher proportion ($\sim 40\%$) compared to **unperturbed** ($\sim 10\%$).

As the dialogue generation is constrained by the predefined list of evaluation criteria, the resulting synthetic doctor-patient interactions remain closely

Station ID	Criteria	Failed *unperturbed*	Failed *perturbed*
113	17	0.0%	29.4%
128	18	5.6%	61.1%
165	19	15.8%	52.6%
		...	
Total	**179**	**10.6%**	**39.7%**

Table 1: Proportion of criteria annotated as *failed* (not done) in the unperturbed and perturbed corpora

tied to those items and do not reflect the broader range of behaviors that real students might exhibit during OSCE encounters. This makes them more idealized than real OSCE interactions, where students sometimes deviate from the expected structure and display spontaneous or unexpected behaviors. Even with perturbed variants, the dialogues rarely deviate beyond the specified criteria. This tendency is further amplified by the inherent bias of large language models toward cooperative or sycophantic responses (Sharma et al., 2023), which often produce well-structured and overly agreeable dialogues. While this affects the realism of the generated data, it also offers advantages: using explicit checklists as generation guides creates a controlled and reproducible environment for our objective of determining the viability of LLM-as-evaluators in OSCEs. Furthermore, we hypothesize that if the evaluation pipeline remains consistently strict under such idealized conditions, it would demonstrate the robustness needed to provide fair and useful feedback to students, making it a valuable tool for helping them articulate their reasoning and communication more clearly during OSCE training.

3.2. LLM-Assisted Silver-Labeling

In addition to the scarcity of French OSCE transcripts, expert-completed evaluation sheets are also difficult to obtain. In practice, during training sessions, the examiner role is often played by other medical students, resulting in indicative rather than fully reliable assessments. To obtain reference evaluation labels for the generated dialogues, we therefore adopted an **LLM-assisted silver-labeling** approach, in which `GPT-4o` was tasked with producing preliminary binary labels for each evaluation criterion, along with a justification text and evidence extracted from the transcript to support subsequent manual verification. The resulting silver-labels were then reviewed manually and, when necessary, adjusted to ensure consistency with the intended evaluation standards, which are further detailed below.

During preliminary experiments, we observed that the strictness of the labeling prompt had a strong influence on the model's decisions. When instructed to "strictly evaluate" whether a student explicitly performed an action, the model tended to assign fewer positive (corresponding to a student passing a criterion) silver labels than when using a more relaxed formulation such as "evaluate whether this criterion is met." When the same data, model, and prompt were kept fixed, repeated runs with random seeds produced highly consistent results (Cohen's $\kappa \approx 0.9$), indicating strong internal reliability. However, when only the prompt phrasing was changed from stricter to more relaxed, the agreement dropped substantially ($\kappa = 0.66$), showing that the model's judgments were very sensitive to the level of evaluation rigor. Given these observations and the absence of expert-provided labels, we experimented with two explicit evaluation standards, referred to as **soft mode** and **strict mode**, to structure subsequent silver-labeling experiments. In the **soft mode**, a criterion is considered passed (doned) once the target information appears in the dialogue, even if introduced by the patient rather than explicitly elicited or confirmed by the student. For example, the criterion "Inquires about smoking frequency" is considered passed if the patient spontaneously says, "I smoke ten cigarettes a day", after a general question like, "Do you smoke?". In the **strict mode**, the student must explicitly elicit or acknowledge the information, for example here, by acknowledging the notion of frequency explicitly with "That seems quite frequent."

Though this labeling step is performed without direct expert involvement, the silver-labeling standards and modes of strictness used here were inspired by French OSCE redaction guidelines provided to us by OSCE organizers, and validated by real training sessions we witnessed. They provide a transparent and reproducible basis for assessing how reliably an LLM can identify and judge behaviors according to predefined criteria in doctor-patient interactions.

4. Experimental Setup

The overall evaluation workflow is illustrated in Figure 2. For each clinical case, the input to the evaluation system consists of the full **transcript**, a single **criterion**, and a default **task description** serving as the evaluation prompt. The task description instructs the LLM to strictly evaluate (similar to **strict mode**, see Section 3.2) the transcript against the given criterion and output the result in JSON format. Given these inputs, the LLM produces three outputs, in the following order: (i) a short **justification** paragraph; (ii) supporting **evidence**, consisting of the transcript segments most relevant to that decision; and (iii) a binary **true/false decision** indicating whether the criterion is satisfied. In addition to acting as explanations for an output, (i) and (ii) are intended to serve as a task-specific form of *chain-of-thought* reasoning (Wei et al., 2022), hence why

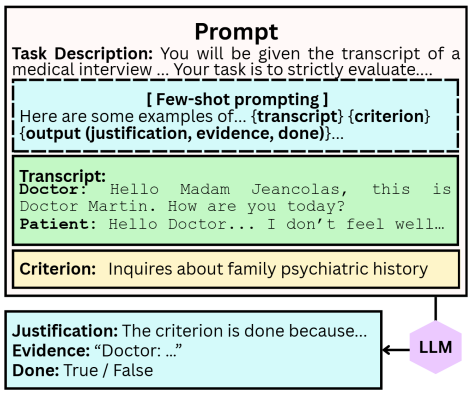

Figure 2: Overview of the evaluation pipeline: an LLM is tasked with judging whether a specific criterion is passed/failed (*Done*), with a *justification* and supporting *evidence*, given a transcript, and optionally some *few-shot* examples.

they precede the binary label decision.

To assess the feasibility of LLM-based automated OSCE assessment, we employ the evaluation workflow introduced above and report the overall binary classification accuracy across all 179 evaluated criteria for both the **perturbed** and **unperturbed** corpora (Table 2). Here, "accuracy" measures agreement with our LLM-generated and human-reviewed (by the authors of this paper) silver labels, rather than performance against a clinician-adjudicated gold standard. We test multiple configurations, combining different LLMs, prompting strategies, and two auxiliary **helper tools** designed to alleviate potential limitations of smaller-scale LLMs. A minimal verification is further conducted on two real tutorial session transcripts to assess consistency with the trends observed on generated data.

4.1. LLM Selection

Due to data privacy concerns, as well as future application as a student training tool, another main focus is on locally deployable open-source models that can be hosted on in-house, reasonably affordable hardware. We thus restrict our selection to quantized models with at most ∼32B parameters for computational feasibility: `qwen3-32b`, `deepseek-r1-32b`, `gpt-oss-20b`, `gemma-3-27b`, `ministral-8b`, `llama3.1-8b`, and `qwen3-8b`, covering both intermediate (20B–32B) and lighter (∼8B) model sizes. For completeness, we also compare these to three state-of-the-art larger LLMs: `GPT-4o` (used for the data generation earlier; see Section 3), `Claude Sonnet 4`, and `Llama-4-Scout`, to provide higher-end points of reference. While none of these models were specifically trained on French and/or medical terminology, preliminary experiments indicated that most of them display adequate usage of French, and even understanding of most medical terms in the training scenarios. To further investigate this point, we also experimented with injecting medical definitions in-context (see Section 4.4).

4.2. Prompting Strategies for Evaluation

We evaluate three prompting strategies. Unlike Geathers et al. (2025), who use specific prompts for 28 station-generic criteria, we adopt a single prompt template designed to handle all 179 station-specific criteria. In the **zero-shot** setting, the model receives as input the **transcript**, the **criterion**, and a **task description** defining the evaluation objective and output format. The **few-shot** setting extends this setup by adding some labeled examples in the task description for guidance (Brown et al., 2020). Figure 2 illustrates the inputs and outputs of each criterion evaluation task. The third strategy is a **multi-step** variant of the zero-shot setup. In this configuration, the process is divided into two stages: a first prompt instructs the model to extract from the full transcript the segments most relevant to the target criterion. Then, a second prompt instructs to evaluate this criterion using only the found transcript segments as context. This design aims to help smaller models maintain focus, as providing the entire transcript may dilute their attention and reduce decision quality.

4.3. Criterion Decomposition (CD)

Many OSCE evaluation criteria are in practice **composite**, containing explicit logical connectors such as *A* OR *B* OR *C*, *A* AND *B*, or *N* OF *A,B,C*. These formulations can be challenging for LLMs, which may misinterpret their intended logic, for instance by treating an OR as an AND, or being overly permissive with chained ANDS. A typical example is the criterion "Asks about the impact of memory (element A) OR concentration problems (element B)," where the model occasionally assumed that both sub-elements had to be validated simultaneously. To mitigate such errors, we propose systematically decomposing composite criteria in evaluation sheets into relatively independent sub-criteria by prompting `GPT-4o` as a preprocessing step. Each sub-criterion obtained is then evaluated as though it were a separate item, and the individual binary decisions are then aggregated programmatically according to the operators, to produce the final decision for the original composite criterion.

4.4. Medical Definitions (MD)

To help the LLMs interpret clinical terminology, we propose to enrich prompts with **medical definitions**. A medical NER model (`medkit/DrBERT-CASM2`, derived from Labrak et al. (2023)) first

identifies relevant medical multi-word expressions, which are then matched to concepts in the Unified Medical Language System (UMLS, National Library of Medicine (US) (2024)), an external biomedical knowledge base. To retrieve definitions for these concepts, we first restrict UMLS to **French** and **English** definition sources, as they can be reliably processed by the selected LLMs. We then select the first available definition for each matched concept from the knowledge base. For example, if "*trouble obsessionnel compulsif*" is identified, it is mapped to the concept *obsessive–compulsive disorder* (OCD) in UMLS, from which a definition such as "An anxiety disorder characterized by ..." is extracted and injected into the prompt.

5. Results and Discussion

To assess the feasibility of LLM-based automated OSCE evaluation, we report the overall binary classification accuracy across all 179 evaluated criteria, for both the **perturbed** and **unperturbed** corpora (Table 2). Some results are omitted for brevity.

Overall model performance: Large industry-leading models such as `GPT-4o` and `Claude 4 Sonnet` achieve consistently robust performance across all datasets, with accuracies of ∼90% in most configurations. `Llama-4-Scout` performed slightly less well, possibly on account of its smaller size and Mixture-of-Experts nature. While previous studies in English OSCE contexts have reported similar observations regarding the promising performance of large models on automated evaluation tasks, those benchmarks are not directly comparable due to linguistic and structural differences. Nonetheless, these results provide encouraging evidence that LLM-based evaluation of French OSCE transcripts is likewise feasible. Though it was used to generate the initial silver labels, some of them were corrected after manual review (see Section 3.2), hence why `GPT-4o` does not obtain close to perfect scores here.

Table 2 shows that models such as `qwen3-32b`, `deepseek-r1-32b`, and `gpt-oss-20b`, together with the lighter `qwen3-8b`, achieve accuracies comparable to `GPT-4o`. `qwen3-32b` is particularly stable, displaying the best accuracy on the **perturbed** corpus for the *zero-shot* and *few-shot* strategies when using the criterion-decomposition tool, and a close second on the **unperturbed** corpus. In contrast, `ministral-8b` displays inconsistent results depending on the strategy and tools used, while `llama3.1-8b` remains consistently worse in all configurations. Results for `gemma-3-27b` were excluded because the model failed to consistently produce the required JSON format. These results may reflect differences in these models' training: `qwen3`, `deepseek`, and `gpt-oss` incorporate enhancement methods such as knowledge distillation from their larger variants, while `llama3.1` and `ministral` were intentionally fine-tuned with more conventional methods, focusing more on data scale and direct alignment techniques.

Effect of prompting strategies: The *zero-shot* setting, while the simplest, proved the most robust, yielding coherent and stable evaluations across cases. Unlike Geathers et al. (2025), we observe no noticeable degradation with the *few-shot* setting compared to zero-shot. A notable exception was `ministral-8b`, where few-shot prompting reduced accuracy, possibly due to a sensitivity to prompt length. In contrast, the *multi-step* strategy degraded performance compared to zero-shot, consistent with prior findings (Geathers et al., 2025). After manual review, we conclude the primary causes were (i) *context loss*, as the evaluation step only sees retrieved partial spans rather than the full transcript, and (ii) *error propagation*, where the retrieval step missing relevant evidence directly biases the downstream binary decision. The latter was especially prevalent, due to the fact criteria-specific information was not necessarily segregated to a few specific spans, but instead often distributed over multiple dialogue turns.

Impact of composite-criteria decomposition: The decomposition tool yielded mixed benefits overall. Improvements appeared mainly in the **perturbed** corpus, where nearly half of the evaluation criteria were transformed into perturbed versions during dialogue generation. Composite criteria ("inquires A OR B") were particularly affected, leading to dialogues in which the virtual doctor executed only one sub-element ("inquires A") while evaluation remained based on the original combined criterion. Some models (e.g., `qwen3-32b` and `llama3.1-8b`) tended to misinterpret OR operators as ANDS, thus mispredicting failures. Decomposition mitigated this by isolating sub-criteria before evaluation. In **unperturbed** dialogues, where criteria were typically fully executed by the simulated doctor, the effect was more negligible. For `gpt-oss-20b`, decomposition slightly reduced accuracy, reflecting a tendency towards over-strict evaluation.

Impact of medical definition injection: The injection of UMLS-based definitions did not yield clear improvements. This is likely because expert-level medical terminology was infrequent, most models already understood common medical terms, and the identification of multi-word medical expressions introduced additional noise. Future gains may come from more selective filtering and evaluations on more jargon-heavy OSCE scenarios.

Preliminary verification on real cases: Previous experiments on synthetic transcripts confirmed that several models were sufficiently performant

perturbed	qwen3-32b				deepseek-r1-32b				gpt-oss-20b				gpt-4o			
	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD
zero-shot	84.36	**90.50**	83.24	**90.50**	83.24	84.92	82.12	**86.03**	**87.71**	86.59	**87.71**	86.59	84.92	87.71	84.36	**88.83**
few-shot	86.59	**90.50**	86.59	89.94	82.68	**85.47**	82.12	84.92	**87.71**	84.62	86.03	84.92	86.03	**86.59**	84.92	**86.59**
multi-step	81.01	**84.92**	82.68	84.83	**79.89**	74.86	78.77	77.65	**83.24**	78.77	81.56	78.21	83.80	**85.47**	80.45	83.24

perturbed	llama3.1-8b				qwen3-8b				ministral-8b			
	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD
zero-shot	48.04	50.84	50.84	**58.66**	83.24	82.12	81.56	**84.36**	71.51	**76.54**	67.04	73.74
few-shot	54.75	**55.87**	**55.87**	54.75	**85.47**	81.56	82.12	82.68	62.01	**65.36**	63.13	62.57
multi-step	53.63	52.51	52.51	**56.98**	69.83	71.51	**72.63**	69.27	59.78	59.78	59.22	**66.48**

unperturbed	qwen3-32b				deepseek-r1-32b				gpt-oss-20b				gpt-4o			
	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD
zero-shot	**90.5**	89.94	**90.5**	89.39	83.8	84.36	**86.59**	82.68	86.59	83.71	**87.71**	83.8	89.94	87.71	**90.5**	85.47
few-shot	**89.89**	88.83	88.76	88.83	**88.83**	78.77	85.47	80.45	**84.92**	83.62	83.24	**84.92**	**92.18**	87.15	**92.18**	88.2
multi-step	**86.03**	79.89	84.36	79.33	**70.95**	69.83	**70.95**	68.16	69.27	73.03	70.95	**74.3**	**87.15**	79.33	86.59	81.94

unperturbed	llama3.1-8b				qwen3-8b				ministral-8b			
	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD	direct	CD	MD	CD+MD
zero-shot	50.84	**53.63**	41.34	51.40	**87.64**	85.39	83.24	83.24	78.77	**84.36**	78.77	78.77
few-shot	**54.91**	54.19	53.07	51.40	**87.71**	85.47	83.24	83.80	**64.80**	56.42	62.01	56.98
multi-step	43.02	**49.16**	39.11	44.13	**65.36**	**65.36**	60.89	61.45	58.66	62.01	**65.92**	58.10

Table 2: Results on both the **perturbed** and **unperturbed** corpora. (direct = no auxiliary tools; CD = criterion decomposition; MD = medical definitions). For each model, the best accuracy score for each prompting strategy is **bolded**. The overall best scores across all models are underlined.

and stable for automated OSCE evaluation. To verify whether these findings extend to real data, we conducted an additional small-scale experiment using two authentic transcripts (38 criteria in total) obtained after-the-fact from tutorial OSCE sessions, where a teacher played the role of the doctor (Table 3). The table reports results for a representative subset of models: the reference model `GPT-4o`, the most performant and stable open-source model `qwen3-32b`, the lighter yet competitive `qwen3-8b`, and the weaker baseline `llama3.1-8b`. Results followed the same overall trends observed on synthetic data, with accuracy levels remaining consistent across settings. Although limited, this small-scale experiment suggests that the relative trends observed on synthetic transcripts may carry over to these tutorial dialogues.

Real case	qwen3-32b	gpt-4o	qwen3-8b	llama3.1-8b
Zero-shot	86.97	83.47	75.77	37.25

Table 3: Average zero-shot accuracy on two real-case OSCE transcripts, without helper tools.

Discussion: Our experiments suggest that general-purpose LLMs show strong potential for automated evaluation in French OSCE settings. While there were initial concerns about their ability to handle French medical terminology, the results indicate that this limitation may be less severe than expected: for most common clinical terms encountered in OSCEs, current models appear encouraging, especially compared to the currently purely indicative grading used in OSCE training sessions, due to the lack of expert examiners. Although the *preliminary verification on real transcripts* exhibited trends similar to those observed on synthetic data, a more systematic validation of the synthetic dataset remains necessary once larger sets of authentic transcripts become available. In particular, stricter validation should compare generated and real dialogues directly—for instance, by aligning criterion labels and applying similarity metrics such as BLEURT (Sellam et al., 2020), or by using *LLM-as-a-judge* approaches.

6. Conclusion

In this work, we developed a controlled pipeline for generating synthetic French OSCE training transcripts alongside an automated evaluation framework for clinical-skills criteria based on locally hostable LLMs. By structuring dialogue generation around evaluation criteria and incorporating perturbations to simulate less idealized student performances, we produced two small datasets for benchmarking LLM-based automated evaluation.

Our experiments indicate that mid-size open-source LLMs (20–32B parameters) can deliver performance comparable to much larger language models, while lighter models ($\sim$ 8B) remain competitive and may even support closer-to-real-time usage. These findings are encouraging, and highlight the feasibility of locally deployable solutions, that would ensure greater control and respect for privacy in this kind of educational setting.

In our binary evaluation task, we also found that although not specifically trained on the medical domain, recent general-purpose LLMs generally handle both French and its medical terminology well, making external definition injection unnecessary

in most cases. However, their handling of composite criteria and logical operators remains less reliable, meaning that supportive external tools are still necessary for some subtasks.

While encouraging, important challenges remain. The synthetic transcripts generated with this framework are likely to reflect more idealized student performances, highlighting the need for larger-scale validation against authentic OSCE data with medical expert annotations. In addition, our evaluation focused primarily on binary labels: future studies should also assess the quality of generated justifications and the relevance of transcript excerpts provided as evidence.

In future work, we plan to expand the diversity of generated scenarios to better reflect the breadth of OSCEs. Medical expert educators will also be involved to calibrate the evaluation strictness modes and validate the automatic labeling of generated transcripts.

Acknowledgements

We would like to thank Yongqiang Yu for proposing the initial idea of adjustable levels of strictness in the evaluation framework and for insightful discussions during the development of this work.

Data and Code Availability Statements

Experiments with local LLMs were performed with a mixture of locally hosted models and, for convenience, OpenRouter-hosted models to parallelize runs, using `Q4_K_M` quantizations as a balance between resource efficiency and quality. All prompts, generated transcripts, and implementation details used in this work will be released as supplementary material upon acceptance.

Limitations

The generated OSCE dialogues, while diverse, remain partly idealized because explicit evaluation criteria guide their structure. As a result, they tend to under-represent spontaneous or "off-script" behaviors typical of real student performances. In particular, the generated dialogues are well-structured and logically coherent but lack hesitations, repetitions, self-corrections, interruptions, and occasional off-topic responses that are common in authentic OSCE conversations. Perturbations of leaf criteria only partially mitigate this effect, and the resulting errors remain comparatively "structured" rather than chaotic, compounding, or unstable as in real student behavior. Larger validation sets, including authentic transcripts with expert annotations, are therefore needed to assess ecological validity; our real-data check is limited to two tutorial-session transcripts (38 criteria) and should not be viewed as a full validation.

Additionally, our synthetic dialogue transcripts do not reproduce the transcription artifacts and conversational noise typically present in real recordings (e.g., disfluencies, interruptions, or incomplete turns). As such, the present study remains theoretical with respect to the robustness of evaluation methods to real-world transcription variability. The silver-labeling procedure used `GPT-4o` to produce reference decisions without expert intervention. Although the labels were manually reviewed for consistency, they did not undergo a formal adjudication process. Accordingly, the labels should be considered a reviewed silver standard rather than definitive ground truth, and the reported accuracies reflect agreement with this reference rather than clinician-adjudicated validity.

Finally, since the synthetic transcripts were generated with `GPT-4o`, they may reflect its linguistic style, which could bias results and partially inflate evaluation performance for models that align better with this style. Future work can mitigate this risk by diversifying generation (e.g., multiple generators / style variation) and by evaluating on larger sets of authentic transcripts when available.

A single evaluation run was performed for each configuration. A temperature of zero was not used because several models still produced nondeterministic and/or incoherent outputs under that setting; instead, following preliminary experiments, a temperature of 0.2 was adopted, balancing determinism and output quality. While averaging over multiple runs could further improve robustness, this was not pursued due to time and resource constraints. Nevertheless, the current results provide a sufficient basis for a first exploration of the feasibility of LLM-based evaluation in French OSCEs.

Finally, the evaluation pipeline focused exclusively on binary decisions. Other aspects, such as the linguistic adequacy of student utterances, non-verbal behavior, or the quality and pedagogical value of model-generated justifications and extracted evidence segments, remain outside the present scope, but may be explored in future work.

Ethics Statement

This study did not involve any collection or processing of personal, clinical, or patient-identifiable data. All dialogues used for experimentation were synthetically generated by large language models (LLMs) based solely on OSCE training scenarios invented by medical educators. No recordings or transcripts from real examinations or students were used at any stage, other than purely as inspiration for defining evaluation standards and understanding the pedagogical context of such practical

exams.

Because all data are artificial, the system presented in this paper is not intended for immediate deployment in real training or examination settings. While our approach provides a controlled framework for studying LLM-based assessment, the synthetic nature of the dialogues implies that unknown model biases may affect both the realism of generated interactions and the reliability of automated judgments. Furthermore, the justifications and evidence segments produced by the models have not yet been systematically evaluated for their appropriateness or pedagogical value; such verification is planned for future research. Accordingly, these outputs should not be used as educational feedback without prior expert validation and ethical oversight. Consequently, any educational use of the proposed system would require prior expert validation, bias auditing, and alignment with ethical guidelines governing medical training and evaluation.

Bibliographical References

Majid Alizadeh and Maryam Jafar Sameri. 2025. Intelligent assessment systems in medical education: A systematic review. *Journal of Advances in Medical Education & Professionalism*, 13(3):173–190.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Krystle K. Campbell, Michael J. Holcomb, Sol Vedovato, Lenora Young, Gaudenz Danuser, Thomas O. Dalton, Andrew R. Jamieson, and Daniel 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):202.

Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic Patient-Physician Dialogue Generation from Clinical Notes Using LLM. arXiv preprint arXiv:2408.06285.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jadon Geathers, Yann Hicke, Colleen Chan, Niroop Rajashekar, Sarah Young, Justin Sewell, Susannah Cornes, Rene F. Kizilcec, and Dennis Shung. 2025. Benchmarking generative ai for scoring medical student interviews in objective structured clinical examinations (osces). In *Artificial Intelligence in Education*, pages 231–245, Cham. Springer Nature Switzerland.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv preprint arXiv:2411.15594.

Iqbal Haider, Aliena Badshah, Arif Raza Khan, and Abid Ullah. 2018. Fatigue level of examiners during objective structured clinical examination (osce). *Journal of Medical Sciences*, 26(3):207–210.

Andrew R. Jamieson, Michael J. Holcomb, Thomas O. Dalton, Krystle K. Campbell, Sol Vedovato, Ameer Hamza Shakur, Shinyoung Kang, David Hein, Jack Lawson, Gaudenz Danuser, and Daniel J. Scott. 2024. Rubrics to prompts: Assessing medical student post-encounter notes with ai. *NEJM AI*, 1(12):AIcs2400631.

S. M. Kurtz and J. D. Silverman. 1996. The Calgary-Cambridge Referenced Observation Guides: An aid to defining the curriculum and organizing the teaching in communication training programmes. *Medical Education*, 30(2):83–89.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper*, Toronto, Canada. Association for Computational Linguistics.

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.

Ministère des Solidarités et de la Santé. 2021. https://sante.gouv.fr/identitovigilance. Accessed: October 2025.

National Library of Medicine (US). 2024. UMLS Knowledge Sources, Release 2024AA. http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html. Bethesda (MD); Released 2024-05-06, Accessed 2024-07-15.

Aimé Nun, Olivier Birot, Gaël Guibon, Frédéric Lapostolle, and Ivan Lerner. 2025. SIM-SAMU - A French medical dispatch dialog open dataset. *Computer Methods and Programs in Biomedicine*, 268:108857.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2023. SOAP Notes. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

H. M. T. W. Seneviratne and S. S. Manathunga. 2025. Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings. *BMC Medical Education*, 25(1):1146.

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. arXiv preprint arXiv:2410.12858.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna M. Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. 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.

Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. 2025. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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.