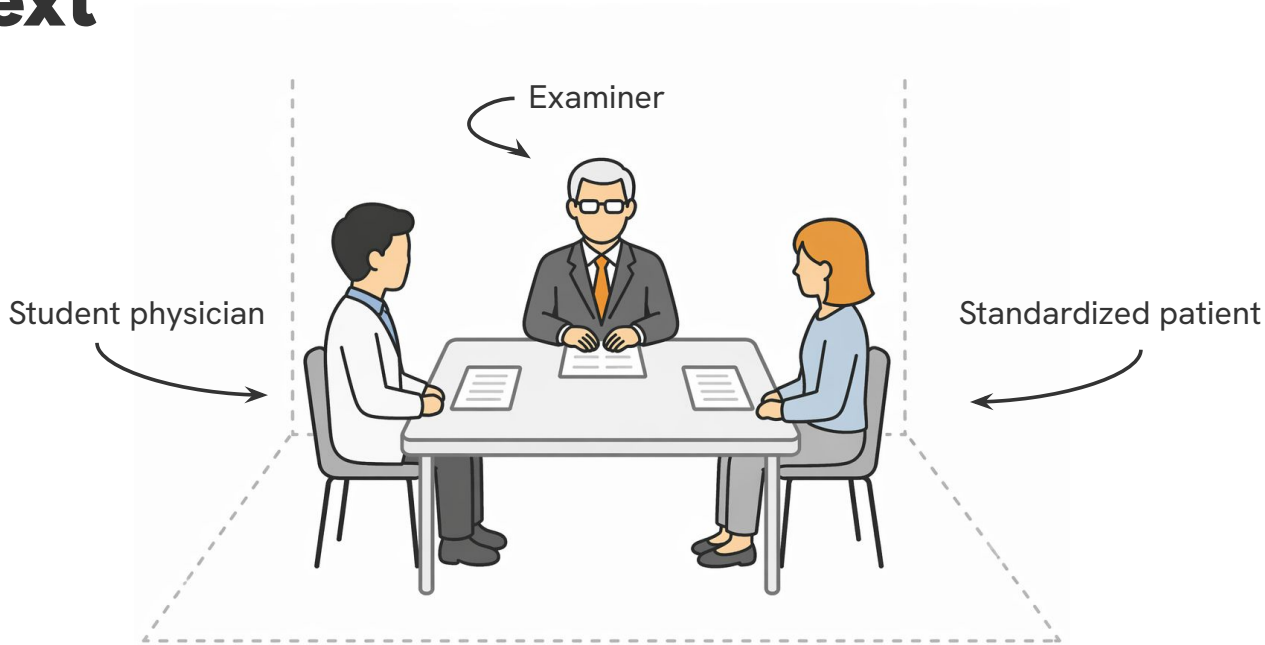
2026 SIGdial Atlanta, GA, USA

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]Université de Lorraine, LORIA, Inria, CNRS
[2]Université d'Avignon, Laboratoire d'Informatique d'Avignon (LIA)

Context



OSCE: **O**bjective **S**tructured **C**linical **E**xaminations

Students are evaluated on their **clinical reasoning** and their **communication skills**

Context

Scenarios based on **text** sheets:

- **Student sheet**: context and objectives of the consultation
- **Patient sheet**: background, medical history, symptoms, state of mind, etc.
- **Examiner sheet**: binary medical criteria, scale-based criteria

Clinical context: You are Samedi Février, a general practitioner.
You are seeing a 17-year-old teenager who was last examined by their general practitioner two months ago.

(...)

Objectives: During this consultation, you should:
- Conduct a medical history interview focused on the patient's sleep disturbances and associated behavioral changes.
- Present your management plan to the examiner at the end of the consultation.

Context

Scenarios based on **text** sheets:

- **Student sheet**: context and objectives of the consultation
- **Patient sheet**: background, medical history, symptoms, state of mind, etc.
- **Examiner sheet**: binary medical criteria, scale-based criteria

Opening statement: "Hello Doctor, I'm here because I can't take it anymore with my mother. I feel suffocated."

Consultation summary: You were brought to your general practitioner by your mother and are being seen by a resident in general practice. Your main goal is for the situation to improve, especially your relationship with your mother.

Identity
Age: 17
Date of birth: January 2, 2009
Sex: Female (...)

Context

Scenarios based on **text** sheets:

- **Student sheet**: context and objectives of the consultation
- **Patient sheet**: background, medical history, symptoms, state of mind, etc.
- **Examiner sheet**: binary medical criteria, scale-based criteria

Binary criteria

- Investigate the patient's sleep (timing, duration, sleep onset, early awakenings) (at least 3)
- Assess the impact of insomnia on daily functioning (daytime sleepiness, performance, memory, concentration) (at least 2)
- Ask about alcohol, recreational drugs, or self-medication related to insomnia
- (...)

Motivation

- **Limited availability of standardized patients and evaluators**
  → limited opportunities for repeated OSCE practice

- **Very few French OSCE resources: datasets and patient systems**
  [1-3]

- **Compared to previous systems, LLM-based virtual patients (VP) improve realism…**
  but still suffer from limited controllability and lack of standardized evaluation [4-6]

[1] Laleye et al. French Medical Conversations Corpus for a Virtual Patient Dialogue System. 2020. LREC.
[2] Campillos-Llanos et al. Virtual patient dialogue system based on terminology-rich resources. 2019. NLE.
[3] Voigt et al. LLM-powered virtual patient agents for clinical skills training. 2025.
[4] García-Torres et al. Enhancing clinical reasoning with virtual patients. 2024. Healthcare.
[5] Cook et al. Virtual patients using large language models. 2025. JMIR.
[6] Li and Lutfi. LLM-based virtual patient systems for history-taking. 2026. Journal of Medical Systems.

Contributions

01

French OSCE dialogue dataset

02

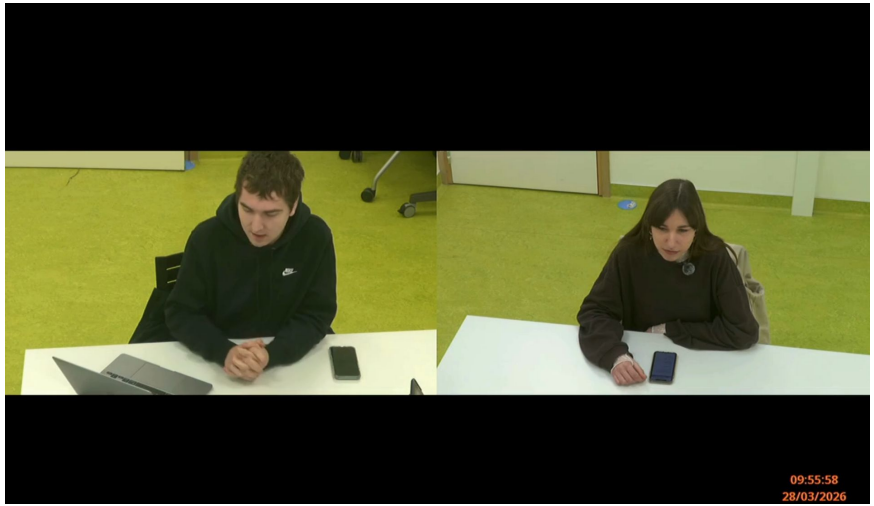
Controllable LLM-based VP pipeline

03

Multi-level evaluation framework

Goal: develop a realistic and controllable virtual patient for French OSCE training

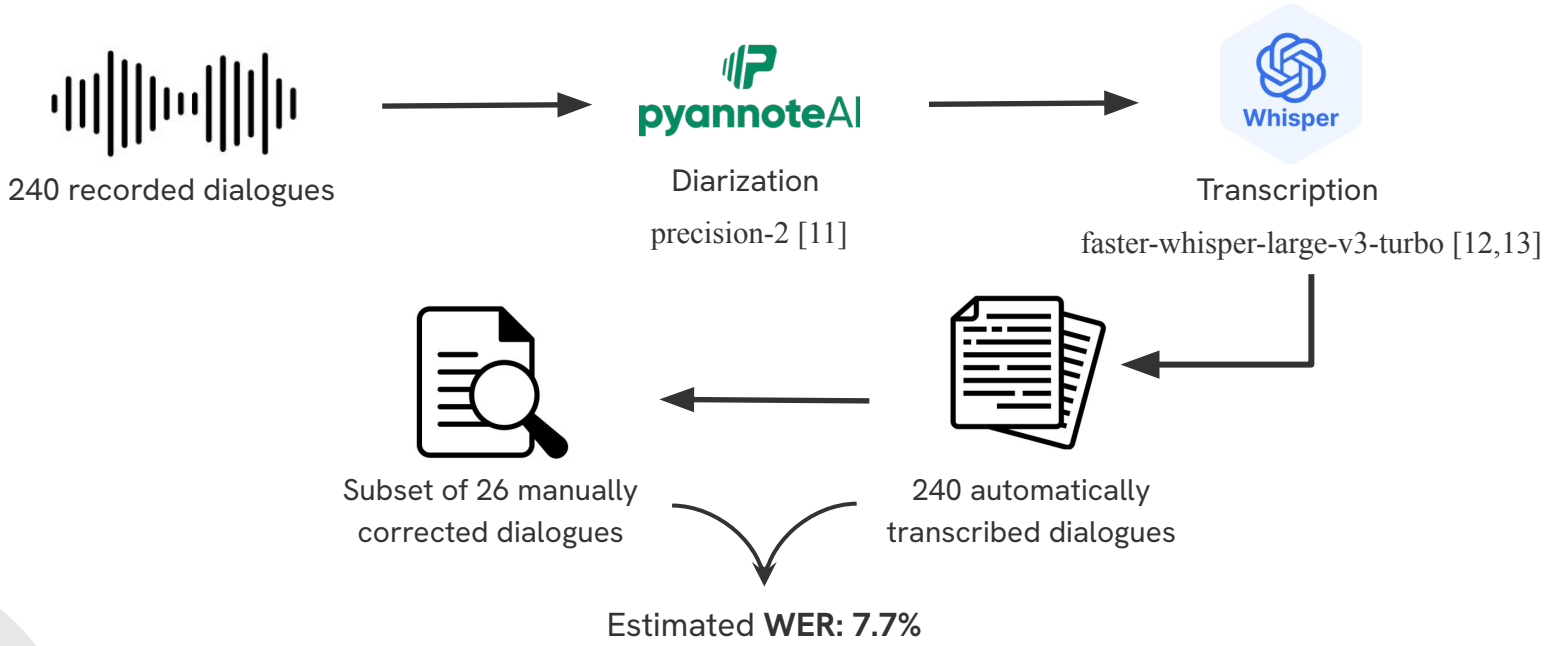
French OSCE dialogue dataset



Component	Count
Recorded dialogues	
Recorded dialogues	240
Total audio duration (hours)	30
Generated dialogues	
Generated dialogues	792

Dataset may be accessed upon request on Zenodo

French OSCE dialogue dataset



[11] PyAnnoteAI. precision-2 diarization model. 2024.
[12] OpenAI. Whisper large-v3-turbo. 2024.
[13] Radford et al. Robust Speech Recognition via Large-Scale Weak Supervision. 2023. ICML.

Controllable virtual patient pipeline

- Focus on **text-based interactions**
- LLM-based generation of **synthetic dialogues for evaluation**

We propose a prompt-controlled LLM pipeline for instruction-based patient response generation

Physician and patient are both simulated
Used for large-scale dialogue generation
Enables controlled experiments

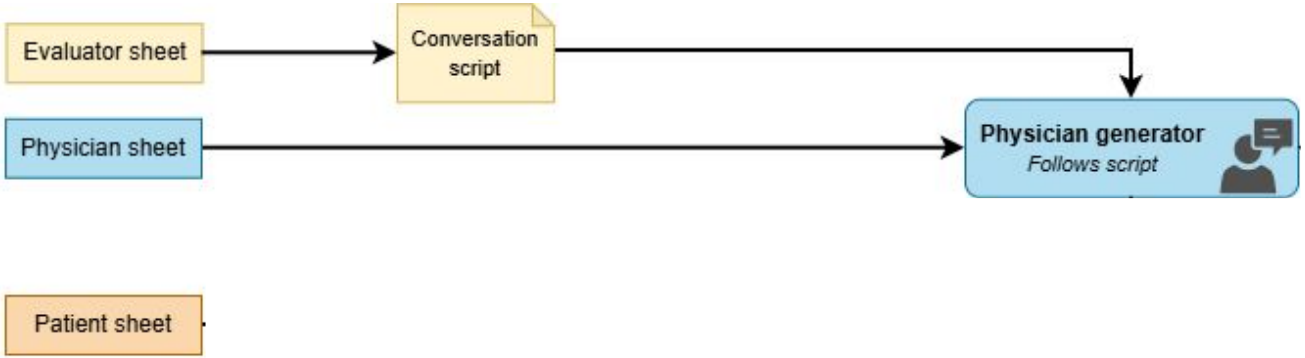


Fig. Overview of the controllable LLM-based dialogue generation pipeline for OSCE simulations.

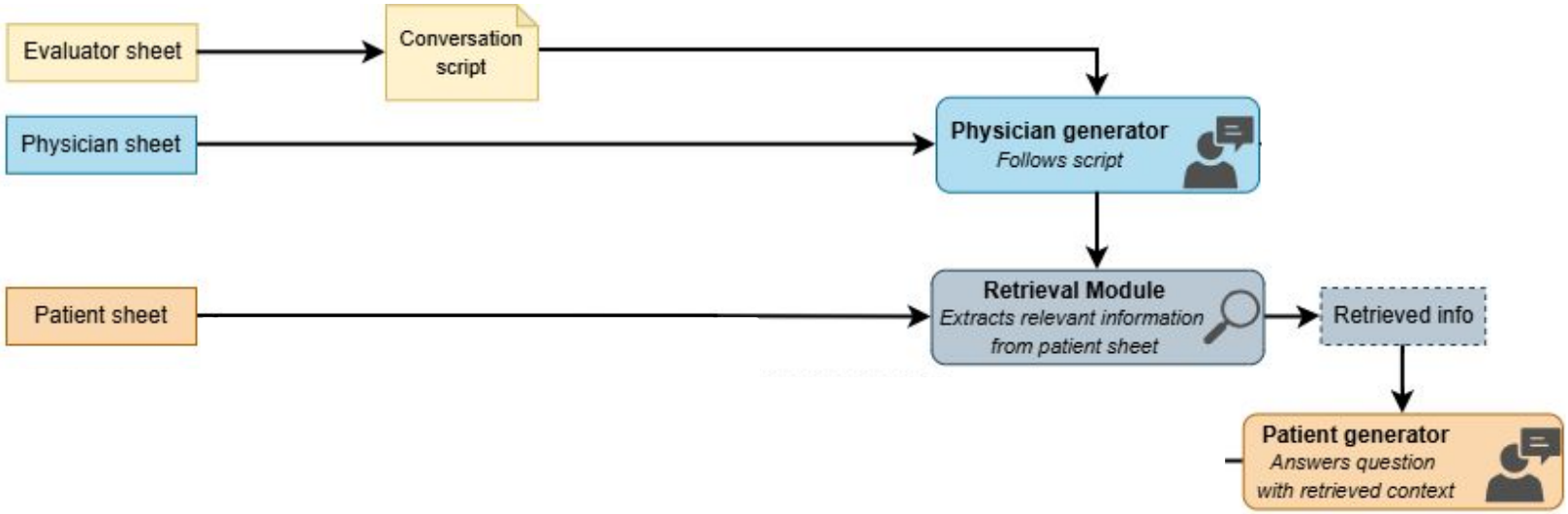


Fig. Overview of the controllable LLM-based dialogue generation pipeline for OSCE simulations.

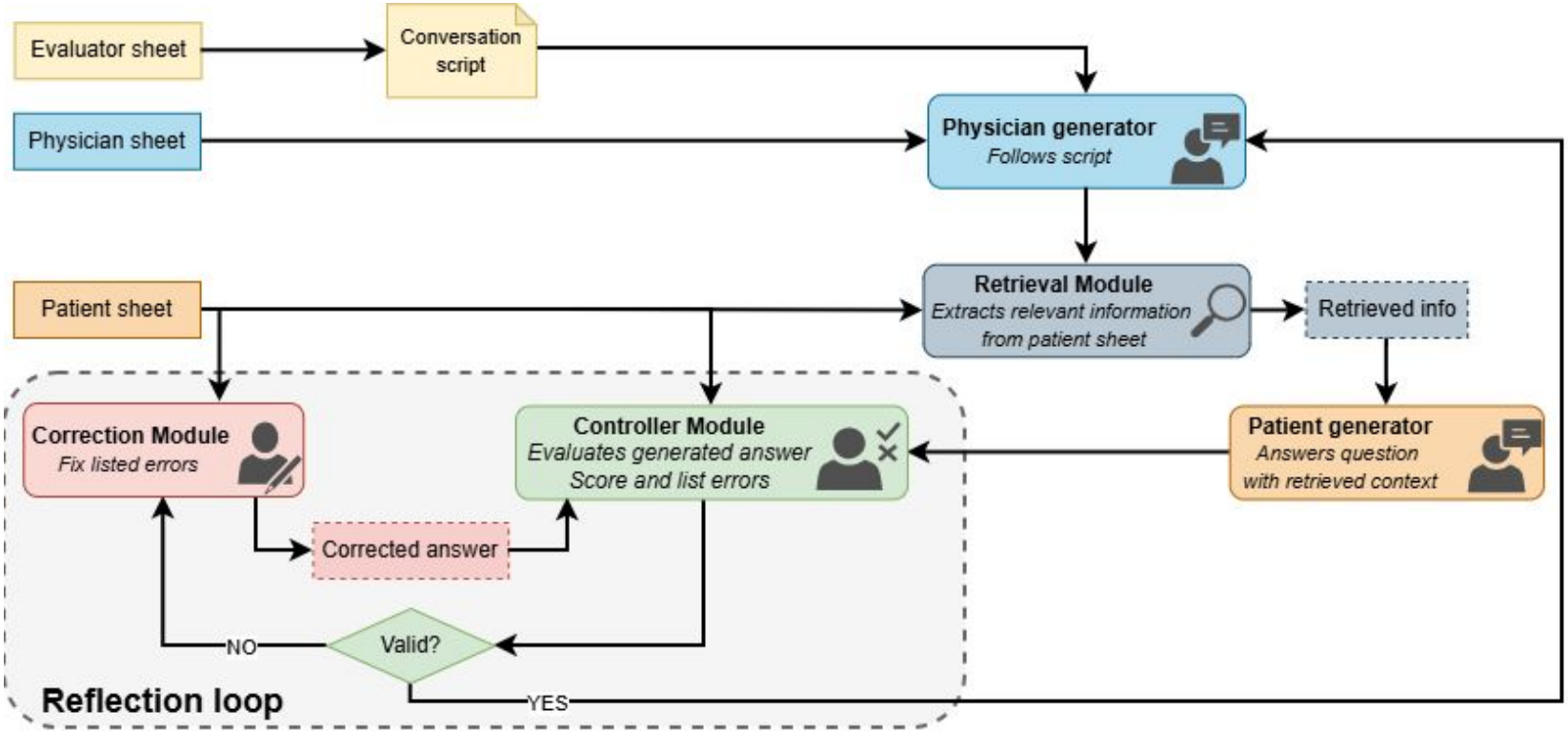


Fig. Overview of the controllable LLM-based dialogue generation pipeline for OSCE simulations.

Evaluation framework

Patient simulation

Information recall (%):
how much patient information is revealed during the dialogue

Response relevance (/5):
how appropriate the patient's answers are to the physician's questions

Physician performance

OSCE evaluation (%):
how many expected OSCE criteria are covered by the student

Role adherence (/5):
how well the physician follows the expected role and consultation

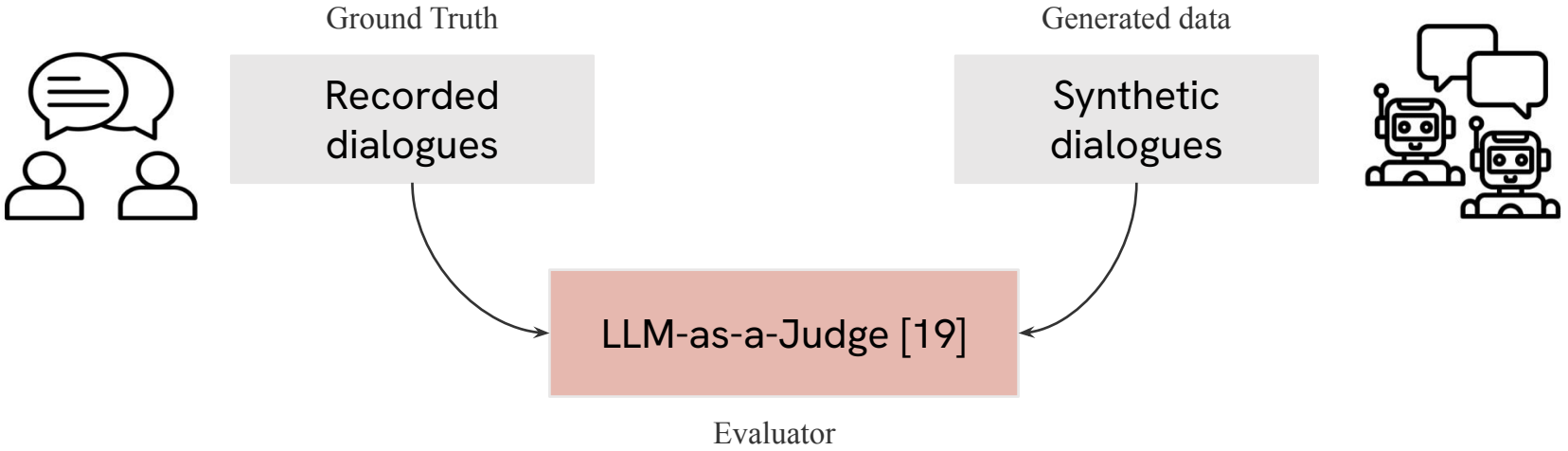
Linguistic quality

Naturalness (/5):
how human-like the dialogue ‘sounds’

Fluency (/5):
how smoothly the conversation flows

Coherence (/5):
how logically the dialogue progresses

Evaluation framework



[19] Gu et al. Survey on LLM-as-a-Judge. 2026. The Innovation.

I. Generated vs. Recorded dialogues

Model	Patient simulation		Linguistic quality	
	Information recall ↑ (%)	Response relevance ↑ (/5)	Fluency ↑ (/5)	Coherence ↑ (/5)
Recorded dialogues	43.58	<u>4.14</u>	3.79	3.91
Gemini-3.1-Flash-Lite	47.26	4.03	<u>4.07</u>	4.20
Ministral-14B	<u>47.44</u>	**4.97**	3.98	<u>4.24</u>
Claude-Haiku-4.5	**52.24**	**4.97**	**4.35**	**4.28**

Table: Results for recorded and generated dialogues on the baseline configuration. Best results in bold, second-best underlined. ↑ Higher is better.

- **Generated dialogues outperform recorded interactions**, including higher information recall across all models
- Claude-Haiku-4.5 consistently achieves the best overall performance

II. Impact of the reflection loop and retrieval module

Experimental setup	Patient simulation		Physician performance		Linguistic quality	
	Information recall ↑ (%)	Response relevance ↑ (/5)	OSCE evaluation ↑ (%)	Role adherence ↑ (/5)	Fluency ↑ (/5)	Coherence ↑ (/5)
baseline	52.24	**4.97**	87.48	4.76	**4.35**	4.28
reflection	54.29	4.83	87.75	4.85	4.30	**4.36**
refl. + retr.	**56.62**	4.89	**89.21**	**4.89**	4.29	**4.36**

Table: Results for Claude-Haiku-4.5 across three experimental setups (baseline with no optional modules, reflection loop, and reflection + retrieval). Best results in bold. Higher is better.

- Reflection loop and retrieval modules sometimes improve results, but their **impact remains moderate**

Conclusions

Main findings

- **French OSCE dialogue dataset** with recorded and synthetic interactions
- **LLM-based simulation of the patient and physician**
- **Generated dialogues score higher than recorded ones**, but this may partly reflect controlled generation and LLM-as-a-Judge bias
- **Reflection, retrieval, and physician-robustness improve fidelity and relevance**, but gains remain moderate.

Future work

- Conduct user studies with medical students
- Evaluate usability, pedagogical value, and feedback quality
- Expand pipeline to support document-based (analyses, radios, pictures, ...) OSCE scenarios

Thank you!

Any questions, comments? doria.bonzi@loria.fr

References

[1] Fréjus A. A. Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 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, May 2020. European Language Resources Association.

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

[3] Henrik Voigt, Yurina Sugamiya, Kai Lawonn, Sina Zarrieß, and Atsuo Takanishi. LLM-powered virtual patient agents for interactive clinical skills training with automated feedback, 2025.

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

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

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

[7] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes. In Findings of the Association for Computational Linguistics ACL 2024, pages 15183–15201, 2024. Association for Computational Linguistics.

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

[9] 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, Yanqiu Xing, Yongfeng Zhang, Themistocles L. Assimes, Danielle S. Bitterman, Xin Ma, Lin Lu, and Lizhou Fan. Simulated Patient Systems Powered by Large Language Model-Based AI Agents Offer Potential for Transforming Medical Education. Communications Medicine, 6(1):27, 2025. Nature Publishing Group.

[10] Laurence Chaby et al. 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, 2022. Frontiers.

[11] PyAnnoteAI. pyannote speaker-diarization precision-2, 2024. Speaker diarization model.

[12] OpenAI. Whisper large-v3-turbo, 2024. Automatic speech recognition model.

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

[14] Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Antoine Neuraz, Sophie Rosset, and Pierre Zweigenbaum. Lessons Learned from the Usability Evaluation of a Simulated Patient Dialogue System. Journal of Medical Systems, 45(7), 2021.

[15] Kyong-Jee Kim. Factors associated with medical student test anxiety in objective structured clinical examinations: A preliminary study. International Journal of Medical Education, 7:424–427, 2016.

[16] Ameer Hamza Shakur, Michael J. Holcomb, David Hein, Shinyoung Kang, Thomas O. Dalton, Krystle K. Campbell, Daniel J. Scott, and Andrew R. Jamieson. Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis. arXiv preprint arXiv:2410.12858, 2024.

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

[18] 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. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36:46595–46623, 2023.

[19] Jiaxin Gu, Xiaojun Jiang, Zhiyu Shi, Hong Tan, Xiaolong Zhai, Cheng Xu, Wenxiang Li, Yifei Shen, Shuang Ma, Haotian Liu, Shuang Wang, Kai Zhang, Zhiyong Lin, Baosheng Zhang, Lei Ni, Wen Gao, Yanfeng Wang, and Jun Guo. A Survey on LLM-as-a-Judge. The Innovation, 2026.

How reliable is our LLM-as-a-Judge?

Linguistic quality
- Naturalness
- Fluency
- Coherence

6 annotators
12 dialogues

→ **Weak agreement** between humans
Limited humans/LLM-as-Judge **correlation**

OSCE evaluation
(e.g. asks about allergies, proposes a treatment)

Average OSCE scores for 14 stations
67 recorded dialogues

→ **Strong agreement** for stations without document analysis

→ **Reliable for checklist-based evaluation, less reliable for subjective dialogue quality**

Is our VP robust to physician performance variability?

Physician performance variability	Patient	Physician	Linguistic quality		
	Response relevance ↑ (/5)	OSCE evaluation ↑ (%)	Naturalness ↑ (/5)	Fluency ↑ (/5)	Coherence ↑ (/5)
100%	4.89	89.21	3.59	4.29	4.36
50%	4.86	45.05	3.46	4.20	4.20
25%	4.75	27.06	3.00	4.00	4.10

Table: Results for Claude-Haiku-4.5 with reflection + retrieval under different physician performance levels. ↑ Higher is better.

- **VP remains robust to physician performance variations**
- Response relevance remains stable
- Linguistic quality only slightly decreases