COGITO: A Dynamic LLM-Driven Virtual Patient Simulation for Speech-language Pathology Training

Loïc Braud*, Alexandra Plonka†, Olivier Cavadenti‡, Maël Addoum*,
* *ISART Digital, R&D Lab, Paris, France*
† *Laboratoire CoBTek*, Université Côte d'Azur, Nice, France
† *Département d'Orthophonie*, Faculty of Medicine, Nice, France
‡ *Intellectuality, Dijon, France*

***Abstract*—This paper presents COGITO, a fully offline LLM-driven virtual patient simulation for speech-language pathology training in a VR environment. The system combines MetaHuman avatars, local large language model inference, and real-time voice interaction to enable dynamic and naturalistic clinical consultations. It simulates patients with mild cognitive impairment through a modular prompt architecture encoding cognitive, linguistic, and behavioral profiles. A preliminary pilot study with 11 medical faculty students showed that the system operates in real time with a latency below 2 seconds, while achieving high perceived cognitive fidelity and strong user engagement.**

***Index Terms*—. Large Language Models; Simulated Patient; Clinical Training; VR environment;**

I. Introduction

Speech-language pathology currently faces several converging challenges: the growing demand for care linked to neurodegenerative disorders in an aging population and the need to train students with clinical and interpersonal competencies [1]. Moreover, internship opportunities in the neurocognitive field are limited, which prevents students from gaining adequate hands-on training. There is a need for a complementary learning tool to develop crucial skills alongside academic courses. The landscape of digital learning tools has expanded considerably [2]. However, most existing tools rely on static, predefined, and non-interactive scenarios, which limit meaningful interaction with a patient's avatar.

Recent advances in large language models (LLMs) have opened new directions for virtual patient simulation in clinical training. LLM-powered virtual patients offer a scalable and alternative to standardized human patients [3]. However, the vast majority of existing systems rely on textual interaction and cloud-based APIs, as illustrated by Cook et al. [4] who explicitly compared *GPT-4.0-Turbo* against *GPT-3.5-Turbo* for clinician–patient dialogue generation. Even VR-integrated systems such as CLiVR implemented an external lightweight backend LLM based on *Gemini 2.0-Flash* [5]. Likewise, VAPS combines LLMs and embodied conversational agents within a VR environment using *Convai* API with *GPT-4o* [6]. Despite the high contextual richness provided by online LLMs, concerns remain about patient data privacy, internet dependency, and deployment in controlled academic settings. Furthermore, speech-language pathology and neurocognitive language disorders remain underrepresented in the existing LLM-based clinical simulations, which generally focus on internal medicine fields or public health [7][8][9].

To address these gaps, we present COGITO, a fully offline, locally deployed LLM-driven Virtual Reality (VR) platform designed specifically for speech-language pathology training. COGITO eliminates any dependency on external APIs while providing naturalistic voice-based interaction in a dynamic simulation that allows training clinical and communication students' skills.

II. METHODOLOGY

A. CoGiTo: Objectives and Design

COGITO is an immersive clinical simulation platform designed to train speech-language pathology students in the conduct of an anamnesis with a virtual patient exhibiting mild memory or language disorders, see Fig.1.

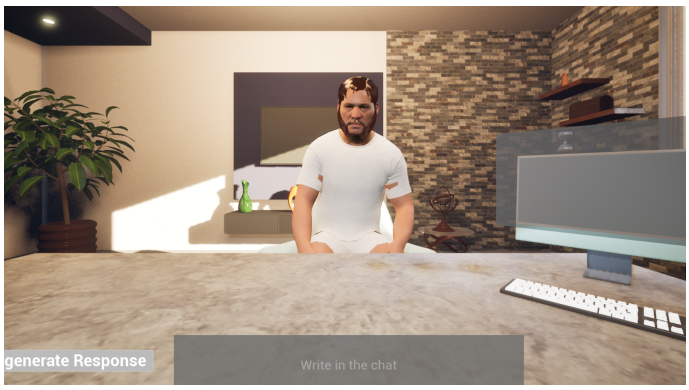


Fig. 1: The virtual environment with the avatar patient

Trainees embody a clinician and can select different patient profiles, cognitive disorders, and new or follow-up patient encounters, see Fig. 2.

The simulation targets two cognitive profiles: Amnestic mild cognitive impairment (aMCI) and logopenic Primary Progressive Aphasia (lPPA) - a syndrome characterized by word-finding difficulties. These disorders were selected for the richness of their verbal manifestations during clinical interviews.

979-8-3315-9476-3/26/$31.00 ©2026 IEEE

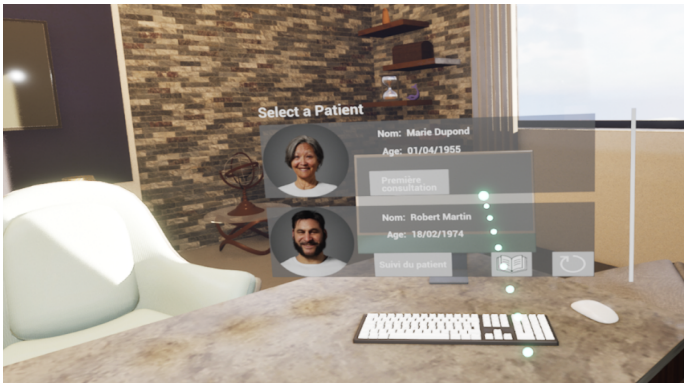


Fig. 2: User interface selecting a new or follow-up patient

COGITO aims to assess two pedagogical dimensions of clinical competence: (i) relational and communicative skills (e.g., empathy, adapted language, patient interaction); and (ii) clinical performances including the logical ordering and relevance of the anamnesis. The system does not evaluate whether the trainee asks a "correct" question, but rather assesses the trainee's overall clinical protocol and ability to adapt communication to the patient's cognitive profile in an open-ended consultation setting.

B. Technical Architecture

COGITO was co-designed by Game engineers and healthcare professionals using Unreal Engine 5, leveraging MetaHuman technology for photorealistic avatar. The system supports two setup configurations: PC and full VR headset (Meta Quest). Interaction is possible both via typed text (PC mode) or voice communication (all modes). Speech recognition is provided by OpenAI's Whisper model, which transcribes the clinician's voice input in real time and forwards the resulting text to the patient LLM pipeline.

C. Patient Simulation Architecture

The patient simulation comprises three core components: the LLM inference backend, the modular patient profile, and the avatar animation layer (facial expression and lipsync).

*1) **LLM Configuration**:* The patient agent is powered by **MistralAI Ministral-3-14B-Reasoning**, quantized at *Q4_K_M* and served locally via LM Studio. This model was chosen thanks to its strong instruction-following and reasoning capabilities for real-time inference. The symptomatic criteria from the DSM-5-TR [10] were incorporated into the patient LLM to replicate the typical verbal characteristics of lPPA and aMCI profiles to ensure clinically accurate results. Our configuration is fully offline, which addresses the privacy constraints inherent of clinical training environments.

*2) **Modular Patient Profile**:* Each virtual patient is defined by a modular patient sheet stored as an XML-text file and injected into the patient LLM's system prompt at initialization. This sheet is structured into semantic categories, enabling the LLM to retrieve information contextually rather than conflating. An example of a patient is shown in section III-B.

This modular schema ensures that the LLM can be prompted consistently across diverse patient profiles. Clinicians and educators can thus add new patients simply by authoring a new text file following this format. The prompt system is dynamically enriched with contextual information (e.g., date and trainee identity), allowing the virtual patient to differentiate first consultations from follow-ups and adapt its responses accordingly.

*3) **Simulating Verbal and Emotional Symptomatology**:* Patient voice synthesis is handled by Piper, an open-source neural text-to-speech (TTS) engine developed by Rhasspy. Piper was selected because it runs entirely offline with low CPU resources, and produces natural-sounding speech with prosody suited to an elderly patient.

The response generated by the LLM model is composed of two distinct elements: (1) the text of the patient's statement, sent to Piper for audio synthesis; and (2) an emotional index, a discrete categorical tag into the model's response that encodes the emotional state associated with the exchange (e.g., anxiety, embarrassment, neutrality, mild confusion). This dual-output format is dictated by the patient system's prompt, which instructs the model to systematically annotate each response with the appropriate emotional label.

The emotion index is used by **NVIDIA Audio2Face** to drive the MetaHuman avatar's facial expressions in real time. Audio2Face receives both the audio stream from Piper and the emotion cue, and generates synchronized lip movements and coherent facial animations (e.g., furrowed brow and downward gaze for anxiety, or hesitant micro-expressions accompanying word-finding pauses). This architecture ensures consistent verbal and non-verbal behavior, all derived from a single LLM generation step.

D. Autonomous Supervisor

After each consultation, a summary is automatically generated and saved, so the trainee can access to their full history across sessions. An autonomous supervisor agent, powered by the same Ministral-3B-14B-Reasoning model as the patient, runs in parallel throughout the session. The supervisor is tasked with evaluating the trainee's clinical and interpersonal competencies along two dimensions based on OSCE evaluation criteria [11].

The first dimension, **communication skills**, assesses whether the trainee: introduced themselves (name and role); used vocabulary and speech rate adapted to the patient's profile; maintained a calm and empathetic attitude; formulated short, clear questions; and explicitly stated the purpose of the consultation, etc.

The second, **anamnesis conduct**, evaluates if the trainee: structured the interview logically, progressing from general to specific; elicited the patient's medical history; explored everyday forgetting episodes and the evolution of difficulties; assessed daily autonomy and domestic independence; noticed word-finding hesitations, circumlocutions, and pauses, etc.

At the end, an evaluation dashboard displays a structured scorecard of all criteria, indicating which skills were demonstrated and which were absent, with a concise feedback report.

III. PRELIMINARY RESULTS

A. Experimental Setup and Participants

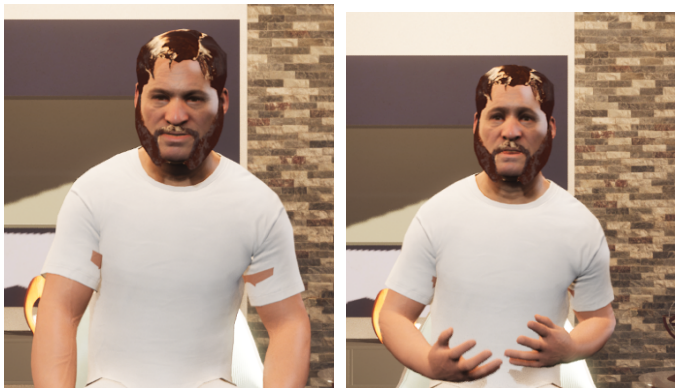
Preliminary trials were carried out with a cohort of 11 master's students of a Speech-Language Pathology program. Participants had limited gaming and VR experience. Sessions were held in an isolated room to provide a controlled and comfortable environment, facilitating natural voice-based interaction with the virtual patient without external distractions. Each session lasted approximately 10-15 minutes, during which the trainee conducted a free-form anamnesis interview. All sessions were run on a single workstation equipped with an NVIDIA GeForce RTX 4090 GPU. Upon completion, participants were invited to fill out a 9-item questionnaire, specifically designed for this study to assess two dimensions: (1) the clinical fidelity of the patient's cognitive and behavioral simulation, and (2) the quality of immersion and acceptability. Responses were collected using a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). The autonomous supervisor module, presented in section II-D, has been implemented within the system; however, it was not evaluated in this preliminary study.

B. Use Case Patient Profile

The experiments were conducted on a single-use case patient. Its profile sheet is implemented as following : ⟨identity⟩ Robert Martin, a 62-year-old male from Nice, France ⟨profession⟩ retired chief financial officer ⟨family⟩ Married to Marie for 30 years, he has two adult children and three grandchildren, and lives with his wife ⟨emotion⟩ slightly anxious, he experiences shame, frustration and hesitation. He shows self-deprecation, occasional irritability, and fear of a serious diagnosis ⟨personality⟩ polite and cooperative, he may ask for repetition, and responds best to short, single-topic questions and appears visibly tired toward the end of the consultation.

For lPPA, the LLM was conditioned via the ⟨language⟩ block to produce moderately frequent hesitation fillers (*"uh…","wait","I had it… no…","how do you say"*), anomia (*"The tool to..."*), phonological approximations (*"pharmachie" for pharmacy*). The ⟨language_memory⟩ sub-tag instructs the model to simulate reduced verbal memory ability and difficulty maintaining complex conversation.

Regarding aMCI, the ⟨general_memory⟩ block was used to condition three distinct sub-categories. ⟨short_term_memory⟩ sub-tag defines in-session working memory behavior. ⟨memory_lapses⟩ specifies episodic memory gaps, such as being unable to recall recent events, names, appointments, etc. ⟨long_term_memory⟩ sub-tag preserves autobiographical memory, allowing Robert to recall old events with emotional richness (*"Our wedding, I remember perfectly — it was in June 1981"*). To avoid clinically implausible outputs, the patient prompt specifies behaviors that must be absent: major disorientation, agrammatism, mutism, or invention of new biographical elements.



(a) embarrassed (b) enthusiastic & engaged

Fig. 3: Emotional reactions of the virtual patient

C. Evaluation

The obtained results are depicted in Table I. From preliminary tests, the simulation ran in real time in VR in a fully local offline configuration, despite the integration of MetaHuman, the audio modules, and the dual instantiation of the LLM pipeline (patient and supervisor agents). The virtual patient responded with a latency under 2 seconds, ensuring a dynamic and fluid interaction.

The mean scores for Q1.1 and Q1.2 ($M_{1.1}$=4.45, $M_{1.2}$=4.27) indicate that the structured field and sub-field format of the patient profile effectively ensured narrative consistency, producing personality outputs aligned with the patient sheet which avoids LLM hallucinations and biographical contradictions. Respondents further found that emotional reactions were relevant to the conversational context ($M_{1.3}$=4.27) and coherent with verbal responses ($M_{1.4}$=4.36), which means that our dual-output format (combining surface text with an explicit emotion index) fostered a consistent association between affective context, verbal output, and emotions.Figure 3 illustrates the patient's facial and gestural emotional reactions during interactions with the clinician. However, the intelligibility of the synthesized French voice ($M_{1.5}$ = 3.36 $\pm$ 1.07) and the naturalness of the patient–clinician interaction ($M_{2.1}$ = 3.36 $\pm$ 0.48) indicate that further refinement of the selected Piper voice model is warranted.

Regarding immersion quality, over 80% of respondents rated the visual realism of the avatar and virtual environment positively ($M_{2.2}$ = 3.81), while only 27.3% considered the facial expressions realistic relative to the patient's emotional state ($M_{2.3}$ = 3.27 $\pm$ 0.45). This discrepancy can be explained by the inherent exaggeration in NVIDIA Audio2Face, which overamplified facial animations and lacked smooth transitions between emotional states. Finally, over 90% of trainees reported full engagement and willingness to repeat the experience ($M_{2.4}$ = 4.90 $\pm$ 0.29), indicating strong acceptance

TABLE I: Evaluation Results (%): Patient Cognition Fidelity and Immersion Quality (5-point Likert Scale, $N = 11$)

Question		1	2	3	4	5	Mean ± SD
1 — Patient Cognitive Fidelity							
Q1.1	Narrative consistency with patient sheet (age, profession, family, daily life)	-	-	-	54,5	45,5	4,45 ± 0,49
Q1.2	Fidelity of personality to patient profile (calm, cooperative, low expressiveness)	-	-	18,2	36,3	45,5	4,27 ± 0,75
Q1.3	Relevance of emotional reactions to conversational context	-	-	-	72,7	27,3	4,27 ± 0,45
Q1.4	Consistency between emotions and verbal responses	-	-	18,2	27,3	54,6	4,36 ± 0,77
Q1.5	Intelligibility of the patient's synthesized voice	-	27,3	27,3	27,3	18,2	3,36 ± 1,07
2 — Immersion Quality & Acceptability							
Q2.1	Naturalness of patient–clinician interaction (speech & gestures)	-	-	63,6	36,4	-	3,36 ± 0,48
Q2.2	Visual realism of the patient avatar	-	-	18,2	81,8	-	3,81 ± 0,39
Q2.3	Realism of facial expressions relative to emotional state	-	-	72,7	27,3	-	3,27 ± 0,45
Q2.4	Trainee engagement and involvement during the simulation	-	-	-	9,10	90,9	4,90 ± 0,29

of this user-friendly tool among a clinical population, despite their lack of prior VR experience.

From qualitative observations, the patient has performed relatively better when the trainee used slow, short, single-topic questions, as parameterized in the ⟨language_memory⟩ sub-field. Furthermore, within a single session, the patient reliably simulated short-term memory deficits as conditioned in the ⟨short_term_memory⟩ category, producing outputs such as (*"You told me your name, didn't you? I'm sorry, I've forgotten"* or *"I couldn't tell you what I ate at breakfast"*) and repeating the same anecdote without awareness as specified in the ⟨forgetting_patterns⟩ sub-field. This demonstrates that the LLM can be guided through a simple modular configuration to generate customized patient profiles with specific cognitive, personality, and behavioral characteristics.

IV. Conclusion & Future perspective

This paper presented COGITO, a LLM-driven virtual patient simulation for clinical training in speech-language pathology, combining MetaHuman avatar rendering and local LLM inference. The system showed the ability of a locally deployed language model to simulate verbal and non-verbal symptoms. Preliminary results highlighted the pedagogical value of this tool, with particularly strong trainee engagement. COGITO thus offers a more natural and dynamic alternative to traditional scripted simulations for practicing communication and anamnesis skills.

Future work will focus on extending the evaluation to a larger cohort including both healthcare professionals. Additionally, emotional realism will be improved by refining facial animation smoothness and embedding affective prosody directly into the TTS pipeline through modulation of pitch, tone, and speech rate in accordance with the patient's emotions.

Acknowledgment

This work was supported by the Faculty of Medicine of Université Côte d'Azur and the NeuroMod Institute. Regarding the use of AI tools, Claude was employed exclusively during the preparation of the Introduction and Methods sections to improve readability and overall language quality. No AI system was used to generate results, perform analyses, or interpret the study findings.

References

[1] P. Chen, H. Cai, W. Bai, et al. Global prevalence of mild cognitive impairment among older adults living in nursing homes: a meta-analysis and systematic review of epidemiological surveys. *Translational Psychiatry*, 13(1):88, 2023.

[2] A. E. J. van Gaalen, J. Brouwer, J. Schönrock-Adema, et al. Gamification of health professions education: a systematic review. *Advances in Health Sciences Education*, 26:683–711, 2021.

[3] D. A. Cook. Creating virtual patients using large language models: scalable, global, and low cost. *Medical Teacher*, 47(1):40–42, 2025.

[4] D. A. Cook, J. Overgaard, V. S. Pankratz, G. Del Fiol, and C. A. Aakre. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486, April 2025.

[5] Akilan Amithasagaran, Sagnik Dakshit, Bhavani Suryadevara, and Lindsey Stockton. Clivr: Conversational learning system in virtual reality with ai-powered patients, 2025.

[6] Xiuqi Tommy Zhu, Heidi Cheerman, Minxin Cheng, Sheri Kiami, Leanne Chukoskie, and Eileen McGivney. Designing vr simulation system for clinical communication training with llms-based embodied conversational agents, 2025.

[7] Dongliang Li and Syaheerah Lebai Lutfi. Large language model–based virtual patient systems for history-taking in medical education: Comprehensive systematic review. *JMIR Medical Informatics*, 14:e79039, 2026.

[8] Young-Woo Jo, Myungeun Lee, and Hyung-Jeong Yang. Large language model-based virtual patient simulations in medical and nursing education: A review. *Applied Sciences*, 15(22):11917, 2025.

[9] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax?, 2025.

[10] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, 5th edition, 2022.

[11] M. Zayyan. Objective structured clinical examination: the assessment of choice. *Oman Medical Journal*, 26(4):219–222, July 2011.