ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251428

VALISE: A Virtual Agent Laboratory for Instruction-Following Simulation and Evaluation of LLM-Powered Digital Health Interventions

Marco Bolpagni [a,b,*], **Simone De Carli** [b], **Leonardo Sanna** [b], **Mauro Dragoni** [b] **and Silvia Gabrielli** [b]

[a]University of Padova
[b]Fondazione Bruno Kessler

Abstract. Digital health interventions often require structured, protocol-driven dialogues delivered with high fidelity. Evaluating whether an agent employing a Large Language Model (LLM) can meet these requirements remains challenging, especially in early development stages. In this work, we present VALISE (Virtual Agent Laboratory for Instruction-Following Simulation and Evaluation), a modular framework for simulating and evaluating LLM agent behavior in delivering structured health interventions. VALISE enables configurable agent–patient simulations using synthetic personas and evaluates protocol adherence through a customizable, automated grid assessed by ensembles of LLM-based judges. We demonstrate its use with Brief Action Planning (BAP), a short intervention promoting behavior change in sedentary individuals. Our results strongly align LLM-based and expert annotations, supporting VALISE's effectiveness for early-stage evaluations. VALISE offers a reproducible, extensible platform for testing instruction-following capabilities of LLM agents in digital health.

1 Introduction

The delivery of digital health interventions often relies on structured, protocol-driven processes designed to promote behavior change, improve self-management, or support clinical decision-making [2, 17]. These interventions are typically guided by well-defined sequences of interactional steps, such as those specified in cognitive behavioral techniques or behavior planning protocols. Ensuring that digital agents can accurately and consistently follow such protocols is essential for maintaining the effectiveness and safety of these interventions [17]. Recent advances in LLMs have enabled the development of conversational agents capable of generating coherent, context-sensitive dialogue [13]. As these models are increasingly adopted in health-related applications, the need to evaluate their capacity to adhere to predefined intervention protocols has become more prominent. Unlike open-domain dialogue, structured intervention delivery imposes specific requirements on conversational turn sequences, content, and logic. A deviation from these requirements may compromise the intended therapeutic or behavioral outcomes. Evaluating LLM agents in this context presents several methodological challenges. Traditional evaluation approaches often depend on expert annotation of dialogue transcripts [21], which is time-consuming, costly, and challenging to scale. Alternative strategies involve human-in-the-loop simulations, which introduce variability and limit reproducibility [12]. Moreover, robust evaluation demands testing agent behavior across a broad range of user profiles [9] that differ in psychological traits, interactional styles, and behavioral readiness, factors that significantly influence dialogue dynamics [14, 10]. To date, there is a lack of standardized, reproducible frameworks to simulate such variability and to assess instruction-following behavior in LLM agents operating within structured dialogue protocols. The ability to perform early-stage, large-scale, protocol-sensitive evaluation is a key requirement for designing and deploying trustworthy digital health agents in the ORBIT framework [6]. This work presents VALISE (Virtual Agent Laboratory for Instruction-Following Simulation and Evaluation), a modular system designed to simulate structured conversations between LLM agents delivering interventions and synthetic personas, and to evaluate the agents' compliance with protocol-specific expectations automatically. VALISE aims to support pre-deployment development and validation of conversational agents for digital health.

2 Related work

This section briefly reviews prior research that informed the development of our system, focusing on synthetic persona generation, simulations of agent–patient interactions in digital health, and automatic evaluation of dialogues using LLMs. The use of synthetic personas in dialogue systems draws inspiration from the practice of standardized patients in medical education. A standardized patient [1] is a trained individual who consistently enacts a predefined clinical case, enabling structured and replicable assessment of clinical and communication skills. This concept has been extended to virtual standardized patients [11], computer-based agents that simulate patient behavior for training and evaluation purposes. Similar approaches have been employed in psychology to support therapeutic skill development through controlled dialogue interactions [18]. Recent advances in LLMs have significantly expanded the potential of such simulations. LLMs enable the generation of dynamic, psychologically diverse personas[1] that vary across dimensions such as emotional reactivity[2], and behavioral readiness [3]. These capabilities support more realistic and scalable multi-turn interaction simulations, moving beyond the limitations of manually scripted sys-

* Corresponding Author. Email: mbolpagni@fbk.eu.

[1] https://arxiv.org/abs/2410.19238
[2] https://arxiv.org/pdf/2310.05418

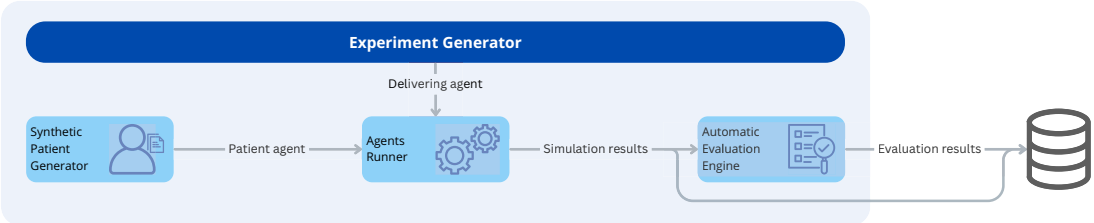

Figure 1. VALISE's System Architecture.

tems. While early applications of LLM-generated personas have primarily appeared in unrelated domains, such as the gaming industry, exemplified by Tencent's Persona Hub[3] for creating LLM-powered characters, there is growing interest in their adoption within verticals such as digital health [20]. In addition to persona generation, LLMs are increasingly adopted as evaluators in a broad range of NLP tasks [23, 4]. Benchmarks on MT-Bench [23] and Topical-Chat[4] demonstrate that models like GPT-4 can approximate with high reliability expert judgments in evaluations related to coherence, instruction adherence, and task completion [23, 5]. Such evaluation pipelines often combine ensemble methods, rubric-based criteria, and prompt-engineered scoring strategies [5]. Although prior work has addressed virtual patients, LLM-based personas, and automated dialogue evaluation, these components are rarely integrated. VALISE combines them into a unified, extensible platform for simulating and evaluating instruction-following agents in digital health.

3 System Architecture

VALISE is built around a modular architecture that supports the end-to-end simulation, execution, and evaluation of instruction-following interactions between LLM agents and synthetic patients. The system is composed of four primary modules: the *Synthetic Patient Generator*, the *Agents Runner*, the *Automatic Evaluation Engine*, and the *Experiment Generator*. Each module is designed for independent configuration but is tightly integrated into a unified experimental pipeline managed by the *Experiment Generator* [5].

3.1 Synthetic Patient Generator

This module enables the generation of diverse synthetic personas to simulate varying user profiles. Each persona is characterized by configurable attributes spanning demographic information (e.g., age, gender, cultural context), psychological traits (e.g., openness to change, affect, resistance), interactional style (e.g., verbosity, cooperativeness), and behavioral domain (e.g., exercise, diet). The system supports both template-based instantiation and stochastic sampling strategies, enabling the creation of realistic, diverse agents that help test the robustness and generalizability of the LLM-powered agent delivering the intervention in structured interactions.

3.2 Agents Runner

The Agents Runner orchestrates the multi-turn interactions between a synthetic patient and a conversational agent operating under a specified intervention protocol. Given the agent prompt (containing the protocol steps), patient prompt, and a specific LLM backend, the module simulates the dialogue in real time, logging each turn for later evaluation. The runner supports flexible parameterization, such as temperature control for each agent. Optional plugins extend functionality, for example by integrating auxiliary agents such as a supervisor or linguistic refinement agents.

3.3 Automatic Evaluation Engine

VALISE includes an automated evaluation module based on an ensemble of LLM judges to assess whether an agent has followed the prescribed protocol. Each judge independently evaluates the dialogue using a predefined protocol grid (modeled as a Pydantic schema), returning judgments per protocol step. Each judge is initialized with a different random seed, which naturally introduces variation in their responses due to the stochastic nature of LLMs. VALISE also supports heterogeneous ensembles, where a different LLM may back each judge. Judges receive both the dialogue transcript and the evaluation grid. Through role-play prompting [19], each LLM assumes the persona of a domain expert (e.g., a clinical supervisor) and systematically completes the grid, mimicking human evaluators in structured assessments. By default, these evaluations are eventually aggregated through majority voting, although the architecture allows for alternative aggregation schemes, such as weighted voting and micro-averaging. Advanced pre-processing (e.g., translation) via configurable LLMs, deterministic tag-matching (keyword spotting), or post-processing tools like summarization of dialogue chunks for downstream evaluation are provided by plugins.

3.4 Experiment Generator

This module automates the creation of experimental setups for large-scale, reproducible studies (see Figure 1). Given user-defined lists of models, temperatures, and synthetic personas, the module computes the Cartesian product of these factors and queues the resulting configurations for simulation. In addition to this, the Experiment Generator supports configuration-driven workflows: a user can specify, via a configuration file, which modules to activate (e.g., running simulations without evaluation), define settings for each module (e.g., which prompts or models to use, temperature values), and include any desired plugins to extend core functionality. It also allows customization of high-level experiment hyperparameters such as the LLM provider[6] and the number of parallel executions. This setup facilitates systematic exploration of agent behavior under varied conditions and supports empirical analysis of factors such as agent sensitivity to temperature or persona traits. The generator provides a foundation for scalable simulations, which are particularly valuable in the

[3] https://arxiv.org/abs/2406.20094
[4] https://arxiv.org/abs/2308.11995
[5] A demo (video) is available here: https://osf.io/7nqpu

[6] VALISE supports any OpenAI-compatible API and local LLMs via Ollama server

pre-piloting stages of the ORBIT framework for developing digital health interventions.

4 Experimental setup

We conducted an experiment using the Brief Action Planning (BAP) [8] protocol as a testbed to evaluate how VALISE facilitates dialogue generation and systematic assessment of instruction-following agents in structured health interactions. The protocol was first translated into a structured system prompt[7] using role-play prompting [19] and assigned to the agent delivering the intervention. Next, we created three distinct patient personas using VALISE's Synthetic Persona Generator. Each persona corresponded to a different stage of the Transtheoretical Model of Change [16]: the *Collaborative patient*, representing the preparation stage and showing motivation to change; the *Confused patient,* in the contemplation stage and seeking guidance; and the *Uncollaborative patient,* aligned with the precontemplation stage and showing resistance. Eventually, we assigned random attributes to each persona, such as gender, age, and occupation, to introduce further variability. In collaboration with domain experts, we then defined an evaluation grid covering the compulsory actions required by the BAP protocol[8]. This grid served both as an annotation rubric for human experts and as a schema for configuring the Automatic Evaluation Engine. For generating conversations we selected Llama 3.3-70B[9] as backbone.

Using VALISE's Experiment Generator, we varied the temperature across five values $(0.1, 0.3, 0.5, 0.7, 0.9)$, and simulated 10 dialogues per temperature–persona combination, resulting in 150 dialogue sessions[10]. To evaluate the adherence to the BAP protocol of the agent delivering the intervention, each generated conversation was scored by VALISE's Automatic Evaluation Engine module backed by Qwen 2.5–72B[11] with $(1, 3, 5)$ judges. We established reference labels by randomly assigning the generated conversations in equal proportions to three digital health experts. Each expert manually annotated their assigned simulations using the evaluation grid, marking for each step whether it was performed correctly. To assess the performance of VALISE's Automatic Evaluation Engine, we treated the task as a multi-label classification problem, where each dialogue includes multiple (binary) labels corresponding to the defined steps in the grid. We quantified the annotation quality of the engine using the Hamming loss [22], computed between the engine's outputs and expert labels. Finally, we investigated whether varying the number of judges significantly impacted VALISE's annotation quality. Since the same set of dialogues was evaluated repeatedly across different ensemble sizes, violating the independence assumption, we applied the Friedman test [7], a non-parametric method appropriate for repeated-measures designs. Pairwise comparisons between ensemble sizes were conducted using Durbin-Conover post hoc tests with Holm-adjusted p-values, implemented in the ggstatsplot library [15].

5 Experimental Results

The evaluation of VALISE's Automatic Evaluation Engine revealed promising performance, showing that expert assessment on protocol adherence is often aligned with LLM judges. Across the 150 dialogue sessions, the system achieved a median Hamming loss of 0.13, a promising result that indicates high agreement with expert labels in this multi-label classification setting. This suggests that the engine effectively captures whether key steps in structured health interventions, such as those defined by the BAP protocol, are correctly executed. There were no statistically significant differences in performance across ensemble sizes of 1, 3, and 5 LLM judges. The Friedman test ($\chi^2_{Friedman}(2) = 4.94$, $p = 0.08$) and post hoc Durbin-Conover pairwise comparisons (all $p > 0.05$) indicate that even a single LLM judge can yield annotations of similar quality to those produced by larger ensembles, offering a cost-effective configuration for scalable evaluation. Nonetheless, as shown in Figure 2, a small subset of conversations benefited from larger ensembles. Specifically, the 3 and 5-judge configurations show fewer dialogues with Hamming losses in the range from 0.4 to 0.7 compared to the single-judge setup. This suggests that while the median performance remains stable, ensemble size may improve robustness when evaluating complex or ambiguous interactions. In addition to the quantitative results, the three digital health experts who annotated the dialogues voluntarily shared qualitative impressions of the generated conversations. They described the interactions between synthetic patients and the agent delivering the intervention as realistic, contextually coherent, and reflective of a meaningful range of patient behaviors and emotional tones. Notably, they highlighted that the variability introduced through persona traits, such as cooperativeness, decisiveness, and readiness to change, significantly contributed to the authenticity of the simulations and their relevance for early-stage intervention design and refinement.

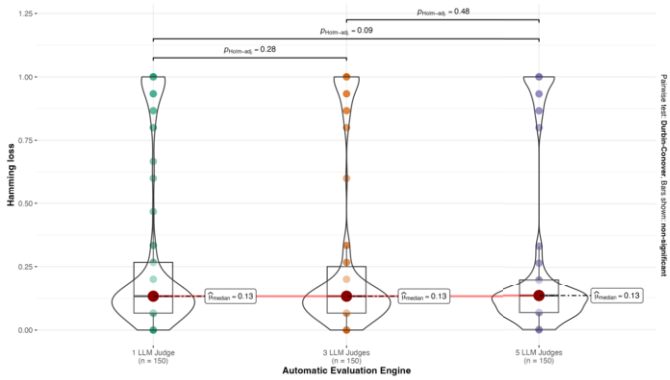

Figure 2. Hamming loss across different LLM ensemble sizes in protocol adherence evaluation performed by VALISE's Automatic Evaluation Engine.

6 Conclusions

This paper presented VALISE, a modular framework for simulating and evaluating instruction-following LLM agents in structured digital health interventions. Our experiments on Brief Action Planning showed that VALISE can generate clinically meaningful dialogues and automate evaluation with high fidelity to expert annotations. Notably, increasing the number of LLM judges improved robustness in edge cases, though single-judge setups often sufficed. Expert feedback further highlighted the realism and behavioral nuance of simulated interactions. These results demonstrate that VALISE supports scalable, reproducible, and protocol-sensitive agent evaluation, which is key for early-stage development under the ORBIT model.

[7] BAP prompt is available here: https://osf.io/8hdcn
[8] Evaluation grid is available here: https://osf.io/6gx3f
[9] Selected due to stable performance in preliminary structured dialogue generation tasks.
[10] A sample simulation is available here: https://osf.io/nfs6j
[11] Selected for its consistent behavior in initial evaluation trials on structured annotation tasks.

References

[1] H. S. Barrows. An overview of the uses of standardized patients for teaching and evaluating clinical skills. aamc. *Academic medicine*, 68 (6):443–51, 1993.

[2] A. Blandford, J. Gibbs, N. Newhouse, O. Perski, A. Singh, and E. Murray. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digital health*, 4:2055207618770325, 2018.

[3] M. Bolpagni, S. De Carli, L. Sanna, S. Gabrielli, and M. Dragoni. Role-play large language models for short behavior change interventions: An exploratory study on brief action planning. In *International Conference on Artificial Intelligence in Medicine*, pages 46–51. Springer, 2025.

[4] M. Boyapati, L. Meesala, R. Aygun, B. Franks, H. Choi, S. Riordan, and G. Modgil. Leveleval: Adaptive pipeline for evaluating llm as a judge-analysis on open llms as judges. In *2024 International Conference on AI x Data and Knowledge Engineering (AIxDKE)*, pages 74–77. IEEE, 2024.

[5] C.-H. Chiang and H.-y. Lee. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, 2023.

[6] S. M. Czajkowski, L. H. Powell, N. Adler, S. Naar-King, K. D. Reynolds, C. M. Hunter, B. Laraia, D. H. Olster, F. M. Perna, J. C. Peterson, et al. From ideas to efficacy: The orbit model for developing behavioral treatments for chronic diseases. *Health Psychology*, 34(10): 971, 2015.

[7] W. W. Daniel. *Applied Nonparametric Statistics*. PWS-Kent, Boston, 2nd edition, 1990. ISBN 978-0-534-91976-4. Chapter: "Friedman two-way analysis of variance by ranks", pp. 262–274.

[8] D. Gutnick, K. Reims, C. Davis, H. Gainforth, M. Jay, and S. Cole. Brief action planning to facilitate behavior change and support patient self-management. *JCOM*, 21(1):17–29, 2014.

[9] A. Hamilton, A. Molzahn, and K. McLemore. The evolution from standardized to virtual patients in medical education. *Cureus*, 16(10), 2024.

[10] D. T. Holt, C. D. Helfrich, C. G. Hall, and B. J. Weiner. Are you ready? how health professionals can comprehensively conceptualize readiness for change. *Journal of general internal medicine*, 25:50–55, 2010.

[11] R. C. Hubal, P. N. Kizakevich, C. I. Guinn, K. D. Merino, and S. L. West. The virtual standardized patient-simulated patient-practitioner dialog for patient interview training. In *Medicine Meets Virtual Reality 2000*, pages 133–138. IOS Press, 2000.

[12] A. B. Kocaballi, E. Sezgin, L. Clark, J. M. Carroll, Y. Huang, J. Huh-Yoo, J. Kim, R. Kocielnik, Y.-C. Lee, L. Mamykina, et al. Design and evaluation challenges of conversational agents in health care and well-being: selective review study. *Journal of medical Internet research*, 24 (11):e38525, 2022.

[13] M. K. Mahto, D. Srivastava, R. Kumar, B. Sah, H. R. Singh, and S. K. Maakar. Personalized user interaction in web applications using adaptive llm model. In *2025 International Conference on Pervasive Computational Technologies (ICPCT)*, pages 962–966. IEEE, 2025.

[14] F. Moors and E. Zech. The effects of psychotherapist's and clients' interpersonal behaviors during a first simulated session: A lab study investigating client satisfaction. *Frontiers in psychology*, 8:1868, 2017.

[15] I. Patil. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, 6(61):3167, 2021. doi: 10.21105/joss.03167. URL https://doi.org/10.21105/joss.03167.

[16] J. O. Prochaska and W. F. Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48, 1997.

[17] G. Recchia, D. M. Capuano, N. Mistri, and R. Verna. Digital therapeutics-what they are, what they will be. *Acta Sci Med Sci*, 4(3): 1–9, 2020.

[18] G. M. Reger, A. M. Norr, M. A. Gramlich, and J. M. Buchman. Virtual standardized patients for mental health education. *Current psychiatry reports*, 23:1–7, 2021.

[19] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[20] U. Smrke, A. Rehberger, N. Plohl, and I. Mlakar. Exploring the feasibility of generative ai in persona research: A comparative analysis of large language model-generated and human-crafted personas in obesity research. *Applied Sciences*, 15(4):1937, 2025.

[21] L. Sun, Y. Liu, G. Joseph, Z. Yu, H. Zhu, and S. P. Dow. Comparing experts and novices for ai data work: Insights on allocating human intelligence to design a conversational agent. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 195–206, 2022.

[22] G. Wu and J. Zhu. Multi-label classification: do hamming loss and subset accuracy really conflict with each other? *Advances in Neural Information Processing Systems*, 33:3130–3140, 2020.

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.