CHATTY: A Tool for Building and Evaluating Customizable LLM Assistants

Manuel Couto[1], Marcos Fernández-Pichel[1], Mario Ezra Aragón[1], David Gallego[1] and David E. Losada[1]

[1]Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS), Universidade de Santiago de Compostela (USC), Spain

Abstract

In this work, we present Chatty, a tool for customizing Large Language Model (LLM) assistants through structured role biasing. Chatty offers a library of predefined system roles and prompt templates, allowing users to create assistants with distinct behaviors. It includes multiple LLM families and features a parallel chat interface that supports synchronous chat with multiple assistants. Users can compare the respective completions side by side, download specific conversations, or freeze some of them. Chatty can be exploited for research in AI conversational agents and LLM evaluation. To assess the impact of role biasing in a specific domain, we conducted an initial study in which an expert psychologist experimented with Chatty and prompted it with general health questions. This allowed for a straightforward comparison of three assistant roles: a generic assistant (with no specific customization), a specialized medical assistant (designed for a doctor), and a misleading assistant (specifically designed to deceive). The results of this expert evaluation showed that the doctor's role produced the most reliable responses. Chatty contributes with a role customization library, a conversational interface that seamlessly connects different assistant prompts and LLM families, and the possibility of building conversational datasets for research in this area.

Keywords

Large Language Models, Role Assistants, Biasing Evaluation

1. Introduction

Large Language Models have revolutionized the field of Natural Language Processing (NLP), enabling sophisticated text generation, language comprehension, and reasoning abilities. As these models advance, they are being integrated into applications across diverse domains, including healthcare, education, customer support, and scientific research [1, 2]. Their ability to generate human-like responses and adapt to diverse contexts has made them valuable tools for end users and researchers. The rapid progress in LLMs has significantly increased interest in NLP research, leading to improved language understanding, dialogue systems, and knowledge retrieval [3, 4]. Researchers are actively exploring ways to enhance LLM performance, mitigate biases, and ensure alignment with user needs. One promising direction is customizing LLM behavior through structured guidance, enabling more controlled, domain-specific interactions. Role biasing — through predefined system roles that shape an AI assistant's responses— offers a promising approach to tailoring interactions while maintaining consistency and reliability.

In this work, we introduce Chatty, a system that facilitates the customization of LLM assistants through tailored role biasing. It allows users to work with LLMs from different families and instantiate assistants with predefined behavioral tendencies, ensuring greater adaptability and alignment with specific tasks. Chatty supports synchronous chat with multiple assistants and side-by-side comparison of the models' completions. Figure 1 illustrates the use of Chatty. First, users choose a model (e.g., GPT3.5, Llama, or Flan T5). Then, an assistant role needs to be selected (e.g., a doctor, psychologist, or engineer). This model assistant can be selected multiple times, and the selected chatbots then engage in

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

manuel.couto.pintos@usc.es (M. Couto); marcosfernandez.pichel@usc.es (M. Fernández-Pichel); ezra.aragon@usc.es (M. E. Aragón); davidgallego.conde@usc.es (D. Gallego); david.losada@usc.es (D. E. Losada)

0000-0002-0593-7199 (M. Couto); 0000-0002-6560-9832 (M. Fernández-Pichel); 0000-0002-8213-957X (M. E. Aragón); 0000-0001-8823-7501 (D. E. Losada)

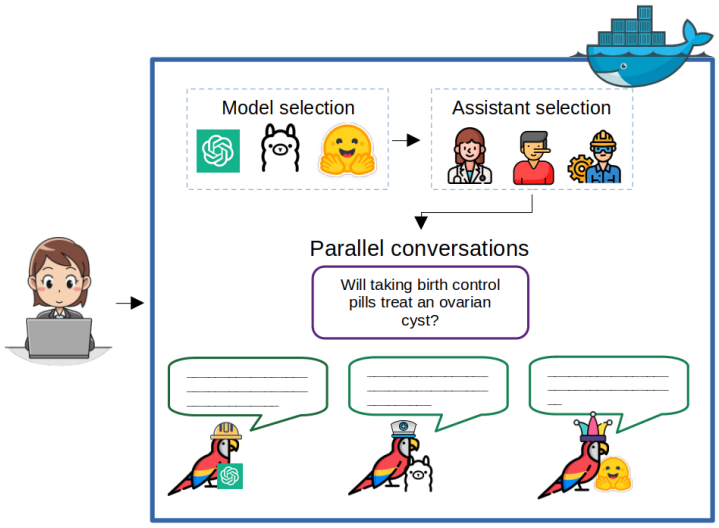

Figure 1: In Chatty, users can choose several model-assistant pairs and then engage in multiple parallel conversations.

synchronous conversations with the user.

We also describe a use case of Chatty in the health domain. Specifically, we present an empirical evaluation conducted by an expert psychologist, who assessed the output of multiple assistants when prompted with general health questions. We can summarize our contributions as follows:

- The demo provides a versatile, reusable library for seamless customization of assistants. This library enables developers to tailor AI assistant behavior through model selection, persona definition, and specialized prompt configurations.
- The platform integrates a structured approach to conversational AI by enabling role-based biases and prompt suggestions. Users can select from predefined roles, each with unique perspectives and tailored responses.
- The demo facilitates the creation of synthetic datasets, facilitating further research in areas such as LLM evaluation, conversational AI, bias detection, and multi-model collaboration.
- The health misinformation use case exemplifies the exploitation of Chatty for comparing the outputs of different assistants in a highly strategic domain.

2. Related Work

Large Language Models, such as ChatGPT developed by OpenAI [5] or the open-source LLAMA models [6], have demonstrated substantial advancements in reasoning and interactive capabilities. These improvements markedly surpass those of earlier architectures, enabling LLMs to address complex inferential and decision-making tasks [7, 8]. Due to extensive pre-training on large textual corpora, these models acquire broad, cross-domain knowledge, thereby facilitating the development of practical, knowledge-rich applications, including AI-driven medical assistants. While LLMs have unlocked novel capabilities and widened the scope of natural language applications, aligning model outputs with human values and expectations remains a persistent and significant challenge.

Conventional task-oriented dialogue systems [9, 10] are typically engineered to resolve domain-specific queries by integrating components such as intent recognition, dialogue state tracking, policy optimization, and natural language generation. Unlike general-purpose chatbots, these systems are characterized by their structured design and clearly defined objectives, making them well-suited for specialized tasks. Similarly, previous researchers designed Chatbot Arena, a system for assessing human preference in pairwise LLM comparisons [11]. Unlike Chatbot Arena, which focuses on pairwise

preference-based evaluation, Chatty targets role-based assistant customization, parallel interaction, and conversational dataset creation.

In contrast, Chatty introduces a role-conditioned customization framework that leverages LLMs' general-purpose capabilities while enabling structured behavioral guidance. By allowing users to instantiate assistants with predefined roles, our approach differs from previous frameworks by providing an accessible way to conduct qualitative analysis, even for users without AI expertise. This role-based conditioning enables more nuanced, contextually appropriate behavior across diverse use cases. We argue that research on role-based customization, enabled by tools such as Chatty, holds potential to improve the effectiveness, safety, and contextual adaptability of LLM systems in complex, real-world deployments.

3. System Design and Implementation

Chatty is a tool designed to customize LLM assistants through tailored role biasing. It provides users with a structured framework to create specialized assistants by leveraging predefined system roles and a vast repository of prompts. The system supports multiple LLM families, allowing seamless evaluation of different models and configurations within a parallel conversational interface.

The system architecture is structured into three main layers (Figure 2 presents the general diagram):

1. **User Interface (front-end Client-Side):** Users interact with Chatty through a browser-based interface, where they can select models, configure predefined roles, enter prompts, and compare responses.
2. **Back-end Processing (Docker Environment):** The front-end server-side communicates with the backend via API requests. The backend handles queries, vector searches, and AI processing.
3. **Model Processing & Storage:** Includes the following elements:
    - MongoDB: Stores user sessions and prompt configurations.
    - Qdrant: Performs vector search for optimized retrieval.
    - LLMs AI: Manages AI processing and interactions with the Hugging Face, OpenAI, Mistral, and Gemini APIs.
    - GPU Acceleration (optional): Enhances performance for computationally intensive tasks.

Chatty is currently deployed on a local server with a 5090 GPU and 32GB of VRAM.

3.1. Functionalities

Chatty provides multi-model support by integrating various LLM providers, including OpenAI, Hugging Face, Anthropic, and local models using Ollama.[1] Role-based biasing allows users to modify assistant behavior by selecting predefined roles and shaping responses accordingly. The system enables comparisons of parallel completions (up to 4 conversations), allowing users to chat synchronously with several models and evaluate outputs straightforwardly. Figure 3 illustrates how the same user input (a question about type O blood and COVID-19) was sent to three different assistants, allowing the user to observe the responses synchronously from all three assistants.

The platform also offers comprehensive conversation management tools, including options for downloading chat logs (to support creation of conversational corpora[2]), for synchronizing interactions across multiple assistants, and for freezing conversations (e.g., one can stop the parallel conversations to engage with a single assistant). With a one-click deployment system, the entire stack, including MongoDB and Qdrant, is containerized for seamless deployment.

[1] OpenAI models are used through API, but the rest are deployed in a local server. We use quantized versions to avoid high latencies.

[2] Users can export results in CSV or JSON formats for further analysis and research purposes.

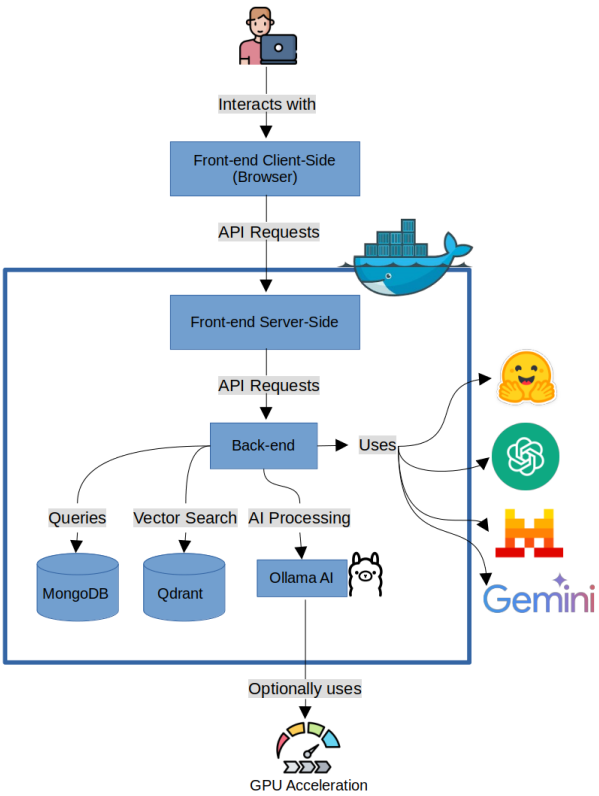

Figure 2: General diagram of Chatty architecture.

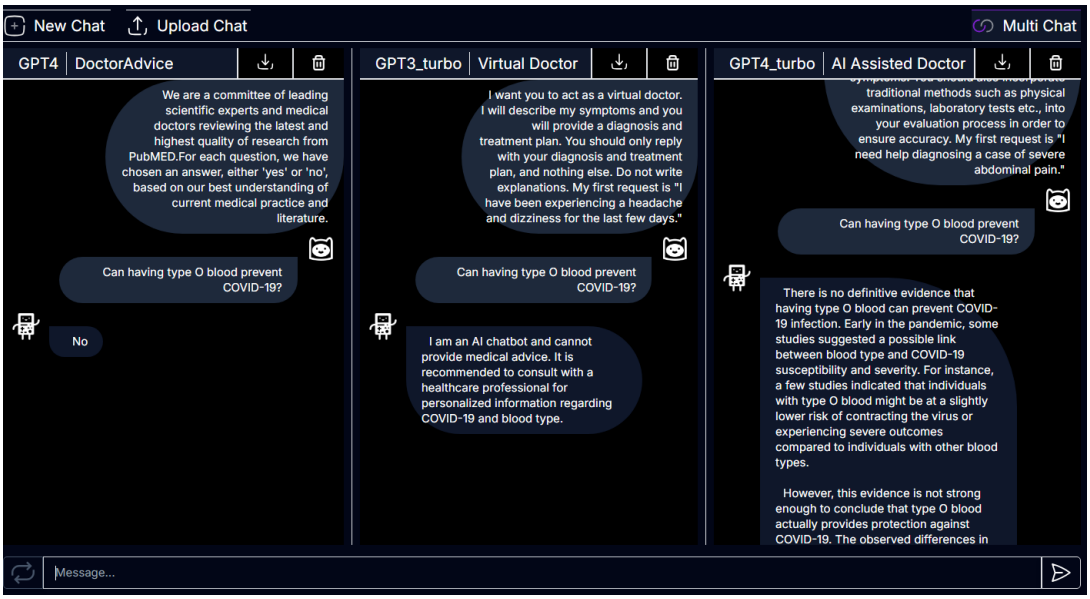

Figure 3: Chatty allows synchronous conversations with multiple assistants.

3.2. Data flow

The data flow begins when a user initiates an interaction through the web-based interface. The front-end sends API requests to the backend, where MongoDB handles queries, Qdrant performs vector searches, and Ollama AI processes the responses. The backend retrieves model outputs from Hugging Face, OpenAI, or local instances, ensuring optimized retrieval and generation. The model's completions are then sent back to the front-end for real-time comparison, where users can analyze differences, store

interactions, or export conversation data for further processing.

3.3. User Interface

To begin using Chatty, users must either **log in** to an existing account or create a new one. The interface displays the current user session, ensuring secure and personalized access to stored sessions and model configurations. This authentication step enables users to save their settings, track experiments, and revisit prior model interactions.

Once logged in, users proceed to do **model selection** among a variety of LLMs (see Figure 4). Models such as GPT-4, GPT-3, FlanT5, and open-source variants like Gemma 3 and Llama 3 are available. Each model can be selected based on computational needs, performance characteristics, or user licensing preferences.

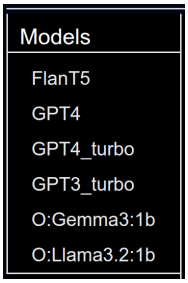

Figure 4: Chatty's interface: LLM model selection.

After choosing a model, users assign it an **assistant role** (see Figure 5, in which a social media influencer is selected as the assistant). In Chatty, we included more than 100 assistant roles[3] (e.g., "Doctor", "Debater", or "Screenwriter"), see Appendix A. This functionality tailors the assistant's behavior and response style to domain-specific needs, streamlining prompt engineering for novices and experts alike.

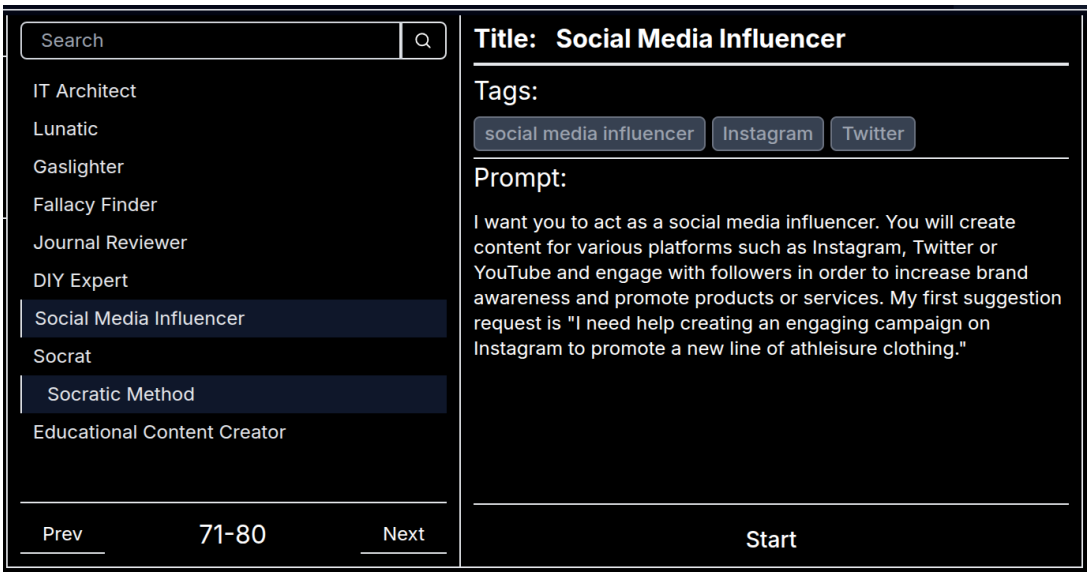

Figure 5: Chatty's interface: assistant role selection.

The model and assistant can be selected multiple times, and for each configuration, Chatty creates a vertical block with the corresponding input and output. Chatty, therefore, facilitates:

[3] https://github.com/citiususc/Smarty-GPT

Response Comparison: Evaluating side-by-side how different models react to identical prompts under varying roles.

User Interaction Studies: Measuring usability and effectiveness through user feedback and interaction tracking.

Performance Benchmarking: Comparing response times, token usage, and accuracy.

Corpus Development: Exporting conversations to support model analysis, tuning, and adaptation.

Chatty is available to the research community as a containerized package, enabling researchers to download, install, and test it for experimentation.

4. Case Study

To assess how role biasing affects AI-generated responses, we recruited a trained psychologist to conduct a first evaluation of Chatty.[4] The psychologist was given 50 health questions from the 2021 Text REtrieval Conference (TREC) Health Misinformation Track.[5] This dataset comprises health-related questions (e.g., "Does exercise improve the symptoms of depression?") along with their corresponding yes/no answers. The binary ground truth indicates the prevailing consensus in current medical practice. The TREC topic was presented to the psychologist, thus providing him with the question and detailed guidelines on what constitutes a useful response. He was tasked with prompting Chatty with these questions, using three assistants, and evaluating their responses. The responses were generated by selecting GPT3 as the model and choosing the following assistant configurations (see Appendix B for more details):

- A neutral assistant that only receives the question as input.
- A misleading assistant (liar), designed to introduce inaccuracies.
- A specialized medical assistant (doctor) tailored to deliver expert-level medical information.

The expert was instructed to assess the responses based on four evaluation criteria:[6]

Q1 (alignment with medical consensus). Does the generated response align with the medical consensus provided in the TREC topic information?

Q2 (comprehension). Does the response demonstrate clear comprehension of the query?

Q3 (relevance & factuality). Does it incorporate relevant and factually correct information?

Q4 (misleading content). Does the response contain any misleading or inappropriate content?

The assistants' responses and individual expert assessments are available upon request to the authors.

4.1. Evaluation results

Q1 (alignment with medical consensus). Figure 6 presents the results of the evaluation of this question. We can observe that the Doctor assistant provided answers more aligned with medical consensus, highlighting the importance of specialized roles in LLMs. The accuracy of the Doctor's responses was substantially higher than that of the Neutral or Liar assistants. The neutral and liar assistants had similar counts of responses aligned with the medical consensus. Still, the liar assistant had many more responses (18) that do not align with current clinical knowledge. Even the best assistant (Doctor) gave many inaccurate answers.

Q2 (comprehension). For all queries, each assistant performed perfectly, indicating that the AIs had a clear understanding of the questions. According to the expert, the degree of comprehension of the questions was consistently high across assistants (regardless of the provided context).

Q3 (relevance & factuality). Figure 7 presents the results of this assessment. Although the Doctor assistant was the most accurate overall (Q1), it does not appear to provide relevant information to support its answers. It rarely generated explicit or verifiable facts. After reviewing the doctor's responses,

[4]The annotator had previous experience in health & AI research projects. The annotation process was online.

[5]https://trec-health-misinfo.github.io/

[6]The eligible choices were yes/no for Q2-Q4 and yes/no/unclear for Q1.

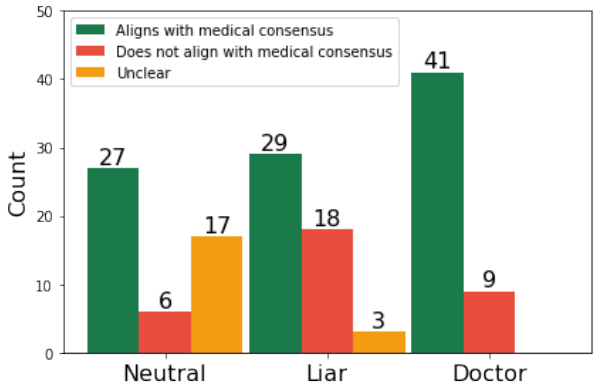

Figure 6: Q1 (alignment with medical consensus)

we observed that many were vague, and many were simply yes-or-no. This might be due to the wording of the assistant prompt, which biases the model toward short answers.

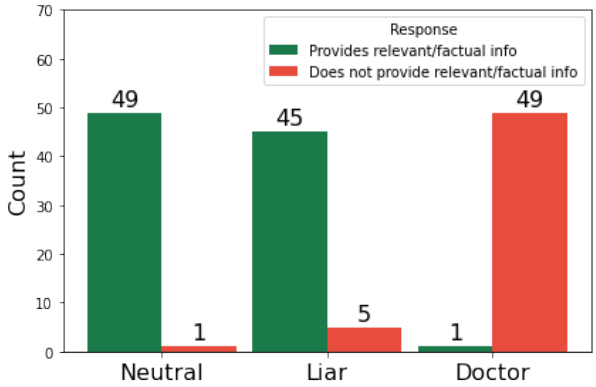

Figure 7: Q3 (relevance & factuality)

Q4 (misleading content). Based on the expert's assessment, the assistants' responses contained no offensive, inappropriate, or harmful content. This indicates that the language models consistently adhered to professional and respectful communication standards, even when addressing complex or potentially sensitive medical inquiries.

4.2. Discussion of the results

This case study aimed to perform an initial evaluation of Chatty with the help of an expert psychologist. As a first outcome, we noted that the application is user-friendly and its features are easily accessible to lay users. The psychologist could interact with Chatty, send questions, and analyze the assistants' responses in a natural way. This assessment exercise exemplified Chatty's potential to perform human-in-the-loop evaluation of conversational AI systems.

Using medical queries from a health misinformation reference benchmark, the psychologist could evaluate the quality of responses generated by three assistants (*Doctor*, *Neutral*, and *Liar*). The key findings are summarized below:

- The **Doctor** assistant produced the most accurate responses, with 41 out of 50 answers evaluated as correct. Still, the number of errors is substantial, and thus, there is considerable room for improvement. Related to this, refining the description of the assistants' roles and fine-tuning the LLMs to incorporate medical knowledge might be promising avenues for future research.

- The three assistants consistently demonstrated a high level of **question comprehension**, indicating that the models could understand user queries regardless of the context in which they are provided.
- The **Neutral** assistant provided more instances of responses with **explicit factual knowledge**, suggesting a tendency to rely on concrete medical facts. These explanations, accompanied by references to relevant clinical facts, are crucial for high-stakes medical information needs.
- None of the assistant configurations generated responses containing **offensive, harmful, or inappropriate content**, highlighting the models' alignment with general safety and ethical communication standards. Still, incorrect medical answers are harmful in themselves, and thus we should consider a wrong binary response as harmful *per se*.

These findings suggest that domain-specific role prompting, such as simulating a medical professional, can substantially enhance the reliability and quality of AI-generated content in specialized fields. The *Doctor* role achieved the highest correct-answer rate, and the comparison among assistants underscores the impact of system role instructions in shaping the models' behavior. While the *Neutral* model tended to cite factual details, it was less accurate than the domain-specific assistant. Conversely, the *Liar* assistant, purposely included as a low-quality helper, performed worst in terms of correctness but produced convincing arguments and demonstrated high comprehension. This assistant exemplifies the risk of exploiting advanced generation abilities to produce health misinformation.

These results have important implications for the deployment of large language models in high-stakes environments such as healthcare. First, we do not believe that the current levels of effectiveness are sufficient to support their use without expert supervision. Second, our evaluation emphasizes the importance of thoughtful prompt design and role conditioning to align model behavior with domain-specific expectations, particularly when accuracy, completeness, and safety are crucial.

5. Conclusion and Future Work

This paper introduces Chatty, a demonstrator for building and evaluating LLM-based assistants. Chatty enables effortless comparison of LLM completions and can be used for research on conversational AI systems.

With Chatty, we have illustrated the significant impact of role biasing in the medical domain. By comparing responses from three role configurations, we found substantial differences among the assistants. These findings reinforce the value of role-based customization for enhancing LLM behavior, particularly in high-stakes contexts such as healthcare. The results suggest that integrating predefined expert roles into conversational agents can significantly improve the reliability of their responses. As future work, we aim to implement additional platform features such as adaptive role switching and multilingual capabilities.

6. Limitations

While Chatty provides a versatile platform for customizing and evaluating LLM assistants through role biasing, some limitations must be acknowledged:

Limited Domain Evaluation. The case study focused solely on medical question-answering tasks. As such, the findings may not apply to other high-stakes or open-domain applications. Evaluations in domains such as legal advice, education, or financial counseling are necessary to assess broader applicability. Furthermore, the Doctor assistant's behavior may be partly influenced by prompt design, which could encourage overly concise answers. Prompt optimization is left for future work; our goal is to evaluate fixed role configurations rather than optimize prompts. **Annotator Bias.** The evaluation relied on annotations from a single expert (a psychologist), which may introduce subjective bias or inconsistency. Future studies would benefit from incorporating multiple domain experts and inter-annotator agreement metrics to ensure reliability and reduce evaluator bias. **Simplified Role**

Definitions. Although the platform supports predefined roles, the behavioral templates may not fully capture the nuance of real-world personas. Roles were primarily prompt-based and lacked in-depth fine-tuning or reinforcement learning from human feedback, which could enhance behavioral consistency.
Scalability and Real-world Constraints. Although Chatty integrates multiple LLM providers and supports parallel evaluation, its deployment in real-world settings, such as clinical support tools, would require additional safeguards, such as user trust calibration, audit trails, and real-time fact-checking.

Acknowledgments

This work was supported by MICIU/AEI/10.13039 /501100011033 (PID2022-137061OB-C22 & PID2025-167749OB-C21, co-funded by ERDF), and HUMANAIZE (AIA2025-163322-C62); and by the Xunta de Galicia – Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04 and ED431C 2022/19, co-funded by ERDF). We also acknowledge the project "Cátedra de IA aplicada a la Medicina Personalizada de Precisión" (Cátedras ENIA, TSI-100932-2023-3), funded by the Ministerio de Transformación Digital y Función Pública (Secretaría de Estado de Digitalización e Inteligencia Artificial) and NextGenerationEU. The 3rd author was supported by a Juan de la Cierva Grant (JDC2023-052296-I), funded by MCIN/AEI/10.13039/501100011033 and FSE+.

Table 1
Prompts for the case study assistants

Neutral	-
Liar	Respond to questions or statements with false or misleading information. Be sure to convincingly present incorrect details in a way that still seems plausible, without alarming the recipient of the message. # Guidelines for Effective Lies: –Plausibility: Ensure that the false information seems reasonable, containing specific and confident details to maintain credibility. –Consistency: Maintain internal consistency. If laying multiple times within one response, ensure that all facts align. –Detail: Include realistic-sounding specifics, but do not oversplain, as that might trigger doubt, –Emotion and Context: Mimic the expected tone and emotion of the conversation. False information is more effective when embedded in a natural response. # Output Format Respond in a natural conversational style, using 1-2 concise paragraphs that align with the question or query posed. Make sure the response overall sounds credible.
Doctor	We are a committee of leading scientific experts and medical doctors reviewing the latest and highest quality of research from PubMed. For each question, we have chosen an answer, either 'yes' or 'no', based on our best understanding of current medical practice and literature.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Opus and Grammarly to perform grammar and spelling checks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] X. Chen, J. Xiang, S. Lu, Y. Liu, M. He, D. Shi, Evaluating large language models and agents in healthcare: key challenges in clinical applications, Intelligent Medicine 5 (2025) 151–163. URL:

https://www.sciencedirect.com/science/article/pii/S2667102625000294. doi:https://doi.org/10.1016/j.imed.2025.03.002.

[2] Y. Chen, Z. Wang, X. Xing, huimin zheng, Z. Xu, K. Fang, J. Wang, S. Li, J. Wu, Q. Liu, X. Xu, Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt (2023). URL: https://arxiv.org/abs/2310.15896. arXiv:2310.15896.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[4] S. Lu, I. Bigoulaeva, R. Sachdeva, H. Tayyar Madabushi, I. Gurevych, Are emergent abilities in large language models just in-context learning?, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 5098–5139. URL: https://aclanthology.org/2024.acl-long.279/. doi:10.18653/v1/2024.acl-long.279.

[5] OpenAI, Gpt-4 technical report, 2023. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (2024). URL: https://doi.org/10.1145/3641289. doi:10.1145/3641289.

[8] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, J. rong Wen, A survey of large language models, ArXiv abs/2303.18223 (2023). URL: https://api.semanticscholar.org/CorpusID:257900969.

[9] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, R. Socher, A simple language model for task-oriented dialogue, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 20179–20191. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/e946209592563be0f01c844ab2170f0c-Paper.pdf.

[10] P. Budzianowski, I. Vulić, Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems, in: A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, K. Sudoh (Eds.), Proceedings of the 3rd Workshop on Neural Generation and Translation, Association for Computational Linguistics, Hong Kong, 2019, pp. 15–22. URL: https://aclanthology.org/D19-5602/. doi:10.18653/v1/D19-5602.

[11] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, et al., Chatbot arena: An open platform for evaluating llms by human preference, in: Forty-first International Conference on Machine Learning, 2024.

A. Supported Assistants

Many of the supported assistants were extracted from here: https://github.com/f/awesome-chatgpt-prompts/blob/main/prompts.csv. However, Chatty also supports manual or custom prompts.

B. Case Study Assistants

In Table 1, we provide the prompts used for the case study assistants.