# Automated testing framework to evaluate multi-agent chat assistants

**Lucas Ramalho, Jose Sousa, Maria Nascimento, Raiza Hanada, Cristian Souza, Eliane Collins**

Instituto de Desenvolvimento Tecnológico (INDT)

{lucas.ramalho, jose.sousa, maria.nascimento, raiza.hanada, cristian.souza, eliane.collins}@indt.org.br

***Abstract.*** *This applied R&D project proposes an automated framework for evaluating multi-agent conversational assistants equipped with retrieval-augmented generation (RAG) capabilities. The solution addresses the high cost and time demands of manual evaluation by introducing a synthetic persona dataset and an automated pipeline that executes large-scale tests on mobile devices. Tests with English and Portuguese personas revealed recurring weaknesses in multi-agent systems, particularly in Portuguese interactions, highlighting the importance of multilingual evaluation. The project is a collaboration between INDT and Motorola Mobility and aims to provide a systematic methodology and testing infrastructure for industry conversational systems.*

## 1. Introduction

The rapid adoption of large language models (LLMs) in personal and mobile applications has enabled increasingly capable conversational assistants. In many real-world deployments, these assistants exhibit multi-agent behavior, relying on multiple interacting components such as RAG retrieval, web search, or location services to deliver context-aware responses [Li et al. 2024, Schick et al. 2023].

In the type of system evaluated in this work, the orchestration process involves multiple steps executed by different agents or modules. For example, if a user asks "Suggest me some restaurant", the assistant may check personal memory for food preferences, use location services to know the user's current city or neighborhood, run a web search for restaurants in that area, possibly filtered by the preferences and then combine all results into a final natural language response.

Each step involves different agents: one to decide which tools to call, another to generate keywords for the web search, another to filter or rank results, and finally the LLM to compose the response. This modular design makes assistants more powerful, but also introduces complexity, since mistakes can happen at any stage of the chain.

Evaluating such systems remains challenging. Existing benchmarks and datasets commonly focus on isolated question–answering tasks or specific application domains, such as healthcare, finance, or customer support. Despite these advances, user-generated data remains underrepresented.

This work presents an ongoing applied R&D project focused on the design and evaluation of an automated framework for multi-agent conversational assistants with RAG capabilities. The framework combines a synthetic persona dataset, an automated pipeline

for realistic mobile interactions, and metrics that assess retrieval behavior, orchestration decisions, and response correctness. Experiments with English and Portuguese personas reveal recurring limitations, particularly in non-English interactions, highlighting the need for systematic multilingual evaluation.

## 2. Project objectives

This project is part of a long-term industrial research and development collaboration between INDT and Motorola Mobility and is conducted by a permanent team of eight engineers and researchers responsible for testing AI models embedded in smartphones.

The primary objective is to develop an automated framework for evaluating multi-agent conversational assistants that use RAG. Manual evaluation is costly and does not scale. By simulating realistic interactions with synthetic personas, the framework reduces evaluation effort while improving coverage and consistency, providing a testing infrastructure that can be integrated into industrial development and QA workflows.

## 3. Proposed Methodology and Technologies

The proposed framework combines synthetic dataset generation with an automated testing pipeline to enable large-scale evaluation of multi-agent conversational assistants on mobile devices.

### 3.1. Dataset Design

The dataset was constructed in two steps: persona profile creation and question–answer (QA) generation. First, we designed 20 synthetic personas representing diverse demographic, socioeconomic, and linguistic characteristics. Attributes such as age, occupation, education, income, and writing style were adapted from prior work and converted into structured prompts [Yukhymenko et al. 2025]. These prompts were processed by the Gemini 2.5 Pro LLM to generate realistic entries in English and Brazilian Portuguese, covering multiple regions and cultural contexts.

The entries include daily activities, preferences, past events, and future plans, as well as non-obvious personal facts (e.g., temporal and nested details) designed to stress memory retrieval and reasoning. In the second stage, 10 QA pairs per persona were generated to simulate memory-recall scenarios, including temporal, comparative, and reasoning-based queries, resulting in 200 QA pairs.

Although both stages rely on automated generation, the dataset was manually curated after creation. The project team reviewed all persona entries and QA pairs to validate internal consistency, linguistic quality, and alignment between questions and ground-truth answers. During this manual evaluation, no strong bias-related issues were identified. At present, the dataset is not publicly available due to internal constraints, but there are plans to release a version of it in the future.

### 3.2. Automated Testing Pipeline

A fully automated pipeline was implemented to replace manual evaluation with reproducible test runs on physical smartphones. The pipeline consists of three stages as described in the Figure 1:

| File_txt | Language | Questions | Answers | Data_entry | Data_entry_id |
|---|---|---|---|---|---|
| **profile_1.txt** | English | What were the two major health recommendations I received from Dr. Evans? | Dr. Evans advised you to consider switching to a lower-impact sport for your knee and to swap your morning bacon for avocado to help lower your cholesterol. | ['My left knee, which I injured playing college basketball, was aching after my morning run in Central Park. Dr. Evans said I might need to consider switching to a lower-impact sport. Getting old is fun.', "Annual physical with Dr. Evans. He's pleased with my blood pressure but wants my cholesterol down. He suggested swapping my morning bacon for avocado. A true sacrifice."] | [8, 13] |
| ... | ... | ... | ... | ... | ... |
| **profile_9.txt** | Portuguese | nome meu ortopedista | Seu ortopedista é o Dr. Mário. | ['Meu joelho esquerdo, aquele que lesionei jogando vôlei na adolescência, doeu o dia todo... Preciso voltar para o pilates. Já marquei uma consulta com o Dr. Mário, meu ortopedista, para a próxima semana.'] | [9] |

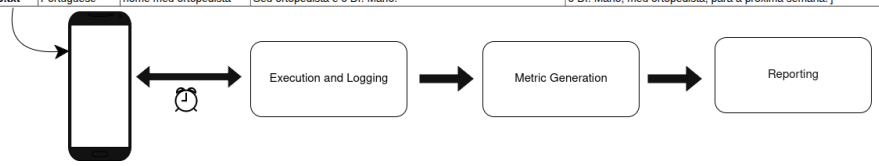


**Figure 1. Automated Testing pipeline**

First for test data ingestion, persona datasets are converted into CSV files and injected into the assistant via UI automation using a framework similar to Appium. Test cases are organized into four categories: sanity, prompt injection, stress, and component validation. Second, in the execution and logging step, tests run on connected devices under the control of an automation server, which captures interaction logs and monitors resource usage (RAM, CPU, GPU).

And finally, for metric generation and reporting, logged outputs are normalized and compared against expected answers using a Python-based engine. Metrics include search memory accuracy, which evaluates whether the assistant correctly invoked the RAG tool, and final answer accuracy, which measures whether the final response matched the expected answer. Final answer accuracy is measured using an LLM-as-a-judge approach via Ollama, employing a Qwen-family model with calibrated prompts for closed-form QA. While occasional misjudgments may occur, this approach enables efficient large-scale evaluation.

The infrastructure includes different tools as shown in Figure 2. It integrates REST APIs for orchestration, Redis for distributed execution, MLflow for experiment tracking, and Nginx for secure access. This design enables scalable, systematic, and language-aware evaluation of conversational assistants in realistic mobile environments.

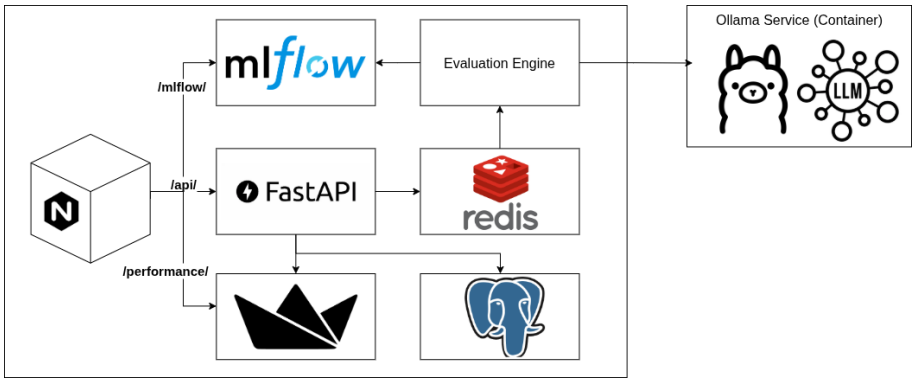


**Figure 2. Metric Generation Pipeline Architecture**

## 4. Results

The evaluation was conducted on a mobile conversational assistant equipped with multiple tools, including general knowledge, web search, location retrieval, device information,

and RAG. The test focused on the assistant's ability to recall personal information from its memory database using the RAG tool.

The automated pipeline processed 20 personas with 40 memory entries and 200 QA queries in under four hours, generating logs and metrics without manual intervention. Results show high tool invocation accuracy but lower final answer correctness, particularly for Portuguese as we can see in the Table 1. This gap reflects the complexity of multi-agent orchestration, where correct responses depend on retrieval, reasoning, and generation, and highlights persistent multilingual limitations stemming from the prioritization of English during model training.

**Table 1. Language group accuracies**

| Language group accuracy | Search memory accuracy | Final answer accuracy |
|---|---|---|
| English Personas | 0.95 | 0.92 |
| Portuguese Personas | 0.91 | 0.83 |

The automated evaluation revealed recurring issues during qualitative analysis. In some cases, the assistant invoked non-existent tools, producing fabricated outputs. Answers sometimes adopted the user's voice, especially in Portuguese, due to copying from memory entries. Retrieval gaps were also observed when relevant entries were missed because of synonyms or phrasing variations. Even when retrieval was correct, reasoning errors occurred, with the assistant misinterpreting relationships between facts and producing inverted or incorrect conclusions.

Finally, the use of synthetic personas enabled large-scale evaluation without exposing real user data, avoiding ethical and privacy concerns that restrict experimentation with personal information. While real user data could provide richer insights, synthetic data allowed controlled, privacy-preserving testing. Future work includes extending the dataset with additional personas, languages, and more complex interaction scenarios.

## Acknowledgements

This work was partially supported by SUFRAMA (Law No. 8387/1991).

## References

Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., Liu, J., Xu, W., Wang, X., Sun, Y., Kong, R., Wang, Y., Geng, H., Luan, J., Jin, X., Ye, Z.-L., Xiong, G., Zhang, F., Li, X., Xu, M., Li, Z., Li, P., Liu, Y., Zhang, Y., and Liu, Y. (2024). Personal llm agents: Insights and survey about the capability, efficiency and security. *ArXiv*, abs/2401.05459.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools.

Yukhymenko, H., Staab, R., Vero, M., and Vechev, M. (2025). A synthetic dataset for personal attribute inference. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.