SmartEdge Assistant: A Voice-Controlled, Privacy-Preserving Home AI Hub

Students

Sylvester Homberger

David Hintermann

Introduction: In the era of the Internet of Things, smart devices from numerous manufacturers across diverse ecosystems are continuously introduced to make our homes more intelligent and connected. Recent advances in artificial intelligence, particularly in natural language processing and large language models, have made it feasible to interact with smart home systems using natural spoken language. However, most voice assistants rely heavily on cloud infrastructure, raising concerns about data privacy, network dependency, and the inability to operate during internet outages.

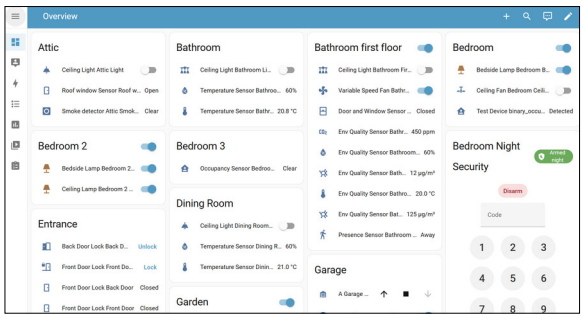
This project addresses these limitations by developing a fully local voice assistant that operates without mandatory cloud connectivity. The assistant is designed to control and query a Home Assistant-based smart home environment while keeping all voice processing, language understanding, and device interactions within the home network. This approach ensures user privacy by preventing sensitive voice data and household information from being transmitted to external servers. The core challenge lies in achieving acceptable accuracy using locally-hosted large language models on consumer-grade hardware while maintaining natural conversational capabilities and reliable device control.

Approach: The solution builds upon the existing Home Assistant "Assist" architecture, leveraging its established pipeline structure which consists of wake word detection, speech-to-text conversion, conversation agent processing, and text-to-speech synthesis. An Agent using a local model and connecting to Home Assistant over the Model Context Protocol (MCP), allows querying information from the home, e.g., temperature in the living room.
To evaluate different LLM capabilities, a comprehensive benchmark was developed. This benchmark consists of categorized test scenarios ranging from simple single-command tasks to complex multi-step reasoning challenges. Each test case defines preconditions (device states and home configurations), expected outcomes, and evaluation criteriay. A weighted sum score was used to compare models. Every executed task is given between 0 and 3 points, multiplied by the category's weight - 1, 1.5 or 2 based on the complexity category of the task. We ran the benchmark across three LLM families—CodeLlama, Falcon H1, and Qwen2.5-Coder—covering model sizes from 0.5B to 34B parameters. OpenAI GPT models were used for comparison.

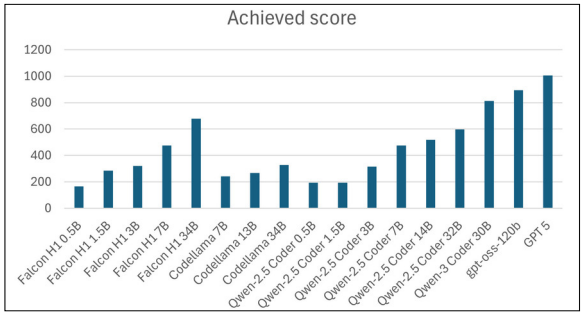
Result: The benchmark demonstrates that larger LLMs achieve better results. Surprisingly, the small flat configuration does not perform significantly better than the more complex house setup with many more rooms and devices. However, running these models locally is challenging: it requires high-end consumer hardware, and latency remains high.
Running the benchmark against GPT-5 showed the best performance, though it did not achieve a perfect score. The results suggest that the LLM is not always the primary bottleneck. We identified opportunities to improve the setup: incorporating additional tools and providing richer context can help the SmartEdge Assistant better understand how to interact with Home Assistant and more accurately follow the user's intentions.

The ATOM Echo ESP32 based smart speaker developer board integrates microphone, speaker and wake word detection
www.brack.ch/m5stack-entwicklerboard-atom-echo-1081521

Home Assistant's dashboard during the execution of the benchmarks
Own presentment



Benchmark scores per model
Own presentment



Advisor
Prof. Dr. Mitra Purandare

Subject Area
Artificial Intelligence, Internet Technologies and Applications