Qeios

Peer Review

Review of: "Large Language Model Interface for Home Energy Management Systems"

Evgenii Arsentev[1]

1. Independent researcher

What the paper claims

The paper puts an LLM agent between a household and a home energy management system. The agent asks questions in plain language, converts loose answers into the eight strictly formatted parameters the optimiser needs, and writes them through function-calling tools (Section 3.1). Two prompting devices are tested on top of that: few-shot examples and ReAct. Because evaluating a conversational interface needs many sessions, the authors replace human testers with a second LLM instructed to behave as a user at three precision levels, measured by cosine similarity at 0.98, 0.94, and 0.84 for Easy, Medium, and Hard (Table 1). Three Mistral-family 7B models, denoted V1 to V3, are run 20 times per difficulty level, and Table 2 reports retrieval accuracy from 53.1% at worst to 97.5% at best. The abstract summarises this as an average of 88%.

The construction is sensible, and the paper is honest about what the optimiser needs. The simulated-user idea is the most transferable part of it, and also the part that most needs defending, which is where I will start.

Does the conclusion hold on the data?

1. The 88% in the abstract is the best model with the best agent type, not the interface. Averaging the ReAct+example row across all three models in Table 2 gives 85.1%; averaging it across difficulty levels for V2 alone gives 87.7%, which rounds to the quoted figure. So the headline describes V2 with ReAct and examples, while the abstract attributes it to "the proposed LLM-based HEMS interface". Say which cells were averaged, and quote the spread alongside the mean, because the same table also contains 53.1%.

doi.org/10.32388/JADXXS

2. The stated success criterion contradicts the numbers. Section 5 says a test counts as a success only if all stored parameters exactly match, which with 20 tests per cell would force every entry in Table 2 to be a multiple of 5. The table instead contains 83.75 and 85.6, which are exactly 134 and 137 out of 160, that is 20 tests times 8 parameters scored per parameter. Per-parameter accuracy is the more informative measure, so keep it, but correct the definition in Section 5 and report the per-session all-eight-correct rate too. For a system that hands its output to an inflexible optimiser, the session-level number is the one a deployer cares about: seven right parameters out of eight still produces a wrong schedule.

3. At 160 scored items per cell, most of the ReAct gains sit inside the noise. The standard error near 85% accuracy with 160 items is about 2.8 percentage points, so a difference between two cells needs roughly eight points before it can be called real. In Table 2, the gap between Act+example and ReAct+example is 1.9 points for V2 Easy, 3.75 for V2 Medium, and 1.3 for V2 Hard, and it reverses for V3 in Medium and Hard. The claim in Section 5.1 that this "highlights the added value of ReAct" is not supported at that resolution. What does survive is the efficiency result from Figure 6: fewer questions and fewer iterations for the same accuracy. That is a real finding, and it deserves to be the headline instead.

4. The user simulator is never validated against people. Section 4 builds the three difficulty levels by rewriting prompts with synonyms, changed formats, and noise, and characterises them with cosine similarity. Nothing checks that these synthetic users resemble the elderly or low-literacy residents named as the motivation in Section 1. There is also a family effect to rule out: when the simulated user and the agent come from related models, they share conventions about dates, units, and phrasing, and the agent is being tested against an interlocutor that fails in the same ways it does. Ten sessions with real people on the Hard setting, compared against the simulator, would settle both points and would not delay the paper much.

5. The ground truth can drift because the user is generative. Scoring compares stored parameters with the personal information handed to the simulated user, but at temperature 0.8, that user may volunteer details it was never given or contradict its own profile. Report how often the simulated user answered inconsistently with its assigned profile, and whether such sessions were discarded or scored as agent failures. Without that number, some of the reported error is the user simulator misbehaving rather than the agent.

6. The 48% saving belongs to the optimiser, not to the interface. Section 5 reports 16 pounds against 31 pounds for one simulated scenario, placed among the interface results where it reads as an outcome of the proposed system. It is a property of price-aware scheduling with one price series over one period.

Move it to Section 2, name the tariff and the dates, and give more than one household profile, or drop the percentage.

Reproducibility

Naming the three checkpoints precisely - OpenHermes-2.5-Mistral-7B, Mistral-7B-Instruct-v0.2, and v0.3 - puts this ahead of most applied LLM papers, and the appendix carries the prompt templates and the parameter table.

Missing for a replication: the agent decoding parameters, since only the simulated user temperature of 0.8 is given and the agent settings are never stated; the serving stack, quantisation, and context window, which matter because Figure 6 makes timing claims and a 7B model quantised to four bits behaves differently from the same weights in half precision; the hardware the timings were taken on; random seeds and whether the 20 sessions per cell differ by anything except sampling; and any code or transcript release. Session transcripts would be particularly valuable here: they are the raw evidence for both accuracy and the question counts, and they cost nothing to publish.

The weather and price data behind the HEMS run should also be cited with source and period, since the 48% figure depends entirely on them.

What to fix

1. Abstract and Section 5.1: state that 88% is V2 with ReAct and examples averaged over the three difficulty levels, and give the range across Table 2.
2. Section 5: correct the success definition to per-parameter scoring over 160 items per cell, and add the per-session all-eight-correct rate.
3. Table 2: add confidence intervals, and restate the ReAct comparison in terms of questions and iterations from Figure 6, where the effect is larger than the noise, rather than in terms of accuracy.
4. Section 4: validate the simulated users against a small human sample, and report whether the agent and the user were drawn from the same model family in each configuration.
5. Section 4 and appendix A.3: report how often the simulated user contradicted its own assigned profile, and how those sessions were handled.
6. Section 5: move the 48% cost result out of the interface results, name the tariff, location, and period, and show more than one household.

7. Appendix A.4: give agent decoding parameters, quantisation, context length, serving stack, hardware, and seeds; release prompts and session transcripts.
8. Table 2: use one decimal convention; 95, 83.75, and 96.9 currently appear in the same table.
9. Section 1: several sentences follow the pattern "In [5] proposed a method", which is missing a subject; "Using LLMs to facilitate the practical implementation of HEMS is a daily topic" appears to mean "a topical question". A language pass would help, and "user stylometry" in Section 3.1 should be defined on first use.
10. Front matter: no corresponding author address is given, only the publisher forwarding address.

Recommendation

Major revisions. The interface works, the ablation is the right one, and the simulated-user harness is a genuinely useful idea for anyone who has to evaluate a conversational system without recruiting a panel. Three things stand between this and a solid paper: the headline number has to name the configuration it came from, the accuracy claims have to be stated at a resolution that 160 items can support, and the simulator has to be checked against at least a handful of real users before it stands in for the elderly and non-technical residents the paper is written for. The efficiency result is stronger than the accuracy result and should carry the abstract.

Evgenii Arsentev, PhD

Declarations

Potential competing interests: No potential competing interests to declare.