

Peer Review

Review of: "Hybrid Memory-Retrieval Model: Enhancing Trust in Medical Chatbots"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper describes a medical question-answering chatbot built from parts that are each well known, assembled in one pipeline. Documents are retrieved from a 216,102-entry corpus (§V.A) by two retrievers in parallel, BM25 and MedCPT, and their top-5 lists are merged with Reciprocal Rank Fusion into a final top 5. In parallel, past query/response pairs are embedded with MiniLM (dimension 384, §V.C), stored in ChromaDB, and the top 2 are pulled back when cosine similarity exceeds 0.4. Documents, memories, and an instruction block are packed into a prompt of at most 1024 tokens (§IV.B, §V.D) and passed to Phi-2 (2.7B parameters). When neither source gives "sufficient context," a fallback instruction asks the model for a "safe, conservative response." Users can see which memories were used and wipe them with a `clear` command (§V.E).

The claims are that this design reduces hallucination, improves multi-turn context retention, and is faster than RAG baselines. The single number that reaches the abstract is BERTScore F1 = 0.8644 on MedQuAD. The body adds a comparison against "Mistral with RAG and fine-tuned Mistral with RAG" on 20 QA samples (Table 1 caption), a memory ablation with 0, 1, and 2 retrieved memories (Table II), and a retrieval-only evaluation with Recall@5, Precision@5, and BERTScore F1 (Table III). The Contributions section (§VIII) goes further and calls the system "significantly faster and more trustworthy than baseline RAG-only systems".

There is real value here, and I would like it to survive revision. The pipeline is described clearly enough that I could sketch it from §IV and §V without the figure. Running a 2.7B model on local hardware is a sensible direction for settings where sending patient questions to a hosted frontier model is not

acceptable. The tie-breaking rule in RRF is spelled out, which is rare. And the user-facing controls in §VE (visible memories, a delete command) show that the author thought about the person on the other side of the screen, not only about the metric.

2. Whether the conclusion holds on the data

A limitation of my own first. As rendered on Qeios, Tables 1, II, and III and Figure 3 are images whose contents did not come through in the text I reviewed, so I could not read the baseline values, the perplexities, the latencies, or the Recall@5 and Precision@5 figures. My comments on those tables are therefore about design, not about the specific cells. The only result I can quote with a coordinate is BERTScore F1 = 0.8644 (Abstract; §I, paragraph 2).

2.1 The evaluation set is inside the retrieval corpus. This is the defect that matters most. §VB states that MedQuAD contributes 17,236 entries to the corpus, and §VA, §VC, and §VD all say that evaluation is performed on MedQuAD. Nothing in the text says the 20 test questions (Table 1 caption; Figure 3 caption) were removed from the index before testing. If they were not, BM25 will very often return the exact reference answer as document rank 1, Phi-2 will paraphrase it, and BERTScore against that same reference will be high by construction. Under that setup, 0.8644 measures copying, not medical correctness. The same applies to Table III: a Recall@5 computed on questions whose gold passages are in the index is a lookup test. I may be wrong about the split, and the author may have excluded the test items. But the paper has to say so explicitly, and ideally show results with the whole MedQuAD source held out.

2.2 Twenty questions cannot carry the claims. Table 1 is labelled "(20 QA samples)". No confidence intervals, standard deviations, or significance tests appear anywhere in §VI, yet the text uses "significantly faster" (Abstract; §VI; §VIII) and "perplexity improved significantly" (§VI, after Table II). With $n = 20$, a paired bootstrap on BERTScore differences is cheap and would show whether the ranking against the baselines is stable. I also could not find how the 20 were chosen from MedQuAD, or whether Table II and Table III use the same 20 or a different sample.

2.3 The baselines do not appear to have been run under the same conditions. The sentence introducing Table 1 attaches reference [4] (Bora and Cuayáhuitl, 2024) to the two Mistral baselines. My reading is that the baseline figures are taken from that paper. If so, the comparison mixes a different corpus, a different test sample, possibly a different BERTScore backbone, and different hardware. Latency, the headline advantage, is meaningless across machines: the system here runs on an unspecified "local CPU/GPU

setup" (§V.F), and a 2.7B model is expected to be faster than a 7B one on the same GPU without any architectural credit to memory or RRF. If the baselines were in fact re-run, the paper should say where, on which hardware, with which retriever, and on which 20 questions.

2.4 Perplexity is not comparable across these models. §VI concedes that the system's perplexity is higher than the fine-tuned baseline and calls this "an acceptable trade-off". But perplexity depends on the tokenizer and on which model scores the text. Phi-2 and Mistral use different vocabularies, so their perplexities are not on one scale unless a single fixed scorer model is applied to all outputs. The paper does not name the scorer. In Table II, the problem is different and, I think, more interesting: adding memories lowers perplexity. Memories in this system contain the chatbot's own previous responses (§V.C). Text that echoes earlier Phi-2 output will naturally look more probable to Phi-2. Lower perplexity there may reflect self-repetition rather than coherence, and §VI itself reports that memory "slightly reduced semantic alignment". Read together, Table II is at least as consistent with "memory makes answers more fluent and less correct" as with the paper's reading.

2.5 Hallucination is never measured. Research goal 1 (§I.B) asks how well the RAG pipeline reduces hallucinations; §VI concludes "low hallucination tendencies" from BERTScore. BERTScore rewards lexical and semantic overlap. An answer that reverses a dose, swaps "contraindicated" for "indicated", or names the wrong drug in the same class can score very close to the reference. To support the word "hallucination," the paper needs a factuality measure: claim-level support checking against retrieved passages, an NLI-based faithfulness score, or, best, clinician annotation of a sample with a stated rubric. None of these is present.

2.6 Context retention is never tested. Research goal 2 concerns multi-turn memory, but MedQuAD consists of single questions with reference answers. §VI acknowledges that "many queries were first-time questions, the memory module's full impact was limited during testing". So the conclusion in §VII that the system shows "improved fluency in multi-turn scenarios" has no multi-turn evaluation behind it. A small scripted set of conversations, where turn 3 depends on a fact stated in turn 1 ("I am pregnant", "I take warfarin"), would test exactly what the architecture promises.

2.7 The fallback is the main safety claim, and it is not evaluated. §VIII calls the fallback "a key contribution" that "ensures safe, non-hallucinatory responses". The paper does not define "insufficient context" (a BM25 score cut-off? an RRF threshold? the 0.4 memory threshold?), does not give the fallback instruction text, and reports no rate at which the fallback fires, no test on out-of-corpus questions, and

no check that Phi-2 obeys the instruction. Small models are known to ignore abstention instructions. "Ensures" is not supported; "is intended to reduce" would be.

2.8 Clinical safety points the paper should address. I read the system as a patient-facing tool, since §I mentions symptom checking and §V.B adds a "symptom and treatment" Kaggle dataset. Several design choices worry me from a clinical standpoint.

- Storing the model's own answers as memory and retrieving them as context creates a loop. One wrong answer, once stored, becomes "evidence" for the next related question. §II names this exact failure ("snowballing"), but the design does nothing to stop it: there is no flag that separates verified source text from past generations inside the prompt.
- A cosine threshold of 0.4 on MiniLM embeddings is permissive. In a medical conversation, that could pull a memory about a different patient context (a child's fever into an adult's chest-pain question) into the prompt. I could not find any measurement of how often irrelevant memories are retrieved.
- MedMCQA, which provides 192,000 of the corpus entries (§V.B), is a bank of Indian postgraduate entrance exam questions. Exam items are written to test discrimination between distractors, not to inform a lay user, and some reflect regional practice. How were they converted into retrievable passages, and was the incorrect-option text excluded?
- There is no red-flag handling. A question describing crushing chest pain or stroke symptoms should trigger an emergency message before any retrieval. §XI says the system is not for triage, but a user cannot be prevented from asking triage questions.
- §XI states that users get random IDs with "no account or identity linkage", and that `clear` deletes entries "associated with their session". §II and §VII describe memory as long-term and "across sessions". Both cannot be true at once. If there is no identity linkage, cross-session memory either does not exist or is shared across users; the second would be a privacy problem.

2.9 Arithmetic check on the corpus. §V.A and §V.B give the total as 216,102. The components in §V.B are 192,000 (MedMCQA) + 17,236 (MedQuAD) + 4,065 (BioASQ Task B) + approximately 801 (Kaggle) = 214,102. That is 2,000 short of the stated total. Probably a typo, but the reader cannot tell which figure is right, and deduplication is not described.

2.10 Smaller methodological points. §V.A says dense embeddings come from MedCPT/Article_Encoder. MedCPT is an asymmetric retriever with a separate query encoder; if queries were also encoded with the article encoder, dense retrieval quality would be degraded, and the RRF

comparison would be unfair to MedCPT. The fusion rule in §V.A ("ensuring that at least one MedCPT-retrieved document is included") is no longer plain RRF, and §IV.C gives k only as "often set as 60", not as the value used. For Table III, the paper does not define what makes a retrieved document "relevant" for Recall@5 and Precision@5. Lastly, §III describes Cache-Augmented Generation [3] as "structured memory modules that persist across sessions"; my understanding of that paper is that it preloads a knowledge collection into the model's KV cache in place of retrieval, which is a different idea. Please recheck that sentence.

Overall: the architecture is plausible, and the paper says what it built. The claims about hallucination, trust, multi-turn coherence, and speed go well beyond 20 in-corpus questions scored with BERTScore against baselines that seem to come from another study. Neither "trust" in the title nor "more trustworthy" in §VIII is measured in any form.

3. Reproducibility

§XII states that code, datasets, preprocessing scripts, and evaluation notebooks are on GitHub. In the Qeios text I reviewed, the link renders only as the word "GitHubRepository" with no visible URL, so I could not inspect the repository. I credit the intent. The following are missing from the paper itself and should be in it, not only in notebooks.

- **Model versions.** Phi-2 is named, but not the Hugging Face repository revision. "MiniLM" (§V.C) covers several checkpoints; dimension 384 narrows it but does not identify it. The MedCPT query encoder is not mentioned. For the Mistral baselines, neither the checkpoint nor the fine-tuning recipe is given here.
- **Decoding.** No temperature, top-p, max new tokens, repetition penalty, or greedy versus sampling setting is reported for Phi-2. As far as I could find in the text, the paper has none of these. No seeds either.
- **Prompts.** The instruction block and the fallback instruction are not reproduced. How the 1024-token limit is enforced (which component gets truncated first when five documents and two memories overflow) is not described.
- **Retrieval parameters.** BM25 k_1 and b , tokenization for `bm25_tokenized.pkl`, the RRF k actually used, and the exact rule that decides "insufficient context".
- **Data.** The Kaggle dataset is not named or linked. The BioASQ year and the MedMCQA split are not stated. The 20 test questions are not listed, and their exclusion from the index is not confirmed.

- **Metrics.** BERTScore backbone model, layer, and whether baseline rescaling was applied (rescaling changes absolute values a lot, so 0.8644 cannot be interpreted without it). The perplexity scorer model. How latency was timed: wall-clock per response including retrieval or generation only, warm or cold, on what device.
- **Compute.** "local CPU/GPU setup" (SV.F) needs the actual GPU or CPU model and memory, since latency is a headline claim.

4. *What specifically to fix*

1. **§V.A, §V.B and §VI (Tables 1, II, III).** State whether the 20 MedQuAD test items were removed from the retrieval index. If not, rerun with the test items, or all of MedQuAD, excluded, and report both sets of numbers side by side.
2. Enlarge the test set. Twenty questions is a demo. Use a few hundred MedQuAD items drawn at random with a stated seed, and report means with 95% bootstrap confidence intervals for every metric in Tables 1, II, and III.
3. **Table 1 and the sentence before it.** Say plainly whether the Mistral baseline values are copied from [4] or were rerun. If copied, drop the latency comparison or rerun both systems on the same machine and the same questions. A fair baseline would also include Phi-2 with BM25-only RAG and Phi-2 with no retrieval, so the reader can see what fusion and memory each add.
4. **§VI, perplexity.** Compute perplexity with one fixed external scorer for all systems, and name it. For Table II, discuss the self-generated-text explanation for lower perplexity.
5. Add a factuality evaluation. Even 50 answers rated by one or two clinicians for correctness and potential for harm, with agreement reported, would be far more informative than BERTScore for a medical system. Until then, replace "low hallucination tendencies" (SVI) with wording that matches what BERTScore measures.
6. **Research goal 2 and §VII.** Build a small multi-turn test set where later turns depend on earlier facts, and measure whether the memory module retrieves them. Without it, remove "improved fluency in multi-turn scenarios" from the Conclusion.
7. **§IV.B and §V.D, the fallback.** Give the trigger rule, the instruction text, the firing rate on the test set, and results on a set of out-of-corpus or unanswerable questions. Change "ensures safe, non-hallucinatory responses" in §VIII to something the data can support.

8. **§VC.** Tag memory entries in the prompt as prior chatbot output, separate from source documents, or store only user-stated facts. Report how often a memory above the 0.4 threshold was irrelevant to the query.
9. Add a red-flag screen for emergency symptoms that bypasses generation, and mention it in §XI.
10. **§XI versus §II and §VII.** Reconcile "no identity linkage" and session-scoped deletion with the claim of long-term, cross-session memory.
11. **§VA and §VB.** Fix the corpus total: the listed components sum to 214,102, not 216,102. Describe how MedMCQA exam items were turned into passages.
12. **§VA.** Confirm which MedCPT encoder embeds queries.
13. **§IV.C.** Report the RRF k used, and note that the "at least one MedCPT document" rule modifies standard RRF.
14. **§III, sentence on [3].** Recheck the description of Cache-Augmented Generation.
15. **Table III.** Define relevance for Recall@5 and Precision@5.
16. **§XII.** Print the full repository URL and a commit hash or release tag in the text.
17. **Title and §VIII.** Either measure trust (a small user study) or retitle around what was measured, for example, retrieval and memory for a compact medical QA chatbot.
18. Tables 1, II, and III should be real tables, not images, so that values can be indexed, quoted, and checked.

5. Recommendation

Major revisions. The system is coherent, clearly described, and aimed at a real need: a small local model with visible memory and a delete button. I do not think the evidence yet supports the main claims. The test questions appear to come from a source that is also in the retrieval index, the sample is 20 items with no uncertainty reported, the baselines seem to be imported from another study, and neither hallucination, multi-turn retention, the fallback, nor trust is directly measured. For a tool meant to answer medical questions, the self-reinforcing memory and the absence of red-flag handling also need attention before any deployment language ("suitable for deployment", §VI) is justified. None of this requires a new architecture. It requires a cleaner evaluation, and I would be glad to read the revised version.

Declarations

Potential competing interests: No potential competing interests to declare.