International Headache Society
IHC 2025
Sep 10 - 13 2025
São Paulo | Brazil

AI Agents Outperform Large Language Models in Making Headache Diagnoses from Free Text

Jim Blythe[1], Vipin Chawla[1], Rob Cowan[2] and Alan Rapoport[3]

BonTriage

(1) BonTriage, Inc., (2) Stanford University, (3) UCLA

Objective

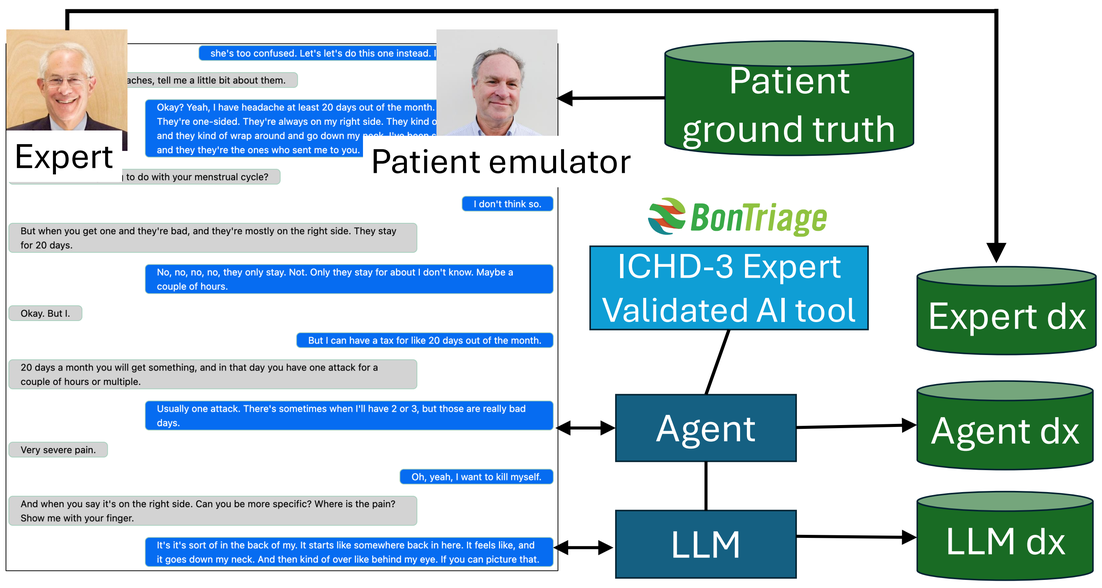
Machine Learning (ML), Large Language Models (LLMs) and more recently AI Agents will make unprecedented contributions to headache diagnosis and care. We compare the performance of an LLM and an AI agent in making a diagnosis from the transcript of a discussion between an MD and a patient. The AI agent uses a validated tool that combines ICHD-3 best practices with expert insight (BonTriage.com).

Method

Twenty conversations were recorded between two MDs. The first played a patient diagnosed with migraine, cluster or tension-type headache. The second asked questions and made a diagnosis.

Transcripts were created using an off-the-shelf speech recognition application (Zoom[4]). The transcripts were used by both the LLM and the agent to make a diagnosis, each asking follow-up questions as needed.

ChatGPT-4o[5] was used as the LLM. The agent was developed using the LangChain[6] platform for building agentes, and also used ChatGPT-4o to extract from the transcript the variables required by the BonTriage diagnostic tool. None of the three judges (the MD, LLM or agent) had access to the diagnoses of the other judges.



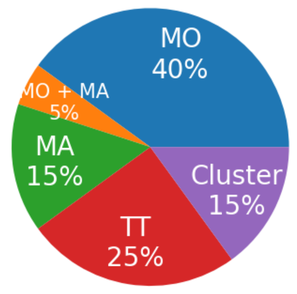
Experiment architecture. All judges make diagnoses based on the conversation between the expert and patient. The agent uses an expert-validated AI tool (Bon Triage) in its diagnosis and uses the LLM to extract data needed by the tool from the conversation.

4. www.zoom.com
5. openai.com/index/hello-gpt-4o
6. www.langchain.com

Results

The agent made a correct diagnosis in 95% of the cases, slightly below the human expert, at 100%. The Large Language Model (LLM) diagnosed 70% of the cases correctly.

Judge	% correct
MD	100%
Agent	95%
Large language model	70%



The proportion of headache types

Migraine without aura comprised 40% of the cases, 20% were migraine with aura, 25% were tension-type and 15% were cluster headaches. All judges correctly diagnosed all cases of migraine without aura. The Large Language Model (LLM) misdiagnosed four of five tension-type cases, and the agent missed one. Although prompted to choose migraine, tension-type or cluster headache, the LLM's misdiagnoses included paroxysmal hemicrania and NDPH.

Conclusion

We report on preliminary results in our ongoing work. The agent's structured approach, using a validated diagnostic tool, is shown to create an important safeguard in comparison with the LLM, reducing the chance of error from 30% to 5% in this experiment, and providing a framework to explain the diagnostic impression.

Approaches based on AI language models must engage humans during diagnosis or other activities to reduce mistakes. Our proposed approach will combine human input from three distinct sources. *First,* expert input is central to the construction of the BonTriage diagnostic tool, which has also been clinically validated [Cowan et al. 2021]. *Second*, the physician/nurse practitioner can review the proposed diagnosis using a clear explanatory structure created by the tool. *Third*, the patient can verify whether the language model has correctly extracted diagnostic variables from the text.

We are developing a new type of diagnostic aid based on this approach.

Please contact us at Jim.Blythe@Bontriage.com for any questions or discussion.