SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Agentic Vision-Language Model for Explainable Stepwise Image Interpretation

David Fussell, MD, University of California, Irvine; Peter D. Chang, MD; Edward Feng

Introduction/Background

Despite recent progress, current vision language foundation models are limited by a single representation that entangles visual and textual features resulting in opaque one-step reasoning and overall poor performance. In this study, we propose a novel paradigm using agentic models with autonomy to interactively probe images via foundational vision tools, yielding high-performing and explainable zero-shot diagnosis.

Methods/Intervention

As a pilot experiment, we apply our method to characterize subtle ventricular abnormalities on head CT in patients with normal age-related atrophy (N=64), hydrocephalus (N=41), and normal pressure hydrocephalus (N=37; NPH). Upon receiving the initial prompt, the LLM agent (Llama3.3-70B) generates an internal representation of key steps required for diagnosis (Figure 1). To extract image findings, the agent generates a tool call to a separate foundation vision model (DeepATLAS) capable of zero-shot characterization of arbitrary target anatomy. For example, to identify ventricular enlargement disproportionate to sulcal widening, the agent translates the request into Python code that quantifies the volume of ventricular compartments and the subarachnoid space. As needed, the agent can repeatedly probe the image with new tool calls based on aggregated information until enough details are available for satisfactory diagnosis.

Results/Outcome

Across all ventricular abnormalities, the agent performed with an accuracy of 91.6% and weighted F1-score of 0.971 (Table 1). Individual prediction of hydrocephalus was most accurate (F1-score, 0.962) followed by atrophy (F1-score, 0.919) and NPH (F1-score, 0.864). The agent required a median of 3 tool calls per exam for diagnosis. A total of 25.4% of exams received a diagnosis after just one finding (accuracy, 0.972), while a total 52% of exams received a diagnosis after three findings (accuracy, 0.987). By contrast the remaining exams required up to five findings with significant deterioration in performance, overall reflecting the inherent uncertainty of borderline cases. Indeed, 9/12 (75%) errors were in differentiating atrophy from NPH, entities with significant overlap in CT appearance.

Conclusion

A novel agentic vision-language model can successfully create and execute a stepwise diagnostic framework for differentiating ventricular abnormalities on CT.

Statement of Impact

By separating visual image search from medical reasoning, the proposed agentic framework enforces explainable stepwise thinking in a manner similar to expert radiologist interpretation.

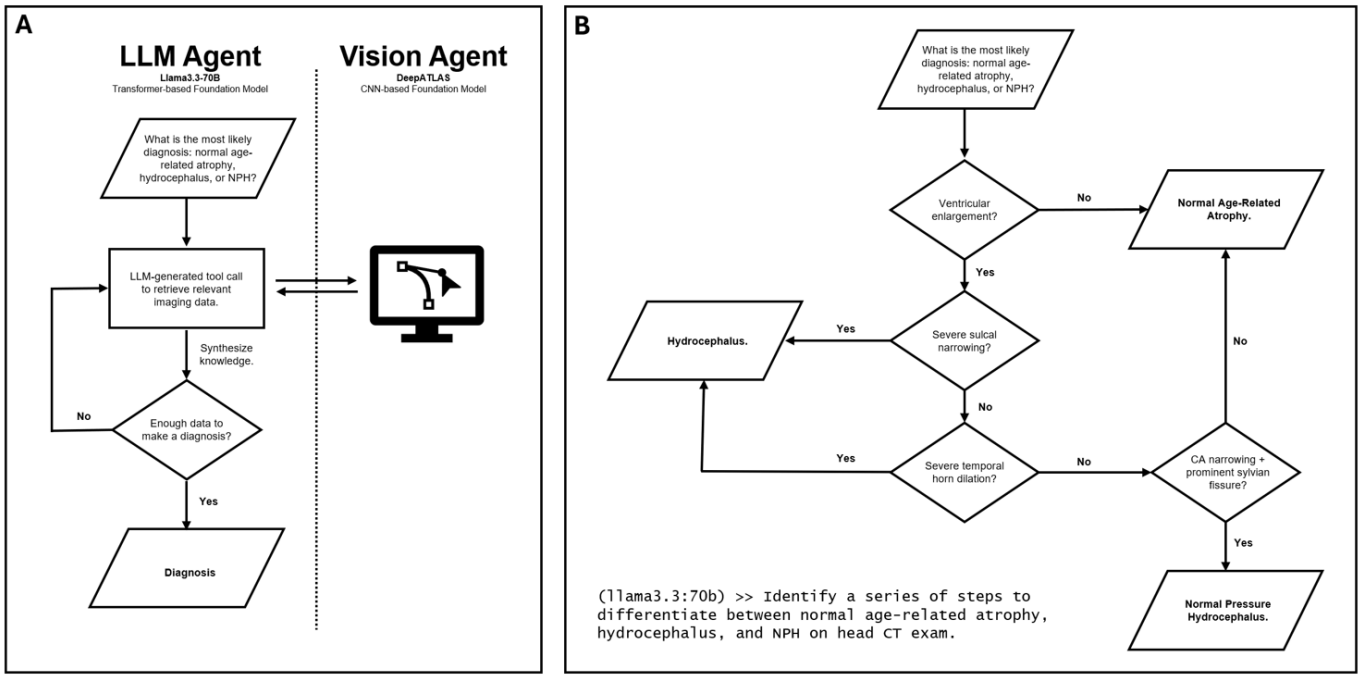

Figure 1. Overview of Agentic Framework. (A) The proposed paradigm separates medical reasoning (LLM agent; Llama3.3-70B) from visual image search (DeepATLAS) into two separate foundation models. Upon receiving a prompt, the LLM agent is equipped with ability to dynamically reason through the problem and generate custom requests to identify relevant imaging findings via queries to the vision agent (tool calls written in Python code). The autonomous agent is allowed to repeatedly probe the image as needed until enough information is gathered to make a diagnosis. (B). This flowchart approximates the stepwise reasoning generated by the LLM agent to work through the diagnostic task. Note that the prompt for generating this response is shown in the bottom left-hand corner of the panel.

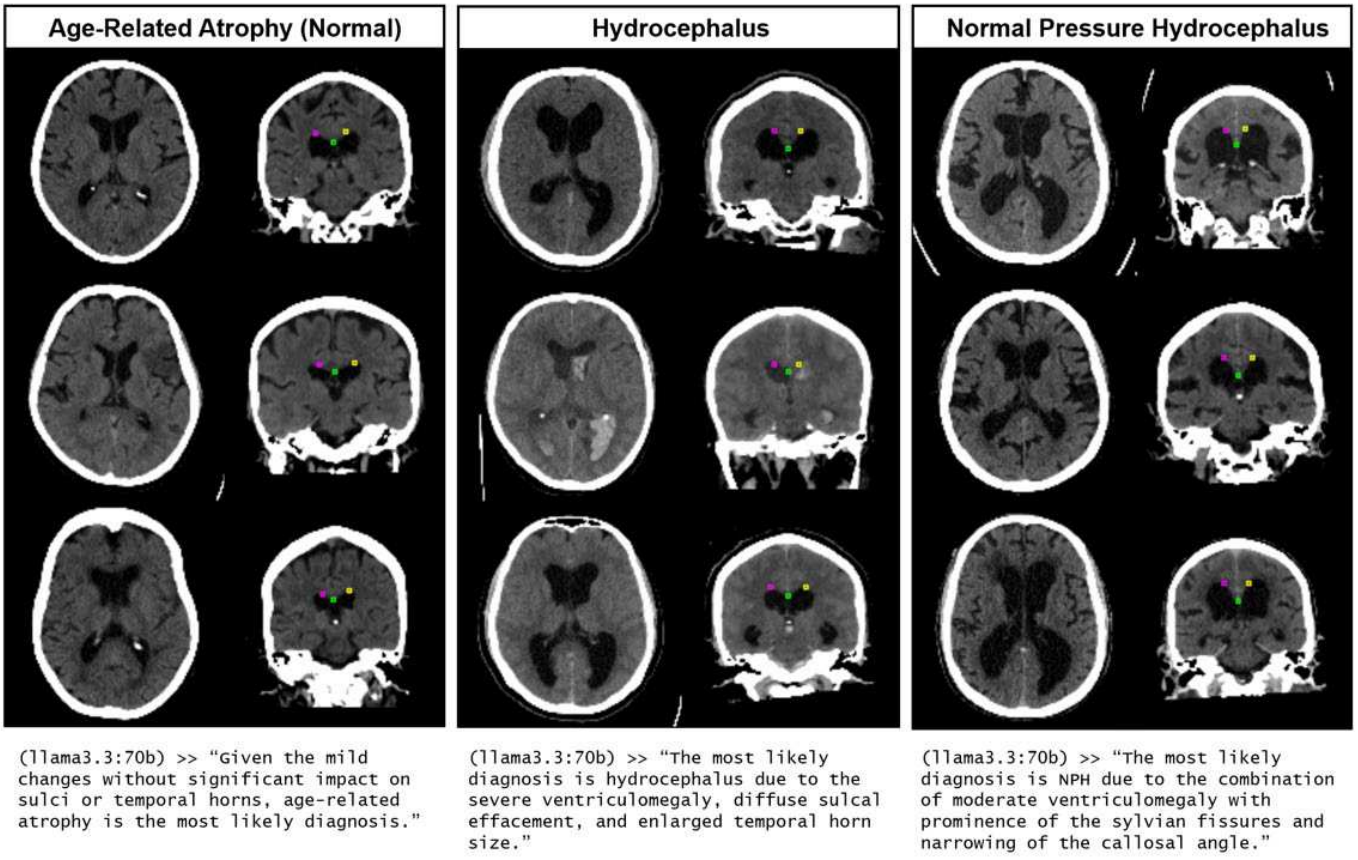

Figure 2. Agent-Generated Predictions. Representative predictions generated by the agentic vision language model are shown across the three diagnoses. In each panel, axial (left) and coronal (right) images are displayed. The coronal images are overlaid

with landmarks generated by the model for calculating the callosal angle. Below each panel is a representative final text generated by the agent which demonstrates synthesis of explainable features for accurate diagnosis.

TARGET / OUTPUT	Atrophy	Hydrocephalus	NPH	SUM
Atrophy	57 40.14%	0 0.00%	7 4.93%	64 89.06% 10.94%
Hydrocephalus	1 0.70%	38 26.76%	2 1.41%	41 92.68% 7.32%
NPH	2 1.41%	0 0.00%	35 24.65%	37 94.59% 5.41%
SUM	60 95.00% 5.00%	38 100.00% 0.00%	44 79.55% 20.45%	130 / 142 91.55% 8.45%

Accuracy	0.9155
Misclassification Rate	0.0845
Macro-F1	0.9152
Weighted-F1	0.9173

Class Name	Precision	1-Precision	Recall	False Negative Rate	F1 score	Specificity (TNR)	False Positive Rate (FPR)
Atrophy	0.9500	0.0500	0.8906	0.1094	0.9194	0.9615	0.0385
Hydrocephalus	1.0000	0.0000	0.9268	0.0732	0.9620	1.0000	0.0000
NPH	0.7955	0.2045	0.9459	0.0541	0.8642	0.9143	0.0857

Table 1. Summary of Model Performance Statistics.

Keywords

Agentic AI; Vision Language Model; Foundation Model; Zero-shot Diagnosis; Head CT; Ventriculomegaly