

IMPERIAL

Self-Directed Multimodal AI: Multi-Turn Reasoning Vision-Language Models for Rare Brain Pathology

From one-shot image answering to auditable MRI reading: what the model says, where it points, and how it got there.

Dr Duaa Alim* Imperial College London · Liam Chalcroft University College London · *dalim@imperial.ac.uk



Preprint
(arXiv)



LinkedIn

906

brain MRI cases
NOVA benchmark

281

rare conditions
long-tail neurology

58.2

mAP @ IoU 0.3
Gemini 3 Flash, Full

34.9%

Top-1 diagnosis
image-only joint protocol

0

task-specific fine-tunes
off-the-shelf model

01 The clinical problem

Rare neurological disease is the long tail of brain MRI. Vision-language models trained on common pathology see rare conditions infrequently.

One-shot VLMs do not read like radiologists. A single forward pass, one answer, no opportunity to revise, no record of where attention went.

GAZE evaluates the workflow, not just the answer. Viewer-level tools and literature retrieval, with schema-validated outputs. Every step is logged for audit.

02 What we measure

Localisation mAP at IoU 0.3, 0.5, and 0.5:0.95. Box overlap with the radiologist reference.

Diagnosis Top-1 and Top-5 against the final NOVA diagnosis, graded by an LLM judge (gpt-4o).

Caption METEOR and clinical-F1 against the NOVA caption.

Joint protocol All three are scored from the image alone. No oracle caption supplied (unlike the original NOVA Task 3).

03 Same model, same MRI, different reading workflow

Case A · Pilocytic astrocytoma · Coronal T1 post-contrast · Gemini 3 Flash (no fine-tuning)

Baseline: one forward pass

Gemini 3 Flash, single-turn



Finding Cystic cerebellar tumour with mural nodule.

Anatomical location Cerebellum (the box is at the lower edge, off the lesion).

Diagnosis Pilocytic astrocytoma (correct)

Localisation IoU 0.00

Workflow Single forward pass · no tools used

GAZE workflow: 9 turns, 6 tools

Same model. Tool sequence reset, zoom, window-level, search_web



Finding Cystic cerebellar lesion with mural nodule, posterior fossa.

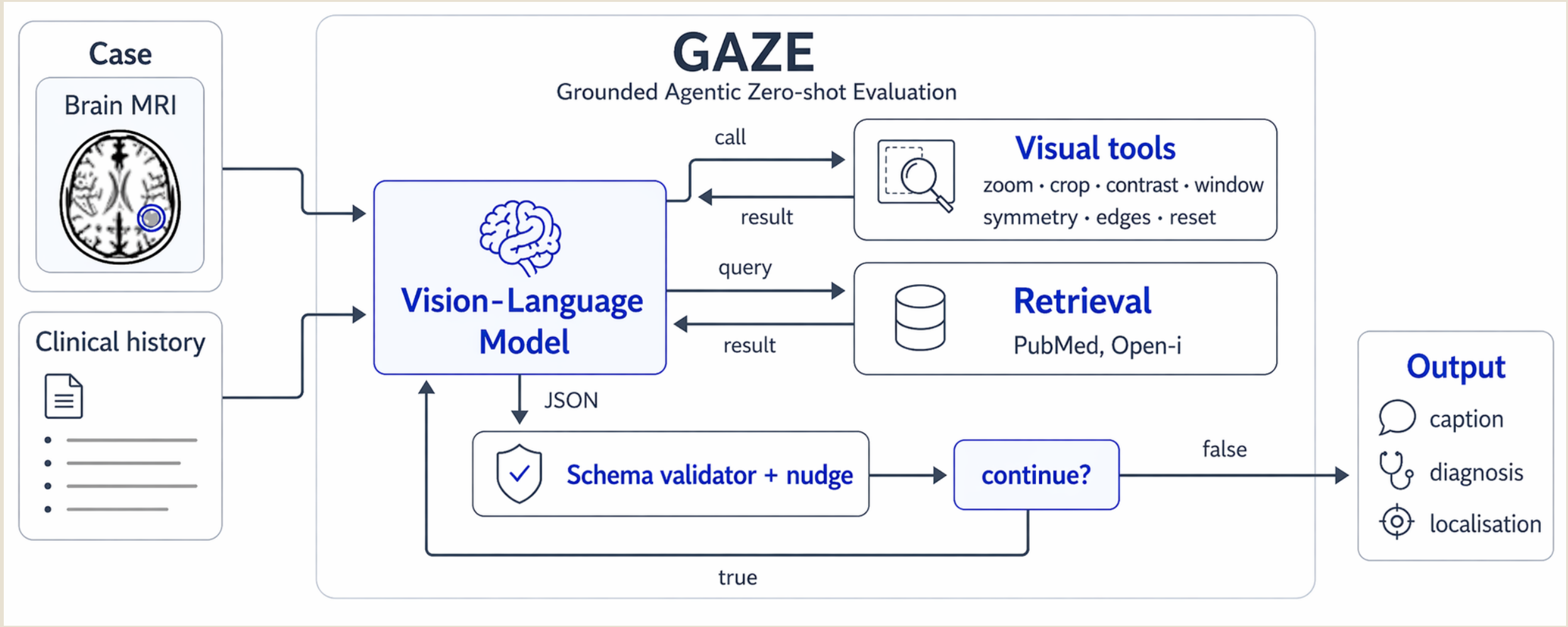
Anatomical location Cerebellum (box now centred on the lesion).

Diagnosis Pilocytic astrocytoma (correct)

Localisation IoU 0.89

Workflow 9 turns · zoom · window-level · search_web · measure · reset

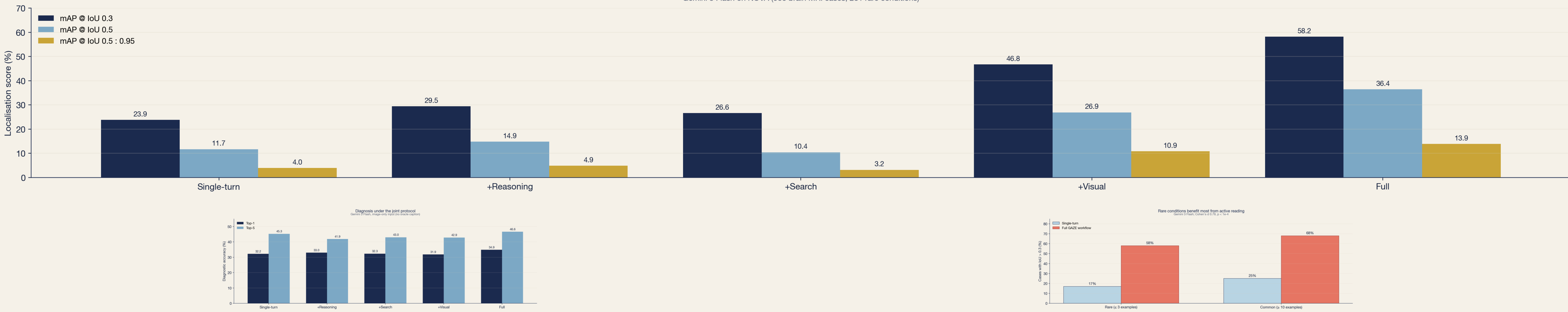
04 How GAZE reads a scan



05 Results

All numbers from results/config/summary.json. 906 brain MRI cases. No task-specific fine-tuning.

Lesion localisation: same model, different reading workflow
Gemini 3 Flash on NOVA (906 brain MRI cases, 281 rare conditions)

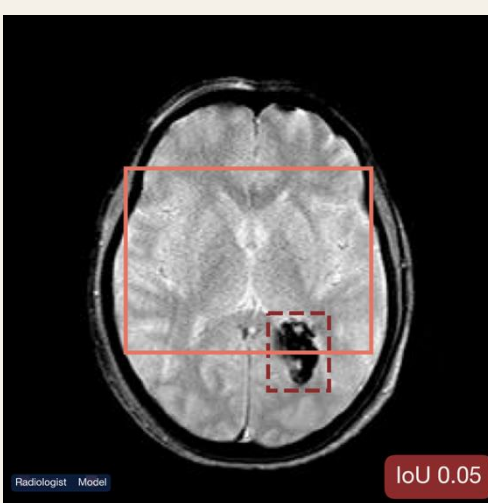


06 Hidden failure mode

Retrieval can lift the diagnostic label while moving the box away from the lesion.

Diagnosis-only scoring would call this case an improvement.

Case B · Calcified intraventricular meningioma · Qwen3.5-72B · before vs after literature search



Localisation IoU 0.58 → 0.05 Diagnosis Cavernoma (wrong) → Meningioma (correct, caption hedges)

SAFETY INSET

07 What this means for clinical AI

Localisation matters.

A correct label is not enough if the model points to the wrong anatomy. RCR-ready evaluation has to score both.

Rare conditions benefit from active inspection.

Tool use lifts IoU > 0.3 cases from 17% to 58% on diagnoses with three or fewer training examples, more than for common conditions.

Audit trails become evaluation data.

Every windowing, zoom, retrieval, and final answer is recorded. The trace can be reviewed after inference.

08 Limitations

Brain MRI only. NOVA spans 281 rare neurological conditions on 906 cases; generalisation to CT or other organs is not evaluated. **Absolute performance is preclinical.** 34.9% Top-1 under the joint protocol sits well below clinical-ready detection. **Diagnosis and localisation can diverge.** The retrieval trade-off (Case B) is real and model-dependent.

Hosted models cannot be fully controlled. Pretraining exposure to NOVA images cannot be ruled out for hosted commercial models; the open-weight Qwen3.5 family is included as an independent check. **Multi-turn inference adds cost and latency.** Full configuration: 11.8 tool calls per case for Gemini 3 Flash.



Preprint



Presenter

Dr Duaa Alim School of Medicine, Imperial College London
Liam Chalcroft Wellcome Centre for Human Neuroimaging, UCL
dalim@imperial.ac.uk

RCR Global AI Conference 2026
Royal College of Radiologists
Imperial College London · UCL