

Evidence-Guided Generation of MRI Reports for Alzheimer's Disease Using Convolutional Neural Networks and Large Language Models

Devesh Singh¹, Dhanush H. Babu², Mathias Manzke¹, Marc-André Weber¹, Majid Ramedani³, Arijana Bohr², Emmanuelle Salin², Björn Eskofier², Stefan Teipel^{3,4}, Martin Dyrba³

Contact: Devesh.Singh@med.uni-rostock.de

¹ Institute of Radiology, Pediatric Radiology & Neuroradiology, University Medical Center Rostock, Germany ² Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

³ German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany ⁴ Clinic for Psychosomatic Medicine, Rostock University Medical Center, Rostock, Germany

Introduction

- Although CNNs effectively detect pathological MRI patterns and LLMs can generate human-readable interpretations, the lack of CNN interpretability and the hallucination risk of LLMs hinder clinical application.
- Large-scale paired MRI and radiology report datasets are scarce, limiting the development of automated report-generation systems in neuroimaging.
- We present a hybrid framework (Figure 1) integrating morphological clinical domain knowledge with LLMs to generate guideline-grounded radiological impressions.
- Goal:** Test whether a rules-plus-RAG approach yields guideline-consistent justifications, and characterise the hallucinations remain.

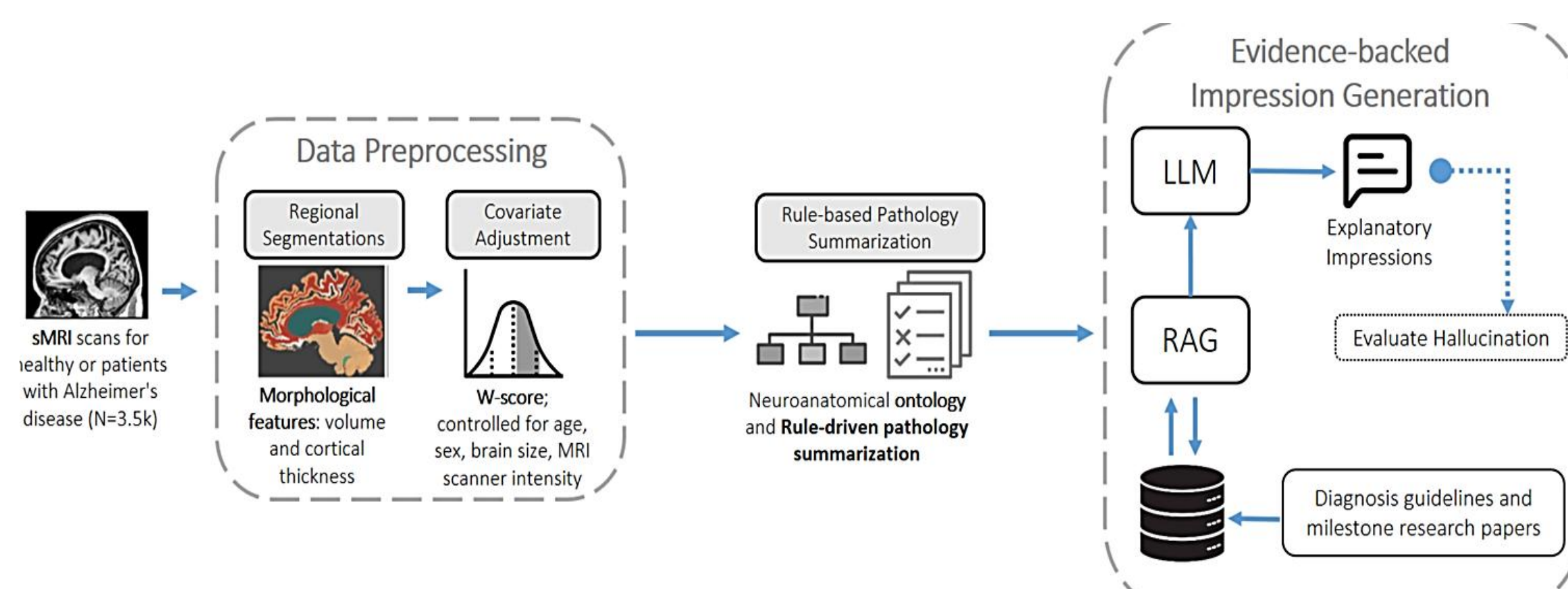
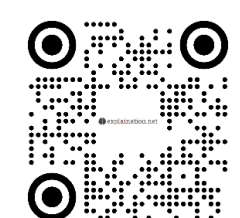


Figure 1. End-to-end framework: CNN-based morphology scoring, rule-based pathology summarization, and a RAG-grounded LLM for guideline-driven impression generation with hallucination evaluation.

Find more details on our website:
[explaination.net](https://www.explaination.net)



Find our source code at:
qr.codes/HQT2cE



Reference: Singh, Babu et al. 2026. MIDL 2026.
<https://openreview.net/forum?id=kEdSwW10G>

Methods

Features: FastSurfer segments scans; gray-matter volume & cortical thickness are adjusted for age, sex and brain size into W-scores (covariate-adjusted z-scores).

Rule-based Pathology Summarization: Retains regions with W-score > 2 SD, collapses subregions, and emits templated text with mild / moderate / strong severity labels.

RAG grounding: An LLM cast as a "radiology trainee" is conditioned on NIA-AA recommendation papers and the German DGN dementia diagnostic guideline to justify findings.

Model: Quantized Llama-3 70B Instruct selected after comparing open-source and medically fine-tuned LLMs.

Clinical correctness: Hallucinations were manually assessed in 30 representative patients.

Example AD impression; claims audited against diagnostic guidelines

Case 3: Alzheimer's disease (AD)

FINDINGS:

- Left Temporal Lobe: Moderate pathology with volume w-score: -2.85, relevance w-score: -4.68, cortical thickness w-score: -2.95. **The pattern of atrophy is moderate, with a significant reduction in volume and cortical thickness, which may be indicative of neurodegenerative changes. (Alzheimer's Dementia - 2011 - Albert - The diagnosis of mild cognitive impairment due to Alzheimer's disease.pdf)**
- Right Temporal Lobe: Mild pathology with volume w-score: -2.79, cortical thickness w-score: -2.77. **The pattern of atrophy is mild, with a moderate reduction in volume and cortical thickness, which may also be indicative of neurodegenerative changes. (Alzheimer's Dementia - 2011 - Albert - The diagnosis of mild cognitive impairment due to Alzheimer's disease.pdf)**

IMPRESSION:

- The observed atrophy patterns in the temporal lobes fit into the category of neuroimaging biomarkers for neurodegenerative diseases. According to the NIA-AA Research Framework, these changes can be classified as indicative of neurodegenerative pathology. (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf)
- **The severity of the atrophy patterns observed in this case may be indicative of an early or preclinical stage of neurodegenerative disease, as the framework suggests that such changes can occur before the onset of symptoms. (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf)**
- The pattern and severity of atrophy observed in this case may be indicative of an underlying neurodegenerative process, which could have implications for cognitive decline and other neurological symptoms. However, it is essential to note that these findings do not necessarily diagnose a specific disease. (Alzheimer's Dementia - 2011 - Albert - The diagnosis of mild cognitive impairment due to Alzheimer's disease.pdf)

Legend: ■ Correct claim ■ Hallucinated claim

Figure 2. Falsifiable RAG taxonomy. Each claim is scored on *citation presence*, *information source*, and *correctness* — yielding seven scenarios from **no hallucination** to **pure fabrication**.

Results

Table1. Hallucination rates by diagnostic group — Llama-3 70B

Hallucination case	AD (N=10) (u=63)	MCI (N=8) (u=43)	CN(N=4) (u=22)	FTD(N=8) (u=37)	Macro Averages
No hallucinations	39.7%	44.2%	77.3%	60%	50.4%
Attribution error ²	14.3%	4.7%	18.2%	5.4%	10.3%
Contradiction hallucinations	23.8%	27.9%	4.5%	27.0%	23.0%
Reference hallucinations	4.8%	20.9%	0.0%	8.0%	9.1%
Extrinsic hallucinations	0.0%	2.3%	0.0%	0.0%	0.6%
False reference hallucinations	3.2%	0.0%	0.0%	0.0%	1.2%
Pure fabrication	14.3%	0.0%	0.0%	0.0%	5.4%

² Attribution errors are uncited but guideline-correct, so not counted as hallucinations.

- Instruction-tuned models consistently followed prompts (Figure 2); smaller medical models (<30B) often failed to generate clinically helpful reports.
- Mistral-Instruct and Llama-Instruct showed strong coherence; Qwen demonstrated strong reasoning but lower controllability. Thus Llama-3 70B Instruct was chosen as the final model.
- Finding:** Hallucinations remain common: 46% AD, and 51% MCI claims were hallucinated. However, only 4.5% CN were hallucinated.

Conclusions

- First end-to-end framework for generating radiological impressions directly from brain MRI using W-scores, rule-based neuroanatomical abstraction, and RAG.
- Rule-based guardrails plus RAG improve auditability, but do not sufficiently mitigate hallucinations; 35–50% of claims in pathological cases were incorrect.**
- Result demonstrates automation potential while underscoring the continued need for human oversight when deploying open-source LLMs for clinical reporting.