

Evidence-Supported Automated Impressions for Alzheimer’s Disease Detection from Brain MRI: A Feasibility Study

Devesh Singh¹ 

DEVESH.SINGH@MED.UNI-ROSTOCK.DE

¹ *Institute and Policlinic of Radiology, Pediatric Radiology and Neuroradiology, University Medical Center Rostock, Rostock, Germany*


Dhanush H. Babu²

DHANUSH.BABU.HAREESH@FAU.DE


² *Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*

Mathias Manzke¹ 

MATHIAS.MANZKE@MED.UNI-ROSTOCK.DE


Marc-André Weber¹ 

MARC-ANDRE.WEBER@MED.UNI-ROSTOCK.DE

Majid Ramedani³ 

MAJID.RAMEDANI@DZNE.DE


³ *German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany*

Arijana Bohr² 

ARIJANA.BOHR@FAU.DE

Emmanuelle Salin² 

EMMANUELLE.SALIN@FAU.DE

Björn Eskofier² 

BJOERN.ESKOFIER@FAU.DE

Martin Dyrba³ 

MARTIN.DYRBA@DZNE.DE

Editors: Under Review for MIDL 2026

Abstract

Convolutional Neural Networks (CNNs) are the standard models for neuroimaging analysis, but their opacity hinders clinical adoption. While Large Language Models (LLMs) offer a potential solution to translate CNN outputs as human-readable impressions, their reliability remains questionable. In the context of Alzheimer’s Disease (AD) detection, we introduce a framework that computes CNN-based brain morphology scores and leverages a rule module for summarization. Using an LLM, conditioned on diagnostic guidelines via Retrieval-Augmented Generation (RAG), we generate explanatory justifications of pathology detected. To assess the impact of hallucinations, we propose a taxonomy that considers generated justifications as falsifiable claims. Manual evaluation on 30 reports shows that hallucinations remain substantial. In the pathological cases, from 34–50% of claims were incorrect. Our feasibility study shows that integrating rule-based guardrails with RAG improves auditability but fails to sufficiently mitigate hallucinations.

Keywords: LLM, Radiological Impressions, Dementia, MRI, RAG, Hallucination

1. Introduction

CNNs are widely used in medical imaging for diagnosing neurodegenerative diseases, and have shown strong performance in identifying pathological MRI patterns (Alsubaie et al., 2024). Correspondingly, LLMs offer promise in generating human-readable explanations (Zhou et al., 2025; Tian et al., 2023), but as they often hallucinate, their usability in clinical practice remains uncertain (Hager et al., 2024). In our feasibility study, we present a hybrid framework that integrates clinical domain knowledge with LLMs to generate radiological impressions of disease detection based on structural MRIs (see Figure 1).

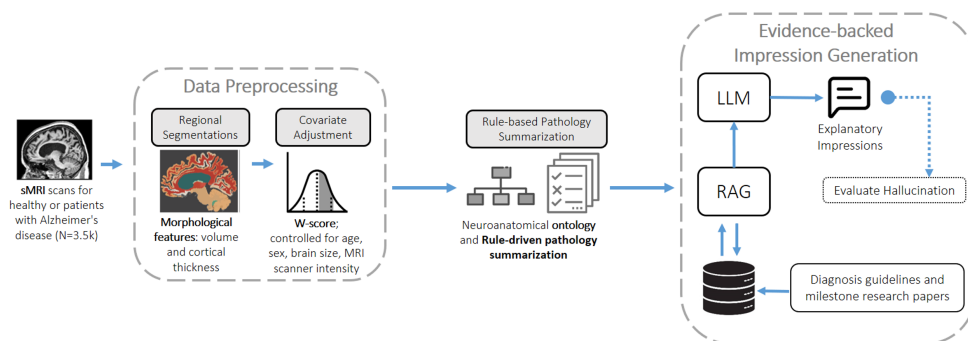


Figure 1: A framework including CNN detection, rule-based templates, and a RAG-grounded LLM for automated guideline-driven impression generation

Since no public dataset jointly contains MRI scans and corresponding radiologist reports, we first generated rule-based pathology summaries using morphological rules. The summary structure was validated by expert radiologists before being used to generate MRI impressions. We assess if the use of a RAG module to ground LLM responses in diagnostic guidelines can improve reliability. Our study evaluates the feasibility of end-to-end MRI-based radiological summary generation and whether a rules-plus-RAG approach produces guideline-consistent diagnostic justifications.

2. Methods and Results

We analyzed 3,433 MRIs from multiple cohorts: 47% cognitively normals (CN), 33% with mild cognitive impairment (MCI), 16% with Alzheimer’s disease (AD), and 4% with frontotemporal dementia (FTD) (App A). FastSurfer segmented scans into neuroanatomical regions and estimated gray-matter volumes and cortical thickness, which were adjusted for age, sex, and brain size to produce W -scores, a covariate-adjusted form of z -score reflecting deviation from the expected distribution. A previously developed rule-based system summarized each patient’s abnormalities by retaining regions with W -scores > 2 SD and collapsing subregions into higher-level structures to reduce verbosity, yielding templated text with graded severity (mild/moderate/strong) labels (Singh et al., 2025). Structured findings were translated into clinically grounded justifications using LLMs via RAG, with prompts engineered to emulate radiologist-style reports. The RAG module was conditioned on four NIA-AA diagnostic recommendation papers (Jack Jr et al., 2024) and the German Neurological Society’s (DGN) diagnostic guidelines (DGN e. V. and DGPPN e. V., 2025). The prompts cast the model as a “radiology trainee” to encourage clear, didactic explanations (App. B). We propose a reporting structure where each impression begins by reiterating findings, followed by guideline-based justifications and then lists limitations (App. C). The source code of our framework is available via GitHub¹.

Clinical correctness was evaluated by manually measuring hallucinations in 30 representative patients across disease spectrum. In a micro-averaged analysis, each claim was

1. GitHub repository: <https://github.com/DhanushBabu18/MRIs-to-Radiological-Dementia-Reports/>

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%

evaluated for citation presence, information source, and overall claim correctness, yielding seven scenarios: (1) *No hallucination*: cited, RAG-grounded, true claim; (2) *Attribution error*: uncited, rag-grounded, true claim; (3) *Contradiction hallucination*: cited, RAG-grounded, false claim (e.g., misinterpretation of the source); (4) *Reference hallucination*: cited, from unknown source, true claim (i.e., fabricated citation); (5) *Extrinsic hallucination*: uncited, from unknown source, true claim; (6) *False reference hallucination*: cited, from unknown source, false claim (i.e., fabricated citation attached to a false claim); (7) *Pure fabrication*: uncited, from unknown source, false claim. Collectively, these categories define our falsifiable hallucination taxonomy for RAG systems (App. D).

We tested several open-source LLMs and performed manual assessments of coherence. Instruction-tuned models best fit our needs, reliably following explicit instructions, whereas medically fine-tuned models under 30B parameters often failed to respond or produced clinically unhelpful outputs (App. E). Among general-purpose models, Mistral-Instruct and Llama-Instruct variants were notably coherent, and Qwen showed strong reasoning but was harder to control. We therefore selected the quantized Llama-3 70B Instruct model for our framework. Using our hallucination taxonomy, we report hallucination rates for representative patients of each disease type and stage, in the table above.

3. Discussion

We combine W -scores for brain morphology analysis, rule-based neuroanatomical abstraction, and RAG to generate pathological impression reports. Despite grounding the generation in medical guidelines, hallucinations² remain common (46% AD, 51% MCI claims). The prevalence of contradictions and reference hallucinations underscores the need to explore more advanced reasoning models and curate more comprehensive RAG sources. Cognitively normal cases showed far fewer hallucinations (4.5%). Existing RAG hallucination mitigation frameworks (Es et al., 2024; Asai et al., 2024), do not systematically characterize how citation behavior, information source, and claim correctness jointly shape hallucination modes (App. F compares it with prior taxonomies). We address this with a RAG-tailored hallucination taxonomy and, to our knowledge, the first end-to-end pipeline for automated radiological impression generation from MRI scans, demonstrating automation potential while underscoring the need for human oversight in using open-source LLMs.

2. Attribution errors were not counted as hallucinations, since the underlying uncited claims still correctly reflected the diagnostic guidelines provided as RAG sources.

References

- Mohammed G Alsubaie, Suhuai Luo, and Kamran Shaukat. Alzheimer’s disease detection using deep learning on neuroimaging: a systematic review. *Machine Learning and Knowledge Extraction*, 6(1):464–505, 2024.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.
- DGN e. V. and DGPPN e. V. S3 guideline on dementia, version 5.2. <https://register.awmf.org/de/leitlinien/detail/038-013>, July 2025. Accessed: 2025-08-01.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In Nikolaos Aletras and Orphee De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.16. URL <https://aclanthology.org/2024.eacl-demo.16/>.
- Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL <https://doi.org/10.1145/3703155>.
- Clifford R Jack Jr, J Scott Andrews, Thomas G Beach, Teresa Buracchio, Billy Dunn, Ana Graf, Oskar Hansson, Carole Ho, William Jagust, Eric McDade, et al. Revised criteria for diagnosis and staging of alzheimer’s disease: Alzheimer’s association workgroup. *Alzheimer’s & Dementia*, 20(8):5143–5169, 2024.
- Devesh Singh, Yusuf Brima, Fedor Levin, Martin Becker, Bjarne Hiller, Andreas Hermann, Irene Villar-Munoz, Lukas Beichert, Alexander Bernhardt, Katharina Buerger, et al. An unsupervised xai framework for dementia detection with context enrichment. *Scientific Reports*, 15(1):39554, 2025.
- Dianzhe Tian, Shitao Jiang, Lei Zhang, Xin Lu, and Yiyao Xu. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 14(1):1108, 2023.
- Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Yi Fang, Sirui Ding, Jiashuo Wang, Kaishuai Xu, et al. Large language models for disease diagnosis: A scoping review. *npj Artificial Intelligence*, 1(1):9, 2025.

Appendix A. Data Cohort Description

In this study, T1-weighted volumetric MRI scans were collected from seven cohorts: the Alzheimer’s Disease Neuroimaging Initiative (ADNI; phases ADNI2 and ADNI3), the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), the DZNE Longitudinal Study on Cognitive Impairment and Dementia (DELCODE), the European DTI Study on Dementia (EDSD), the DZNE Clinical Registry Study on Frontotemporal Dementia (DESCRIBE-FTD), and the Frontotemporal Lobar Degeneration Neuroimaging Initiative (FTLDNI), also referred to as the Neuroimaging Initiative in Frontotemporal Dementia (NIFD). In total, the study comprised 3,433 MRI scans. Demographic characteristics across diagnostic groups are summarized in Table 1.

Table 1: Demographic and clinical characteristics by diagnostic group. *Values are mean \pm SD. CN, cognitively normal; MCI, mild cognitive impairment; AD, Alzheimer’s disease dementia; FTD, Frontotemporal dementia; MMSE, Mini-Mental State Examination.*

Variable	CN (N = 1625)	MCI (N = 1132)	AD (N = 549)	FTD (N = 127)
Age (years)	70.4 \pm 7.6	72.5 \pm 7.4	74.1 \pm 7.7	62.8 \pm 8.2
MMSE score	29.1 \pm 1.1	27.6 \pm 2.0	22.2 \pm 4.2	23.7 \pm 5.7
Education (years)	15.7 \pm 3.0	15.3 \pm 3.2	13.8 \pm 3.9	14.7 \pm 3.3
Sex (F/M)	922 / 703	504 / 628	280 / 269	45 / 82

Appendix B. Prompt Structure

This appendix shows the prompt used (see Fig.2) to generate structured radiology summaries from rule-based template pathology summaries (see Fig.3) as input text. In this setup, the model is constrained to act as a trainee radiology assistant with access only to the diagnostic guideline documents and a clinical significance file (a json file highlighting each regions clinical and functional use). The LLM is instructed to (1) produce a structured FINDINGS section summarizing regional abnormalities and W-scores, (2) generate a guideline-based, multi-paragraph IMPRESSION relating patterns and severity to established biomarker and staging frameworks, and (3) attach explicit filename-based citations to every claim, marking unsupported statements with “Not supported in available sources” and forbidding any use of outside knowledge.

```

prompt-f'''
You are a trainee radiology assistant.

You have access ONLY to the following documents:
1. Alzheimer s Dementia - 2011 - Jack - Introduction to the recommendations from the National Institute on Aging-Alzheimer s.pdf
2. Alzheimer s Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer s disease Recommendations from the.pdf
3. Alzheimer s Dementia - 2011- Albert - The diagnosis of mild cognitive impairment due to Alzheimer s disease.pdf
4. Alzheimer s Dementia - 2010 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer s disease.pdf
5. DGM Guidelines Diagnosis.pdf
6. Clinical_Significance.json

If a claim is not explicitly supported in one of these documents, say:
"[Not supported in available sources]".
Do NOT use outside knowledge or make assumptions.

---
PATIENT FINDINGS:
(Patients_findings)
---
|
TASK:
1. Write **FINDINGS**: summarize abnormalities with W-scores. For each region,specify the pattern, severity, and w-scores, and briefly discuss the potential clinical and pathophysiological relevance, citing one of the six sources by filename.
2. Write **IMPRESSION**:
   - Provide a **multi-paragraph, guideline-based summary**:
     - Describe how the observed atrophy patterns fit into the categories of neuroimaging biomarkers and how the framework classifies such changes .
     - Relate the severity to how the framework discusses early, preclinical, or symptomatic stages.
     - Discuss the pattern and severity of atrophy and its implications for neurodegenerative diseases but do not diagnose any disease.
     - Add a short paragraph explaining the clinical significance about each regions in the findings in detail based on this document Clinical_Significance.json. If no document supports it, clearly state: "Not supported in available sources."
3. Every claim must be followed by a citation in the format: (Filename, page/section if available).

Do NOT cite anything outside the five listed documents. Do NOT invent facts.

OUTPUT FORMAT:
FINDINGS:
- -
IMPRESSION:
- -
[IMPORTANT]Show step by step before giving the output
[REMEMBER]Only use the documents to claim and do not claim anything outside the documents
'''

```

Figure 2: Prompt template used in the study, constraining the model to predefined diagnostic guidelines while generating structured *findings* and *impression* sections with instructions to cite references.

Moderate pathology in **atrophied** Left Temporal Lobe (volume w-score: -2.86, CNN relevance w-score: -4.69, cortical_thk w-score: -2.95)
Mild pathology in **atrophied** Right Inferior Temporal (volume w-score: -2.21, CNN relevance w-score: -2.65 , cortical_thk w-score: -2.20)
Mild pathology in **enlarged** Left Inf-Lat-Vent (volume w-score: 3.02, CNN relevance w-score: -2.39)

Figure 3: Example of the rule-based template text provided to the LLM. The template comes from a rule-based system that summarizes abnormalities by keeping regions with W -scores $> 2SD$ and merging subregions into higher-level structures, producing concise, severity-graded pathology descriptions.

Appendix C. LLM impression responses for each disease type

Figures 4-7 present LLM-generated impression summaries for representative Alzheimer’s Disease (AD), Mild Cognitive Impairment (MCI), Cognitively Normal (CN), and Frontotemporal Dementia (FTD) cases, with claims marked as non-hallucinated (green) or hallucinated (red).

Case 1: Cognitively Normal (CN)

FINDINGS:

- No abnormalities or pathology present.

IMPRESSION:

The patient’s neuroimaging results do not show any evidence of atrophy or abnormalities. According to the NIA-AA Research Framework, neuroimaging biomarkers are used to assess the presence and severity of Alzheimer’s disease (Alzheimer’s Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer’s disease.pdf). However, in this case, there are no biomarkers present to indicate any neurodegenerative changes.

The framework classifies neuroimaging biomarkers into three stages: preclinical, early symptomatic, and late symptomatic (Alzheimer’s Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer’s disease.pdf). Since there are no abnormalities present, the patient’s results do not fit into any of these categories.

The absence of atrophy or abnormalities in the patient’s neuroimaging results suggests that there is no evidence of neurodegenerative disease at this time. However, it is essential to note that the absence of evidence does not necessarily rule out the possibility of disease.

Clinical significance of the regions cannot be discussed as there are no abnormalities present in the patient’s neuroimaging results.

No Hallucination
Hallucination

Figure 4: Cognitively Normal

Case 2: Mild Cognitive Impairment (MCI)

FINDINGS:

- Moderate pathology in atrophied Left Temporal Lobe: (volume w-score: -2.18, relevance w-score: -5.13, cortical thickness w-score: -2.27)

IMPRESSION:

The observed atrophy patterns in the Left Temporal Lobe can be classified as a neuroimaging biomarker for neurodegeneration (Alzheimer’s Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer’s disease.pdf).

The severity of atrophy in the Left Temporal Lobe may indicate a preclinical or symptomatic stage of neurodegenerative disease, as discussed in the Framework (Alzheimer’s Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer’s disease.pdf).

The pattern and severity of atrophy in the Left Temporal Lobe may be indicative of neurodegenerative diseases, such as Alzheimer’s disease (Alzheimer’s Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer’s disease Recommendations from the pdf).

The Left Temporal Lobe is involved in processing auditory information and is also important for memory and language functions (Clinical_Significance.json). Atrophy in this region may be associated with cognitive decline and memory impairments.

No Hallucination
Hallucination

Figure 5: Mild Cognitive Impairment

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)

No Hallucination
Hallucination

Figure 6: Alzheimer’s Disease

Case 4: Frontotemporal dementia (FTD)

OUTPUT:

FINDINGS:

- Left Middle Temporal: mild atrophy, volume w-score: -2.51, cortical thickness w-score: -2.29, potentially related to mild cognitive impairment (Albert, 2011).

IMPRESSION:

The observed atrophy patterns in the Left Middle Temporal region can be classified as a neuroimaging biomarker for Alzheimer’s disease (Jack, 2018). The mild severity of the atrophy may indicate an early or preclinical stage of neurodegenerative disease (Jack, 2018). The pattern of atrophy in the Left Middle Temporal region, combined with its mild severity, may indicate a risk for neurodegenerative diseases such as Alzheimer’s disease (McKhann, 2011). The clinical significance of the atrophy in the Left Middle Temporal region is related to cognitive decline, as this region is involved in memory and cognitive functions (Clinical_Significance.json).

No Hallucination
Hallucination

Figure 7: Frontotemporal Dementia

Appendix D. Hallucination Scenarios

This appendix provides illustrative examples of various hallucination type defined in our falsifiable RAG taxonomy. The following samples demonstrate how citation absence, information source, and claim correctness interact to produce distinct error categories. We highlight non-hallucinated (green) or hallucinated (red) claims. See Figures 8-13.

Type 2 : Attribution Error (uncited, rag-grounded, true statement)

FINDINGS:

- Right Temporal Lobe: significant atrophy (volume w-score: -4.08, cortical thickness w-score: -5.05), potentially indicating neurodegenerative changes (McKhann, 2011).

IMPRESSION:

- The observed atrophy patterns in the Right Temporal Lobe can be classified as a neuroimaging biomarker, specifically a measure of neurodegeneration (Jack, 2018).
- The severity of the atrophy, as indicated by the w-scores, suggests that the changes are significant and may be indicative of a symptomatic stage (Jack, 2018).
- The pattern of atrophy in the Right Temporal Lobe, combined with the severity of the changes, suggests that there may be underlying neurodegenerative processes at play. However, we cannot diagnose any specific disease based on this information alone.

Figure 8: Attribution Error

Type 3: Contradiction Hallucination (cited, rag-grounded, false statement, i.e., misinterpreted cited source)

FINDINGS:

- Left Amygdala: Moderate pathology with volume w-score: -3.39, relevance w-score: -2.96, cortical thickness w-score: not available. The atrophy in the left amygdala may be related to emotional processing and memory, as the amygdala is involved in the processing of emotional information (Alzheimer's Dementia - 2011 - Jack - Introduction to the recommendations from the National Institute on Aging-Alzheimer's.pdf).
- Left Hippocampus: Moderate pathology with volume w-score: -5.32, relevance w-score: -2.55, cortical thickness w-score: not available. The atrophy in the left hippocampus may be related to memory formation and spatial navigation, as the hippocampus is involved in the formation of new memories (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf).

IMPRESSION:

- The observed atrophy patterns in the left amygdala and left hippocampus fit into the category of neuroimaging biomarkers for Alzheimer's disease, as described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf). The framework classifies such changes as indicative of neurodegenerative disease.
- The severity of the atrophy in the left amygdala and left hippocampus is moderate, which may indicate an early or preclinical stage of neurodegenerative disease, as described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).
- The pattern and severity of atrophy in the left amygdala and left hippocampus may be indicative of neurodegenerative disease, such as Alzheimer's disease. The atrophy in these regions may be related to cognitive decline and memory impairment (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf).

Figure 9: Contradiction Hallucination

Type 5: Reference hallucination (cited, from unknown source, true statement, i.e., fabricated citation attached to a true claim)

FINDINGS:

- Left Amygdala: Moderate pathology with volume w-score: -3.39, relevance w-score: -2.96, cortical thickness w-score: not available. The atrophy in the left amygdala may be related to emotional processing and memory, as the amygdala is involved in the processing of emotional information (Alzheimer's Dementia - 2011 - Jack - Introduction to the recommendations from the National Institute on Aging-Alzheimer's.pdf).
- Left Hippocampus: Moderate pathology with volume w-score: -5.32, relevance w-score: -2.55, cortical thickness w-score: not available. The atrophy in the left hippocampus may be related to memory formation and spatial navigation, as the hippocampus is involved in the formation of new memories (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf).

IMPRESSION:

- The observed atrophy patterns in the left amygdala and left hippocampus fit into the category of neuroimaging biomarkers for Alzheimer's disease, as described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf). The framework classifies such changes as indicative of neurodegenerative disease.
- The severity of the atrophy in the left amygdala and left hippocampus is moderate, which may indicate an early or preclinical stage of neurodegenerative disease, as described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).
- The pattern and severity of atrophy in the left amygdala and left hippocampus may be indicative of neurodegenerative disease, such as Alzheimer's disease. The atrophy in these regions may be related to cognitive decline and memory impairment (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf).

Figure 10: Reference hallucination

Type 6: Extrinsic hallucination (uncited, from unknown source, true statements)

FINDINGS:

- Right Temporal Lobe: Mild pathology with volume w-score: -2.72, cortical thickness w-score: -2.11. The observed atrophy in the right temporal lobe may be indicative of neurodegenerative changes, which can be associated with cognitive decline (Alzheimer's Dementia - 2011 - Jack - Introduction to the recommendations from the National Institute on Aging-Alzheimer's.pdf).
- Left Temporal Lobe: Mild pathology with volume w-score: -2.47, relevance w-score: -2.30, cortical thickness w-score: -1.99. The atrophy in the left temporal lobe, particularly with a relevance w-score of 2.30, may suggest a more significant impact on cognitive functions, potentially related to Alzheimer's disease (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf).

IMPRESSION:

- The observed atrophy patterns in the temporal lobes can be classified as neuroimaging biomarkers for neurodegenerative diseases, specifically Alzheimer's disease (Alzheimer's Dementia - 2011 - Jack - Introduction to the recommendations from the National Institute on Aging-Alzheimer's.pdf). According to the Framework, these changes can be indicative of early or preclinical stages of the disease (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).
- The mild severity of the atrophy in both temporal lobes suggests that the individual may be in an early or preclinical stage of neurodegenerative disease, as the Framework suggests that more severe atrophy is associated with symptomatic stages (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).
- The pattern of atrophy in the temporal lobes, particularly the right temporal lobe, may be indicative of Alzheimer's disease, as this region is commonly affected in the disease (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf). The mild severity of the atrophy suggests that the individual may not yet be experiencing significant cognitive decline, but the presence of atrophy in this region may indicate an increased risk of developing neurodegenerative disease.

Figure 11: Extrinsic hallucination

Type 7: False reference Hallucination (cited, from unknown source, false statements, i.e., fabricated citation attached to a false claim)

FINDINGS:

- Left Temporal Lobe: Mild pathology with volume w-score: -2.97, relevance w-score: -3.66, cortical thickness w-score: -2.02. The pattern of atrophy is mild, with a moderate level of atrophy, and may be associated with mild cognitive impairment (Alzheimer's Dementia - 2011 - Albert - The diagnosis of mild cognitive impairment due to Alzheimer's disease.pdf).
- Right Temporal Lobe: Strong pathology with volume w-score: -4.57, relevance w-score: not available, cortical thickness w-score: -4.70. The pattern of atrophy is strong, with a severe level of atrophy, and may be associated with more severe cognitive decline (Alzheimer's Dementia - 2011 - McKhann - The diagnosis of dementia due to Alzheimer's disease Recommendations from the pdf).

IMPRESSION:

- The observed atrophy patterns in the left and right temporal lobes fit into the category of neuroimaging biomarkers for Alzheimer's disease, as described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).
- The severity of the atrophy in the right temporal lobe, with a volume w-score of -4.57, suggests a more advanced stage of neurodegenerative disease, potentially corresponding to the symptomatic stage described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf). Structural MRI atrophy alone definitively diagnoses Alzheimer's disease and recommends the clinical stage (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).
- The pattern and severity of atrophy in the left and right temporal lobes suggest a neurodegenerative disease process, with the right temporal lobe showing more severe atrophy. This pattern is consistent with the progression of Alzheimer's disease, as described in the NIA-AA Research Framework (Alzheimer's Dementia - 2018 - Jack - NIA-AA Research Framework Toward a biological definition of Alzheimer's disease.pdf).

Figure 12: False reference Hallucination

Type 8: Pure fabrication (uncited, from unknown source, false statements)

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).
- Unfortunately, the provided document Clinical_Significance_jon does not contain information about the clinical significance of the observed atrophy patterns in the temporal lobes. Therefore, it is not possible to provide a detailed explanation of the clinical significance of these findings based on the document. (Not supported in available source)

Figure 13: Pure fabrication

Appendix E. Hallucinations with medically fine tuned LLM

In this section, we present the hallucination rates for Qwen-3 32B Medical Reasoning (4-bit quantized), a medically fine-tuned large language model (See Table 2). These results allow us to evaluate how domain-specific tuning affects reliability across multiple failure modes.

For context, we compare these outcomes to those of the Llama-3 70B Instruct model, which we selected as our default baseline. Relative to this baseline, Qwen-3 32B Medical Reasoning shows a substantial increase in hallucination rates. For the same set of representative patients, hallucinations³ rose to 52.4% (a 6% increase) for AD patients, 68.9% (a 17.8% increase) for MCI patients, and 72.3%, representing the largest increase of 37.7%, for FTD patients. These findings indicate that, contrary to our expectations, the medical fine-tuning did not yield performance improvements for our use case. Instead, the results suggest that model size remains a critical determinant of reliability, and in our experience, larger-parameter models consistently perform better.

Table 2: Measured hallucination rates for Qwen-3 32B Medical Reasoning LLM. N denotes sample size; u denotes evaluated claims.

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 hallucination	46.0%	29.5%	70.0%	27.7%	40.8%
Attribution error ³	1.6%	1.6%	30.0%	0.0%	5.0%
Contradiction hallucination	42.9%	24.6%	0.0%	30.8%	29.7%
Reference hallucination	1.6%	36.1%	0.0%	32.3%	17.3%
Extrinsic hallucination	4.8%	3.3%	0.0%	1.5%	3.0%
False reference hallucination	0.0%	0.0%	0.0%	1.5%	0.3%
Pure fabrication	3.2%	4.9%	0.0%	6.2%	3.9%

3. Attribution errors were not counted as hallucinations, since the underlying uncited claims still correctly reflected the diagnostic guidelines provided as RAG sources

Appendix F. Validating Hallucination Taxonomy

To contextualize and validate our hallucination taxonomy, we compare it with the framework proposed by Huang et al (Huang et al., 2025). Their survey offers a broad categorization of hallucinations in LLMs, largely assuming a stand-alone generation setting. In contrast, our taxonomy defines hallucinations as falsifiable scenarios arising under RAG, yielding a more operational framework for high-risk, evidence-grounded applications.

Several of Huang et al.’s categories map directly onto our taxonomy. Their *contradiction errors* correspond to our *contradiction* and *contextual* hallucinations, while *factual fabrication* decomposes into *false reference hallucinations* and *pure fabrication*, depending on citation behavior. This correspondence indicates conceptual convergence, with our framework resolving these errors into RAG-specific, operationally distinct cases.

A key conceptual difference lies in the treatment of extrinsic hallucinations. While Huang et al. acknowledge that LLMs may generate true statements sourced outside the provided context, this case is not included in their final taxonomy. Our framework explicitly models uncited, externally sourced true claims as grounding violations and further distinguishes reference hallucinations, where correct statements are paired with fabricated citations—an error type not addressed by Huang et al. These distinctions show that factual correctness alone is insufficient for evaluating corpus-grounded generation.

Our taxonomy also has some limitations. Our taxonomy focuses exclusively on factual grounding under RAG and does not explicitly measure faithfulness hallucinations, such as instruction inconsistency or logical incoherence as response relevance metrics, which are addressed in prior work (Huang et al., 2025; Es et al., 2024). Instead, these aspects were handled indirectly through prompting strategies and rule-based constraints in our pipeline. Additional limitations include reliance on time-consuming, manual claim-level evaluation and residual ambiguity in uncited statements, where distinguishing RAG-grounded content from pretrained knowledge remains challenging.

In summary, our taxonomy complements prior hallucination frameworks by offering a RAG-specific, falsifiable view of grounded generation errors, while highlighting open challenges in modeling instruction-level and reasoning-based failures.