Topic: Framework for Explainable AI (XAI): Textual Explanations for Deep Learning Models using Domain Knowledge and Large Language Models in the context of Dementia Detection from MRI scans

Abstract

Explainable AI (XAI) in dementia detection is essential as it enhances transparency, allowing healthcare professionals to understand and trust AI-driven diagnoses predictions. With the public availability of large language models (LLMs) - such as the ones meant for daily tasks such as GPT3 and GPT4, or the LLMS more tuned for medical tasks such Med-PaLM, there is an increasing trend of testing LLMs suitability for explaining medical AI methods [1,2] or creating new LLM augmented AI methods [3,4]. However, current research has yet to critically understand aspects of these LLMs, such as their robustness, reliability, interpretability, and logical consistency [5,6,11], particularly in high-stakes medical environments [7,8,9]. Infusing medical domain knowledge into LLMs has demonstrated interesting results in terms of improved expert alignment and reduced clinically harmful responses [18], and also have been used in several studies to improve reliability [10,12].

This thesis aims to perform a feasibility study by developing and evaluating a hybrid explanation framework that integrates rule-based logic with Large Language Models (LLMs) to generate textual explanations for dementia diagnosis predictions made by deep learning models [14]. The goal is to see whether it is possible to enhance the description of pathology data in a way that is clinically relevant.

Introduction

Background and context:

The existing deep learning based segmentation model [13] helps to produce a metric called the W-score, which reflects volumetric deviations in brain regions based on MRI scans. It is computed through volumetric analysis and indicates how much a subject's brain structure differs from a healthy baseline, while controlling for common covariates such as age, sex and brain size. These W-scores are generated for multiple brain regions and are used to assess atrophy patterns associated with dementia. By using W-scores, the model aims to determine whether a subject shows signs of dementia.

As part of the framework, we will also incorporate Retrieval-Augmented Generation (RAG) techniques to enrich LLM outputs with domain-specific knowledge. In particular, clinical guidelines from the Deutsche Gesellschaft für Neurologie (DGN) will be used to ground the explanations in validated medical context [14], thereby improving their reliability, interpretability, and practical utility in diagnostic settings. In addition, established diagnostic guidelines for Alzheimer's disease from the NIA-AA criteria and research framework will be included in RAG knowledge base [15][16][17].

The rule-based method would apply predefined logical conditions to volume w-scores. For instance, regions are marked abnormal if w-scores exceed a threshold of two standard deviations. By using an anatomical hierarchy, the rule-based method would collapse multiple affected sub-regions into higher-level areas and classifies pathology severity as mild, moderate, or strong based on the w-scores [13].

This thesis will assess if combining rule-based methods with LLMs is capable of producing superior explanations for dementia detection. The model's performance will be quantitatively assessed, using methods such as perturbation analysis, on ability to understand anatomical regions and pathological process of brain atrophy. The qualitative evaluations will be conducted by neurologists who will give feedback on the clinical usefulness of explanations and give recommendations on how to possibly improve the explanations. The goal is to determine whether this combined approach can produce sufficiently reliable explanations to support clinical usage.

Research questions:

1. How do purely data-driven LLMs, LLMs augmented with rule-based logic, and purely rule-based systems compare regarding their explanations of dementia detection from MRI scans?
2. Quantitative evaluation of the feasibility study - Can LLMs be controlled via feedback from the ruled based logic? Does this enable them to recognize and correct their own mistakes? We want to compare the explanations in terms of accuracy, consistency and clinical relevance.
3. Qualitative Evaluation - Explore whether LLMs enhance the logical consistency of explanations, and evaluate their clinical usefulness and trustworthiness by presenting them to neurologists and collecting their expert feedback.

Experiment order	**Generic LLM**	**LLM fine-tuned on medical use cases**
1	Base LLM [single shot prompting]	Medical text finetuned LLM [single shot prompting]
2	Base LLM + RAG on The German Neurological Society (DGN)'s disease guidelines)	Medical text finetuned LLM + RAG on DGN
3	Base LLM + RAG on DGN + ontology's hierarchical rules	Medical text finetuned LLM + RAG on DGN + ontology's hierarchical rules)

Table 1: Research objective for comparison of different framework

Tasks

- Literature Research: A PRISMA (inspired) systematic study, with explicitly defined inclusion and exclusion criteria. Measurable outcome: Identifying 20 relevant papers of LLMs for explanation generation and reporting the model types, datasets, and identifying the metrics used for LLM evaluation.
- Develop a proof-of-concept framework using two open source LLMs - one generic (for e.g. Mistral AI) and another finetuned for medical use-case (for e.g. BioMistral AI) to create rule-based explanations using prompt engineering by incorporating Retrieval-Augmented Generation (RAG) over the German Neurological Society's (Deutsche Gesellschaft für Neurologie - DGN) disease guidelines [14] and other medical text [15-17] (refer Fig 1).

The framework will be evaluated against the medical accuracy of the explanation by the metrics identified by the literature research. Additionally, the framework will be benchmarked against standalone LLM and rule-based approaches [13] to evaluate the added value of their integration.

- Quantitative framework evaluation, using public datasets – ADNI [19], AIBL [20], and DELCODE [21] using metrics identified earlier by literature survey:

    1. Automate the similarity measurement on the full dataset (N=~4000 samples) using perturbation analysis - comparing LLM outputs before and after perturbation, incorporating similarity in embedding space and N-gram metrics.
    2. Measure rate for hallucination on a subset of samples (N=~30) with proxy ground truth [7].

- Qualitative framework evaluation. Collective feedback from neurologists on (N=3-5) MRI scans with subjects on the AD continuum.

- Documentation and Thesis Writing.

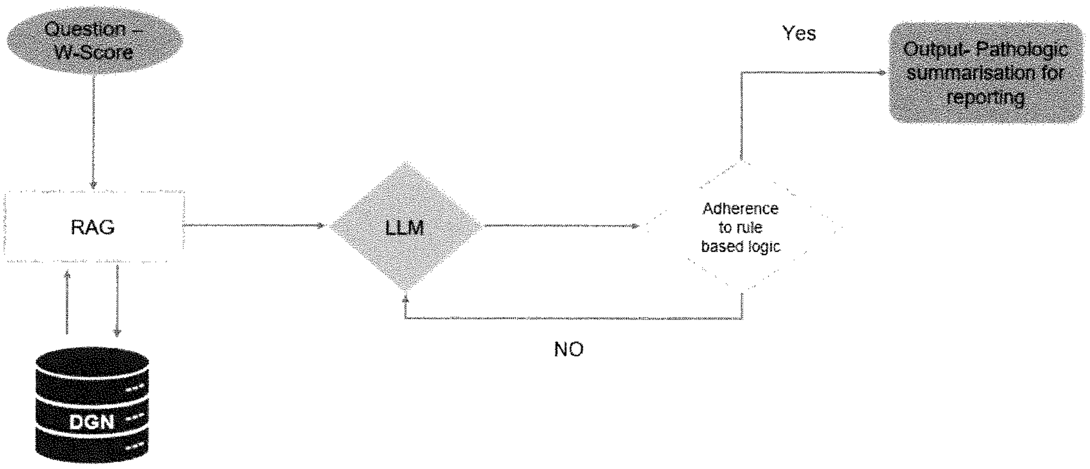
Question – W-Score
Yes
Output- Pathologic summarisation for reporting
RAG
LLM
Adherence to rule based logic
NO
DGN

Fig 1: Rule based augmented prompt engineering framework. In this setup, W-score measures volumetric deviations in brain regions while controlling for covariate. RAG refers to Retrieval-Augmented Generation. DGN denotes the clinical disease guidelines provided by the Deutsche Gesellschaft für Neurologie to be used as context for the RAG.

Expected research outcomes:

The expected outcome of this thesis is to evaluate the feasibility of LLM models being controlled using domain based rules. We could benchmark the performance of the developed LLMs augmented with a rule-based logic system against existing rule-based logic and purely data-driven LLMs. Additionally, we would test the outcomes of using a general-purpose (for e.g. Mistral AI) LLM versus a medical-specific (for e.g. BioMistral AI). Overall based on feasibility analysis, our study aims to deliver a well-analysed, interpretable system that enhances clinician trust and supports real-world application in dementia diagnosis by evaluating how effectively domain knowledge can be integrated into LLMs to generate meaningful explanations for deep learning model decisions based on MRI scans.

The thesis must contain a detailed description of all developed and used algorithms as well as a profound result evaluation and discussion. The implemented code has to be documented and provided. Extended research on literature, existing patents and related work in the corresponding areas has to be performed.

Advisors: Prof. Dr. Eskofier, Dr. Martin Dyrba, Dr. Emmanuelle Salin, Arijana Bohr M. Sc., Devesh Singh M.Sc.

Student: Dhanush Hareesh Babu

Start -End: 16.06.2025 – 16.12.2025

References

[1] Mo, Tingyu, et al. "Leveraging Large Language Models for Identifying Interpretable Linguistic Markers and Enhancing Alzheimer's Disease Diagnostics." medRxiv (2024): 2024-08.

[2] BT, Balamurali, and Jer-Ming Chen. "Performance Assessment of ChatGPT versus Bard in Detecting Alzheimer's Dementia." Diagnostics 14.8 (2024): 817.

[3] Feng, Yingjie, et al. "Large language models improve Alzheimer's disease diagnosis using multi-modality data." 2023 IEEE International Conference on Medical Artificial Intelligence (MedAI). IEEE, 2023.

[4] Li, Victor OK, Jacqueline CK Lam, and Yang Han. "LMP-TX: An AI-driven Integrated Longitudinal Multi-modal Platform for Early Prognosis of Late Onset Alzheimer's Disease." medRxiv (2024): 2024-10.

[5] Ye, Wentao, et al. "Assessing hidden risks of LLMs: an empirical study on robustness, consistency, and credibility." arXiv preprint arXiv:2305.10235 (2023).

[6] Talukdar, Wrick, and Anjanava Biswas. "Improving Large Language Model (LLM) fidelity through context-aware grounding: A systematic approach to reliability and veracity." arXiv preprint arXiv:2408.04023 (2024).

[7] Pal, Ankit, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. "Med-halt: Medical domain hallucination test for large language models." arXiv preprint arXiv:2307.15343 (2023).

[8] Ullah, Ehsan, et al. "Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review." Diagnostic pathology 19.1 (2024): 43.

[9] D'Antonoli, Tugba Akinci, et al. "Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions." Diagnostic and Interventional Radiology 30.2 (2024): 80.

[10] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

[11] Li, Ming, et al. "Ruler: Improving llm controllability by rule-based data recycling." arXiv preprint arXiv:2406.15938 (2024).

[12] Liu, Yang, et al. "Trustworthy llms: a survey and guideline for evaluating large language models' alignment." arXiv preprint arXiv:2308.05374 (2023).

[13] Vertsel, Aliaksei, and Mikhail Rumiantsau. "Hybrid LLM/Rule-based Approaches to Business Insights Generation from Structured Data." arXiv preprint arXiv:2404.15604 (2024).

[14] Singh, Devesh, and Martin Dyrba. "Computational Ontology and Visualization Framework for the Visual Comparison of Brain Atrophy Profiles." BVM Workshop. Wiesbaden: Springer Fachmedien Wiesbaden, 2024.

[15]Deutsche Gesellschaft für Neurologie - https://dgn.org/leitlinie/demenzen

[16] Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 270–279.

[17] McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., ... & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263–269.

[18] Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., ... & Silverberg, N. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562

[19] ADNI - https://adni.loni.usc.edu/

[20] Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 2009; 21: 672–687.

[21] Jessen F, Spottke A, Boecker H, et al. Design and first baseline data of the DZNE multicenter observational study on predementia Alzheimer's disease (DELCODE). Alzheimers Res Ther 2018; 10: 15.