A General Paradigm for Fine-Tuning Large Language Models in Alzheimer's Disease Diagnosis

Marcus Zhan[1], **Kun Zhao**[2], **Guodong Liu**[2], **Haoteng Tang**[3*]

[1]Sewickley Academy, Pittsburgh
[2]Electrical and Computer Engineering, University of Pittsburgh
[3]Computer Science, University of Texas Rio Grande Valley
marcusgzhan@gmail.com, {kun.zhao, guodong.liu.e}@pitt.edu, haoteng.tang@utrgv.edu

Abstract

Alzheimer's disease (AD), a complex neurodegenerative disorder, presents significant challenges for early and accurate diagnosis due to its multifactorial nature. This study introduces a novel approach to fine-tuning large language models (LLMs) for classifying AD-related dementia stages, using genetic and contextual demographic data. By harnessing the unique ability of LLMs to capture complex relationships in high-dimensional data, we developed a prompt structure that integrates genetic information, such as single nucleotide polymorphisms (SNPs), with patient-specific factors like age, sex, and clinical scores. Extensive experiments on the ADNI dataset demonstrate the superior performance of LLM-based methods. Our findings highlight the crucial role of high-quality prompts and carefully curated data in improving model accuracy. This research lays the groundwork for applying LLMs in precision medicine, providing a scalable and interpretable framework to address complex biomedical challenges, extending beyond AD.

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder and the leading cause of dementia worldwide, affecting over 55 million people globally and posing significant challenges to healthcare systems and society at large (Better 2023). Characterized by cognitive decline, memory loss, and functional impairment, AD is a complex and multifaceted disease with both genetic and environmental underpinnings. Despite decades of research(Tang et al. 2024), early and accurate diagnosis of AD remains a formidable challenge, primarily due to its heterogeneity and the interplay of various risk factors, including genetics, demographics, and clinical presentations. Current diagnostic approaches(Tang et al. 2023), while improving, often rely on labor-intensive neuroimaging, invasive cerebrospinal fluid analysis, or subjective clinical assessments, underscoring the urgent need for computational tools that can integrate multimodal data for improved diagnostic precision.

Genetic factors play a pivotal role in the etiology of AD. Variations in several genes, most notably APOE, have been strongly associated with disease risk, with carriers of the APOE-ϵ4 allele exhibiting a significantly higher likelihood of developing AD (Liu et al. 2013). However, genetic predispositions alone do not fully explain disease onset or progression, necessitating the consideration of additional factors such as age, sex, and comorbid conditions. For example, age remains the strongest risk factor for AD, with prevalence doubling approximately every five years after age 65 (Sperling et al. 2011). Sex differences in AD prevalence and progression are also well-documented, with women being disproportionately affected (Mielke 2018). Furthermore, clinical characteristics such as depressive symptoms, often assessed through tools like the Geriatric Depression Scale (GDS), have been identified as potential contributors to cognitive decline and dementia risk (Saczynski et al. 2010). These complex interactions between genetic, demographic, and clinical factors necessitate integrative analytical frameworks capable of uncovering subtle patterns and relationships in multidimensional datasets.

Recent advances in artificial intelligence (AI) and machine learning (ML) have opened new avenues for addressing such challenges, particularly in leveraging vast amounts of biomedical data for disease prediction and classification. Large language models (LLMs), originally designed for natural language processing tasks, have emerged as transformative tools with broad applicability across domains, including biomedical research (Brown et al. 2020; Lee et al. 2020). These models, pre-trained on massive text corpora, possess the ability to generalize knowledge and learn domain-specific patterns when fine-tuned on specialized datasets. While LLMs have primarily been applied to tasks involving text-based data, their flexibility and capacity for contextual reasoning position them as promising candidates for analyzing structured biomedical data, such as genetic and phenotypic information. However, applying LLMs to structured data in the context of disease diagnosis remains an underexplored area.

This study introduces a general paradigm for fine-tuning LLMs to diagnose AD using genetic data in combination with contextual prompts such as age, sex, and GDS scores. By leveraging the inherent capacity of LLMs to model complex interactions, our approach enables the extraction of meaningful patterns across genetic and phenotypic domains. The inclusion of contextual prompts further enhances the model's ability to provide nuanced predictions that reflect

*Corresponding author.

individual variability. This paradigm not only builds on the growing body of work demonstrating the utility of LLMs in biomedical applications but also addresses critical gaps in applying these tools to structured data analysis for disease diagnosis.

The contributions of this study are threefold. First, it demonstrates the feasibility of adapting general-purpose LLMs to structured biomedical data, specifically for the diagnosis of AD. Second, it highlights the importance of incorporating demographic and clinical context to enhance model performance and relevance to real-world scenarios. Third, it establishes a methodological foundation for future research at the intersection of AI and precision medicine. By bridging the gap between general-purpose AI tools and domain-specific biomedical challenges, this work aims to advance the use of AI in healthcare and contribute to the development of more personalized and accurate diagnostic tools. Through this effort, we aim to address key limitations in existing diagnostic methods for AD and demonstrate the potential of LLMs as powerful allies in the fight against neurodegenerative diseases. Beyond AD, the paradigm presented in this study has broader implications for applying LLMs to other complex diseases, fostering innovation in AI-driven precision medicine.

Related Works

Large Language Models

The evolution of natural language processing (NLP) and AI models has followed a transformative trajectory, progressing from rule-based systems to statistical approaches, and culminating in neural network architectures (Josh Achiam 2024). A pivotal shift occurred with the introduction of self-attention mechanisms and Transformer-based architectures (Vaswani 2017), which catalyzed the rise of pre-trained language models (PLMs). These models learn generalized linguistic patterns from vast corpora through unsupervised training, enabling robust performance across diverse NLP tasks such as multiple-choice question answering (Robinson, Rytting, and Wingate 2023), narrative generation (Cao et al. 2023), and commonsense reasoning (Yang et al. 2023), while reducing overfitting risks.

Recent years witnessed rapid advancements in large language models (LLMs), exemplified by GPT-3 (Brown et al. 2020), PaLM (Aakanksha Chowdhery 2022), LLaMA (Touvron et al. 2023), Megatron-Turing NLG (Smith et al. 2022). This growth has been driven by exponential increases in training data volume and computational power, with empirical studies confirming that model performance scales predictably with parameters and dataset size—a phenomenon formalized as scaling laws (Kaplan et al. 2020). LLMs now represent a cornerstone of AI research, surpassing smaller models in text comprehension and generation fidelity. Their ability to streamline scientific inquiry, accelerate discovery, and bridge interdisciplinary gaps positions them as transformative tools for both technical and social sciences.

AD Prediction Based on LLM with Genetic Data

The application of large language models (LLMs) to genetic data for predicting brain dementia states (e.g., AD, mild cognitive impairment or MCI) has gained significant attractions in recent years. A few existing studies focus on encoding high-dimensional genetic data, such as single nucleotide polymorphisms (SNPs), into meaningful embeddings through LLMs (i.e., DNA-BERT (Ji et al. 2021) and SNP2Vec (Alipanahi et al. 2015)), which have demonstrated improved disease classification accuracy by leveraging contextual information from genetic sequences. Meanwhile, multimodal fusion methods integrate genetic data with other modalities, such as neuroimaging, to capture the complementary information across data sources (Feng et al. 2023; Liu et al. 2024). Some other interaction-based approaches analyze genetic interactions, such as SNP-SNP or gene-gene networks, using LLMs with graph neural networks to uncover complex patterns underlying AD progression (Permana, Beatson, and Forde 2023; Xiao, Wang, and Wan 2024) Additionally, some progression prediction methods employ LLMs for longitudinal modeling of genetic and clinical data, focusing on predicting long-term transitions between cognitive states (e.g., from cognitively normal or CN to AD through different cognitive impairment stages) (Machado Reyes et al. 2024; Danter 2024).

<Instruction>: Given a series of gene segments and patient demographic information, tell me the label of the patient.

<Input>:
0000, 10, 0, 01110011111, 001, 0111101112111111001000000010112100002010, 0002000010002010000000001001110020, 20112, 0, 00, 01, 2121200000011110 01010, 11200, 10111110000000002l101, 01000, 01, 112, 00201201100100, 22, 1111, 1, 000000, 1111, 221, 0, 10, 0100001000, 01100001111111000101, 1001110, 2201, 21011101, 22, 1, 0, 10, 000, 2110001, 2, 20020, 000, 0100, 2, 1, 0, 01, 10, 10000000, 022000020000001111001001, 00111011002000100001010111011000 10001111000110010010101101101100010111011100011000200100121011221 00, 00, 1100, 0102010100000011020111011110101011110100101111000000001000 00200200220111, 00010011001000211, 111011220, 111000222001, 01, 1, 0, 0, 111111, 0100100120101101001100011001021120120201011120000101001100110001, 1, 010, 021, 0020, 111122101000, 2, 0000, 121, 20000000000, 0012001020000, 01002121, 0001, 1200101111011111000000100, 2100, 011110001000, 211000000 0000121020200000012001110110000100001011100100100020011000001211 0 00111001110122221020111112002020112100100010101111211010022210110 1 1101100000011011001110000000020000011001100010211011 10, 0, 0, 000, 010, 0, 2110, 00, 111, 02, 110101120201001111, 000110201101000010000101 00, 00, 2, 000, 0, 2, 0020000, 0100000010122000000000102001001000000000010020020 0, 0, 012, 00101, 201001, 0011121010, 2, 01001010111000002011, 0, 0.
Patient information: The individual is a 68.1-year-old and sex is F. GDS score is 1.0.

<Output>: The label is MCI.

Figure 1: An example of our designed prompt consists of three parts. The <*Instruction*> part is presented in the violet box, the <*Input*> part is presented in the red box, and the <*Output*> part is presented in the blue box. The <*Input*> part includes SNP segments and subject demographic information, where commas are used to separate different genes, as presented in the green block.

Methodology

In this section, we start with presenting the elaborated prompt specifically designed for genetic data to fine-tune Large Language Models (LLMs). Subsequently, we detail the fine-tuning process by utilizing the crafted prompts.

Prompt Design

The Alpaca format (Taori et al. 2023), an instruction-response dataset format, is utilized to standardize our constructed prompts. Specifically, the constructed prompt con-

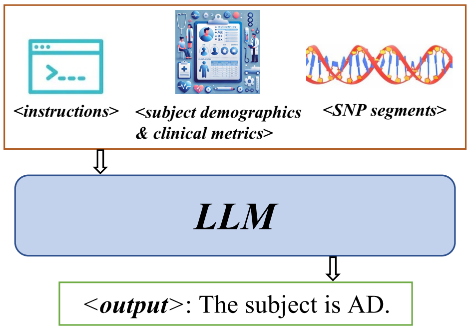

Figure 2: Diagram illustrating the architecture of LLM with the proposed prompt.

sists of three parts including *instruction*, *input* and *output*. For the *instruction* part, we design a generic prompt that guides the LLM to classify input data. The *input* part includes a series of gene segments with corresponding SNP values. To ensure that different gene segments are not combined during tokenization, we use commas to distinctly separate these segments associated with different gene names. Additionally, the *input* part also includes a dedicated field to provide the demographic information of each subject, appended at the end of the gene segments. The *output* section describes the classification labels assigned to each subject. We present an example of our designed prompt in the Figure 1. The specific task of LLMs involves utilizing the *instruction* and *input* parts to produce the corresponding *output* section. During the inference phase, the generated text is parsed to extract the predicted label (i.e., different disease states) for each test sample, which is then compared against the groundtruth label to calculate the accuracy of the generated output texts.

LLM Fine-tune

The objective of fine-tuning a Large Language Model (LLM) is to adapt the pre-trained model (i.e., LLM_θ) for a specific downstream task by updating its parameters θ using task-specific annotated data. As shown in Figure 2, the input of the LLMs are the *instruction* and *input* sections of our designed prompts, and the target is the *output* section of the prompts. We first tokenize the input and output of LLM as follows:

$$\begin{aligned} X &= \text{Tokenizer}(< \textit{instruction}, \textit{input} >) \\ Y &= \text{Tokenizer}(< \textit{output} >), \end{aligned} \tag{1}$$

where $X = \{x_1, x_2, ..., x_n\}$ is a sequence of input tokens and $Y = \{y_1, y_2, ..., y_m\}$ is the target tokens. The LLM learns to predice the conditional probability of the target given the input:

$$P_\theta(Y|X) = \prod_{t=1}^{m} P_\theta(y_t|y_{<t}, X), \tag{2}$$

where y_t is the token at position t in the target token sequence. $y_{<t}$ represent all tokens before position t. The fine-tuning process aims to optimize the model by minimizing the negative log-likelihood (NLL) loss on the target token sequence, which is equivalent to minimize the following cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{m_i} \log P_\theta\left(y_t^{(i)}|y_{<t}^{(i)}, X^{(i)}\right), \tag{3}$$

where N is the number of training samples, m_i is the length of the target token sequence for the $i-th$ sample. $X^{(i)}$ and $Y^{(i)}$ are the input and target token sequences for the $i-th$ sample. In the inference phase, the fine-tuned model LLM_{θ^*} generates the target sequence Y given the input sequence X by maximizing the likelihood of the output token sequence. Particularly, the LLM generates tokens position by position. At each position t, the next token y_t is predicated as follow:

$$y_t = \arg\max_{y \in \mathcal{V}} P_{\theta^*}(y|y_{<t}, X) \tag{4}$$

Where $\mathcal{V}$ is the vocabulary of the model, and $P_{\theta^*}(y|y_{<t}, X)$ denotes the conditional probability of the token y_t given the input token sequence and the previously generated tokens $y_{<t}$.

Experiments and Results

Dataset

The genetic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database were used to evaluate our framework, consisting of 1998 subjects (mean age: 73.43 ± 7.17 years, 927 women). Genotype data were obtained from 1079 Single Nucleotide Polymorphisms (SNPs) generated using Illumina genotyping platforms. Each SNP represents the genotype encoding at the given locus, indicating the allele dosage for each subject. Specifically, **SNP = 0** indicates homozygosity for the reference allele, **SNP = 1** indicates heterozygosity, and **SNP = 2** indicates homozygosity for the alternate allele at the given locus. Details regarding the genotyping protocols and quality control can be found in (Saykin et al. 2010). SNPs were annotated with their corresponding genes using their chromosomal positions based on Ensembl gene annotations (Yates et al. 2016). Subjects were categorized into five groups based on dementia progression: cognitively normal (CN), significant memory concern (SMC), early mild cognitive impairment (EMCI), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI), and Alzheimer's Disease (AD). Meanwhile, Alzheimer's disease (AD)-related clinical metrics, including the Mini-Mental State Examination (MMSE) scores (Folstein, Folstein, and McHugh 1975), the Clinical Dementia Rating (CDR) scores (Morris 1993) and the Geriatric Depression Scale (GDS) scores (Yesavage et al. 1982), were also provided for each subject.

Implementation Details

Data Augmentation. Since the information provided by each SNP sequence is independent of the order of the associated genes, we augmented our dataset by randomly permuting the gene order while keeping the corresponding SNPs unchanged. For instance, if $SNP_{i_1}, ..., SNP_{i_K}$ are associated with $Gene_i$ and $SNP_{j_1}, ..., SNP_{j_K}$ are associated

Model	MLP	Llama3	Qwen2.5	Mistral
Acc.(%)	26.23	88.33	**91.74**	84.31

Table 1: Classification accuracy (%) of MLP and different LLMs on the six-class dementia states.

with $Gene_j$, the order of $Gene_i$ and $Gene_j$ was shuffled in the augmented data. However, the order of SNP sequences within each gene (e.g., $SNP_{i_1}, ..., SNP_{i_K}$) remained unaltered. Utilizing this augmentation strategy, the dataset was substantially expanded including $50,000$ training samples and $10,000$ testing samples.

Experimental Setting. Three pretrained models are selected to conduct our experiments, including Llama3-8B-instruct (Touvron et al. 2023), Qwen2.5-7B (Yang et al. 2024) and Mistral-7B (Jiang et al. 2023). We train them on $8\times$ A100 GPUs with full parameters, using LLAMA-Factory. The number of epochs is 5 and the consumption time to fine-tune each model is about 5 hours.

Differences Between Three Selected LLMs Llama 3 (up to 405B parameters) excels in scalability and multimodal integration with a decoder-only transformer and Grouped-Query Attention (GQA), making it ideal for enterprise NLP. Qwen 2.5 specializes in mathematical reasoning and multilingual processing (29+ languages), leveraging YaRN-enhanced RoPE embeddings for a 128K-token context. Its domain-specific variants (e.g., Qwen2.5-Math/Coder) enhance structured data analysis. Mistral, with a 7B Sparse Mixture-of-Experts (MoE) design, achieves 6× faster inference via top-2 expert routing, optimizing for real-time and edge deployment. Each model reflects distinct priorities: Llama 3 for scalability, Qwen 2.5 for domain-specific tasks, and Mistral for efficiency.

Comparative Experiment

In this experiment, we compare the performance among different LLMs for a 6-class classification task including CN, SMC, EMCI, MCI, LMCI, and AD. Additionally, we also evaluate the effectiveness of LLMs in comparison to traditional deep learning approach (i.e., Multilayer Perceptron or MLP) on this task. Since the MLP is not able to embed the patient demographic information, we only include SNP segments in this comparative experiment for a fair comparision. The classification results, summarized in Table 1, demonstrate a significant performance gap between traditional deep learning methods and Large Language Models (LLMs) for the six-class classification task. The MLP network achieves an accuracy of 26.23%, indicating its limited capability in handling the complexity of genetic data and demographic features. In contrast, LLMs exhibit superior performance, with Llama3-8B-instruct achieving an accuracy of 88.33%, Mistral-7B obtaining the accuracy of 84.31%, and Qwen2.5-7B reaching the highest accuracy of 91.74%. This improvement highlights the ability of LLMs to leverage their pretrained knowledge and capture intricate patterns in the input data. Furthermore, among the LLMs, Qwen2.5 achieves the best classification performance.

Impact of Prompt Quality on Dementia Prediction

To investigate the impact of prompt quality on model performance, we designed three experiments with different prompt variants. The first prompt variant included only SNP segments, representing a baseline that focuses solely on genetic information. The second prompt variant extended this by incorporating subject demographic information including age and sex, which represents a high-quality prompt. The last prompt variant further included dementia-related clinical metrics such as MMSE scores, GDS scores, and CDR scores. However, these clinical scores are incomplete, where part of them are missing and set to "NaN" during the data collections.

The results (see Table 2) demonstrate that prompt quality plays a pivotal role in improving model performance. Compared with the first prompt variant, the second prompt variant, which combined SNP segments with age and sex information, achieved the best classification performance, suggesting that demographic context enhances the predictive power of genetic data. The third prompt variant, despite including the most comprehensive information, yielded the poorest results due to the presence of missing data. These results underscore that incomplete or noisy data in prompts may hinder learning process and reduce performance of the LLMs, which highlights the importance of designing well-curated, high-quality prompts for fine-tuning LLMs in genetic and clinical prediction tasks.

Model	Llama3			Qwen2.5			Mistral		
	$*$	†	‡	$*$	†	‡	$*$	†	‡
Acc.(%)	88.33	94.89	87.91	91.74	**96.64**	89.78	84.31	82.89	94.24

Table 2: Classification accuracy for dementia states with different prompt variations. $*$ indicates the results using prompts with only SNP segments. † indicates the results using high-quality prompts that include SNP segments, sex, and age information. ‡ indicates the results using prompts that include SNP segments, sex, age, and incomplete clinical metrics (i.e., MMSE, GDS, and CDR).

Impact of Prompt Components on Dementia Prediction

To identify which components of the prompt have the greatest impact on classification accuracy, we experimented with different prompt variants incorporating different clinical metrics. Building on the findings from the previous section, which highlight the importance of prompt quality, we first excluded all data samples with incomplete clinical metric scores. Demographic information, including sex and age, was shown to significantly enhance the performance of our LLM models, and thus we used SNP segments combined with sex and age as the basic prompt. To explore additional influences, we introduced the AD-related allele, APOE-ϵ4, into the basic prompt to create the first prompt variant. Additionally, we incorporated a depression-related metric (i.e., GDS scores) into the basic prompt to form another prompt variant. The results (see Table 3) demonstrate that adding GDS information significantly improved classification performance, however, the information of APOE-ϵ4 decreased

the classification accuracy.

The negative impact of introducing the APOE-ϵ4 allele may arise from the dataset containing a disproportionate number of individuals with specific APOE allele combinations, potentially introducing bias into the model's predictions. This is particularly relevant as APOE-ϵ4 carriers are overrepresented within the AD population. Depression is a well-established risk factor for both cognitive decline and dementia (Ownby et al. 2006; Steffens and Potter 2008). Incorporating depression-related metrics, such as GDS scores, into the prompt enhances classification accuracy for several reasons. First, depression-related metrics capture both emotional and cognitive symptoms, which often overlap with early manifestations of Alzheimer's Disease, offering additional predictive signals. Second, unlike APOE alleles that primarily represent genetic predisposition, GDS scores provide dynamic, symptom-based insights into disease progression. This complementary information enriches the context available for LLMs to learn from, leading to more nuanced and accurate classifications.

Model	Llama3			Qwen2.5			Mistral		
	$*$	†	‡	$*$	†	‡	$*$	†	‡
Acc.(%)	91.57	87.91	96.91	96.65	95.80	**97.92**	90.79	88.41	92.90

Table 3: Classification accuracy for dementia states with different prompt variations. $*$ indicates the results using basic prompts including SNP segments, seg and age information. † indicates the results of basic prompts with APOE allele. ‡ indicates the results of basic prompts with GDS scores.

Conclusions

In this study, we demonstrated the potential of fine-tuning LLMs for AD diagnosis using genetic, demographic, and clinical data. By integrating SNP sequences with patient-specific contextual information such as age, sex, and clinical metrics, our approach achieved significant improvements in classification accuracy across multiple dementia stages. Moreover, our analysis revealed the critical influence of prompt quality on model performance. Incorporating demographic and clinical context substantially enhanced predictive accuracy, while incomplete or noisy data components, such as missing clinical scores, hindered the effectiveness of the models. These findings underscore the importance of designing well-curated, high-quality prompts tailored to the specific requirements of biomedical tasks.

Beyond AD, the proposed framework is flexible for broader applicability to other complex diseases, paving the way for advancements in AI-driven precision medicine. Future research should explore additional data modalities, such as neuroimaging, to further enrich the diagnostic capabilities of LLMs and address remaining challenges in multimodal data integration. By bridging the gap between general-purpose AI and domain-specific applications, this work contributes to the development of scalable and personalized healthcare solutions.

Acknowledgments

This study was supported by the Presidential Research Fellowship (PRF) in the Department of Computer Science at the University of Texas Rio Grande Valley (UTRGV), and the UTRGV seed grant. We acknowledge the UTRGV High Performance Computing Resource, supported by NSF grants 2018900 and IIS-2334389, and DoD grant W911NF2110169. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1] funded by NIH grant U19AG024904. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: [2]

References

Aakanksha Chowdhery, e. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311.

Alipanahi, B.; Delong, A.; Weirauch, M. T.; and Frey, B. J. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8): 831–838.

Better, M. A. 2023. Alzheimer's disease facts and figures. *Alzheimers Dement*, 19(4): 1598–1695.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P. S.; and Sun, L. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. arXiv:2303.04226.

Danter, W. R. 2024. Tracing Alzheimer's Genetic Footprints: A Pioneering Longitudinal Study Using Artificial Intelligence to Unravel Mutation-Driven Risks and Progression in Virtual Patients; Part 1 The APOE genotypes. *medRxiv*, 2024–04.

Feng, Y.; Xu, X.; Zhuang, Y.; and Zhang, M. 2023. Large language models improve Alzheimer's disease diagnosis using multi-modality data. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, 61–66. IEEE.

Folstein, M. F.; Folstein, S. E.; and McHugh, P. R. 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3): 189–198.

Ji, Y.; Zhou, Z.; Liu, H.; and Davuluri, R. V. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

[1] http://adni.loni.usc.edu

[2] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Josh Achiam, e. a. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Liu, C.-C.; Kanekiyo, T.; Xu, H.; and Bu, G. 2013. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology*, 9(2): 106–118.

Liu, Q.; Zeng, W.; Zhu, H.; Li, L.; Wong, W. H.; and Initiative, A. D. N. 2024. Leveraging genomic large language models to enhance causal genotype-brain-clinical pathways in Alzheimer's disease. *medRxiv*, 2024–10.

Machado Reyes, D.; Chao, H.; Hahn, J.; Shen, L.; Yan, P.; and Initiative, A. D. N. 2024. Identifying Progression-Specific Alzheimer's Subtypes Using Multimodal Transformer. *Journal of Personalized Medicine*, 14(4): 421.

Mielke, M. M. 2018. Sex and gender differences in Alzheimer's disease dementia. *The Psychiatric times*, 35(11): 14.

Morris, J. C. 1993. The Clinical Dementia Rating (CDR) current version and scoring rules. *Neurology*, 43(11): 2412–2412.

Ownby, R. L.; Crocco, E.; Acevedo, A.; John, V.; and Loewenstein, D. 2006. Depression and risk for Alzheimer disease: systematic review, meta-analysis, and metaregression analysis. *Archives of general psychiatry*, 63(5): 530–538.

Permana, B.; Beatson, S. A.; and Forde, B. M. 2023. GraphSNP: an interactive distance viewer for investigating outbreaks and transmission networks using a graph approach. *BMC bioinformatics*, 24(1): 209.

Robinson, J.; Rytting, C. M.; and Wingate, D. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. arXiv:2210.12353.

Saczynski, J. S.; Beiser, A.; Seshadri, S.; Auerbach, S.; Wolf, P.; and Au, R. 2010. Depressive symptoms and risk of dementia: the Framingham Heart Study. *Neurology*, 75(1): 35–41.

Saykin, A. J.; Shen, L.; Foroud, T. M.; Potkin, S. G.; Swaminathan, S.; Kim, S.; Risacher, S. L.; Nho, K.; Huentelman, M. J.; Craig, D. W.; et al. 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer's & Dementia*, 6(3): 265–273.

Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; Zhang, E.; Child, R.; Aminabadi, R. Y.; Bernauer, J.; Song, X.; Shoeybi, M.; He, Y.; Houston, M.; Tiwary, S.; and Catanzaro, B. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. arXiv:2201.11990.

Sperling, R. A.; Aisen, P. S.; Beckett, L. A.; Bennett, D. A.; Craft, S.; Fagan, A. M.; Iwatsubo, T.; Jack Jr, C. R.; Kaye, J.; Montine, T. J.; et al. 2011. Toward defining the preclinical stages of 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): 280–292.

Steffens, D. C.; and Potter, G. G. 2008. Geriatric depression and cognitive impairment. *Psychological medicine*, 38(2): 163–175.

Tang, H.; Guo, L.; Fu, X.; Wang, Y.; Mackin, S.; Ajilore, O.; Leow, A. D.; Thompson, P. M.; Huang, H.; and Zhan, L. 2023. Signed graph representation learning for functional-to-structural brain network mapping. *Medical image analysis*, 83: 102674.

Tang, H.; Liu, G.; Dai, S.; Ye, K.; Zhao, K.; Wang, W.; Yang, C.; He, L.; Leow, A.; Thompson, P.; et al. 2024. Interpretable Spatio-Temporal Embedding for Brain Structural-Effective Network with Ordinary Differential Equation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 227–237. Springer.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Xiao, H.; Wang, J.; and Wan, S. 2024. WIMOAD: Weighted Integration of Multi-Omics data for Alzheimer's Disease (AD) Diagnosis. *bioRxiv*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Yin, B.; and Hu, X. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv:2304.13712.

Yates, A.; Akanni, W.; Amode, M. R.; Barrell, D.; Billis, K.; Carvalho-Silva, D.; Cummins, C.; Clapham, P.; Fitzgerald, S.; Gil, L.; et al. 2016. Ensembl 2016. *Nucleic acids research*, 44(D1): D710–D716.

Yesavage, J. A.; Brink, T. L.; Rose, T. L.; Lum, O.; Huang, V.; Adey, M.; and Leirer, V. O. 1982. Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of psychiatric research*, 17(1): 37–49.