Adapting Large Language Models for Applications in Alzheimer's Disease BioMarkers

Abreham Tadesse[1] Gaurab Subedi[1]
Department of Computer Science, University of Nevada, Las Vegas
abreham.tadesse@unlv.edu gaurab.subedi@unlv.edu

Abstract

The advent of Large Language Models (LLMs) has advanced the use of Natural Language Processing within the Alzheimer's disease domain. Current strategies for adapting general purpose LLMs primarily rely on performing supervised fine-tuning on pre-trained LLMs with in-domain text. We introduce *Domain Aware Span Corruption (DASC)* - an augmented self-supervised training objective. In this work, we employ *DASC* as a semi-self-supervised fine-tuning objective on the FLAN-T5 XL model. We analytically show that DASC marginally increases the semantic richness of hidden-state representations of in domain texts, a proxy of domain understanding. However, DASC shows a significant improvement in the transferability of deeper domain understanding in a downstream Question Answering task within the Alzheimer's disease domain.

Keywords:

1. Introduction

Biomarkers for Alzheimer's Disease (AD) span multiple categories, including neuroimaging markers (e.g., amyloid-PET, tau-PET, structural MRI), fluid biomarkers (e.g., cerebrospinal fluid (CSF) and blood-based markers such as amyloid- , tau phosphorylation, and neurofilament light chain), digital biomarkers, and cognitive assessments (Zetterberg and Bendlin, 2021). The rapidly expanding volume of research literature on these biomarkers presents significant challenges for researchers and clinicians attempting to synthesize and apply this knowledge effectively. The complexity and heterogeneity of AD biomarker data require sophisticated computational approaches to organize, analyze, and interpret this information (Winchester et al., 2023). Currently, an estimated 6.9 million Americans aged 65 and older are living with Alzheimer's dementia, with projections indicating this number will increase to 13.8 million by 2060 (Better, 2023). AD represents one of the most pressing public health challenges in the United States, placing substantial burdens on patients, caregivers, healthcare systems, and the national economy (Castro et al., 2010).

Recent advances in artificial intelligence, particularly Large Language Models (LLMs), offer promising new approaches to accelerate biomarker research and clinical applications in the Alzheimer's Disease domain (Wilczok, 2025). LLMs such as Google's Flan-T5 family (Chung et al., 2024), OpenAI's GPT series (Achiam et al., 2023), and Meta's LLaMA models (Touvron et al., 2023) have demonstrated remarkable capabilities in understanding and generating human language, as well as processing complex medical and scientific literature.

However, the application of LLMs to specialized domains such as Alzheimer's Disease Biomarkers (ADBM) presents unique challenges and opportunities that have not been fully explored. While general-purpose LLMs possess broad knowledge across many domains, they often lack the depth of understanding required for specialized biomedical applications, particularly in areas requiring nuanced interpretation of disease-specific biomarker information. Moreover, the technical language, domain-specific terminology, and complex relationships between biomarkers and disease mechanisms in AD literature require models with enhanced domain expertise (Wang et al., 2023).

Current strategies for adapting general-purpose LLMs to specialized domains primarily rely on performing supervised fine-tuning on pre-trained LLMs with in-domain text. In this research we propose *Domain Aware Span Corruption* (DASC), a novel self-supervised fine-tuning objective inspired by the Span Corruption pretraining objective on which the Google T5 LLM family was trained (Raffel et al., 2020). The T5 team employs a span-corruption training objective for their pre-training stage. This objective corrupts and drops contiguous, randomly spaced spans of tokens. After which, all corrupted spans are replaced by sentinel tokens with a token ID unique to them. The target sequence is then made up of all the dropped sequences, delimited by their corresponding sentinel tokens. (Raffel et al., 2020) Inspired by this method, DASC employs a domain aware corruption strategy in which mutli-token spanning annotations are used to drop tokens from the input sequence and replaces them with unique sentinel tokens. The target sequence is then made of these dropped spans delimited by their corresponding sentinel tokens. Our approach employs DASC as a semi self-supervised fine-tuning objective on the FLAN-T5 XL model, followed by supervised fine-tuning on a text classification task. To evaluate the contribution of DASC, we also fine-tune FLAN-T5 on the text classification task without the preceding DASC-based self-supervised fine-tuning stage.

The significance of this work extends beyond technological advancement. Effective text classification systems can rapidly categorize vast volumes of Alzheimer's research, helping researchers identify relevant studies and trends in biomarker research (Counts et al., 2017). Enhanced domain adaptation enables automated extraction of biomarker-relevant information from the literature, potentially uncovering relationships not previously recognized. Models built on domain-adapted foundations could serve as assistive tools for researchers navigating complex biomarker literature and for clinicians interpreting biomarker results, ultimately accelerating knowledge discovery and clinical translation.

Our methodological approach employs Google's Flan-T5 XL model as the baseline foundation (Chung et al., 2024). We then implement domain-adaptive finetuning using DASC on a carefully curated corpus of ADBM literature (Gururangan et al., 2020; Hiwarkhedkar et al., 2023). To address the challenge of limited domain-specific training data (approximately 260,000 tokens, equivalent to roughly 195,000 words or about 650 pages of text, compared to standard pretraining datasets of 100-300 million tokens), we incorporate additional biomedical research papers and employ domain-aware masking strategies

during pretraining to prioritize biomarker-relevant terms. Performance is evaluated using comprehensive metrics including F1 scores (Hand et al., 2021), recall, precision, and confusion matrices for the classification task, alongside embedding similarity analyses to measure domain understanding.

This paper contributes to the growing field of domain-specific LLM adaptation, with particular emphasis on biomedical applications. By explicitly addressing challenges such as limited domain-specific data and the introduction of domain-aware finetuning objectives, we provide insights that extend beyond Alzheimer's research to the broader application of Language Modeling in specialized medical domains. The results presented here demonstrate the potential of DASC-adapted LLMs to serve as powerful tools in accelerating Alzheimer's biomarker research, with implications for both research efficiency and clinical translation.

The remainder of this paper is organized as follows: Section 2 reviews related work in LLM adaptation, domain-specific pretraining, and AI applications in Alzheimer's research. Section 3 details our methodology, including data collection, DASC pretraining procedures, fine-tuning approaches, and evaluation metrics. Section 4 presents our experimental results, while Section 5 discusses these findings, their implications, and limitations. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

The application of Large Language Models (LLMs) to specialized biomedical domains represents a rapidly evolving research area with significant implications for healthcare. This literature review examines relevant work across three interconnected areas: (1) domain-specific adaptation of language models for biomedical applications, (2) alternative fine-tuning approaches, and (3) unsupervised and semi-supervised adaptation strategies. These areas collectively inform our approach to adapting LLMs for Alzheimer's Disease biomarker applications through Domain Aware Span Corruption (DASC) pretraining followed by supervised fine-tuning.

2.1 Domain-Specific Adaptation for Biomedical NLP

The adaptation of language models to specialized biomedical domains has demonstrated substantial performance improvements over general-purpose models. BioBERT (Lee et al., 2020) pioneered this approach by adapting BERT with additional pre-training on PubMed abstracts and PMC full-text articles, achieving significant performance gains across biomedical NLP tasks including named entity recognition, relation extraction, and question answering. Building on BioBERT's foundation, this work demonstrated that continued pretraining on domain-specific text significantly improves model understanding of biomedical terminology and relationships compared to using general-domain models directly.

Further advancing this direction, (Han et al., 2021) challenged the conventional wisdom that starting with general-domain pre-trained models is optimal for biomedical NLP. Their comprehensive study demonstrated that pretraining from scratch on biomedical data can yield superior performance compared to continual pretraining of general-domain models. They introduced BLURB (Biomedical Language Understanding and Reasoning Benchmark), a comprehensive evaluation suite that includes six diverse biomedical NLP tasks:

named entity recognition on disease mentions (NCBI-Disease) and chemical-protein relations (BC5-chem, BC5-disease), relation extraction for chemical-disease and gene-disease interactions (ChemProt, DDI), document classification for scientific abstracts (HoC), sentence similarity assessment (BIOSSES), and question answering (PubMedQA). BLURB provides standardized evaluation protocols and datasets, enabling systematic comparison of different pretraining strategies across the full spectrum of biomedical NLP applications. Their findings showed that domain-specific pretraining strategies outperform general approaches across multiple tasks, and revealed that some complex NLP practices, such as elaborate tagging schemes for NER, become unnecessary when using domain-adapted models. This work established important principles for domain adaptation that inform our DASC approach.

More recently, (Wu et al., 2024) introduced PMC-LLaMA, an open-source large language model based on the LLaMA 7B architecture specifically designed for medical applications. By training on 4.8 million biomedical papers and 30,000 medical textbooks, along with a 202 million token instruction-tuning dataset covering medical question-answering, reasoning, and dialogues, they developed a model that outperformed even ChatGPT on specialized medical tasks. This work demonstrates the potential of domain-adaptive pretraining combined with task-specific instruction tuning for creating powerful domain-specific models, directly supporting our hypothesis that self-supervised domain adaptation (DASC) followed by supervised fine-tuning can enhance performance on specialized tasks. In the context of literature review applications, (Panagides et al., 2024) developed FusBERT, a fine-tuned Bio-ClinicalBERT model for classifying scientific abstracts related to focused ultrasound therapies. Their model achieved high accuracy (0.91) and recall (0.99) in distinguishing relevant papers, demonstrating how domain-adapted models can streamline the literature review process in specialized medical fields. This work is particularly relevant to our text classification objectives for Alzheimer's biomarker literature, as it shows that targeted domain adaptation can enable accurate categorization of specialized research papers.

These studies collectively highlight the significant performance benefits of domain-specific adaptation for biomedical NLP tasks. However, most prior work has focused on general biomedical domains rather than disease-specific applications like Alzheimer's biomarkers. Our research addresses this gap by introducing DASC, a domain-aware pretraining objective specifically designed to adapt LLMs to the highly specialized ADBM domain before supervised fine-tuning on downstream classification tasks.

2.2 Alternative Fine-Tuning Approaches

While our research employs full fine-tuning to establish baseline performance with DASC pretraining, it is worth noting alternative approaches developed for computational efficiency. Parameter-efficient fine-tuning techniques, such as adapter modules (Houlsby et al., 2019) and Low-Rank Adaptation methods like QLoRA (Dettmers et al., 2023), enable model adaptation by updating only a small subset of parameters (typically 3-4%) compared to full fine-tuning (100% of parameters). These approaches reduce computational requirements significantly—for instance, quantization reduces model precision from 32-bit to 4-bit, typically resulting in 1-2% accuracy loss but enabling dramatically faster computation and reduced memory usage. While these methods offer practical advantages for resource-constrained environments, we chose full fine-tuning for our experiments to iso-

late and measure the impact of DASC pretraining without introducing additional variables from parameter-efficient techniques. These alternative approaches represent promising directions for future work to make DASC-based domain adaptation more accessible and scalable.

The computational resources required for fine-tuning large language models present significant challenges, particularly for specialized domains like Alzheimer's research where computational resources may be limited. Parameter-efficient fine-tuning approaches offer promising solutions to this challenge.

2.3 Unsupervised and Semi-Supervised Adaptation Strategies

Given the challenges of obtaining large labeled datasets in specialized domains like Alzheimer's research, unsupervised and semi-supervised adaptation strategies offer promising alternatives for improving model performance without extensive manual annotation. (Zhang et al., 2024) explored unsupervised prompt learning for classification with black-box language models, leveraging pseudo-labeled data generated by the LLM itself to fine-tune the model without requiring external labeled data. This self-supervision approach is particularly relevant for adapting models to new domains where labeled data is scarce. Similarly, (Zeng et al., 2024) demonstrated how unsupervised fine-tuning via instruction-tuning can improve the performance of dense retrieval systems, particularly in low-resource settings, showing how models can learn effective text representations without supervised training signals. In the scientific domain specifically, (Shi et al., 2024) enhanced prompt tuning for language models in scientific text classification through data augmentation with L2 regularization. Their experiments on scientific citation datasets showed significant improvements in accuracy and robustness, even with reduced labeled data, demonstrating how augmentation strategies can enhance performance in scientific domains with limited supervision. Most relevant to our DASC approach, (Kimura et al., 2024) introduced L3Masking for multi-task fine-tuning of language models. This approach improves performance by selectively masking tokens with low likelihood based on prior knowledge from the base model, enhancing task-specific adaptation. This technique directly informs our domain-adaptive pretraining strategy, where we employ domain-aware masking to prioritize biomarker-relevant terms during the self-supervised DASC phase. These unsupervised and semi-supervised adaptation strategies provide valuable approaches for adapting LLMs to specialized domains with limited labeled data. Our research combines these insights with domain-specific self-supervised pretraining techniques to develop effective models for Alzheimer's biomarker applications.

2.4 Research Gaps and Our Contributions

Based on this literature review, we identify several important research gaps that our work addresses:

- **Domain-Specific Self-Supervised Pretraining for Alzheimer's Biomarkers:** While significant research has explored biomedical adaptations of LLMs through continued pretraining on general biomedical corpora, few studies have developed domain-aware pretraining objectives specifically for highly specialized subdomains like Alzheimer's Disease biomarkers. The unique terminology, biomarker nomencla-

ture, and disease-specific relationships in ADBM literature require targeted adaptation strategies beyond general biomedical pretraining.

- **Domain-Aware Masking for Biomarker Literature:** Current pretraining approaches for biomedical LLMs typically use random masking or general span corruption objectives derived from broad scientific text. There is limited investigation into domain-aware masking strategies that prioritize disease-specific and biomarker-relevant terms during self-supervised pretraining, particularly for specialized subdomains within biomedicine.
- **Evaluation of Self-Supervised Pretraining Impact:**Most domain adaptation studies in biomedical NLP apply either direct supervised fine-tuning or general continued pretraining without systematically isolating and measuring the contribution of domain-aware self-supervised pretraining objectives. There is a need for controlled experiments that compare models adapted with domain-specific self-supervised pretraining against those trained with supervised fine-tuning alone.

Our research addresses these gaps through the following contributions:

- **ntroduction of Domain Aware Span Corruption (DASC):**We propose DASC, a novel self-supervised pretraining objective inspired by the T5 span corruption objective, specifically adapted for the Alzheimer's biomarker domain. DASC employs domain-aware masking that prioritizes biomarker-relevant terminology during pretraining.
- **Systematic Evaluation of DASC's Impact:** We conduct controlled experiments comparing FLAN-T5 XL models fine-tuned with and without prior DASC pretraining on a text classification task. This enables us to isolate and measure the specific contribution of domain-aware self-supervised pretraining to downstream task performance.
- **Application to Alzheimer's Biomarker Domain:** We demonstrate the effectiveness of DASC for adapting LLMs to the highly specialized ADBM subdomain, providing insights into domain adaptation strategies for niche biomedical areas with limited training data.

These contributions advance our understanding of how domain-aware self-supervised pretraining objectives can enhance LLM adaptation for specialized biomedical subdomains, with specific application to Alzheimer's Disease biomarker research.

3. Methodology

Using the *Domain Aware Span Corruption* method, we experimentally fine-tune the FLAN-T5 XL model, then analytically evaluate the model's hidden state representations of in-domain text against out-of-domain texts. Additionally, we evaluate our DASC trained model against a base model on a Question Answering (QA) task related to Alzheimer's disease.

3.1 Base model

In this work the Flan-T5 XL LLM serves as the base model for domain adaptation. Flan-T5 XL is a 3-billion parameter encoder-decoder transformer model that extends the original T5 model. Similar to T5 it is a sequence-to-sequence model, however unlike it, all versions of Flan-T5 have been instruction tuned on a diverse set of tasks to improve generalization and instruction following capabilities. (Chung et al., 2022; Raffel et al., 2020). Our fine-tuning leverages this generalization ability for our QA task. The structure for the QA task dataset is described in later sections. The 3-billion parameter model was chosen as the base model as it offered a great trade-off between computational requirements and performance.

3.2 Data Acquisition

3.2.1 Automated annotated data

In order to perform semi self-supervised training; raw Pubmed articles from the Pubmed database were downloaded using the BioC API. After which tools provided by the National Center for Biotechnology Information (NCBI) were used to tag gene and protein names from the pubmed articles. Specifically the GNormPlus system was used to identify gene/protein names from the Pubmed articles. (Wei et al., 2015).

3.2.2 QA testing data

For testing purposes, annotated QA pairs from the PubmedQA dataset (Jin et al., 2019) and synthetic QA pairs data generated from ChatGPT-4o were used(OpenAI, 2024).

3.3 Training

The semi self-supervised fine-tuning we performed is inspired by the *span corruption* denoising pretraining objective on which both the base T5 models and Flan-T5 models were trained. (Raffel et al., 2020; Chung et al., 2022)

3.3.1 Span Corruption Denoising objective

The T5 team employs a span-corruption training objective during the pre-training stage. This objective corrupts and drops contiguous, randomly spaced spans of tokens. After which, all corrupted spans are replaced by sentinel tokens with a token ID unique to that span. The target sequence is then made up of all the dropped sequences, delimited by their corresponding sentinel tokens. (Raffel et al., 2020) It is our belief that favoring span masking of domain relevant tokens improves the domain understanding of an LLM. DASC takes mutli-token spanning annotations and replaces their occurrences in the input sequence with a unique sentinel token. If a 15% corruption rate has not been achieved after DASC, more span corruption is applied following a *Poisson distribution* with a mean span length of 3.

3.4 Tests

We perform two tests to measure the effects of DASC. The first test uses a linear probe to measure if and how the model encodes Alzheimer's disease knowledge. This by itself is not descriptive of the model's performance on downstream NLP tasks within the AD domain. Therefore, we perform a Question Answering test in order to assess if DASC has any effect on the model's downstream performance.

3.4.1 Linear Probing

Linear probes have been used a tool for understanding the interpretability of deep neural networks. An earlier work that introduced them uses probes as a diagnostic tool to measure the linear separability of features in a Convolutional Neural Network (Alain and Bengio, 2016). However, recently (Park et al., 2023) have shown that linear probes recover the linear directions LLMs encode abstract concepts. By training a logistic regression model on the hidden state representations of input sequences, we analyzed how well a DASC fine-tuned model encodes these concepts compared to a base model. Input sequences include texts that are within the AD domain (*i.e* positive labels) and without (negative labels).

EXPLAIN HOW YOU TRAINED THE PROBE HERE. I.E. MEAN POOLING THE HIDDEN STATE REPR

3.4.2 QA testing

DON'T FORGET TO EXPLAIN THIS LATER ON!!!!

4. Results

The linear probe test was performed across the 24 layers of our DASC finetuned model vs a base flan T5-XL model. The probes were then evaluated using *precision, recall and f1-score*. Table 1.0 shows these results in detail. From the results it is clear that DASC has made slight improvements in the models internal representation of ADBM concepts, leading to a greater linear distinction between in-domain vs out-of domain concepts.

Additionally, for the QA task, DASC shows an approximate (≈10%) reduction in the model's perplexity, indicating that the model assigns about 10% higher average probability to the correct next tokens. Perplexity is used as it provides a reliable measure of how well the model's generations align with the expected text distribution. See table 3.0 for results.

	Accuracy	F1	Precision
Layer x	N/A	N/A	N/A
Layer y	N/A	N/A	N/A
Layer Last	0.91	0.88	0.90

Table 1: Linear Probe on DASC. Performance metrics across layers

	Accuracy	F1	Precision
Layer x	N/A	N/A	N/A
Layer y	N/A	N/A	N/A
Layer z	0.91	0.88	0.90

Table 2: Linear Probe on base model. Performance metrics across layers

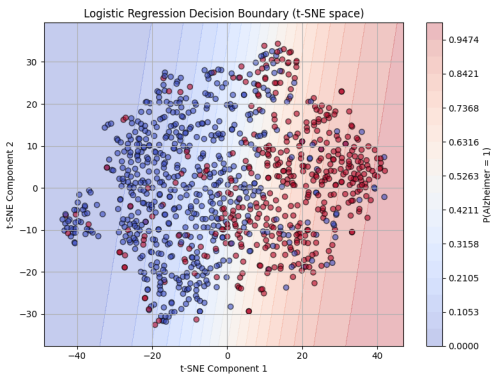
Model	**Training Objective**	**Domain**	**Test PPL ↓**
Base T5	Span corruption	General	41.46 ± 0.19
Domain-T5 (ours)	Domain corruption	Alzheimer's	**37.33 ± 0.20**

Table 3: Perplexity comparison between base and domain-finetuned models. Lower is better (↓).

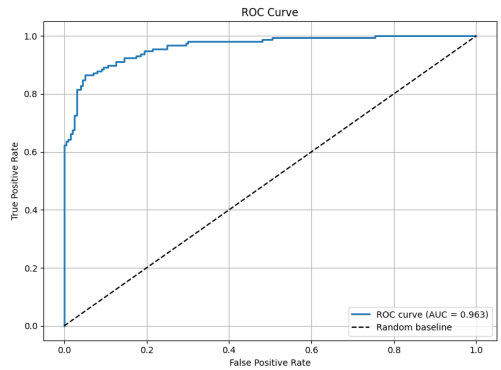
5. Conclusion

This paper studies the impact of DASC, a novel semi self-supervised training method, on Alzheimer's Disease biomarker text. DASC was applied to fine-tune the FLAN T5-XL Large Language Model, which was selected as it offered a great tradeoff between size and speed. While preliminary results suggest that DASC enhances the model's domain understanding, further large scale training with higher quality dataset and different models. Additionally, the promises of DASC should be explored in different domains to explore it's adaptability in a range of contexts.

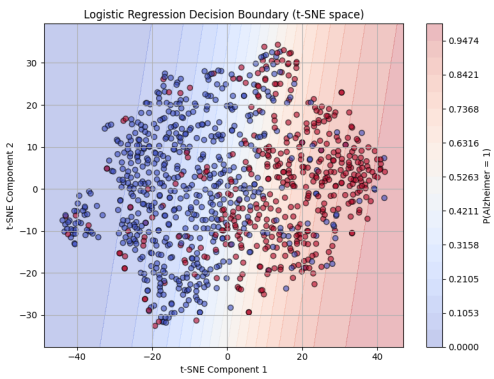
Although linear probing provides a useful measure of domain-specific representation, the evaluation structure can be expanded to include more complex tasks that demand deeper domain reasoning. For instance, distinguishing between different biomarkers or inferring whether a text indicates abnormality in specific biomarkers would offer stronger evidence of domain understanding. These experiments were not conducted here due to the lack of a sufficiently large and curated dataset at the time of this study.



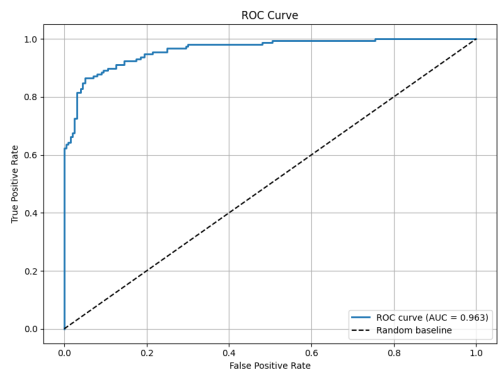
(a) Adapted Model t-SNE



(b) Adapted Model ROC



(c) Base Model t-SNE



(d) Base Model ROC

Figure 1: Left: T-SNE maps of last hidden state representation. Right: ROC Curve of Logistic Regression Probe

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.

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

Castro, D. M., Dillon, C., Machnicki, G., and Allegri, R. F. (2010). The economic cost of alzheimer's disease: Family or public-health burden? Dementia & neuropsychologia, 4(4):262–267.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022). Scaling instruction-finetuned language models.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70):1–53.

Counts, S. E., Ikonomovic, M. D., Mercado, N., Vega, I. E., and Mufson, E. J. (2017). Biomarkers for the early detection and progression of alzheimer's disease. Neurotherapeutics, 14(1):35–53.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36:10088–10115.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., et al. (2021). Pre-trained models: Past, present and future. AI Open, 2:225–250.

Hand, D. J., Christen, P., and Kirielle, N. (2021). F*: an interpretable transformation of the f-measure. Machine Learning, 110(3):451–456.

Hiwarkhedkar, S., Mittal, S., Magdum, V., Dhekane, O., Joshi, R., Kale, G., and Ladkat, A. (2023). Textgram: Towards a better domain-adaptive pretraining. In International Conference on Speech and Language Technologies for Low-resource Languages, pages 161–173. Springer.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In International conference on machine learning, pages 2790–2799. PMLR.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. (2019). Pubmedqa: A dataset for biomedical research question answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577.

Kimura, Y., Komamizu, T., and Hatano, K. (2024). L3masking: Multi-task fine-tuning for language models by leveraging lessons learned from vanilla models. In Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 53–62.

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.

OpenAI (2024). gpt-4o (april 29 version).

Panagides, R. K., Fu, S. H., Jung, S. H., Singh, A., Eluvathingal Muttikkal, R. T., Broad, R. M., Meakem, T. D., and Hamilton, R. A. (2024). Enhancing literature review efficiency: A case study on using fine-tuned bert for classifying focused ultrasound-related articles. AI, 5(3):1670–1683.

Park, K., Choe, Y. J., and Veitch, V. (2023). The linear representation hypothesis and the geometry of large language models. arXiv preprint arXiv:2311.03658.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Shi, S., Hu, K., Xie, J., Guo, Y., and Wu, H. (2024). Robust scientific text classification using prompt tuning based on data augmentation with l2 regularization. Information Processing & Management, 61(1):103531.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Wang, H., Sun, M., Li, W., Liu, X., Zhu, M., and Qin, H. (2023). Biomarkers associated with the pathogenesis of alzheimer's disease. Frontiers in Cellular Neuroscience, 17:1279046.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2015). Gnormplus: An integrative approach for tagging gene, gene family and protein domain. BioMed Research International, page Article ID 918710. Text Mining for Translational Bioinformatics special issue.

Wilczok, D. (2025). Deep learning and generative artificial intelligence in aging research and healthy longevity medicine. Aging (Albany NY), 17(1):251.

Winchester, L. M., Harshfield, E. L., Shi, L., Badhwar, A., Khleifat, A. A., Clarke, N., Dehsarvi, A., Lengyel, I., Lourida, I., Madan, C. R., et al. (2023). Artificial intelligence for biomarker discovery in alzheimer's disease and dementia. Alzheimer's & Dementia, 19(12):5860–5871.

Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., and Wang, Y. (2024). Pmc-llama: toward building open-source language models for medicine. Journal of the American Medical Informatics Association, 31(9):1833–1843.

Zeng, Q., Qiu, Z., Hwang, D. Y., He, X., and Campbell, W. M. (2024). Unsupervised text representation learning via instruction-tuning for zero-shot dense retrieval. arXiv preprint arXiv:2409.16497.

Zetterberg, H. and Bendlin, B. B. (2021). Biomarkers for alzheimer's disease—preparing for a new era of disease-modifying therapies. Molecular psychiatry, 26(1):296–308.

Zhang, Z.-Y., Zhang, J., Yao, H., Niu, G., and Sugiyama, M. (2024). On unsupervised prompt learning for classification with black-box language models. arXiv preprint arXiv:2410.03124.