GENERATIVE LARGE LANGUAGE MODELS FOR TRANSPARENT ARTIFICIAL INTELLIGENCE IN CLINICAL RESEARCH

GENERATIVE LARGE LANGUAGE MODELS FOR TRANSPARENT ARTIFICIAL INTELLIGENCE IN CLINICAL RESEARCH: ENHANCING INTERPRETABILITY THROUGH APPRAISAL AND EXPLANATION

By FANGWEN ZHOU, B.H.Sc

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the

Requirements for the Degree Master of Science in eHealth

McMaster University MASTER OF SCIENCE (2025) Hamilton, Ontario (eHealth)

TITLE: Generative large language models for transparent artificial intelligence in clinical research: Enhancing interpretability through appraisal and explanation AUTHOR: Fangwen Zhou, B.HSc. (McMaster University) SUPERVISOR: Dr. C. Lokker NUMBER OF PAGES: xxi, 182

LAY ABSTRACT

Artificial intelligence (AI) is increasingly used in clinical research to help automate the classification and evaluation of scientific studies. However, understanding how complex AI models make decisions, known as interpretability, is important for ethical use, and it remains a major challenge. This thesis explores how generative language models, particularly GPT from OpenAI, can enhance the interpretability. Two approaches were tested: 1) using GPT to classify medical research articles by explaining its reasoning, and 2) using GPT to interpret decisions made by another advanced model by assigning a numerical importance value, called feature attribution, to each word. Results showed GPT was effective in classifying articles and explaining its own decisions, but it was not able to effectively explain other models using feature attributions. These results support the use of GPT to improve the transparency and accessibility of automated medical text classification and highlight potential future research in this field.

ABSTRACT

Background

The rapid growth of medical literature necessitates effective, transparent automation tools for classification. Generative large language models (LLMs), including the Generative Pre-trained Transformer (GPT), have the potential to provide transparent classification and explain other black box models.

Objective

This sandwich thesis evaluates the performance of GPT in 1) classifying biomedical literature compared with a fine-tuned BioLinkBERT model, and 2) explaining the decision of encoder-only models with feature attributions compared to traditional eXplainable AI (XAI) frameworks like SHapley Additive exPlanations (SHAP) and integrated gradients (IG).

Methods

Randomly sampled, manually annotated clinical research articles from the Health Information Research Unit (HIRU) were used along with a top-performing BioLinkBERT classifier. In Chapter 2, GPT-4o and GPT-o3-mini were used either alone or with BioLinkBERT's predictions in the prompt to classify article methodological rigour based on HIRU's criteria. Either the title and abstract or the full text was provided to GPT. Performance was compared to the BioLinkBERT model and assessed primarily using Matthew's correlation coefficient (MCC). In Chapter 3, GPT-4o was used to generate feature attributions for the BioLinkBERT model through masking perturbations and was

compared to SHAP and IG using a modified area under the perturbation curve (AOPC) metric which gives a measure of performance.

Results

GPT-4o alone, using full text (MCC 0.429), achieved comparable classification performance to BioLinkBERT (MCC 0.466). Performance was worse with other models and inputs. As a perturbation explainer, GPT-4o's (AOPC 0.029) performance was poor and significantly underperformed compared to SHAP (AOPC 0.222) and IG (AOPC 0.225). The identified important tokens by GPT did not align with the manual appraisal criteria.

Conclusion

GPT has potential in appraising biomedical literature, even without explicit training. GPT's transparency through textual explanations improves interpretability. GPT's poor performance in generating faithful feature attributions warrants future research. The inherent variability and stochasticity of GPT outputs necessitate careful prompting and reproducibility measures.

ACKNOWLEDGEMENT

I would like to express my gratitude to my eHealth supervisor and mentor, Dr. Cynthia Lokker, for your consistent support throughout the program and during my time at HIRU. Your encouragement, understanding, and willingness to accommodate uncertainties surrounding my health have meant a great deal to me. You continue to inspire me to stay curious and kind.

I am also thankful to the members of my supervisory committee, Dr. Muhammed Afzal and Dr. Ashirbani Saha, for generously sharing their expertise in artificial intelligence and for providing timely, constructive feedback that shaped this work.

A special thank you goes to Rick Parrish and the entire team at HIRU for your continued support and valuable feedback throughout this work.

Thank you to Margaret Leyland, Sheila Richardson, and the rest of the eHealth faculty and staff for your guidance throughout the program.

To all my friends, and in particular, Jiawen Deng, Jasdeep Dhillon, Kiyan Heybati, Kevin Hu, Benjamin Miralles, Julia Spafford, Samson Wu, and Qi Kang Zuo, thank you for your constant reassurance and support during both this academic chapter and a personally challenging time.

To Diya Li, thank you.

Last but not least, I want to extend my most sincere appreciation to my mother, Yongchang Yu, and the rest of my family for your unwavering, unconditional support throughout my academic and personal journey.

To everyone, thank you. This work and the next step in my journey as an MD/PhD student at Queen's University would not have been possible without your support and guidance.

I was funded by Mitacs's Business Strategy Internship grant (IT42947) with matching funds from EBSCO. The use of GPT was graciously supported by OpenAI's Researcher Access Program (0000014443). The computational resources used to fine-tune the encoder transformer were kindly provided by the Digital Research Alliance of Canada. None of the above-mentioned parties was involved in the conceptualization, design, conduct, and dissemination of this thesis.

TABLE OF CONTENTS

FIGURES AND TABLES

ABBREVIATIONS AND SYMBOLS

AI Artificial intelligence

AP Average precision

API Application programming interface

AOPC Area over the perturbation curve

AUPRC Area under the precision-recall curve

AUROC Area under the receiver operating characteristic curve

BERT Bidirectional Encoder Representations from Transformers

CI Confidence interval

COVID-19 Coronavirus Disease 2019

CPU Central processing unit

DDR Double data rate

DL Deep learning

EBM Evidence-based medicine

ELECTRA Efficiently Learning an Encoder that Classifies Token Replacements Accurately

FN False negative

GBM Gradient Boosted Machine

GDDR Graphics double data rate

GPU Graphics processing unit

GPT Generative Pre-trained Transformer

HBM High bandwidth memory

HIRU Health Information Research Unit

IG Integrated gradients

JSON JavaScript Object Notation

LIME Local Interpretable Model-agnostic Explanations

LLM Large language model

MCC Matthew's correlation coefficient

ML Machine learning

NLP Natural language processing

NN Neural network

PDF Portable Document Format

PLUS McMaster Premium LiteratUre Service

PMID PubMed Identifier

RAM Random access memory

RCT Randomized controlled trial

RNN Recurrent neural network

ROB Risk of bias

SHAP SHapley Additive exPlanations

SL Shallow learning

SVM Support vector machine

TIAB Title and abstract

WSS – Work saved over sampling

XAI Explainable artificial intelligence

DECLARATION OF ACADEMIC ACHIEVEMENT

I, Fangwen Zhou, hereby declare that the work presented in this thesis is the result of my own original research conducted under the supervision of Dr. Cynthia Lokker at the Health Information Research Unit (HIRU), Department of Health Research Methods, Evidence, and Impact, Faculty of Health Sciences, McMaster University. This research was undertaken as part of the requirements for the completion of the MSc in eHealth.

The study was designed and conceptualized by me, Dr. Cynthia Lokker, and the two members of my thesis committee, Drs. Muhammed Afzal and Ashirbani Saha. I was solely responsible for the review of relevant literature, data preparation, model development, implementation of prompting strategies, execution of experiments, and statistical analysis. I was responsible for developing the methodology involving large language models (LLMs) and evaluating their performance in biomedical literature classification and explanation. All programming, data management, and statistical analyses were carried out by me using Python and associated libraries.

The datasets used in both experiments were sourced from existing databases curated by HIRU (Premium LiteratUre Service and Clinical Hedges). The feature attribution frameworks (SHapley Additive eXpalantions and integrated gradients) and evaluation metrics were adapted and implemented by me to assess and compare model explainability. The primary encoder-only transformer model used in both experiments for comparison and explanation, BioLinkBERT, was primarily developed by me with inputs from HIRU and colleagues (F. Zhou, Parrish, et al., 2025). All adaptations, experimental designs, LLM interactions, and critical analyses presented in this thesis are my own work.

All collaborators, tools, and sources that supported this research have been appropriately credited and referenced throughout the thesis.

This thesis contains no material previously published or written by another person except where due reference is made. The research has not been submitted for any other degree or qualification at any other academic institution.

Chapter 2 CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Fangwen Zhou: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, resources, software, validation, visualization, writing – original draft, writing – review and editing. **Muhammad Afzal:** Conceptualization, investigation, methodology, supervision, validation, writing – review and editing. **Ashirbani Saha:** Conceptualization, investigation, methodology, supervision, validation, writing – review and editing. **Rick Parrish:** Data curation, investigation, methodology, software, validation, writing – review and editing. **R. Brian Haynes:** Validation, writing – review and editing. **Alfonso Iorio:** Investigation, resources, writing – review and editing. **Cynthia Lokker:** Conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing – original draft, writing – review and editing.

Chapter 3 CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Fangwen Zhou: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, resources, software, validation, visualization, writing –

original draft, writing – review and editing. **Ashirbani Saha:** Conceptualization, investigation, methodology, supervision, validation, writing – review and editing. **Muhammad Afzal:** Conceptualization, investigation, methodology, supervision, validation, writing – review and editing. **Rick Parrish:** Data curation, investigation, methodology, software, validation, writing – review and editing. **R. Brian Haynes:** Validation, writing – review and editing. **Alfonso Iorio:** Investigation, resources, writing – review and editing. **Cynthia Lokker:** Conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing – original draft, writing – review and editing.

CHAPTER 1

INTRODUCTION

1 | 1 Health Information Research Unit

Evidence-based medicine (EBM) is a framework that informs clinical decisions according to the best available literature, clinical knowledge, experience, and patient preferences (Tenny & Varacallo, 2022). According to the hierarchy of evidence, randomized controlled trials (RCTs) are considered the strongest form of primary evidence and is able to establish causality. To aid in the search for relevant, rigorous literature, the Health Information Research Unit (HIRU) at McMaster University has been maintaining the Premium LiteratUre Service (PLUS) (Haynes et al., 2006). The service involves retrieving, categorizing, and appraising articles from 123 journals indexed in PubMed, a leading repository of biomedical literature, by research methods experts on a daily basis. Specifically, articles are classified by human reviewers into one of four mutually exclusive classes: 1) original study, 2) review, 3) evidence-based guideline, and 4) non-experimental. Articles classified as 1), 2), or 3) can then be labelled with eight non-mutually exclusive labels: 1) treatment, 2) primary prevention, 3) diagnosis, 4) prognosis, 5) etiology, 6) quality improvement, 7) economics, and 8) other. Subsequently, the studies are appraised for their methodological rigour as being either sound or unsound. Methodologically sound articles are then sent to practicing clinicians worldwide to determine their clinical relevance to practice and newsworthiness. Those that are deemed both clinically relevant and newsworthy are ultimately delivered to clinician subscribers and other services, such as Evidence Alerts. The methodological criteria have been modified from those used when developing Clinical Hedges, a database curated by HIRU

in 2000, comprising 49,028 unique records published in 161 journals indexed in MEDLINE (Wilczynski et al., 2005).

1 | 2 The Increasing Burden in Medical Knowledge Translation

In the healthcare field, the sheer volume of clinical research has become a major challenge. The rapid expansion of digital information has resulted in an overwhelming amount of unstructured text data. For example, PubMed indexed over 36 million articles by 2025, with nearly one million new publications added each year (*MEDLINE PubMed production statistics*, 2018). Although this surge in literature offers tremendous opportunities to advance medical science, it also poses significant difficulties for clinicians and researchers who must navigate vast amounts of information to stay up to date. This issue is further compounded by the accelerating pace of medical knowledge, which was estimated in 2020 to double every 73 days—an enormous leap from the seven-year doubling rate observed in 2010 (Densen, 2011). As a result, there is an urgent need for powerful, automated tools that can efficiently classify and retrieve text-based information to support timely and accurate clinical and research decision-making.

1 | 3 Artificial Intelligence (AI) and Natural Language Processing (NLP)

NLP is a field of AI that involves using machines to understand, interpret, process, and generate human language (Stryker & Holdsworth, 2025). Text classification is an extensively explored NLP task, and it involves assigning predefined categories to free, unstructured text and is critical for information retrieval, content organization, and

decision support across multiple domains (Taha et al., 2024; Wan, 2023). Prominent examples of text classification include sentiment analysis (S. Kumar et al., 2023) and spam email detection (AbdulNabi & Yaseen, 2021). Several major AI architectures have been extensively explored for the task of text classification.

1 | 3.1 Rule-based methods

The most rudimentary AI systems for NLP leverage rule-based, deterministic methods, such as pattern matching with regular expressions (Kowsari, Meimandi, et al., 2019; Markov et al., 2021). One example of this is the Bing Liu Lexicon, a dictionary of 6,790 words mapped to either a positive or a negative sentiment (n.d.-a). Sentiment analysis with a lexicon-based approach may leverage this dictionary and count the number of words with either a positive or negative sentiment (Haddaoui et al., 2025). However, the rules require substantial manual labour to develop and maintain. They often scale poorly as texts become longer with complex contextual dependencies, as exponentially more rules are required (Chatla et al., 2024; Kotelnikova et al., 2021; B. Kumar et al., 2024; X.-L. Li, n.d.; Michael et al., 2023).

1 | 3.2 Shallow learning (SL) and traditional deep learning (DL) models

As opposed to rule-based AI, supervised ML is a branch that involves automatically creating a functional model from a set of inputs and outcomes, which then can be applied to another set of inputs. ML is further categorized into SL and DL (Sarker, 2021; Xu et al., 2021). Specifically, SL does not involve complex hierarchies or layers of features and

transformations. On the contrary, DL most often refers to neural networks (NNs) and their variants, which almost always involve numerous layers of computations.

SL techniques, such as Naïve Bayes and Support Vector Machine (SVM), are relatively simple architectures that leverage well-defined statistical principles or geometric decision boundaries and have been explored in text classification (Kowsari, Meimandi, et al., 2019; Q. Li et al., 2020). Compared to rule-based systems, SL does not require the manual definition of rules but rather learns relationships among the input features. This is a process called training, and the trained model can then be applied to new pieces of text for prediction. Common features include bag-of-words, where the feature represents the frequency of each word in the English vocabulary (Taha et al., 2024), and term frequency-inverse document frequency, which is the product of how often a word appears in the document and the inverse of how often the word appears in a set of documents (Taha et al., 2024). While SL techniques are typically more effective and scalable than rule-based systems (Muliono & Tanzil, 2018; Taha et al., 2024), they require meticulous feature engineering and data augmentation to extract meaningful representations from free text (Q. Li et al., 2020; Oleynik et al., 2019; Yin et al., 2014). Furthermore, due to their simplicity, their scalability is often limited, and they are unable to fully capture complex semantic dependencies and relationships (Le et al., 2017; Lokker, Abdelkader, et al., 2024; Nassif et al., 2021). Nevertheless, SL methods remain relevant today due to their computational efficiency and often superior performance when training data is limited (Aphinyanaphongs et al., 2005; Q. Li et al., 2020; Oleynik et al., 2019; Pasupa & Sunhem, 2016).

In contrast to SL methods, DL marked a significant advancement. The multilayered nature of DL enables more complex relationships to be learned, often precluding the need for sophisticated feature engineering or data augmentation (Q. Li et al., 2022; Wan, 2023; Xu et al., 2021). Recurrent NNs (RNNs) and their variants, such as bidirectional long-short-term memory and gated recurrent units, have demonstrated improved performance in a variety of NLP tasks due to their ability to model sequential relationships within text (L. Guo et al., 2018; Shahid et al., 2020; Sunagar & Kanavalli, 2022; Y. Zhou, 2020). However, RNNs suffer from issues such as vanishing and exploding gradients (Noh, 2021; Wu et al., 2024), unidirectional context interpretation (Cui et al., 2018), and challenges with rare vocabulary (Mienye et al., 2024; Ravi et al., 2020). RNN variants mitigate these concerns to a certain extent. However, compared to vanilla RNN, these variants suffer from a substantial increase in model complexity and longer computational times (Shiwei Liu et al., 2021).

1 | 3.3 Transformer models

The transformer architecture was initially introduced in 2017 and typically consists of either an encoder, decoder, or both (Aitken et al., 2021; Nielsen et al., 2024; Vaswani et al., 2017). The introduction of transformer-based pre-trained language models has revolutionized text classification by addressing many of the limitations inherent in earlier approaches (Pu et al., 2024; Vaswani et al., 2017). These are typically end-to-end models that require little data preprocessing. Compared to sequential recurrent NNs, transformers are able to process all inputs in parallel with positional encodings that map the relative

location of each textual token, resulting in drastically improved training times (H. Zhang & Shafiq, 2024). Tokens are pieces of text that transformer models use as input features and are generated by model-specific tokenizers (Zimmerman et al., 2024). Among the tokenizers, subword tokenizers are one of the most commonly used, which break down words into subwords for processing (Velayuthan & Sarveswaran, 2024). For instance, the word "prespecified" would be processed as three separate inputs, "pres," "##pec," and "##ified." Transformer's multihead self-attention mechanism models how each token would be affected by all other tokens in the input (Hernández & Amigó, 2021), mitigating issues around vanishing and exploding gradients and resulting in more accurate representations of long-range dependencies (Devlin et al., 2018). In other words, self-attention allows a model to better interpret each word's meaning in consideration of its context, especially over long pieces of text (Vaswani et al., 2017). In addition to parallelization and self-attention, transformers respond well to transfer learning, a technique where a model developed for a particular task is reused as the starting point for a model on a related task (Hosna et al., 2022). This method significantly reduces the amount of time and resources required when compared with training a model from scratch, resulting in domain-specific variants that achieve better performance in their respective fields (Chalkidis et al., 2020; J. Lee et al., 2019).

1 | 3.3.1 Encoder-only transformers

Encoder-only transformer architectures, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), consist solely of the encoder stack. The

input text is processed through subword tokenization, and the output is an embedding (a numerical vector) of fixed size that is the sum of the token, positional, and segment embeddings. For text classification specifically, a special token at the beginning, "[CLS]", is inserted, processed, and passed to a traditional NN for classification. This token can also be used as a feature in other types of classifiers (Afzal et al., 2024).

Encoder-only transformers are typically pre-trained on large corpora of text data. BERT, for instance, is trained on BookCorpus with masked language modelling, where a portion of tokens are masked and predicted, as well as next sentence prediction, where the model determines whether a pair of sentences is consecutive (Devlin et al., 2018). Another architecture, Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), utilizes replaced token detection where a generator-discriminator pair is trained, in which the generator replaces tokens with plausible alternatives and the discriminator attempts to detect replaced tokens (Clark et al., 2020). This approach, compared to BERT's masking, enables learning from all input positions as opposed to masked ones and is typically more efficient computationally (Cortiz, 2021).

These architectures are commonly adapted to domain-specific texts, as previously mentioned. BioBERT and BiomedBERT are prominent examples of BERT variants that focus on biomedical text, similar to BioELECTRA (Gu et al., 2022; Kanakarajan et al., 2021; J. Lee et al., 2019). These models may be pre-trained from scratch using domain-specific texts (Gu et al., 2022; Kanakarajan et al., 2021) or continually pre-trained from

general corpora (J. Lee et al., 2019). These domain-specific variants typically outperform their general counterparts in domain-specific NLP tasks.

Pre-trained models are often fine-tuned for specific downstream tasks via supervised learning (Lokker et al., 2023; F. Zhou, Parrish, et al., 2025). Compared to training from scratch, pre-training and fine-tuning with a low learning rate require substantially less data and computational resources and often achieve better performance, especially for the biomedical corpora that often involve complex vocabulary and contextual dependencies (Devlin et al., 2018; J. Lee et al., 2019; Lokker et al., 2023).

1 | 3.3.2 Decoder transformers and generative large language models (LLMs)

In contrast to encoder-only models, decoder-based transformers focus on generative language modelling with autoregressive self-attention, where each token attends only to previous tokens (Roberts, 2023). This directionality allows the model to predict, or generate, future tokens sequentially based on what is already present (Jiawen Deng, Heybati, Park, et al., 2024). Decoder models are typically pre-trained using causal language modelling, where a piece of text is given, and the model is asked to predict subsequent tokens (*Decoder models - Hugging Face NLP Course*, n.d.).

This architecture gave rise to generative LLMs, including the Generative Pretrained Transformer (GPT) from OpenAI, that have garnered significant attention. LLMs, as the name suggests, are massive models trained on diverse internet texts using a combination of causal language modelling and reinforcement learning. One of the most advanced models by OpenAI, GPT-4, is speculated to have approximately 1.8 trillion

parameters by some sources (Bastian, 2023; Howarth, 2024), while BERT comprises only 110 million (*bert: TensorFlow code and pre-trained models for BERT*, n.d.). Furthermore, these models can be fine-tuned with domain-specific text, similar to encoder-only transformers (Luo et al., 2022).

These models are typically used by providing them with a "prompt" or instruction that starts the autoregressive sequence (Schulhoff et al., 2024). The model then generates subsequent tokens sequentially based on the prompt and previous generations. The combination of the autoregressive nature and extensive training enables the models to learn complex representations and is typically considered task-agnostic. While the primary task is text generation, LLMs can be easily adapted to a myriad of NLP tasks, including classification, through prompting or fine-tuning (E. Guo et al., 2024; *Holistic Evaluation of Language Models (HELM)*, n.d.; *Open LLM Leaderboard - a Hugging Face Space by open-llm-leaderboard*, n.d.; Jin et al., 2019).

In addition to architecture, the sophistication of the prompt can drastically affect an LLM's performance (J. He et al., 2024; J. Kim et al., 2023). Consequently, prompt engineering has been an active area of research, and methods such as role prompting, chain-of-thought, decomposition, and few-shot prompting have shown promise in most circumstances (Schulhoff et al., 2024).

1 | 4 Automated Biomedical Literature Processing

Clinical practitioners and researchers rely on article retrieval from medical databases to inform clinical decisions and synthesize the most up-to-date evidence. However,

healthcare research has become increasingly complex and abundant, making it challenging for readers to remain current with all available evidence in their respective fields (Gai et al., 2021; Markey et al., 2024; Mendlovic et al., 2022; *Number of clinical trials by year, country, WHO region and income group (1999-2022)*, n.d.; J. Sun et al., 2021; Zhao et al., 2022). Empirical search systems, such as the Medical SubHeadings (MeSH) indexing terms and author-identified keywords, provide tags for articles to improve the efficiency of literature retrieval from medical databases. These approaches, however, are not without issues. For example, it may take up to one year for an article to be fully indexed in MEDLINE (Irwin & Rackham, 2017), and keywords may vary from author to author despite similar contexts (Dhammi & Kumar, 2014).

Since 2002, the Medical Text Indexer-Automatic has been used to automate MeSH indexing with rule-based methods (*MEDLINE 2022 initiative: Transition to automated indexing*, 2021). The MTI-NeXt Generation adopts an NN model and significantly outperforms its predecessor with regard to recall (*MTIX: The next-generation algorithm for automated indexing of MEDLINE*, 2024). Additionally, studies have also used ML to classify articles based on their topic or type (Qingyu Chen et al., 2022; Thushari et al., 2023), as well as to evaluate ML models based on standardized corpora, such as the Hallmarks of Cancer (Baker et al., 2016; Shicai Liu et al., 2021; Verma et al., 2024). Additionally, several experiments have used ML to classify articles based on methodological soundness (Aphinyanaphongs et al., 2005; Kilicoglu et al., 2009; Lokker, Abdelkader, et al., 2024; Lokker et al., 2023; Marshall et al., 2016) or relevance for systematic reviews (Aum & Choe, 2021; Chernikova et al., 2024; Lange et

al., 2021; Qin et al., 2021; Shekelle et al., 2017; van den Bulk et al., 2022). Software platforms aimed at supporting systematic reviewers, such as Covidence (*Covidence - Better systematic review management*, 2020) and Rayyan (Shanaa, 2021), are leveraging multiple model architectures to rank titles and abstracts (TIABs) based on their potential relevance to a review topic or classify whether they are RCTs.

More recently, studies have examined the role of LLMs in traditional biomedical NLP tasks, such as question-answering and relation extraction (Ateia & Kruschwitz, 2023; Bousselham et al., 2024; Rehana et al., 2023; J. Zhang et al., 2024). With regards to text classification, studies used GPT to conduct TIAB screening for systematic reviews (E. Guo et al., 2024), identify health-related tweets (Y. Guo et al., 2024), classify health advice in the scientific literature (S. Chen et al., 2024), as well as critical appraisal (Hasan et al., 2024; Lai et al., 2024; Pitre et al., 2023). In general, while GPT often performed worse than other methods, it precludes the need for lengthy training using a large, annotated dataset. Therefore, the convenience and flexibility are of particular interest to clinicians and researchers alike.

In response to advancements in machine automation, HIRU has conducted several experiments using ML for medical text classification. For example, Lokker et al. fine-tuned BERT and four variants for binary classification of the methodological soundness of clinical articles and reported 63% of work saved in a literature surveillance process while maintaining sensitivity >99% (Lokker et al., 2023). Another experiment leveraging AutoML and LightGBM has been conducted and published as well (Lokker, Abdelkader, et al., 2024). More recently, two additional experiments have been conducted, leveraging

a series of domain-specific BERT models to classify literature based on study design (F. Zhou, Lokker, et al., 2025) and methodological soundness specifically for RCTs (F. Zhou, Parrish, et al., 2025) and systematic reviews (F. Zhou, Afzal, et al., 2025). These experiments demonstrate satisfactory performance of BERT models, with the best models achieving an area under the receiver operating characteristic curve (AUROC) >94% and Matthew's correlation coefficient (MCC) ranging from 0.75 to 0.90.

1 | 5 Explainability in Artificial Intelligence

Important aspects of decisions in the clinical context are transparency and explainability, as they are crucial for correctness, reproducibility, ethical practice, and knowledge translation (Fukami, 2024). Similarly, decisions made on clinical literature must be supported with direct evidence as well (Gannot et al., 2017). For instance, Cochrane Risk of Bias (ROB) tools involve a series of clear, unambiguous decisions to ensure that the final result is justified (J. P. T. Higgins et al., 2011; J. A. C. Sterne et al., 2019). In high-stakes domains like biomedicine, the explainability of NLP models is crucial for building trust. A model's predictions should ideally be backed by rationales that align with human domain knowledge (Talebi et al., 2024).

Considering this, a major pitfall of DL, including transformers, is the lack of explainability of model decisions (Price, 2018). While DL can consider the vast array of variables and relationships in biomedical information, its sophistication comes at the cost of interpretability. In other words, as a model becomes more complicated, less is understood about its decision-making process. For instance, decisions from logistic

regression models can be traced to individual features' odds ratios and weights. In an NN, each feature is processed through multiple layers of neurons with individual weights, activations, and biases, rendering it much more difficult to interpret the impact of a certain feature. This is known as the infamous black box problem (Fomin & Astromskis, 2023).

Consequently, a model may achieve satisfactory performance yet provide minimal insights into how these decisions were reached. Therefore, it is unfeasible to predict how well a model can generalize to another dataset, to examine systematic biases in the model's decisions, or to further improve its performance in a constructive way other than trial-and-error. Due to the aforementioned reasons, the applicability of black-box models, especially those without external validation, is limited in medicine and clinical research (Wadden, 2021).

One solution to the black box issue is using a framework to assign importance attributions to each input feature with respect to how much they contribute to the model's output (M. Mersha et al., 2024). These frameworks can be broadly categorized as either model-specific or model-agnostic. They can also be categorized based on their scope, where local models attempt to explain individual decisions provided by an algorithm, while global models attempt to explain the entire workings of the algorithm.

1 | 5.1 Perturbation-based explainable artificial intelligence (XAI) frameworks

Perturbation-based frameworks work by systematically modifying input instances to assess how these alterations affect model output, thus identifying the relative importance

and contribution of each feature to the model's final decision (Hsieh et al., 2024). For instance, by changing or removing certain tokens from the input and observing how the model's predictions change, it can be inferred to a certain extent which tokens are important for the decision.

1 | 5.1.1 Local Interpretable Model-agnostic Explanations (LIME)

LIME is a framework used to explain the decisions of any classifier locally (M. T. Ribeiro et al., 2016). It approximates individual predictions with a linear model, which is simpler and interpretable. Specifically, it generates perturbed instances of an input instance and examines how changes to the instance affect the model's predictions. Then, weights to the perturbed instances are generated, and a linear model is fitted. However, LIME is strictly local, and each explanation would not be able to be generalized to other instances or to the model as a whole. LIME is also computationally expensive for large feature spaces, such as BERT models with a max token length of 512, and the fitted linear model is limited in complex models aimed at mapping non-linear relationships. For language tasks, LIME also struggles to generate meaningful perturbed instances, as small changes in words can result in significant shifts in semantic meaning or render the input grammatically incorrect.

1 | 5.1.2 Anchor

Anchor, developed by the same researchers behind LIME, builds upon the local, model-agnostic approach with a rule-based system (Molnar, 2024). Similar to LIME, Anchor is a

model-agnostic, local explanation framework. It uses reinforcement learning and a graph search algorithm. The framework outputs IF-THEN rules called anchors that can be applied to one or several local instances. Along with the rule, Anchor returns the coverage and precision of the rule, indicating how much of the perturbation space the rule applies to and the precision of the rule in the coverage. When generating these rules, those that are more specific (i.e., more AND statements) offer better precision but less coverage. A balance must be struck considering the precision and coverage of these rules.

1 | 5.1.3 SHapley Additive eXplanations (SHAP)

SHAP differs from LIME and Anchor in that it draws from cooperative game theory, specifically Shapley values, to attribute prediction contributions across all features (Lundberg & Lee, 2017). The Shaley value represents the average contribution of a feature to the prediction for a specific instance. These values can be aggregated over multiple instances due to Shapley's additive nature for a better representation of a feature's impact. To calculate the Shapley value, a baseline prediction value is first set at the probability when no features are present, which is the prevalence of the class. Then, the model's prediction with each feature subset combination is calculated, and the marginal contribution—the difference in the model's prediction—for the feature in each subset is calculated by subtracting the predicted value of the subset without the feature from the predicted value of the subset with the feature. A weighted average of all marginal contributions is used to calculate the Shapley value.

SHAP suffers from similar limitations as LIME, where significant computational effort is required, and its complexity scales factorially with the size of the feature space. This makes SHAP infeasible for typical DL text models with hundreds of features. To address this concern, approximation methods have been developed (Lundberg & Lee, 2017). One notable variant is the SHAP partition explainer, also known as partition SHAP (*shap.PartitionExplainer — SHAP latest documentation*, n.d.). Partition SHAP leverages a hierarchical tree-based partition by iteratively grouping the text using the cosine similarity of token embeddings, where close tokens are grouped first. This enables the capture of feature interactions (H. Chen et al., 2020; Lundberg & Lee, 2017). Owen value, a generalization of the Shapley value for hierarchical structures, is computed by masking or removing features in clusters together at each node of the hierarchical tree, thereby reducing the computational cost to quadratic complexity (López & Saboya, 2009).

1 | 5.2 Gradient-based explanation frameworks

As opposed to perturbation-based frameworks that rely on modifications to input instances, gradient-based explanations systematically examine model parameters to explain how the model makes predictions (Y. Wang et al., 2024). As the name implies, gradient-based frameworks leverage the model's own gradients during backpropagation and, therefore, are only applicable to differentiable architectures, including NNs. Consequently, they are typically more computationally efficient as well.

1 | 5.2.1 Vanilla gradients

Vanilla gradients are the baseline gradient method and involve computing the gradient of the model's output relative to the input features (Y. Wang et al., 2024). A larger gradient would indicate that the output would change more relative to changes in the input feature, indicating that the feature is relatively more important. This method is extremely efficient computationally (*28 saliency maps – interpretable machine learning*, n.d.). However, the method only captures the importance in a single iteration and may lead to overfitted attributions (Adebayo et al., 2018). Specifically, a well-converged NN would be saturated and have smaller gradients for confident predictions (Miglani et al., 2020), and in contrast, some instances can be unstable (Ghorbani et al., 2017).

1 | 5.2.2 SmoothGrad

SmoothGrad builds on top of vanilla gradients by averaging gradients over multiple noisy versions of the input, similar to perturbation-based methods (Smilkov et al., 2017). The addition of noisy inputs mitigates gradients with local spurious fluctuations, resulting in more consistent, reliable explanations.

1 | 5.2.3 Integrated Gradients (IG)

A more sophisticated gradient-based framework compared to vanilla and smooth gradients is IG (Sundararajan et al., 2017). As opposed to one-shot or few-shot gradients around the input, IG uses gradients on multiple steps along the path between a defined baseline and the instance input to establish feature importance. In other words, input

features are slowly added, step by step, from an empty baseline to the input instance, and the gradient at each step is considered. By accumulating infinitesimal gradients along this trajectory, IG produces attributions that sum up to the difference in model output between the input and the baseline. This results in more intuitive explanations, analogous to Aumann-Shapley values, which are a continuous analogue of Shapley values (Kirman et al., 1976). This approach also better mitigates issues surrounding gradient saturation and noise (Enguehard, 2023; Kirman et al., 1976; M. Ribeiro et al., 2024; Sikdar et al., 2021).

1 | 6 The Role of Generative LLMs for Interpretable Text Classifications

While the aforementioned ML model architectures and XAI frameworks have been examined in numerous studies, several limitations hinder their applicability in medical text classification (M. A. Mersha et al., 2025; Minaee et al., 2020; Taha et al., 2024). First, their explanations are limited to numeric attributions, and the interpretation remains a challenge for clinicians and clinical researchers (Zeng, 2024; Zytek, Pido, et al., 2024; Zytek, Pidò, et al., 2024). Second, the training and fine-tuning of traditional ML models require a substantial amount of high-quality annotations, which is often infeasible for tasks like systematic reviews (Golestaneh et al., 2024). Third, computational cost remains a significant consideration for both the training and the explanation of text classifiers, especially for deep NNs (Justus et al., 2018). Lastly, the development and implementation of these architectures require a substantial amount of technical knowledge. The effort of communication between researchers and software engineers adds another barrier to the workflow.

The rise of generative LLMs offers an interesting opportunity to tackle the above limitations. LLMs are able to provide plain-text rationale surrounding a decision using techniques such as chain-of-thought prompting, similar to how a critical appraiser would explain their ratings for the methods criteria of an article during conflict resolution (Wei et al., 2022). Additionally, LLMs that are pre-trained on merely general corpora have the potential to achieve satisfactory zero-shot performance (Ateia & Kruschwitz, 2023; Kojima et al., 2022). This flexibility allows LLMs to be used in a variety of tasks, including critical appraisal (Hasan et al., 2024; Lai et al., 2024; Pitre et al., 2023). Furthermore, LLMs can be queried remotely at a relatively accessible cost. For example, in April 2025, GPT-4o has input and output costs of USD$2.50 and USD$10.00 per one million tokens, respectively (*Precios*, n.d.). This mitigates concerns about local computational hardware requirements. LLM use also involves free text prompts, making them accessible for stakeholders without an extensive technical background.

1 | 7 Objective

The overarching objective of this thesis was to explore the applications of LLMs, specifically GPT models by OpenAI, to improve the interpretability of automated medical text classifications. GPT was used to 1) provide zero-shot interpretable classifications of medical literature, and 2) generate feature attributions using perturbations, similar to perturbation-based XAI frameworks. Original studies with the purpose of treatment, primary prevention, or quality improvement from HIRU's PLUS and Clinical Hedges datasets were leveraged for both experiments. For experiment 1), the ability of GPT to

explain model decisions was compared to the best-performing encoder-only model fine-tuned in a previous study (F. Zhou, Parrish, et al., 2025), and for experiment 2) performance of GPT was compared to the SHAP partition explainer and IG.

CHAPTER 2

ZERO-SHOT INTERPRETABLE BIOMEDICAL LITERATURE APPRAISAL WITH GENERATIVE LARGE LANGUAGE MODELS

2 | 1 Abstract

2 | 1.1 Background

The automation of clinical literature appraisal has garnered wide attention as an increasing number of articles are being published every year. While encoder-only transformer models have demonstrated strong performance, their development requires a large, high-quality dataset for training, and their decisions are opaque. Task-agnostic, pre-trained large language models (LLMs) have the potential to conduct zero-shot appraisal with supporting rationales.

2 | 1.2 Objective

To assess the performance of Generative Pre-trained Transformer (GPT) -4o and GPT-o3-mini in automating the methodological appraisal of randomized controlled trials (RCTs) compared to a fine-tuned encoder-only BioLinkBERT model.

2 | 1.3 Methods

A stratified random sample of 800 articles from the McMaster Premium LiteratUre Service and Clinical Hedges databases was appraised using two prompting schemes: 1) classifier (independent assessment) and 2) verifier (validation of BioLinkBERT predictions). Both GPT models utilized title and abstract (TIAB) or full text. Performance was primarily evaluated against human assessments using Matthew's correlation coefficient (MCC). Bootstrapping over 1,000 iterations was used to estimate 95% confidence intervals (CIs).

2 | 1.4 Results

Using full text, GPT-4o demonstrated comparable performance (MCC 0.429; 95% CI 0.387 to 0.470) to BioLinkBERT (MCC 0.466; 95% CI 0.409 to 0.519), drastically outperforming GPT-o3-mini (MCC 0.272; 95% CI 0.211 to 0.334). GPT-4o's verifier scheme showed similar performance (MCC 0.391; 95% CI 0.335 to 0.444). GPT models provided transparent criterion-specific justifications. Performance using TIAB alone markedly decreased for GPT models (MCC $\leqslant$0.100), highlighting dependency on detailed methodological information.

2 | 1.5 Conclusion

GPT-4o effectively automates RCT critical appraisal with comparable performance to specialized fine-tuned models when provided full text, enhancing interpretability and transparency through explicit justifications. Limitations in abstract-level detail suggest complementary roles for fine-tuned models when full texts are unavailable. Future studies should optimize goal-specific prompting to further facilitate adoption in clinical knowledge translation workflows.

2 | 2 Introduction

Randomized controlled trials (RCTs) are generally considered the gold standard primary evidence in informing clinical practice. However, they often suffer from poor methodological design or small sample sizes due to resource constraints (Hariton & Locascio, 2018). The critical appraisal process ensures that study limitations are assessed and findings are interpreted within the appropriate context (Al-Jundi & Sakka, 2017). For RCTs, critical appraisals are often conducted as a part of the knowledge synthesis and transfer process when preparing systematic reviews (J. A. C. Sterne et al., 2019) or in knowledge transfer workflows such as the McMaster Premium LiteratUre Service (PLUS) (Haynes et al., 2006). Given the complexity of critical appraisal, it often necessitates expertise in both clinical practice and research methodology and is typically performed in duplicate to enhance reliability (J. Higgins & Welch, 2011). The growing volume of biomedical literature further exacerbates this challenge, increasing the burden on clinical researchers and systematic reviewers (*Number of clinical trials by year, country, WHO region and income group (1999-2022)*, n.d.).

To address these difficulties, the use of artificial intelligence (AI) for automation has been an area of ongoing investigation (Santos et al., 2023). Previous machine learning (ML) methods, including naïve Bayes, SVMs and NNs, were widely applied to natural language processing (NLP) tasks (Aphinyanaphongs et al., 2005; Hassan et al., 2012; Kilicoglu et al., 2009; Lokker, Abdelkader, et al., 2024; Lokker et al., 2023; Marshall et al., 2015; Millard et al., 2016-2). However, each suffers from architecture-specific limitations, including the need for feature engineering (Kowsari, Jafari Meimandi, et al.,

2019; Scott, 1999) and poor scalability (X. He et al., 2019; Lokker, Abdelkader, et al., 2024) for shallow learning (SL) methods, and the requirement for substantial computational resources for neural networks (NNs) (Thompson et al., 2020). Their development and implementation require substantial technical knowledge and a large, robust training set, often precluding smaller research groups from leveraging these models (Golestaneh et al., 2024; Hestness et al., 2017). Moreover, these traditional models often lack transparency in their decision-making and generalizability across different appraisal tools or criteria (Harish et al., 2022; K. Li et al., 2022; Wadden, 2021).

The introduction of large language models (LLMs) utilizing decoder transformers, including the GPT by OpenAI, revolutionized NLP (Yenduri et al., 2023). Their pre-training on vast, diverse datasets allows them to perform a wide range of language tasks with natural language prompts and minimal task-specific training and fine-tuning (Jiawen Deng, Heybati, Park, et al., 2024). LLMs have demonstrated the ability to classify medical text, identify key study limitations, and generate structured summaries, making them potentially viable tools for automating critical appraisal in systematic reviews and knowledge translation (Qijie Chen et al., 2023; Ghosh et al., 2024; E. Guo et al., 2024; Tang et al., 2023; Van Veen et al., 2023). Moreover, advanced prompting techniques, such as chain-of-thought reasoning and stepwise decomposition, allow LLMs to articulate their decision-making process transparently, improving the interpretability and confidence (E. Guo et al., 2024; Schulhoff et al., 2024). Despite this, few studies have yet examined using GPT to appraise biomedical articles, and existing literature suffers from

limited sample size or replicability concerns (Hasan et al., 2024; Lai et al., 2024; Pitre et al., 2023).

The Health Information Research Unit (HIRU) at McMaster University is a pioneer in clinical knowledge transfer, constructing the Clinical Hedges database and spearheading PLUS (Haynes et al., 2006; Lokker, McKibbon, et al., 2024; Wilczynski et al., 2005). In the exploration of automated methods for biomedical literature classification and appraisal, HIRU has conducted and published several experiments utilizing various forms of traditional ML, utilizing Microsoft AutoML and LightGBM (Lokker, Abdelkader, et al., 2024) and encoder-only transformer architectures (Lokker et al., 2023; F. Zhou, Parrish, et al., 2025). While these studies have been largely successful, obtaining high-quality labels for training or fine-tuning remains a significant challenge for independent researchers and smaller groups. Additionally, the interpretability of automated predictions poses difficulties for human-in-the-loop workflows that require duplicate assessments or verification.

Due to the aforementioned concerns, LLMs utilizing autoregressive decoders can be a potential solution. For this study, the performance of two state-of-the-art LLMs from OpenAI, Generative Pre-trained Transformer (GPT) -4o and GPT-o3-mini, in appraising RCTs based on methodological rigour alone and in justifying and verifying a top-performing encoder-only transformer model was examined.

2 | 3 Methods

2 | 3.1 Dataset description

This study uses the same dataset and a classifier fine-tuned in a previous study (F. Zhou, Parrish, et al., 2025). In brief, encoder-only transformers were fine-tuned to classify rigour (see **Appendix A1** for the 9-item tool) (*Methodological Criteria*, n.d.) of primary articles on treatment, prevention, and/or quality improvement from the PLUS and the Clinical Hedges database associated with the McMaster HIRU (**Appendix A2**). Details regarding the development of these two databases are published elsewhere (Haynes et al., 2006; *MCMASTER+*, n.d.; Wilczynski et al., 2005; F. Zhou, Parrish, et al., 2025).

Briefly, the dataset included 53,219 (31,928 rigorous; 60.0%) articles from the PLUS database, spanning from its inception in 2003 to 2023. These were used for model development and evaluation. Specifically, 42,575 (25,561; 60.0%) were randomly allocated for training, 5,322 (3,203; 60.2%) for validation, and 5,322 (3,164; 59.5%) for testing. Additional external testing was conducted using 1,011 (575; 56.9%) articles from PLUS published in 2024 and 6,572 (1,587; 24.1%) articles from the Clinical Hedges dataset.

For the current study, a random sample of 800 PLUS articles was selected stratified by the four evaluation datasets (PLUS-validate, PLUS-test, PLUS-2024, and Clinical Hedges), resulting in 200 articles from each. Each of the 200 articles was also stratified by the predicted rigour probability by the encoder-only transformer model into 10 bins. This resulted in 200 articles from each dataset and 20 articles per probability bin per dataset.

2 | 3.2 Classifier description

The encoder-only transformer model chosen for comparison was a BioLinkBERT-base model configured with a learning rate of 3E-5, a batch size of 64, a random seed of 2, with class weight adjustments to address class imbalance, and trained using only the TIAB (F. Zhou, Parrish, et al., 2025). Fine-tuning was conducted for five epochs, but early stopping led to the selection of weights from epoch 2, which corresponded to the lowest validation cross-entropy loss. On the original PLUS-validate set (n=5,322), the BioLinkBERT model achieved the best loss of 0.291, an AUROC of 0.941, and an accuracy of 0.879 using the default threshold of ≥0.50.

Two GPT models, GPT-4o (gpt-4o-2024-11-20) and GPT-o3-mini (o3-mini-2025-01-3), were used for rigour classification. For both models, the presence and frequency penalty were set to the default value of 0. To ensure reproducibility, the temperature was set to 0 for GPT-4o and the seed was set to 1 for GPT-o3-mini. The reasoning effort for GPT-o3-mini was set at the default value of medium to achieve a balance between performance and practicality.

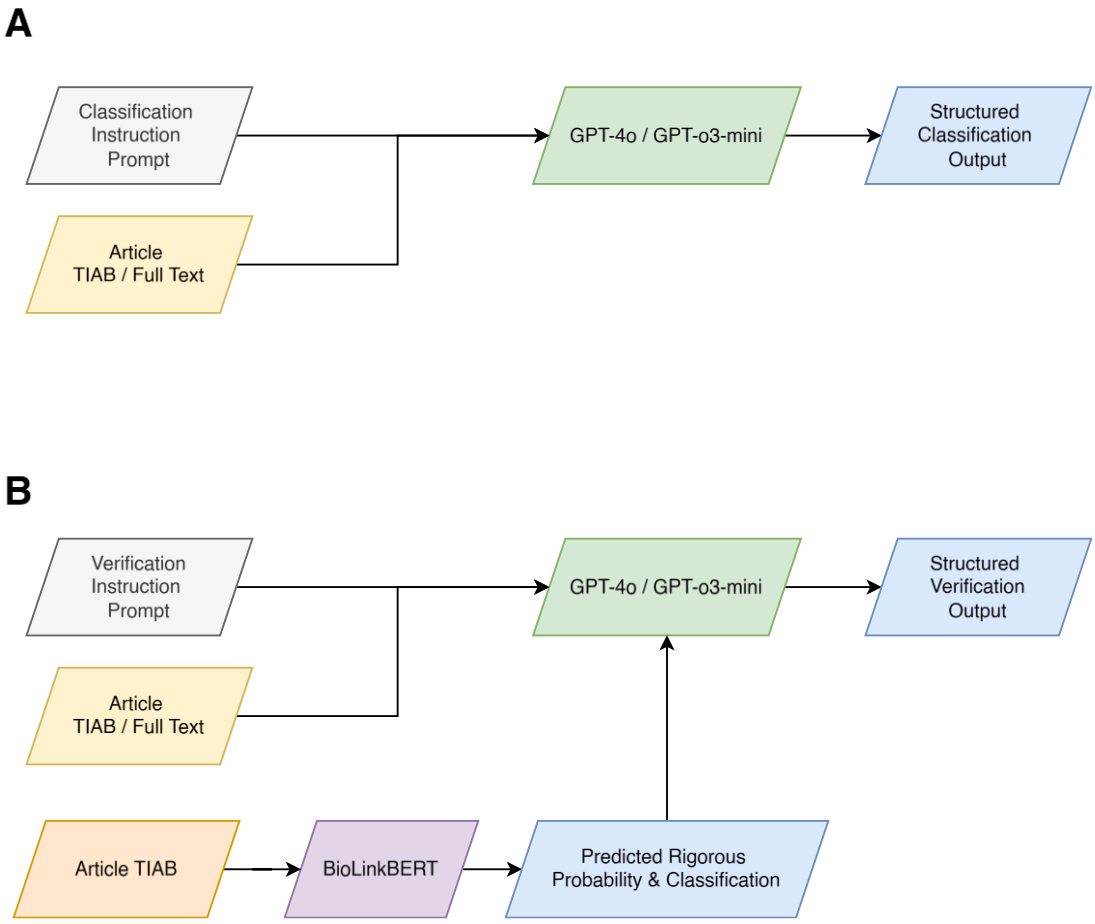
2 | 3.3 Prompting

Two different prompting schemes—classifier and verifier—were used to assess the performance of the GPT models (**Figure 2-1**), and structured output was leveraged to ensure consistency. Each of the two schemes was tested with both the TIAB and the full text of the article. Details regarding the prompts can be found in **Appendix B**.

In the classifier scheme, GPT was directed to independently assess the methodological rigour of each article by generating a justification followed by a rating for each criterion individually. Subsequently, a final assessment was made by GPT, where an article was rated rigorous only if all 9 criteria were met.

In the verifier scheme, GPT was provided with the probability-based rigour classification predicted by the BioLinkBERT model and was tasked with explaining the rationale behind this decision. As in the classifier scheme, GPT generated justifications and assigned ratings for each criterion. However, in this scheme, GPT was additionally required to explicitly state whether it concurred with the classification determined by BioLinkBERT.

Figure 2-1. Flowchart for the classifier and verifier schemes



A Classifier; **B** Verifier.

2 | 3.4 Full text preprocessing

PDFs of the articles were retrieved and converted to plain text using the *pdfminer* package in Python, in which columnized formatting in typeset documents was able to be recognized and extracted properly. Optical character recognition was used to convert Portable Document Format to an extractable format for those that could not initially be extracted properly. For full texts with multiple chapters or which exceeded the 128,000 token limit, including Kiru et al. (Kiru et al., 2016) and Hopewell et al. (Hopewell et al., 2021), the scientific summary was used as the input.

2 | 3.5 Evaluation and statistical analysis

The following metrics were used to assess classification performance: sensitivity, specificity, accuracy, F1 score, Matthew's correlation coefficient (MCC), and work saved over sampling (WSS) (Cohen et al., 2006). MCC was used as the main metric to establish relative performance, as it is a balanced measure that is particularly useful for imbalanced classification tasks when the cost of false positives and false negatives is equally important (Chicco & Jurman, 2020; Jiawen Deng, Moskalyk, et al., 2024). The formulae for confusion matrix metrics are in **Table 2-1**, and detailed interpretations of these metrics can be found in our previous publication (F. Zhou, Parrish, et al., 2025). Bootstrapping over 1,000 iterations was utilized to estimate the 95% confidence interval (CI).

Table 2-1. Evaluation metric formulae

Metric	Formula
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
F1 score	$\frac{2TP}{2TP + FP + FN}$
WSS	$\frac{TN + FN}{N} - (1 - \frac{TP}{TP + FN})$

MCC Matthew's correlation coefficient; **WSS** Work saved over sampling.

2 | 3.6 Example outputs from GPT

GPT outputs of four articles from the PLUS-test set, one in each quadrant of the confusion matrix for BioLinkBERT, were presented.

2 | 3.7 Software and hardware

Software development was conducted using Visual Studio Code and Python 3.11.9. Pre-trained models were sourced from the Hugging Face *transformers* library, while model evaluation was performed using *torch*. The *openai* library was used to query GPT-4o and -o3-mini. Data and statistical analyses were conducted using *pandas*, *numpy*, and *scikit-learn*, while *matplotlib* and *seaborn* were employed for data visualization. The software environments can be found in **Apppendix C**.

The initial fine-tuning of encoder-only transformer models was performed utilizing computing resources from the Cedar cluster, provided by the Digital Research Alliance of Canada. Each model was trained on a single NVIDIA V100 Volta graphics processing unit (GPU) (32GB memory) with an allocation of eight central processing unit (CPU) cores and 40GB of random access memory.

2 | 4 Results

The stratified sampling of 800 articles resulted in 340 (42.5%) rigorous articles. The performance metrics of the models can be found in **Table 2-2**, and confusion matrices for all models can be found in **Figure 2-1.** Using TIAB as inputs, the fine-tuned BioLinkBERT model achieved an MCC of 0.466 (95% CI 0.409, 0.519) with 64.4% of errors as FNs (**Figure 2-1A**).

2 | 4.1 GPT classifier performance

When TIAB were provided, both GPT models classified most instances as non-rigorous, resulting in MCCs ≤0.1 (**Table 2-2**) and >94% of errors as FNs (**Figure 2-2B**, **2-2C**). When the full text was provided, GPT-4o had an MCC of 0.429 (95% CI 0.387, 0.470) with 4.8% (17; **Figure 2-2D**) of errors as FNs. GPT-o3-mini had an MCC of 0.272 (95% CI 0.211, 0.334) with 67.1% (232; **Figure 2-2E**) of errors as FNs.

2 | 4.2 GPT verifier performance

When TIAB were provided, the MCC of GPT-4o increased markedly (0.427 [95% CI 0.372, 0.483]), with 73.9% (204; **Figure 2-2F**) of errors as FNs. The performance of GPT-o3-mini remained poor with an MCC of 0.087 (95% CI 0.025, 0.147) and 94.6% (389; **Figure 2-2G**) of errors as FNs. When the full text was provided, GPT-4o had an MCC of 0.391 (95% CI 0.335, 0.444) and 71.9 % (210; **Figure 2-2H**) of errors as FNs. GPT-o3-mini had an MCC of 0.149 (95% CI 0.084, 0.209) with 65.8% (264; **Figure 2-2I**) of errors as FNs.

2 | 4.3 Example outputs from GPT

Of the four articles, the two that were manually labelled rigorous had a predicted rigour probability of 41.7% (Leong et al., 2006) and 94.3% (H.-H. Kim et al., 2023) by BioLinkBERT. The two that were non-rigorous had a predicted probability of 60.8% (Issler et al., 2009) and 28.8% (Moe et al., 2011). The outputs from GPT-4o classifier using full text are tabulated in **Table 2-3**, and the output for other GPT schemes can be found in **Appendix D**.

Table 2-2. Classifier Performance

Model	Scheme	Input	MCC	Sensitivity	Specificity	Accuracy	F1 Score	WSS
BioLinkBERT (reference standard)	N/A	TIAB	0.466 (0.409, 0.519)	0.773 (0.732, 0.809)	0.699 (0.662, 0.738)	0.730 (0.701, 0.756)	0.707 (0.675, 0.738)	0.273 (0.239, 0.306)
GPT-4o - Classifier with TIAB	Classifier	TIAB	0.097 (0.038, 0.154)	0.043 (0.025, 0.061)	0.988 (0.979, 0.997)	0.589 (0.559, 0.620)	0.081 (0.048, 0.113)	0.018 (0.006, 0.030)
GPT-o3-mini - Classifier with TIAB			0.083 (0.022, 0.144)	0.073 (0.049, 0.100)	0.964 (0.948, 0.978)	0.588 (0.558, 0.619)	0.131 (0.088, 0.173)	0.021 (0.005, 0.039)
GPT-4o - Classifier with Full Text		Full Text	0.429 (0.387, 0.470)	0.960 (0.939, 0.977)	0.422 (0.386, 0.462)	0.649 (0.620, 0.677)	0.698 (0.667, 0.728)	0.221 (0.193, 0.248)
GPT-o3-mini - Classifier with Full Text			0.272 (0.211, 0.334)	0.450 (0.404, 0.499)	0.803 (0.767, 0.834)	0.654 (0.624, 0.685)	0.523 (0.478, 0.566)	0.146 (0.112, 0.183)
GPT-4o - Verifier with TIAB	Verifier	TIAB	0.427 (0.372, 0.483)	0.517 (0.471, 0.563)	0.875 (0.847, 0.902)	0.724 (0.697, 0.751)	0.612 (0.571, 0.654)	0.227 (0.194, 0.260)
GPT-o3-mini - Verifier with TIAB			0.087 (0.025, 0.147)	0.078 (0.055, 0.107)	0.962 (0.945, 0.977)	0.589 (0.558, 0.621)	0.138 (0.099, 0.184)	0.023 (0.007, 0.041)
GPT-4o - Verifier with Full Text		Full Text	0.391 (0.335, 0.444)	0.502 (0.453, 0.553)	0.858 (0.829, 0.884)	0.708 (0.679, 0.734)	0.592 (0.549, 0.634)	0.208 (0.174, 0.242)
GPT-o3-mini - Verifier with Full Text			0.149 (0.084, 0.209)	0.374 (0.330, 0.419)	0.763 (0.728, 0.795)	0.599 (0.569, 0.627)	0.441 (0.392, 0.485)	0.079 (0.045, 0.111)

Note: All numerical values are shown as score (95% CI from bootstrapping).

BERT Bidirectional Encoder Representations from Transformers; **GPT** Generative Pre-trained Transformer; **MCC** Matthew's correlation coefficient; **TIAB** Title and abstract; **WSS** Work saved over sampling.

Table 2-3. Example output for the four articles in the PLUS-test set using GPT-4o - Classifier with Full Text

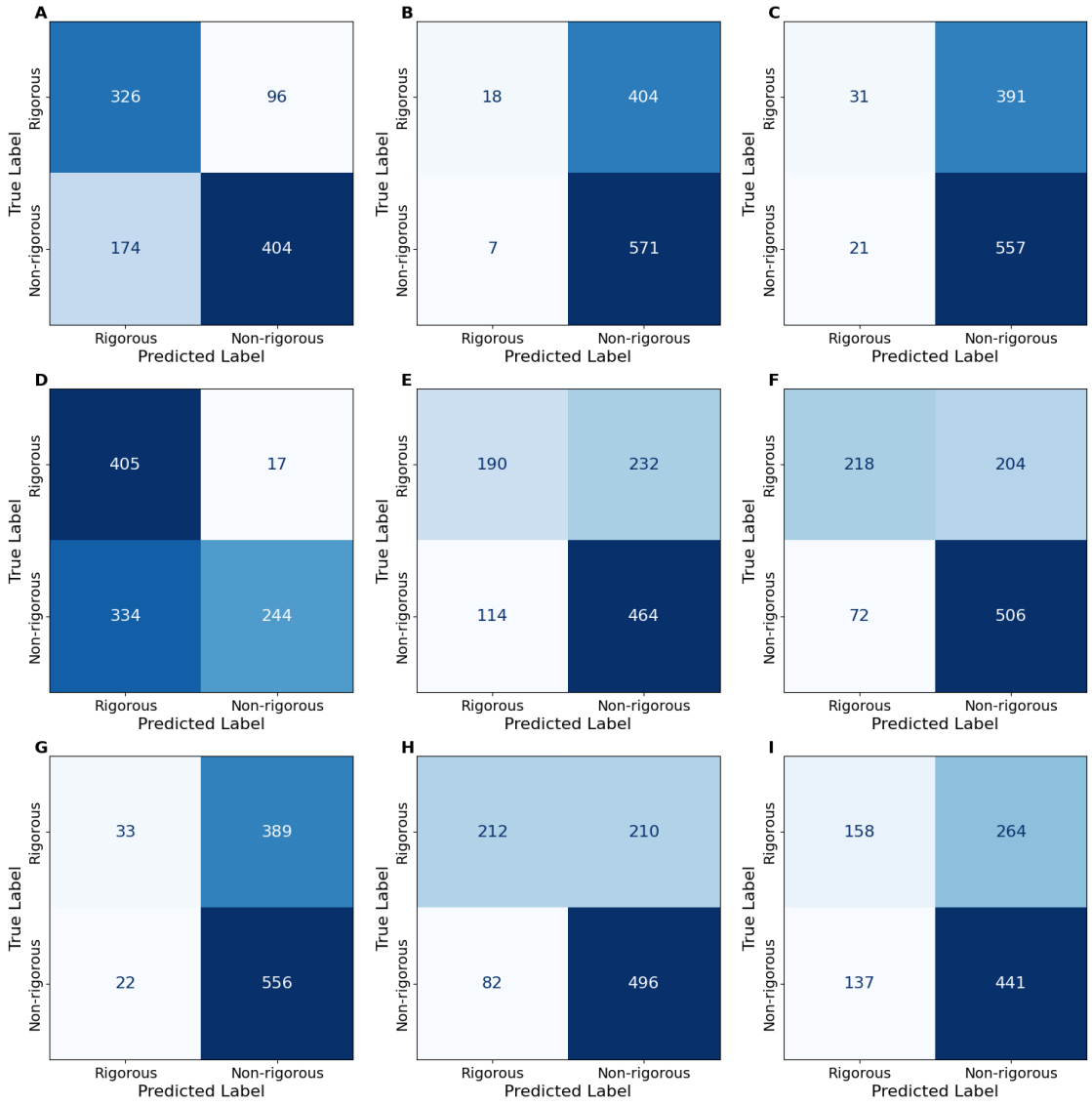
PMID	Label	Predicted Rigorous Probability (%)†	Criterion	Justification	Assessment
19489805	Non-rigorous	60.8	1	The article is written in English.	TRUE
			2	The study is about human participants, specifically mothers and their infants.	TRUE
			3	The study addresses an important clinical topic, namely the prevention of sudden infant death syndrome (SIDS) through educational interventions.	TRUE
			4	The analysis in the article is consistent with the study question, focusing on the effect of the educational intervention on infant sleep position.	TRUE
			5	Participants were randomly allocated to intervention and control groups.	TRUE
			6	The study had more than 10 participants per group completing the primary outcome assessment at the 3-month follow-up.	TRUE
			7	The primary outcome was assessed in more than 80% of those randomized at the 3-month follow-up.	TRUE
			8	The primary outcome, infant sleep position, is clinically important as it relates to the prevention of SIDS.	TRUE
			9	No subgroup analyses were reported in the study.	TRUE
			Final	All criteria for rigour were met based on the provided information.	TRUE
21183586	Non-rigorous	28.8	1	The article is written in English.	TRUE
			2	The study involves human participants with chronic kidney disease.	TRUE

			3	The study addresses dietary interventions in chronic kidney disease, which is relevant to clinical practice.	TRUE
			4	The analysis aligns with the study's objective to compare the effects of vegetarian and meat diets on phosphorus homeostasis.	TRUE
			5	Participants were randomized to the order of the two diets in the crossover design.	TRUE
			6	Only 8 participants completed the study, which is less than the required 10 per group.	FALSE
			7	The primary outcomes were assessed in all participants who completed the study.	TRUE
			8	The primary outcome, phosphorus homeostasis, is clinically important for CKD management.	TRUE
			9	No subgroup analyses were reported in the study.	TRUE
			Final	The study does not meet the criterion of having at least 10 participants per group completing the primary outcome assessment.	FALSE
16916871	Rigorous	41.7	1	The article is written in English.	TRUE
			2	The study is conducted on human participants.	TRUE
			3	The study addresses the effectiveness of text messaging reminders in improving attendance in primary care, which is relevant to clinical practice.	TRUE
			4	The analysis aligns with the study's objective to evaluate the effectiveness and cost-effectiveness of text messaging reminders.	TRUE
			5	Participants were randomly allocated to the intervention and control groups.	TRUE
			6	Each group had more than 10 participants completing the primary outcome assessment.	TRUE

			7	The primary outcome (attendance rate) was assessed in more than 80% of those randomized.	TRUE
			8	The primary outcome (attendance rate) is clinically important.	TRUE
			9	Subgroup analyses were preplanned and conducted appropriately.	TRUE
			Final	All criteria for rigour were met based on the provided information.	TRUE
36395875	Rigorous	94.3	1	The article is written in English.	TRUE
			2	The study is conducted on human participants.	TRUE
			3	The study investigates the effects of different antiplatelet therapies after coronary bypass surgery, which is relevant to clinical practice.	TRUE
			4	The analysis aligns with the study's objectives, comparing outcomes between two treatment groups.	TRUE
			5	Participants were randomly allocated to the two treatment groups.	TRUE
			6	Each group had 102 participants completing the primary outcome assessment.	TRUE
			7	The primary outcome was assessed in 97.5% of participants, meeting the 80% threshold.	TRUE
			8	The primary outcome, major adverse cardiovascular events, is clinically significant.	TRUE
			9	Subgroup analyses were preplanned and adhered to the randomization groups.	TRUE
			Final	All criteria for rigour have been met based on the provided information.	TRUE

† Predicted by the BioLinkBERT model.

Figure 2-2. Confusion matrices for classifiers



A BioLinkBERT; **B** GPT-4o - Classifier with TIAB; **C** GPT-o3-mini - Classifier with TIAB; **D** GPT-4o - Classifier with Full Text; **E** GPT-o3-mini - Classifier with Full Text; **F** GPT-4o - Verifier with TIAB; **G** GPT-o3-mini - Verifier with TIAB; **H** GPT-4o - Verifier with Full Text; **I** GPT-o3-mini - Verifier with Full Text.

2 | 5 Discussion

In response to the advancements of ML and the need for transparency and usability in medical NLP, two state-of-the-art LLMs from OpenAI were assessed in assisting clinical knowledge translation and critical appraisal workflows. To our knowledge, this is the largest study on LLMs for critical appraisal to date and the first to assess a reasoning LLM for this task.

2 | 5.1 Summary of findings

Our findings demonstrate that GPT models have the potential to assist in classification and appraisal workflows, even without additional task-specific training. Notably, both GPT-4o and GPT-o3-mini achieved performance comparable to BioLinkBERT when full texts were available, despite BioLinkBERT being fine-tuned on over 40,000 labeled articles. This is particularly significant given that the dataset used in this evaluation included a higher proportion of articles that BioLinkBERT struggled to classify with high confidence (i.e., a predicted rigorous probability close to 0% or 100%), making it a challenging benchmark. Moreover, GPT's ability to generate concise justifications for its classifications enhances transparency and provides convenience for manual verification, offering a practical advantage in real-world applications. These results highlight the potential of LLMs in biomedical literature appraisal, particularly in scenarios where labeled training data is limited or unavailable.

Abstracts, however, seldom contained enough information for GPT to make confident decisions which resulted in erroneous negative classifications due to insufficient reporting of methodological details. Conversely, BioLinkBERT and other

encoder-only transformer models remain relevant when access to full text articles is restrictive, as their performances were satisfactory when only the TIAB was used as input for fine-tuning and assessment (Lokker et al., 2023; F. Zhou, Parrish, et al., 2025).

Based on this experiment, providing the predictions from BioLinkBERT to GPT did not significantly improve performance compared to using either alone. Nevertheless, GPT's natural language justifications may still serve as a valuable reference for human researchers.

2 | 5.2 Development and implementation considerations

While GPT's overall performance was comparable to the best models from our previous experiments (Lokker et al., 2023; F. Zhou, Parrish, et al., 2025), careful consideration is required before adopting it as a ROB assessment tool. A key challenge is ensuring reproducibility, a fundamental principle of evidence-based medicine (National Academies of Sciences, Engineering et al., 2019). The inherent stochasticity of LLM outputs, which is influenced by parameters such as temperature (n.d.-b), must be accounted for in both research and deployment. Future studies should implement measures to enhance reproducibility, such as setting deterministic parameters or conducting sensitivity analyses to assess variability in model outputs.

The importance of evaluating a model in the context of its intended use case was also highlighted. For HIRU, deployed models must prioritize sensitivity for confident negative classifications and traditional deep-learning models achieve this through threshold tuning. For LLMs, it is important to adjust the instructions to align with the

specific objective—whether optimizing for balanced classification, maximizing sensitivity for negative classifications, or enhancing specificity for positive classifications. Future research should explore prompting techniques to optimize goal-specific performance while maintaining interpretability and reliability.

In addition, it is known that prompt engineering can significantly affect the model's performance (Schulhoff et al., 2024). Techniques such as role prompting, decomposition, and chain-of-thought should be considered and evaluated on a small validation set before large-scale implementation. While GPT-o3-mini had worse performance, the built-in chain-of-thought mechanism of reasoning models may reduce the need for complex, sophisticated prompting (G. Wang et al., 2024). Therefore, reasoning-optimized models may offer a distinct advantage in deployment settings where ease of implementation and interpretability are crucial.

2 | 5.3 Comparison with existing literature

Three studies have examined the use of LLMs for biomedical critical appraisal. Lai et al. (Lai et al., 2024) utilized ChatGPT and Claude to assess 30 RCTs using the original Cochrane ROB tool (J. P. T. Higgins et al., 2011), where each RCT was assessed twice by each LLM. Performance was measured using accuracy, and ChatGPT and Claude achieved a mean accuracy of 0.845 (95% CI 0.815, 0.873) and 0.895 (95% CI 0.870, 0.918). Despite this, numerous domains had a sensitivity at or close to 0 due to models predominantly classifying instances as negative. Given the high prevalence of negative cases in the dataset, this led to an inflated overall accuracy but resulted in poor sensitivity, reflecting the model's limited ability to correctly identify positive instances. Future

research should include MCC as a primary classification metric or normalize accuracy based on class prevalence.

Pitre et al. (Pitre et al., 2023) leveraged GPT-4 to appraise 157 RCTs identified from Cochrane reviews using the ROB-2 tool (J. A. C. Sterne et al., 2019). They analyzed three prompts with limited, extensive, and optimized instructions. Despite the existence of ground truth in the form of ratings from Cochrane reviews, performance was measured with weighted Cohen's κ. Across all the prompts and ROB assessment domains, the level of agreement was poor, ranging from 0.11 to 0.17. The authors suggested that GPT had issues understanding the implications of randomization and allocation concealment methods, making reasonable assumptions about attrition, and comprehending the effect of blinding on outcomes. Token limits were also identified as a potential challenge.

Hasan et al. (Hasan et al., 2024) presented a framework for using GPT-4 to assess 307 articles using Risk Of Bias In Non-randomized Studies of Interventions (J. A. Sterne et al., 2016). Similar to Pitre et al. (Pitre et al., 2023), only agreement metrics, including Cohen's κ, were used to assess performance. Overall, the agreement between GPT-4 and Cochrane reviewers was poor with $\kappa \leq 0.15$ in 6 of 7 domains and 0.13 for the overall assessment. The authors highlighted challenges with file handling, token limits, and the quality of prompt engineering.

Overall, these studies demonstrated mixed performance of current LLMs in critical appraisal. While the performances are not directly comparable to our results, future studies may wish to include large datasets, leverage API functions to mitigate

concerns surrounding output stochasticity, and consider appropriate classification metrics during evaluation.

2 | 5.4 Limitations

Several important limitations must be considered when interpreting our findings. First, the findings are limited to PLUS's criteria and articles from 123 journals indexed by PubMed. Additionally, neither the Clinical Hedges nor the PLUS database documented ratings for each criterion, and therefore, performance per criterion was not available (except that articles that were labelled rigorous met all criteria). Second, due to current API limitations, the time and cost associated with GPT compared to BioLinkBERT or manual evaluation could not be analyzed. However, it was not unreasonable to observe that GPT was drastically more efficient resource-wise compared to other methods due to its ease of implementation and relatively cheap cost per text token (USD$2.50 and USD$10.00 per one million input and output tokens at time of analysis, respectively). Lastly, while the κ for PLUS's criteria was assessed to be ≥ 0.80 (Wilczynski et al., 1993), there are still likely inaccuracies in manual article labelling. If feasible, future studies should consider direct labelling to minimize the biases introduced by noisy labels and having standardized, open access datasets would benefit the community of researchers.

2 | 6 Conclusion

This study demonstrated the promising capabilities of LLMs, particularly GPT-4o and GPT-o3-mini, in automating the critical appraisal of RCTs. GPT models achieved

comparable performance to a fine-tuned BioLinkBERT model, with GPT-4o being notably effective when utilized both independently and in verifying BioLinkBERT predictions. Methods to mitigate the stochastic nature of LLMs, advanced prompting techniques, and a wide range of classification metrics were utilized. LLM performance was constrained by the level of methodological detail available in abstracts alone, indicating the continued utility of fine-tuned encoder models when full text access is limited. Ultimately, GPT models showed enhanced interpretability through their ability to generate transparent, criterion-specific justifications, facilitating easier integration into clinical knowledge translation workflows. Future research should prioritize enhancing reproducibility, further optimizing prompting strategies, and evaluating cost-effectiveness and efficiency in real-world applications.

CHAPTER 3

UNDERSTANDING TRANSFORMER-BASED CLASSIFICATIONS OF MEDICAL TEXT: USING AN LLM FOR ATTRIBUTION OF FEATURE IMPORTANCE

3 | 1 Abstract

3 | 1.1 Background

Deep learning has demonstrated excellent performance in biomedical literature classification. However, the opacity of these models' decision-making processes limits their interpretability and adoption. Explainable artificial intelligence (XAI) methods, including SHapley Additive exPlanations (SHAP) and integrated gradients (IG), have been proposed to address this issue, yet computational complexity remains high. Generative large language models (LLMs) may offer a novel approach for generating interpretable, context-aware explanations.

3 | 1.2 Objective

To investigate the effectiveness of Generative Pre-trained Transformer (GPT) -4o as a perturbation-based explainer for a BioLinkBERT text classifier by comparing its explanations to SHAP partition explainer and IG in terms of faithfulness.

3 | 1.3 Methods

A stratified sample of 200 articles from McMaster PLUS and Clinical Hedges databases was classified by BioLinkBERT. GPT-4o, SHAP partition explainer, and IG were used to generate token-level feature attributions. GPT-based explanations were derived through iterative masking perturbation. Explanations were evaluated using a modified version of the area over the perturbation curve (AOPC), correlation analyses, and qualitative assessment of feature importance attribution.

3 | 1.4 Results

SHAP (AOPC 0.222; 95% confidence interval [CI] 0.200 to 0.244) and IG (AOPC 0.225; 95% CI 0.202 to 0.247) provided consistent and faithful explanations, effectively identifying tokens relevant to study rigour (e.g., "randomized," "blind"). Conversely, GPT-4o explanations were poor (AOPC 0.029; 95% CI 0.014 to 0.043) with nonsensical token attributions. Correlation analysis showed moderate alignment between SHAP and IG (Pearson's r 0.367), whereas GPT-4o had minimal (Pearson's r ≤0.032) correlation with these established methods.

3 | 1.5 Conclusion

GPT-4o, despite its advanced contextual capabilities, performed poorly as a standalone explainer compared to established methods like SHAP and IG. These findings highlight the need for further research into specialized prompt engineering and potential hybrid methods integrating LLMs with traditional XAI techniques to improve interpretability without sacrificing computational efficiency or explanation quality.

3 | 2 Introduction

The rapid growth of biomedical literature has driven the development of automated classification systems to facilitate knowledge synthesis and translation (*MEDLINE PubMed production statistics*, 2018). Deep learning, particularly encoder-only transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT), has gained significant attention in biomedical text classification (*BLURB leaderboard*, n.d.). These models excel due to their ability to capture contextual information, leverage transfer learning, and minimize the need for extensive data preprocessing and feature engineering, making them highly effective for biomedical applications (Devlin et al., 2018; Vaswani et al., 2017).

However, the complex, multi-layered nature of BERT models undermines their interpretability, posing challenges in understanding their decision-making processes (Wadden, 2021). Explainable artificial intelligence (XAI) techniques aim to address this limitation by providing insights into feature importance (Gohel et al., 2021). One widely used XAI framework is SHapley Additive exPlanations (SHAP), which is grounded in game theory and utilizes Shapley values to systematically estimate feature contributions by perturbing inputs (Lundberg & Lee, 2017). Despite its theoretical robustness, SHAP has substantial computational overhead. It requires summing marginal contributions across feature subsets, which leads to an exponential increase in complexity as the feature space grows (Bertossi et al., 2020). Consequently, computing SHAP values becomes impractical for BERT models that process long sequences of up to 512 tokens.

To mitigate this challenge, a partition explainer groups features into structured partitions, which reduces complexity while preserving interactions. By approximating Shapley values using Owen values (López & Saboya, 2009), the partition explainer enhances scalability, making it particularly suitable for high-dimensional text classification tasks. Another widely used method is integrated gradients (IG) based on the Aumann-Shapley method, which ensures axiomatic fairness and path-integrated attribution of feature importance (Enguehard, 2023; Kirman et al., 1976; M. Ribeiro et al., 2024; Sikdar et al., 2021). It offers a computationally efficient approach to estimating feature importance by measuring the accumulated gradients along the path between a baseline and the instance input. IG has been widely applied in natural language processing (NLP) tasks, providing a balance between interpretability and computational feasibility (Enguehard, 2023; M. Ribeiro et al., 2024; Sikdar et al., 2021). However, these methods face challenges in explaining text classifiers due to significant multicollinearity between input tokens and high-dimensional feature spaces (H. Chen et al., 2020; Enguehard, 2023; Mosca et al., 2022).

More recently, pre-trained generative large language models (LLMs) leveraging transformer decoders have garnered wide attention in NLP due to their performance and flexibility (Minaee et al., 2024). Previous studies explored LLMs in model explanation, such as Zytek et al. (Zytek, Pido, et al., 2024) and Zeng (Zeng, 2024), who investigated using LLMs to convert SHAP explanations into plain-text descriptions to improve human interpretability. Unlike perturbation- or gradient-based XAI methods, LLMs can generate explanations while incorporating token-level contextual relationships, potentially leading

to more faithful feature attributions. More recently, LLMs now support structured JavaScript Object Notation (JSON) output and function calling, providing a convenient way to integrate model predictions (*Build with Claude*, n.d.; *Llama AI*, n.d., n.d.-c).

Despite these advances, no prior studies have explored the usage of LLMs as standalone explainers for deep learning models in biomedical text classification. To address this gap, this study developed and validated a methodology to investigate Generative Pre-trained Transformer (GPT) -4o by OpenAI as a perturbation explainer for a BERT-based biomedical text classifier (BioLinkBERT identified as the top model from our previous study). Performance of GPT-4o was compared against SHAP's partition explainer and IG explanations.

3 | 3 Methods

3 | 3.1 Classifier and dataset description

This study builds upon the work from a previous study (F. Zhou, Parrish, et al., 2025) where 630 encoder-only transformer models were fine-tuned using grid search. The data came from McMaster Premium LiteratUre Service (PLUS) and the Clinical Hedges database associated with the McMaster Health Information Reseearch Unit (HIRU). Detailed descriptions of these two databases are published elsewhere (Haynes et al., 2006; Lokker et al., 2023; *MCMASTER+*, n.d.; Wilczynski et al., 2005; F. Zhou, Parrish, et al., 2025). In short, both databases include primary treatment, prevention, and/or quality improvement studies that had been manually appraised using custom criteria for randomized controlled trials (RCTs) (*Methodological Criteria*, n.d.) as methodologically rigorous or non-rigorous (**Appendix A1**). Articles in the PLUS database from inception (2003) to 2023 (n=53,219) were used for training (n=42,575), validation (n=5,322), and testing (n=5,322). Articles from 2024 in PLUS (n=1,011) and the Clinical Hedges (n=6,572) were used for external testing (**Appendix A2**). The top-performing models were identified on the validation set and subsequently tested.

For this study, a stratified random sample of 200 articles was selected, 40 from each data subset. For each of the 5 data subsets, articles were placed into 10 bins based on their predicted probability for rigour and a random sample of 4 articles per probability bin per dataset was selected. The probability scores were generated by the model that had the lowest validation loss, which was a BioLinkBERT-based model with a learning rate of 3E-5, a batch size of 64, a random seed of 2, and included class weight adjustments. The

model was fine-tuned for 5 epochs before premature termination by early stopping, and weights from epoch 2 were used as it achieved the lowest validation loss. Other relevant configurations can be found in our previous publication (F. Zhou, Parrish, et al., 2025). The model achieved cross-entropy loss of 0.291, an area under the receiver operating characteristic curve (AUROC) of 0.941, and an accuracy of 0.879 on the full validation set.

3 | 3.2 SHAP partition explainer

The SHAP partition explainer was used (*shap.PartitionExplainer — SHAP latest documentation*, n.d.) to compute an Owen value for each token in each prediction. The partition explainer was chosen due to its efficiency in high-dimensional text classification and its ability to capture feature interactions more effectively than standard Shapley value approximations (Bitton et al., 2022). SHAP values were calculated using logits back-transformed from SoftMax probabilities.

3 | 3.3 IG

IG was employed to estimate token-level feature attributions for each prediction. A padded empty sequence was used as the baseline input, ensuring the absence of semantic content while preserving the tokenization structure. Attributions were derived by computing gradients with respect to the input embeddings across 30 interpolation steps. The total IG attribution per token was calculated by aggregating gradients across all embedding dimensions.

3 | 3.4 GPT

GPT-4o-2024-11-20 with a temperature of 0 was used to ensure deterministic outputs, and both presence and frequency penalties were set to 0. The objective was to evaluate GPT's ability to estimate token-level feature attributions through perturbation-based explanations, similar to SHAP. Two prompting schemes, GPT-index and GPT-token, were designed to systematically mask tokens and assess their influence on classifier predictions. Tokens were obtained by processing the original input through BioLinkBERT's word-piece tokenizer. Both schemes received the number of input tokens, predicted logits for both classes and the probability of the positive class. Additionally, GPT-token was provided with the complete list of input tokens in a comma-separated format and the manual appraisal criteria. The full prompts used for both schemes are available in **Appendix E**. A flow diagram can be found in **Figure 1**.

3 | 3.4.1 Developer Prompt

In the developer prompt, GPT was provided with 1) the role as a machine learning model explainer, 2) the task to explain a binary encoder-only transformer text classifier's prediction via perturbations by masking input tokens, 3) the scheme-specific information that will be provided in the user prompts, 4) step-by-step instructions on defining importance, masking, function calling, and generating importance values that would be executed subsequently (**Figure 3-1A**). The manual appraisal criteria (*Methodological Criteria*, n.d.) for GPT-token was included in the developer prompt.

3 | 3.4.2 Initial User Prompt

In the initial user prompt, both prompting schemes were provided with the number of tokens, the predicted logit of the positive and negative classes, and the probability of the negative class (**Figure 3-1B**). The input tokens in the format of a comma-separated list were provided to GPT-token only in the initial user prompt.

3 | 3.4.3 Subsequent User Prompts

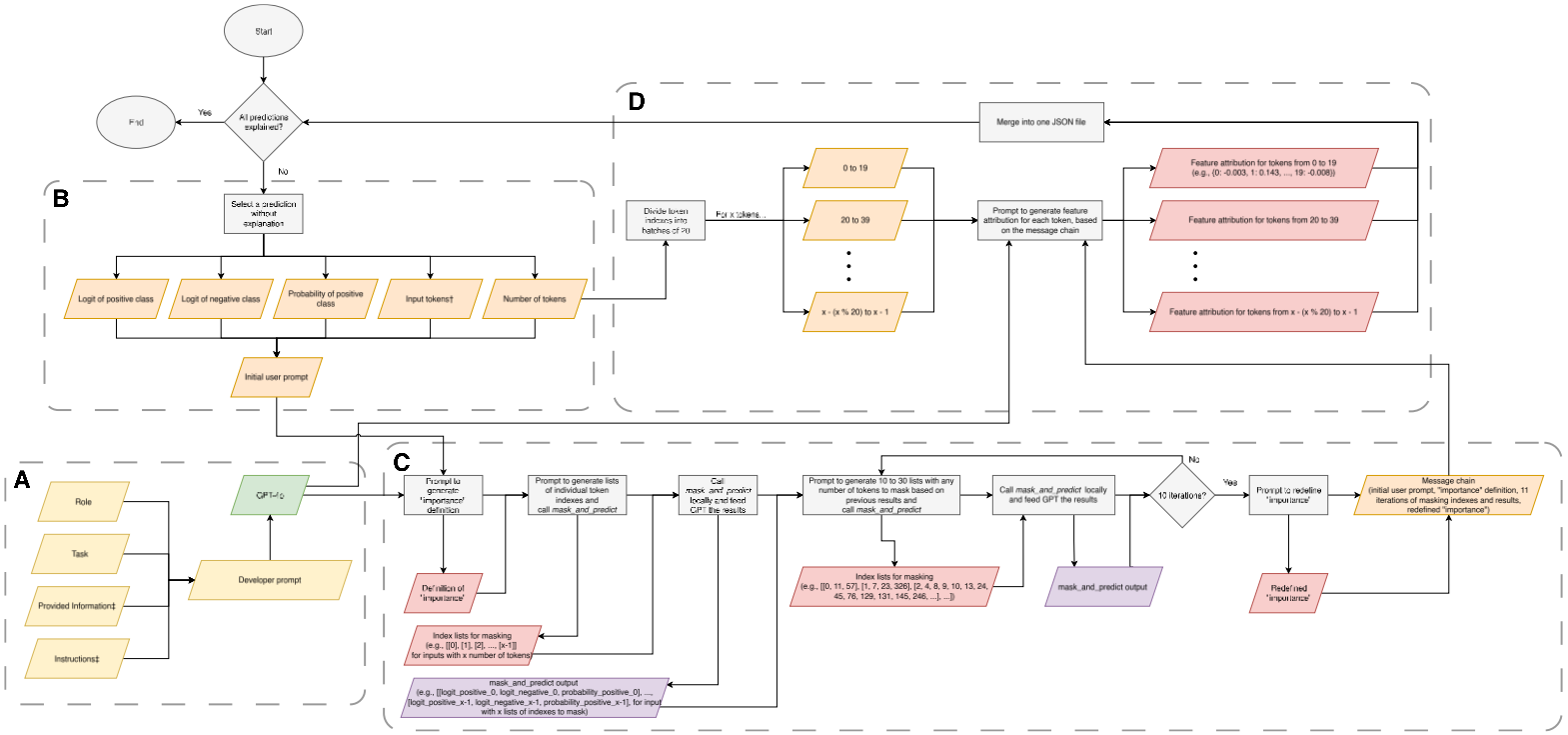
The model was first instructed to generate the definition of "importance" for itself and then to call *mask_and_predict* with lists of individual indices (e.g., [[0], [1], … [x-1]], for an input with x tokens), echoing the instructions provided in the developer prompt. To call *mask_and_predict*, function-calling (n.d.-d) in OpenAI's API was utilized (**Figure 3-1C**). The function, in general, takes lists of integers as input and returns the logits for both classes and the probability of the positive class for each list of indices to mask, with every token at the integer replaced with "[MASK]".

Subsequently, the model was prompted 10 times to generate any number of lists with any number of indices to mask and call *mask_and_predict*, where each iteration had the results of all previous iterations. The model was explicitly instructed to avoid generating the same combinations of indices and to adapt future maskings based on prior iteration results. Lastly, the model was asked to redefine "importance" based on the initial definition and the results of all masking iterations.

3 | 3.4.4 Feature Attribution Calculations

The model was prompted, with the final message chain including the initial user prompt, all iterations of perturbations, and both iterations of importance definition, to generate the feature importance for each token, 20 tokens per batch (**Figure 3-1D**). The model was not provided with the feature attributions of other batches. This batched approach was taken as the model often had issues with generating longer sequences. The structured output function (n.d.-e) of the API was leveraged to generate a list of dictionaries of token indices and their corresponding feature attributions.

Figure 3-1. Flowchart for the generation of GPT explanations



A Developer prompt creation; **B** Initial user prompt creation; **C** Subsequent, iterative user prompts for token masking and prediction; **D** Feature attribution calculations.

† Input tokens are only included in the initial user prompt for GPT-token.

‡ The provided information and instructions in the developer prompt would differ for GPT-index and GPT-token, as GPT-index was not provided with the input tokens. Detailed prompts can be found in the Supplementary.

3 | 3.5 Evaluation

3 | 3.5.1 Area over the perturbation curve (AOPC)

To establish feature attribution performance, the AOPC metric used in previous literature was modified (H. Chen et al., 2020; Nguyen, 2018; Samek et al., 2016). The AOPC was calculated for each explanation individually and then averaged across all 200 instances.

The original AOPC is calculated using the following formula:

Equation 3-1. Original AOPC

$$AOPC = \frac{1}{K}\sum_{i=1}^{K}\left(P(x) - P(x^{(i)})\right)$$

where $P(x)$ is the predicted probability for the positive class with the original input x, $x^{(i)}$ is the perturbed input with top i important features removed or masked, and K is the number of perturbation steps. This formula assumes that features contribute to the positive class, hence their removal would result in a decrease in the predicted probability, and $P(x) - P(x^{(i)})$ would be positive.

For binary text classification, feature attributions could be associated with a negative value indicating more support for the negative class. Under such circumstances, their removal would lead to an increase in the probability of the positive class. For this reason, the AOPC formula was adapted to the following:

Equation 3-2. Modified AOPC

$$AOPC = \frac{1}{K_p + K_n}\left(\sum_{i=1}^{K_p}\left(P(x) - P\left(x^{(i)}\right)\right) + \sum_{j=1}^{K_n}\left(P\left(x^{(j)}\right) - P(x)\right)\right)$$

where $x^{(i)}$ and $x^{(j)}$ is the perturbed input with top i positive features and top j negative features masked, respectively. K_p and K_n are the number of perturbation steps for the positive features and negative features, respectively, which in this case would respectively equal to the number of positively and negatively attributed tokens. Similar to the original metric, a larger value would indicate higher attribution faithfulness. Note that the operands corresponding to the '+' operation must be computed separately (to enable the removal of positive features and negative features separately) before the final summation is performed.

3 | 3.5.2 Correlation Analysis

The pairwise correlation between feature attributions for each of the four methods (SHAP, IG, GPT-index, GPT-token) was assessed using Pearson's r, Spearman's ρ, and Kendall's τ. Distribution similarity was measured with Wasserstein distance. A p-value ≤0.05 is indicative of statistical significance. The distributions of feature attributions were visualized using scatter plots.

3 | 3.5.3 Feature Importance Attributions

The most important 10 features that had an occurrence of ≥1, ≥10, and ≥100 for each explainer were examined using bar graphs.

3 | 3.6 Hardware and software

Initial fine-tuning of encoder-only transformer models was conducted using the resources from the Cedar cluster of the Digital Research Alliance of Canada. Each model was trained using one NVIDIA V100 Volta (32GB memory), as well as an allocation of 8 central processing unit cores and 40 GB of random access memory (RAM). IG, SHAP, and GPT-4o feature attribution calculations were conducted locally using one NVIDIA RTX 2070 (32GB memory), as well as an AMD 9950x with 64GB RAM.

Visual Studio Code and Python 3.11.9 was used for all software development. The *transformers* library by Hugging Face was used to obtain pre-trained models, and *torch* was used for evaluation purposes. The *shap* and *capum* libraries were used to calculate feature attributions via partition explainer and IG, respectively. The *openai* library was used to query GPT-4o. Data management and statistical analysis were conducted using *pandas*, *numpy*, and *scikit-learn*. Data visualization was done with *matplotlib* and *seaborn*. The full list of libraries used on the Digital Research Alliance of Canada and local environment can be found in the supplementary **Appendix B**.

3 | 4 Results

3 | 4.1 Characteristics of the dataset and classifier

The original dataset contained 60,802 instances, of which 34,090 (56%) are rigorous. After stratified sampling, the 200 instances contain 83 (41.5%) rigorous articles. Within this dataset, the BioLinkBERT model achieved a cross-entropy loss of 0.527, an AUROC of 0.812, and an accuracy of 0.705 using the default threshold of ≥ 0.50.

3 | 4.2 Importance definitions by GPT

GPT, in both prompting schemes, were instructed to define 'importance' after being provided with the initial user prompt and subsequently redefine importance after all iterations of masking have been completed. Both GPT-index and -token initially defined 'importance' as the change in the predicted probability of the positive class before and after masking for all (200/200; 100%) instances.

After redefinition for GPT-index, the definition remained consistent as the change in predicted probability in 199 instances (99.5%). Of these, 3 (1.5%), 37 (18.5%), and 16 (8.0%) instances normalized the change by logits, initial predicted probability, and number of masked tokens in a perturbation, respectively. The remaining instance utilized the change in the difference between the positive and negative logit as the definition of importance.

For GPT-token, the definition for all (200/200; 100%) instances remained consistent, as the change in predicted probability. Among these, 67 (33.5%) and 9 (4.5%)

instances normalized the change by the initial predicted probability and number of tokens masked, respectively.

3 | 4.3 AOPC analysis

SHAP and IG explanations achieved similar faithfulness, with a mean (95% confidence interval [CI]) of 0.222 (0.200, 0.244) and 0.225 (0.202, 0.247), respectively (**Table 3-1**). SHAP was better at identifying negative tokens, while IG was better at identifying positive tokens. GPT-index and GPT-token had significantly worse AOPC of 0.025 (0.012, 0.038) and 0.029 (0.014, 0.043), respectively. Both had negative AOPC for negative tokens.

Table 3-1. AOPC performance

Explainer	AOPC	AOPC (Positive Tokens)	AOPC (Negative Tokens)
SHAP	0.222 (0.200, 0.244)	0.277 (0.249, 0.306)	0.037 (0.030, 0.044)
IG	0.225 (0.202, 0.247)	0.326 (0.293, 0.359)	0.026 (0.019, 0.033)
GPT-index	0.025 (0.012, 0.038)	0.045 (0.028, 0.063)	-0.021 (-0.034, -0.008)
GPT-token	0.029 (0.014, 0.043)	0.049 (0.029, 0.068)	-0.021 (-0.031, -0.010)

Note: All values are shown as the mean (95% CI) across the 200 instances.

3 | 4.4 Correlation analysis

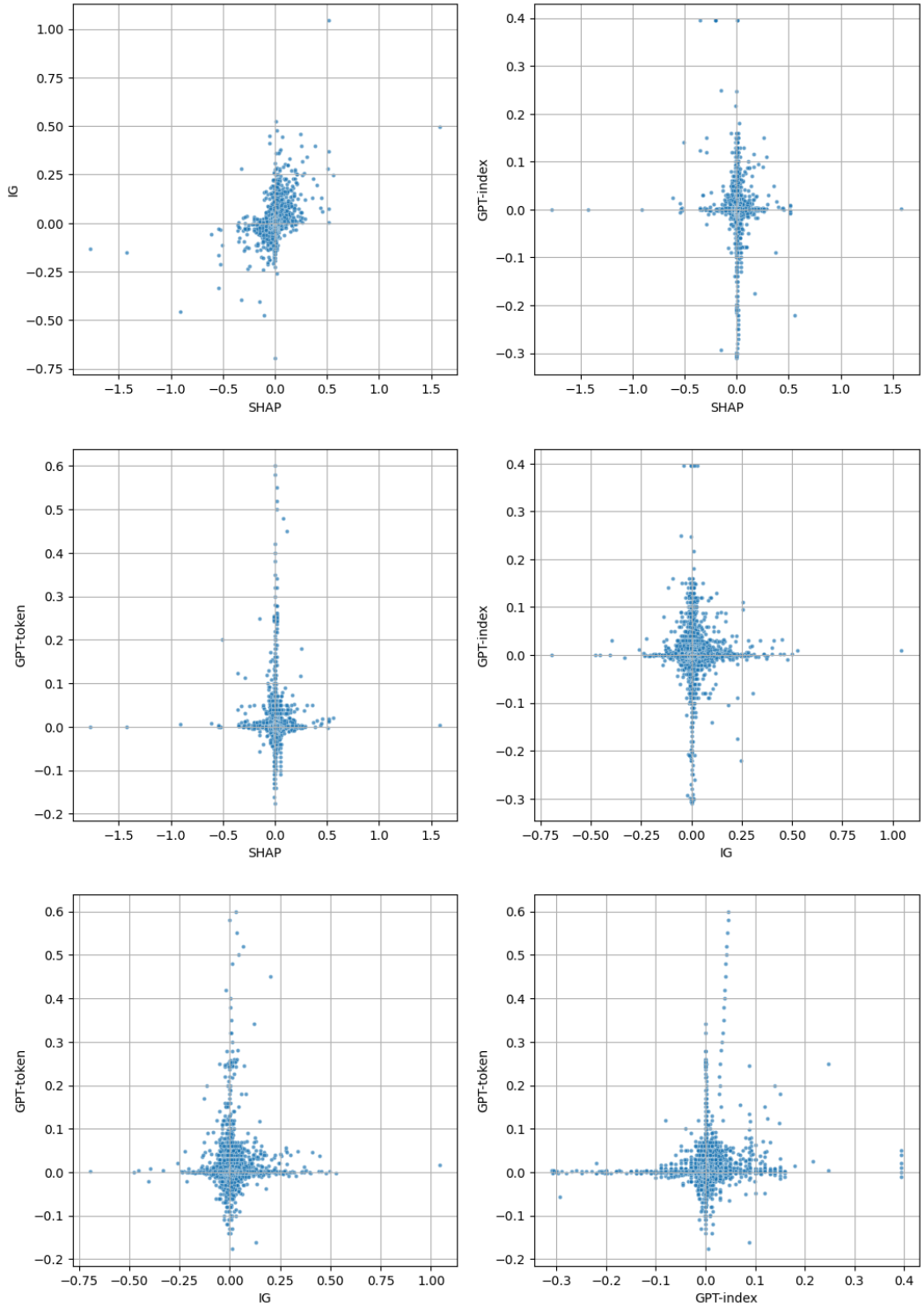
Feature attributions from SHAP and IG exhibit moderate correlation with each other, with a Pearson's r of 0.367 (**Table 3-2**, **Figure 3-2**). No notable correlation was evident between other pairs of explainers. Wasserstein distances reveal that the distributions of feature attributions are similar across all explainers.

Table 3-2. Correlation of feature attributions

Explainer A	Explainer B	Pearson's r	Spearman's ρ	Kendall τ	Wasserstein Distance
SHAP	IG	0.367*	0.275*	0.192*	0.002
SHAP	GPT-index	-0.031*	0.061*	0.041*	0.003
SHAP	GPT-token	0.004	0.037*	0.025*	0.003
IG	GPT-index	0.003	0.038*	0.026*	0.004
IG	GPT-token	0.032*	0.029*	0.020*	0.005
GPT-index	GPT-token	0.083*	0.096*	0.071*	0.001

*Statistical significance ($P<0.05$).

Figure 3-2. Distribution of feature attributions



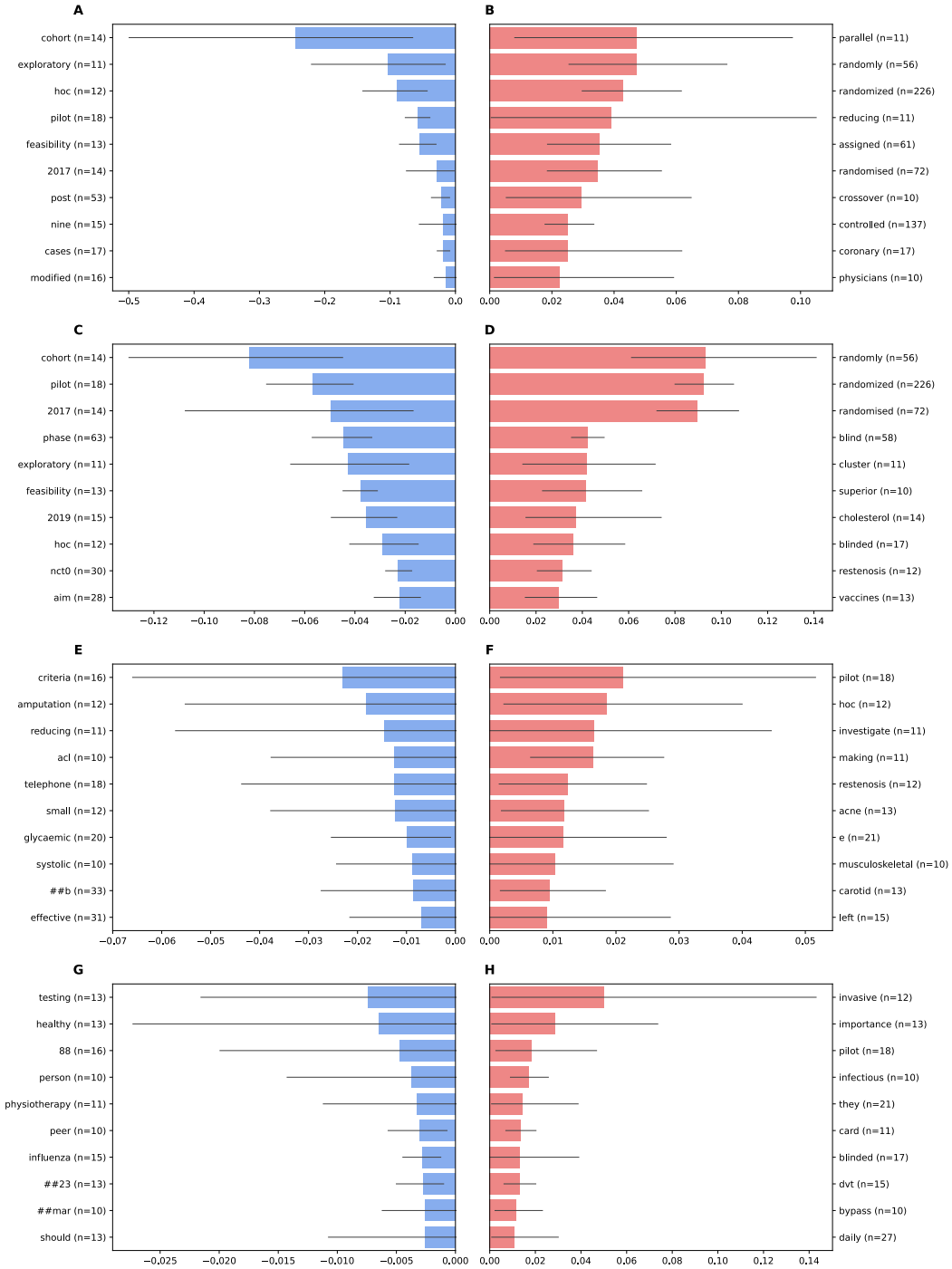
3 | 4.5 Feature importance attributions

The 200 instances contained a total of 80,901 tokens. Of these, 6,369, 1,073, and 87 unique tokens had an occurrence of ≥1, ≥10, and ≥100. The most important unique tokens with ≥10 occurrences can be found in **Figure 3-3**. Those with occurrences ≥1 and ≥100 can be found in **Appendix F**.

Among those with ≥10 and ≥100 occurrences, both SHAP and IG identified tokens that were associated with study designs, including “cohort”, “pilot”, “exploratory”, “randomly”, and “blind”, among others. For the two GPT explainers, there is no consistent pattern among tokens with ≥10 occurrences. Among those with ≥100 occurrences, GPT-index was able to identify terms such as “controlled” and “trial” as being positively associated with higher probability, consistent with SHAP and IG. GPT-token was able to identify key tokens such as “trial”, “randomized”, and “clinical”, but could not identify any negative tokens with ≥100 occurrences.

Important tokens with ≥1 occurrence for SHAP and IG primarily consisted of terms related to study design, year, or topic. There is no consistent pattern between the two GPT explainers.

Figure 3-3. Important tokens with ⩾10 occurrences



A Negative tokens for SHAP; **B** Positive tokens for SHAP; **C** Negative tokens for IG; **D** Positive tokens for IG; **E** Negative tokens for GPT-index; **F** Positive tokens for GPT-index; **G** Negative tokens for GPT-token; **H** Positive tokens for GPT-token.

3 | 5 Discussion

To our knowledge, this is the first experiment that attempts to leverage decoder transformers to establish feature attributions for text classifiers by perturbation. While our results do not indicate GPT to be a potential substitute for conventional explanation methods in this context, this study nevertheless serves as a valuable exploratory analysis that could inspire future research in this area.

3 | 5.1 Summary of findings

While AOPC does not establish absolute faithfulness, it is a common method to compare the relative performance of explainers on the same model (Edin et al., 2024). Our results demonstrate that the SHAP partition explainer and IG were similar in their overall performance. SHAP was shown to better identify negative tokens, while IG was better for positive tokens. Our results also demonstrated that GPT was able to generate reasonable definitions of importance when provided the task of generating feature attributions as an explainer. In spite of this, neither GPT explainer was able to establish useful feature attributions. In particular, the negative AOPC for negative tokens indicate that GPT explainers mistakenly associated negative attributions to features that increased rigour probability. These findings were echoed by the correlation analyses, where attributions by SHAP and IG had a moderate correlation with each other, while the two GPT explainers had weak or no correlation with the others.

While methods to examine the global attributions for transformer models are an area of active research (Covert et al., 2020), the accumulated local attributions across all

200 instances were examined. SHAP and IG indicate that the BioLinkBERT model generally aligned with the manual appraisal criteria (*Methodological Criteria*, n.d.), with terms such as “cohort”, “pilot”, “randomized”, and “blind”, among others, being identified as the most important. The tokens identified by GPT did not align with SHAP or IG and seemed to be nonsensical. For instance, both GPT-index and GPT-token identified “pilot” to be a positive contributor, contrary to manual appraisal and SHAP and IG explanations. The poor faithfulness of GPT for this could be limitations in pre-training, where there is a lack of similar tasks the models could mimic from the training corpora (Gordon, n.d.; Yan et al., 2025).

3 | 5.2 Prompting

A challenge of this experiment was the development of prompts for GPT, considering the complex nature of generating feature attributions from perturbations. It is known that sophisticated prompting techniques can improve GPT’s performance in NLP (Schulhoff et al., 2024; Sivarajkumar et al., 2023, n.d.-f). In this study, numerous established techniques in prompt engineering were tested in an attempt to improve performance, including role prompting, decomposition by providing instructions step by step, as well as chain-of-thought with multiple iterations of perturbations and the redefinition of importance (Schulhoff et al., 2024). GPT was also limited in responding with long, quantitative sequences despite explicit instructions and structured output restrictions (Z. Yang et al., 2023; Yuan et al., 2023). This concern was mitigated by explicitly instructing GPT to respond with a certain number of lists as parameters to the *mask_and_predict*

function, utilizing structured outputs and function calling, and decomposing the calculation steps to 20 tokens per batch. Despite this, GPT was not able to generate faithful attributions. Furthermore, a potential advantage of LLMs would be the ability to recognize likely important tokens before any quantitative explanations have been generated considering their ability to understand and encode contextualized information from plain text (BehnamGhader et al., 2024). Therefore, the two prompting schemes, namely GPT-index and GPT-token, were tested. However, the results show that there was no meaningful difference regardless of the inclusion of input tokens in the initial user prompt.

3 | 5.3 Resource requirements

A challenge with traditional XAI methods is the significant computation resources required. As previously mentioned, the exhaustive nature of calculating SHAP values from all possible perturbations is infeasible, resulting in the rise in numerous methods to approximate SHAP values (*shap.LinearExplainer — SHAP latest documentation*, n.d.; J. Yang, 2021), including the partition explainer (*shap.PartitionExplainer — SHAP latest documentation*, n.d.). The computational requirement for IG is associated with integration steps. While more steps result in higher precision, 30 steps was feasible on GPUs with 32GB of memory and more efficient than the SHAP partition explainer.

Significant computational cost was required due to the iterative approach with the GPT explainers. Similar to SHAP, the BioLinkBERT model must be queried to obtain predictions for the perturbed instances. Additionally, each subsequent prompt in the chain

results in increased inference and response times and financial costs from OpenAI's servers. Ultimately, both prompting schemes incurred a direct API cost of approximately USD$1.00 per instance.

3 | 5.4 Deployment and research implications

Explainability and interpretability in biomedical and clinical ML are key areas of research (Marcus & Teuwen, 2024; G. Yang et al., 2022). As a pioneer in evidence-based medicine and knowledge translation, HIRU aims to not only automate biomedical literature classification and appraisal (Lokker, Abdelkader, et al., 2024; Lokker et al., 2023), but also to ensure that the process is transparent and reproducible to facilitate trust among clinicians who subscribe to PLUS and PLUS-associated services. Based on the results of this experiment, both SHAP and IG may be suitable for deployment alongside a top-performing model. More recently, studies (Dias et al., 2025; Marshall et al., 2017-7) and systematic review support systems (Chelsea, 2023; *DistillerSR AI*, 2023; Rayyan, n.d.) have begun to leverage supervised or active learning extensively to support knowledge translation and synthesis by relevance ranking or automatic classification. For these reasons, systems should attempt to integrate XAI frameworks alongside any black box models for better transparency.

While the performance of GPT as an end-to-end approach for feature attributions was poor, this work nevertheless serves as a foundation for future research. Given the sensitivity of GPT-based explanations to prompt design, future studies could explore more sophisticated, domain-tailored prompting strategies and iterative prompt refinement

using techniques such as few-shot learning to better align GPT's output with domain-specific interpretability criteria (Schulhoff et al., 2024). Fine-tuning LLM explainers on biomedical corpora could also improve their understanding of specialized terminology and context (Luo et al., 2022). Hybrid explanation frameworks, such as leveraging LLMs to establish partition hierarchy (H. Chen et al., 2020; *shap.PartitionExplainer — SHAP latest documentation*, n.d.), or integrating model-internal signals, such as attention weights, with LLM-based explanation methods, may also be of interest (Feng et al., 2023; Ntrougkas et al., 2025; Waghela et al., 2024).

3 | 5.5 Strength and limitations

Our study has several strengths. First, a concern of leveraging LLMs in medical research is reproducibility, as evidence-based medicine is founded upon concepts of transparency, reliability, and the ability to validate findings through rigorous, repeatable methodologies (Davis et al., 2024; Jiawen Deng, Heybati, & Shammas-Toma, 2024; Mete & Özmen, 2024; National Academies of Sciences, Engineering et al., 2019). This concern was addressed using a temperature of 0, making the outputs of the LLM deterministic and replicable. Second, the original AOPC metric was modified to separately consider negative and positive features. This allowed for the better capturing of the faithfulness of the explanations. Third, sophisticated prompting techniques for GPT were used. This indicates that the poor results from GPT are likely an inherent limitation of the pre-training and the model architecture rather than the prompt.

Nevertheless, important limitations must be considered when interpreting our results. First, there has been no known method to establish ground truth in black-box models, and explaining text models with a high feature space remains a challenge (Edin et al., 2024; Melamed & Caruana, 2023). For this reason, it is important to note that our findings are context-, dataset-, and model-specific. Second, due to resource restraints, only a subset of the original dataset could be used. While sampling bias was minimized through stratified sampling, a larger dataset would further increase our confidence. Third, word-piece tokenization often separates words into fragments, potentially affecting how feature attributions are assigned (X. Song et al., 2020). The explanations may not correspond to human-interpretable linguistic units, especially for numerical texts. Lastly, GPT's performance on a task is prompt-specific. While our prompts were sophisticated, it is unknown whether GPT would show promise with a different set of prompts and inputs.

3 | 6 Conclusion

A comprehensive exploration into the application of GPT-4o as a perturbation-based explainer was conducted for a BiolinkBERT biomedical text classifier. Our investigation compared the performance of GPT-driven explanations with the SHAP partition explainer and IG. The results demonstrated that while SHAP and IG provided consistent and relatively faithful feature attributions, the GPT-based approaches yielded poor or counterintuitive explanations as measured by AOPC and an examination of the most important identified tokens. Several challenges were identified. Notably, the sensitivity of GPT outputs to prompt design and the computational overhead associated with iterative perturbation schemes may limit its current applicability in clinical and biomedical settings. Despite these limitations, our work is the first in this area and offers valuable insights and establishes a foundation for future research aimed at integrating LLMs into the explainability framework. Future investigations may wish to explore advanced prompt engineering strategies, domain-specific fine-tuning, and hybrid methods that combine internal model signals with LLM-based explanations, or foundational advances in AI to enhance interpretability without sacrificing performance.

CHAPTER 4

DISCUSSION

4 | 1 Summary of Findings

This work explored two ways GPT can improve the interpretability of medical literature appraisal. In Chapter 2, GPT-4o and GPT-o3-mini were used, alone or with BioLinkBERT, to provide free-text justifications and rationales for each of the nine steps of HIRU's critical appraisal tool. The results showed that GPT-4o alone could achieve comparable performance to a fine-tuned BioLinkBERT model without additional training when the full text was provided. When the results of BioLinkBERT were provided, GPT-4o saw an improvement in performance when only TIABs were provided. Otherwise, there were no notable improvements in providing BioLinkBERT results for GPT, as the performance was not better than using GPT or BioLinkBERT alone. Additionally, GPT-o3-mini's performance was worse than that of GPT-4o in every aspect. These results demonstrated GPT-4o's practical potential to classify medical literature without significant resource and data constraints while providing justifications for human validation in a critical appraisal workflow.

In Chapter 3, GPT-4o was utilized as an explainer that defined importance, perturbed the inputs, and calculated feature attributions. Compared to the SHAP partition explainer and IG, GPT-4o fell short in the faithfulness of their explanations despite sophisticated prompting schemes and the fact that GPT-4o generated reasonable definitions for importance. Regardless of whether the original input tokens were provided in the prompt or not, GPT-4o tended to assign negative attributions to positive token features, resulting in nonsensical explanations. The faithfulness of positive attributions was also relatively low. Both SHAP and IG performed similarly as measured by a

modified version of the AOPC metric. Considering the positive and negative attributions separately, SHAP was better at properly assigning negative attribution to negative tokens, while the opposite was true for IG. Based on these results, unfortunately, GPT-4o is not currently suitable as a standalone perturbation explainer for this type of task.

4 | 2 LLM Mechanisms

It is important to explore potential reasons for the polarizing findings in Chapters 2 and 3. In our experiments, LLMs excelled at language-based classification but struggled with calculating numerical feature attributions. This divergence can be explained by the fundamental design of decoder language models. Specifically, while their transformer-based architecture and extensive language training make them excel at reasoning over unstructured text, they lack the ability to understand numbers and perform complex arithmetic (Shakarian et al., 2023).

Recently, state-of-the-art LLMs have been scaled up drastically, with many consisting of hundreds of billions or even a trillion parameters (Naveed et al., 2023). The multihead self-attention mechanism helps LLMs understand contextual relationships, even over long ranges of text and mitigates issues surrounding vanishing or exploding gradients associated with recurrent NNs (Pascanu et al., 2012). In the context of biomedical literature, GPT can ingest entire articles and reason about their content without domain-specific pre-training and has been shown to match or even surpass human-level performance on complex language tasks (Brin et al., 2023; Divya Venkatesh et al., 2024; E. Guo et al., 2024; Spitale et al., 2023). Overall, the architectural and

training characteristics of LLMs make them particularly powerful for language-driven reasoning and classification tasks even without fine-tuning or retraining.

On the other hand, LLMs have struggled with tasks requiring mathematics and arithmetic (Yuan et al., 2023). This may partly be due to the lack of accurate numerical text feature representations due to byte-pair encoding (Bostrom & Durrett, 2020; F. Zhou, Parrish, et al., 2025). Furthermore, studies have suggested that LLMs heavily rely on similar tasks in the original training data. In other words, LLMs often attempt to match patterns when approaching numerical tasks, as opposed to reasoning through the solutions (Gordon, n.d.; Yan et al., 2025). Considering the novelty of Chapter 3, it is unlikely that GPT was trained on similar tasks. Due to the complex nature of the task and prompts, GPT may have also experienced performance degradation due to the long contexts (N. F. Liu et al., 2023; Ye et al., 2025). Compared to SHAP and IG, which are built on theoretically rigorous mathematical frameworks (Lundberg & Lee, 2017; Sundararajan et al., 2017), GPT was unable to replicate the same level of sophistication despite reasonable definitions of "importance". Unlike IG or other gradient-based explainers, GPT did not have access to model parameters directly. GPT also has no guarantee of satisfying important axioms, including conservation or additivity, fundamental to XAI frameworks.

4 | 3 Practice and Research Implications

Considering the limitations of current ML classifiers, including computational cost, ease of use, and transparency, the practical implications of these findings are significant for biomedical research workflows.

4 | 3.1 LLM development and deployment

Based on the results, sophisticated LLMs, such as GPT-4o, should be considered in knowledge synthesis and translation workflows, especially for organizations with limited access to data or computational resources. For zero- or few-shot classification tasks, LLMs are relatively inexpensive and efficient, offering potential mitigation towards increasing medical literature. Additionally, the ability for LLMs to provide step-by-step explanations for every decision is a key strength, allowing human researchers to easily verify its decisions. This is echoed by several other studies that utilized LLMs for systematic review screening (E. Guo et al., 2024; Oami et al., 2024) or critical appraisal (Hasan et al., 2024; Lai et al., 2024; Pitre et al., 2023). In the real world, LLMs can drastically reduce the human workload in identifying relevant or high-quality articles, thereby significantly reducing time and financial costs (Cao et al., 2025).

Currently, no known evidence translation or synthesis systems, such as Covidence and DistillerSR, utilize LLMs, despite their integration of ML to aid in relevance ranking or RCT identification (Aliani, 2024; Chelsea, 2023). These systems could benefit from the deployment of LLMs to further aid researchers in labour-intensive processes. For instance, LLMs can highlight relevant texts or make classification suggestions based on the inclusion and exclusion criteria and support their own decisions with plain text. Additionally, they may also be able to refine inclusion criteria and identify potential areas of ambiguity (Delgado-Chaves et al., 2025). These potential uses warrant exploration by literature translation systems.

One significant limitation of LLMs is the lack of flexibility in terms of threshold tuning. In text classification tasks, models may be tuned to prioritize inclusions, exclusions, or a balanced approach using different probability thresholds (F. Zhou, Parrish, et al., 2025). For instance, most workflows identifying relevant articles would prioritize the automated exclusion of irrelevant articles, and deployed models would have a high sensitivity to maximize the quality of negative classifications (Lokker et al., 2023; F. Zhou, Parrish, et al., 2025). The lack of threshold tuning for LLMs necessitates experimentation with different prompts to tailor LLMs' classifications or to use them as a second reviewer (Hasan et al., 2024).

Many studies examining LLMs primarily focus on the public-facing graphical user interface and fail to leverage developer tools, such as the API (Qingyu Chen et al., 2023; E. Guo et al., 2024; Hasan et al., 2024; Lai et al., 2024; Oami et al., 2024; Pitre et al., 2023; Tang et al., 2023; Yanagita et al., 2023). For this reason, current research is limited in reproducibility due to the inherent stochasticity in LLMs. Since reproducibility is a fundamental concept of scientific research, a temperature of 0, a set seed, or averaging over multiple repetitions may be necessary to ensure that the results are replicable by others. In the current state, selective reporting of optimistic findings cannot be easily identified. In addition, future studies may wish to explore the effects of parameters, such as frequency and presence penalty, to assess how they affect classification performance.

In general, reasoning models have demonstrated superior performance in complex, reasoning tasks but offer limited benefits otherwise (*LiveBench*, n.d.; G. Wang

et al., 2024). How they perform in medical literature classification has seldom been explored. This work demonstrates that they are slightly worse but remain relatively potent compared to general LLMs. However, this result may not be generalizable to another dataset or reasoning model. Several other factors may need to be considered during development. OpenAI suggests that reasoning models may only require high-level guidance as opposed to specific instructions (*OpenAI platform*, n.d.), and this may be a key advantage. On the other hand, there is no in-depth exploration of prompt techniques for reasoning LLMs, and the additional cost may be prohibitive (*Precios*, n.d.). Ultimately, the choice between general-purpose and reasoning LLM needs to be considered holistically.

4 | 3.2 Explainability and confidence in research and practice

Explanation faithfulness metrics are an area of active research, and there is no known established metric to measure the faithfulness of explanations. While the original AOPC was used in several previous studies (H. Chen et al., 2020; Nguyen, 2018; Samek et al., 2016), it has important limitations, including the lack of consideration for negative features and the inability to generalize the results of one model and dataset to another. An important contribution of Chapter 3 is the modified version of AOPC that considers positive and negative features separately. Explanation studies involving tabular or text classification should consider this metric, as tokens may significantly affect the predicted probability in either direction, unlike image classification. For generalizability, a study recently introduced a normalized version of AOPC that allows for direct comparison

across different datasets and models (Edin et al., 2024). The normalized AOPC may be considered if generalizability is a primary consideration.

Notwithstanding the importance of interpretability, previous research and evidence-synthesis platforms seldom integrate explanations with ML models (Chaddad et al., 2023; Q. Sun et al., 2024). Considering the findings of Chapter 3, it is recommended that SHAP, IG, or both be deployed alongside ML models. Furthermore, externally validating models on a different or future dataset is paramount to establishing the generalizability. External validation is as important in clinical modelling studies as in medical text NLP applications to improve researchers' confidence in model predictions and gauge performance degradation due to potential data drift over time (Heßler et al., 2020; Markey et al., 2024; Sahiner et al., 2023). While LLMs may not be able to generate faithful attributions, they should be used in combination with current XAI frameworks. For instance, generating plain-text explanations with feature attributions from XAI is a promising use, allowing non-technical stakeholders to better understand the explanations (Zeng, 2024; Zytek, Pido, et al., 2024; Zytek, Pidò, et al., 2024). Additionally, there are preferences for using interpretable models as opposed to black box models with XAI (Rudin, 2019). Therefore, the inherent interpretability of LLM outputs for text classification is a considerable strength.

4 | 3.3 Implications for HIRU

Currently, HIRU's automation is rudimentary, with an overarching BioBERT classifier excluding potentially irrelevant articles. Articles that pass BioBERT proceed with human

evaluations as normal. With the development of HIRU's Critical Appraisal Process 2.0, several implications can be drawn from this work.

First, it is important that specific models are used to automate each step of the process, as opposed to an overarching model applied to all article types, purposes, and rigour criteria. While sophisticated NNs may not necessarily suffer from performance degradation, the interpretation of a model with mixed tasks is obscure, as it would be difficult to trace the contribution of features or patterns to the prediction. The previous experiment that this work extended upon (F. Zhou, Parrish, et al., 2025) is a step in this direction, exploring the performance of models specific to the rigour classification of original treatment, primary prevention, and quality improvement studies. Furthermore, two additional experiments detailing the multiclass classification of study type (F. Zhou, Lokker, et al., 2025), as well as the rigour classification of reviews (F. Zhou, Afzal, et al., 2025), are also complete. Reports of other models are under development and will be completed and submitted for publication in the near future.

In addition to encoder-only transformers, HIRU should also consider including LLMs in its workflow. Using LLMs to provide classifications alongside encoder transformer models can provide additional warranty in case of specific errors. For instance, during the initial exclusion phase, an article with disagreeing classifications may warrant attention from a human reviewer. Subsequently, articles that are passed and manually evaluated should be compared with both models and in case of disagreements, an arbiter can utilize highlighting based on SHAP and IG feature attributions and justifications provided by LLMs to efficiently resolve the conflict.

Lastly, it is evident that drift in medical literature may degrade model performance over time (Abdelwahab et al., 2023). This concern is present in HIRU as well, as models consistently performed worse on Clinical Hedges and future datasets (Lokker et al., 2023; F. Zhou, Parrish, et al., 2025). Studies investigated using feature attribution XAI frameworks to track changes in feature importance over time (Duckworth et al., 2021; Haug et al., 2022; Y. Lee et al., 2023). Duckworth et al. specifically monitored SHAP value changes to assess the presence of drift in a model that predicted the risk of hospital admission for patients attending the emergency department (Duckworth et al., 2021). This approach may be implemented by HIRU as well to periodically retrain models with updated annotations to mitigate performance degradation over time.

4 | 4 Future Research Directions

4 | 4.1 Prompting

Since the introduction of LLMs, prompting has been the centre of exploration. The work presented here leveraged several comprehensive surveys that detailed current best practices in prompting (W. Li et al., 2025; Sahoo et al., 2024; Schulhoff et al., 2024; Vatsal & Dubey, 2024). Nevertheless, future work should aim to explore how prompting affects LLM preferences for inclusion versus exclusion in text classification, allowing for additional flexibility tailored toward specific workflows. Additionally, another area of interest is how the effectiveness of prompts differs between general LLMs and reasoning models in this context, as previous work has suggested that prompt engineering

techniques developed for earlier LLMs may hinder performance on newer models (G. Wang et al., 2024). To further improve performance, few-shot prompting with classification examples may be promising, incorporating a few gold-standard examples for in-context learning (Ge et al., 2022). Additionally, injecting domain-specific knowledge into LLMs trained on general corpora may be worth exploring (Z. Song et al., 2025). However, research in this area is limited, potentially due to a lack of data or standardized benchmarks (Ge et al., 2023). A promising area of research to further reduce manual burden is automated prompt generation. Unlike manual prompt engineering, which is time-consuming and often reliant on domain expertise and trial-and-error, automated generation seeks to programmatically generate effective prompts for downstream tasks. Several recent approaches, such as AutoPrompt (Shin et al., 2020), utilize gradient-guided search to discover token sequences that elicit desired model behaviour, enabling interpretable and label-efficient prompt creation. RLPrompt and PromptAgent (M. Deng et al., 2022; X. Wang et al., 2023) are other notable frameworks that leverage reinforcement learning or multi-agent planning to iteratively optimize prompts based on task feedback.

4 | 4.2 Standardization for explanations

Unlike model performance, there is no accepted standardized benchmark and metric for the explanation of faithfulness. Consequently, the comparison of XAI methods has largely been infeasible. However, Edin et al. made a meaningful contribution by introducing the normalized AOPC, allowing AOPC values to be compared across

separate datasets and models (Edin et al., 2024). Furthermore, two benchmarks were introduced relatively recently as well. The Evaluating Rationales And Simple English Reasoning (ERASER) benchmark is a framework used to evaluate NLP XAI using seven datasets, including but not limited to sentiment analysis, evidence inference, and question-answering (DeYoung et al., 2019). Each of these datasets was annotated by humans with decisions and rationales, and several baseline models were provided for explanation. Explanation performance was measured by agreement with human rationales using Intersection-Over-Union and F1 scores, as well as token rankings and AUPRC. Additionally, faithfulness was measured using comprehensiveness, sufficiency, and AOPC. The M4 benchmark focuses on evaluating XAI feature attribution faithfulness (X. Li et al., 2023). The benchmark encompasses image classification and sentiment analysis using ImageNet (Jia Deng et al., 2009) and MovieReview (Zaidan & Eisner, 2008), and several publicly available classifiers were recommended. Regarding metrics, four metrics are used when no ground truth is present: the most relevant first, the least relevant first, the area between the perturbation curves and Infidelity. The PScore and SynScore were proposed for cases with pseudo ground truth and synthetic ground truth, respectively.

Despite this, there is no known guideline for the development and evaluation of explainers. This is partially due to the fact that it is unknown how the task, dataset, model performance or model architecture would affect the performance of an explainer, and whether a well-performing explainer on standardized benchmarks would generalize to bespoke environments. There is also no benchmark focused on medical literature

classification. As transparency in AI becomes an increasing concern, future research on these gaps may be warranted.

4 | 4.3 GPT for feature attributions

The task-agnostic nature of this work enabled it to examine GPT's capabilities in generating feature attributions, an area that had not been previously explored. While GPT did not obtain satisfactory performance, several interesting areas for continued research may be warranted. First, since the task is iterative by nature, reasoning LLMs, such as GPT-o3, warrant further exploration as they are designed with automatic chain-of-thought and are particularly suitable for complex, numerical tasks (*Learning to reason with LLMs*, n.d.). While this work leveraged long chains of prompts, this may not necessarily be beneficial to model performance (Kusano et al., 2024). The reasoning steps built into reasoning LLMs may be better suited as the model can decide and adjust the length and sophistication for itself. Past studies have found promising results using LLMs to translate feature attributions from traditional XAI frameworks to plain-text rationales (Zeng, 2024; Zytek, Pido, et al., 2024; Zytek, Pidò, et al., 2024). Considering this, LLMs may be used in combination with XAI tools to improve faithfulness. For instance, the contextualized understanding of input tokens may allow LLMs to generate hierarchical partitions more effectively than default methods relying on spatial distance. This could lead to more semantically meaningful input segmentations, particularly for complex documents where spatial locality does not necessarily reflect conceptual relationships. Future work could explore hybrid frameworks where LLMs are leveraged not to generate

final attributions directly, but to assist with preprocessing steps in perturbation-based XAI methods, such as feature grouping, salience filtering, or baseline selection (Salih et al., 2024).

4 | 5 Strengths and Limitations

This work has several unique strengths compared to recent literature. First, the stochasticity of LLMs was considered and addressed by setting temperature or seed equal to 0. This ensures that the model outputs are replicable with the same prompt and configurations, allowing other researchers to validate and build on the findings reliably. Second, the dataset used by both experiments was sourced from gold-standard databases and randomly stratified to mitigate sampling bias. This ensured that the cost associated with the experiments was practical while ensuring that the results could be trusted. Third, valid comparisons were used in both experiments to provide a clear context of GPT performance. For classification, a strong BioLinkBERT model was used as a baseline, and for explanations, both the SHAP partition explainer and IG are widely accepted (M. Mersha et al., 2024). Lastly, sophisticated prompting techniques were used in an attempt to improve the performance of GPT. It is evident through past research that an LLM's performance can be heavily influenced by the content and the structure of the prompts (J. He et al., 2024; J. Kim et al., 2023). This research utilized several techniques, including role prompting, chain-of-thought, and decomposition, and leveraged relevant developer tools such as structured outputs and function calling to ensure performance and consistency.

There are also several limitations that must be considered. First, it is important to note that the results of both experiments are limited to HIRU's datasets, criteria, and models, and may not necessarily generalize to other contexts, such as observational studies or clinical notes. For feature attributions, the proprietary nature of AOPC precludes any meaningful comparison to other contexts. Future studies may wish to extend this research to other datasets and tools to better assess the potential of LLMs for these tasks. The development of standardized benchmarks in this field may also be warranted, as previously discussed. Second, while the prompts were sophisticated, it is unknown how techniques such as few-shot prompting affect performance. Furthermore, due to cost and time constraints, the results are limited to OpenAI's GPT models with specific configurations and may not translate to other configurations or LLMs, such as LLaMA and Claude. LLM performance may vary considerably based on configurations and prompting strategies (Vatsal & Dubey, 2024). Therefore, studies focusing on other techniques and models may be of interest. Third, the computational and financial costs were barriers in this work, and consequently, random samples of the original datasets were used. The training of encoder-only transformers and the calculation of feature attributions and AOPC require access to powerful central and graphical processing units. The access to GPT, especially for Chapter 3, was considerably costly at approximately $1 per instance. This raises concerns relating to the scalability and accessibility of these solutions. Nevertheless, these limitations may become less apparent as more efficient LLMs are developed.

REFERENCES

28 saliency maps – interpretable machine learning. (n.d.). Retrieved April 3, 2025, from https://christophm.github.io/interpretable-ml-book/pixel-attribution.html

Abdelwahab, H., Martens, C., Beck, N., & Wegener, D. (2023). Investigation of drift detection for clinical text classification. In *Studies in Computational Intelligence* (pp. 43–56). Springer Nature Switzerland.

AbdulNabi, I., & Yaseen, Q. (2021). Spam email detection using deep learning techniques. *Procedia Computer Science*, *184*, 853–858.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1810.03292

Afzal, M., Hussain, J., Abbas, A., Hussain, M., Attique, M., & Lee, S. (2024). Transformer-based active learning for multi-class text annotation and classification. *Digital Health*, *10*, 20552076241287356.

Aitken, K., Ramasesh, V. V., Cao, Y., & Maheswaranathan, N. (2021). Understanding how encoder-decoder architectures attend. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2110.15253

Aliani, R. (2024, March 14). *From manual to machine: How covidence's ML is streamlining systematic reviews*. Covidence. https://www.covidence.org/blog/from-manual-to-machine-how-covidences-ml-is-streamlining-systematic-reviews/

Al-Jundi, A., & Sakka, S. (2017). Critical appraisal of clinical research. *Journal of Clinical and Diagnostic Research: JCDR*, *11*(5), JE01–JE05.

Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D., & Aliferis, C. F. (2005). Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association: JAMIA*, *12*(2), 207–216.

Ateia, S., & Kruschwitz, U. (2023). Is ChatGPT a biomedical expert? -- exploring the zero-shot performance of current GPT models in biomedical tasks. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2306.16108

Aum, S., & Choe, S. (2021). srBERT: automatic article classification model for systematic review using BERT. *Systematic Reviews*, *10*(1), 285.

Baker, S., Silins, I., Guo, Y., Ali, I., Högberg, J., Stenius, U., & Korhonen, A. (2016). Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics (Oxford, England)*, *32*(3), 432–440.

Bastian, M. (2023, March 25). *GPT-4 has more than a trillion parameters - Report*. The-decoder.com. https://the-decoder.com/gpt-4-has-a-trillion-parameters/

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., & Reddy, S. (2024). LLM2Vec: Large language models are secretly powerful text encoders. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2404.05961

bert: TensorFlow code and pre-trained models for BERT. (n.d.). Github. Retrieved April 2, 2025, from https://github.com/google-research/bert

Bertossi, L., Li, J., Schleich, M., Suciu, D., & Vagena, Z. (2020, June 14). Causality-based explanation of classification outcomes. *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*. SIGMOD/PODS '20: International Conference on Management of Data, Portland OR USA. https://doi.org/10.1145/3399579.3399865

Bitton, R., Malach, A., Meiseles, A., Momiyama, S., Araki, T., Furukawa, J., Elovici, Y., & Shabtai, A. (2022). Latent SHAP: Toward practical human-interpretable explanations. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2211.14797

BLURB leaderboard. (n.d.). Retrieved August 13, 2024, from https://microsoft.github.io/BLURB/leaderboard.html

Bostrom, K., & Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.03720

Bousselham, H., Nfaoui, E. H., & Mourhir, A. (2024). Fine-tuning GPT on biomedical NLP tasks: An empirical evaluation. *2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, 1–6.

Brin, D., Sorin, V., Vaid, A., Soroush, A., Glicksberg, B. S., Charney, A. W., Nadkarni, G., & Klang, E. (2023). Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports*, *13*(1), 16492.

Build with Claude. (n.d.). Retrieved February 24, 2025, from https://www.anthropic.com/api

Cao, C., Sang, J., Arora, R., Chen, D., Kloosterman, R., Cecere, M., Gorla, J., Saleh, R., Drennan, I., Teja, B., Fehlings, M., Ronksley, P., Leung, A. A., Weisz, D. E.,

Ware, H., Whelan, M., Emerson, D. B., Arora, R. K., & Bobrovitz, N. (2025). Development of prompt templates for large language model-driven screening in systematic reviews. *Annals of Internal Medicine*, *178*(3), 389–401.

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors (Basel, Switzerland)*, *23*(2), 634.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2010.02559

Chatla, S. K., Geeda, S. C. R., Pavani, K., Kokkiligadda, R. R., & Shaik, M. G. (2024). Comparative analysis of sentiment analysis models on twitter data using machine learning. In *Lecture Notes in Networks and Systems* (pp. 729–738). Springer Nature Singapore.

Chelsea. (2023, February 7). *Machine learning – the game changer for trustworthy evidence*. Covidence. https://www.covidence.org/blog/machine-learning-the-game-changer-for-trustworthy-evidence/

Chen, H., Zheng, G., & Ji, Y. (2020). Generating hierarchical explanations on text classification via feature interaction detection. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.02015

Chen, Qijie, Sun, H., Liu, H., Jiang, Y., Ran, T., Jin, X., Xiao, X., Lin, Z., Chen, H., & Niu, Z. (2023). An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics (Oxford, England)*, *39*(9). https://doi.org/10.1093/bioinformatics/btad557

Chen, Qingyu, Du, J., Allot, A., & Lu, Z. (2022). LitMC-BERT: Transformer-Based Multi-Label Classification of Biomedical Literature With An Application on COVID-19 Literature Curation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *19*(5), 2584–2595.

Chen, Qingyu, Hu, Y., Peng, X., Xie, Q., Jin, Q., Gilson, A., Singer, M. B., Ai, X., Lai, P.-T., Wang, Z., Keloth, V. K., Raja, K., Huang, J., He, H., Lin, F., Du, J., Zhang, R., Zheng, W. J., Adelman, R. A., … Xu, H. (2023). A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.16326

Chen, S., Li, Y., Lu, S., Van, H., Aerts, H. J. W. L., Savova, G. K., & Bitterman, D. S. (2024). Evaluating the ChatGPT family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association: JAMIA*, *31*(4), 940–948.

Chernikova, O., Stadler, M., Melev, I., & Fischer, F. (2024). Using machine learning for continuous updating of meta-analysis in educational context. *Computers in Human Behavior*, *156*(108215), 108215.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2003.10555

Cohen, A. M., Hersh, W. R., Peterson, K., & Yen, P.-Y. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association: JAMIA*, *13*(2), 206–219.

Cortiz, D. (2021). Exploring transformers in Emotion Recognition: A comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2104.02041

Covert, I., Lundberg, S., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2004.00668

Covidence - Better systematic review management. (2020, June 11). Covidence. https://www.covidence.org/

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1801.02143

Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The temperature feature of ChatGPT: Modifying creativity for clinical research. *JMIR Human Factors*, *11*, e53559.

Decoder models - Hugging Face NLP Course. (n.d.). Retrieved April 2, 2025, from https://huggingface.co/learn/nlp-course/en/chapter1/6

Delgado-Chaves, F. M., Jennings, M. J., Atalaia, A., Wolff, J., Horvath, R., Mamdouh, Z. M., Baumbach, J., & Baumbach, L. (2025). Transforming literature screening: The emerging role of large language models in systematic reviews. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(2), e2411962122.

Deng, Jia, Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A largescale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Deng, Jiawen, Heybati, K., Park, Y.-J., Zhou, F., & Bozzo, A. (2024). Artificial intelligence in clinical practice: A look at ChatGPT. *Cleveland Clinic Journal of Medicine*, *91*(3), 173–180.

Deng, Jiawen, Heybati, K., & Shammas-Toma, M. (2024). When vision meets reality: Exploring the clinical applicability of GPT-4 with vision. *Clinical Imaging*, *108*(110101), 110101.

Deng, Jiawen, Moskalyk, M., Shammas-Toma, M., Aoude, A., Ghert, M., Bhatnagar, S., & Bozzo, A. (2024). Development of machine learning models for predicting the 1-year risk of reoperation after lower limb oncological resection and endoprosthetic reconstruction based on data from the PARITY trial. *Journal of Surgical Oncology*, *130*(8), 1706–1716.

Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E. P., & Hu, Z. (2022). RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2205.12548

Densen, P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*, *122*, 48–58.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. In *arXiv [cs.CL]*. https://doi.org/10.48550/ARXIV.1810.04805

DeYoung, J., Jain, S., Rajani, N., Lehman, E. P., Xiong, C., Socher, R., & Wallace, B. C. (2019). ERASER: A benchmark to evaluate rationalized NLP models. *Annual Meeting of the Association for Computational Linguistics*, 4443–4458.

Dhammi, I. K., & Kumar, S. (2014). Medical subject headings (MeSH) terms. *Indian Journal of Orthopaedics*, *48*(5), 443–444.

Dias, A. C., Moreira, V. P., & Comba, J. L. D. (2025). RoBIn: A Transformer-based model for risk of bias inference with machine reading comprehension. *Journal of Biomedical Informatics*, 104819.

DistillerSR AI. (2023, December 22). DistillerSR; DistillerSR Inc. https://www.distillersr.com/products/distillersrai

Divya Venkatesh, J., Jaiswal, A., & Nanda, G. (2024). Comparing human text classification performance and explainability with large language and machine learning models using eye-tracking. *Scientific Reports*, *14*(1), 14295.

Duckworth, C., Chmiel, F. P., Burns, D. K., Zlatev, Z. D., White, N. M., Daniels, T. W. V., Kiuber, M., & Boniface, M. J. (2021). Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Scientific Reports*, *11*(1), 23017.

Edin, J., Motzfeldt, A. G., Christensen, C. L., Ruotsalo, T., Maaløe, L., & Maistro, M. (2024). Normalized AOPC: Fixing misleading faithfulness metrics for feature attribution explainability. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2408.08137

Enguehard, J. (2023). Sequential Integrated Gradients: a simple but effective method for explaining language models. In *arXiv [cs.CL]*. arXiv. https://aclanthology.org/2023.findings-acl.477.pdf

Feng, H., Lin, Z., & Ma, Q. (2023). Perturbation-based self-supervised attention for attention bias in text classification. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.15684

Fomin, V. V., & Astromskis, P. (2023). The Black Box Problem. In *Future Law, Ethics, and Smart Technologies* (pp. 112–125). BRILL.

Fukami, T. (2024). Enhancing healthcare accountability for administrators: Fostering transparency for patient safety and quality enhancement. *Cureus*, *16*(8), e66007.

Gai, N., Aoyama, K., Faraoni, D., Goldenberg, N. M., Levin, D. N., Maynes, J. T., McVey, M. J., Munshey, F., Siddiqui, A., Switzer, T., & Steinberg, B. E. (2021). General medical publications during COVID-19 show increased dissemination despite lower validation. *PloS One*, *16*(2), e0246427.

Gannot, G., Cutting, M. A., Fischer, D. J., & Hsu, L. J. (2017). Reproducibility and transparency in biomedical sciences. *Oral Diseases*, *23*(7), 813–816.

Ge, Y., Guo, Y., Das, S., Al-Garadi, M. A., & Sarker, A. (2023). Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, *144*(104458), 104458.

Ge, Y., Guo, Y., Yang, Y.-C., Al-Garadi, M. A., & Sarker, A. (2022). *Few-shot learning for medical text: A systematic review*. https://doi.org/10.48550/arXiv.2204.14081

Ghorbani, A., Abid, A., & Zou, J. (2017). Interpretation of Neural Networks is Fragile. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1710.10547

Ghosh, M., Mukherjee, S., Ganguly, A., Basuchowdhuri, P., Naskar, S. K., & Ganguly, D. (2024). AlpaPICO: Extraction of PICO frames from clinical trial documents using LLMs. *Methods (San Diego, Calif.)*, *226*, 78–88.

Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2107.07045

Golestaneh, P., Taheri, M., & Lederer, J. (2024). How many samples are needed to train a deep neural network? In *arXiv [math.ST]*. arXiv. http://arxiv.org/abs/2405.16696

Gordon, R. (n.d.). *Reasoning skills of large language models are often overestimated*. MIT News | Massachusetts Institute of Technology. Retrieved April 9, 2025, from https://news.mit.edu/2024/reasoning-skills-large-language-models-often-overestimated-0711

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23.

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024). Automated paper screening for clinical reviews using large language models: Data analysis study. *Journal of Medical Internet Research*, *26*, e48996.

Guo, L., Zhang, D., Wang, L., Wang, H., & Cui, B. (2018). CRAN: A hybrid CNN-RNN attention-based model for text classification. In *Conceptual Modeling* (pp. 571–585). Springer International Publishing.

Guo, Y., Ovadje, A., Al-Garadi, M. A., & Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association: JAMIA*, *31*(10), 2181–2189.

Haddaoui, B. E., Chiheb, R., Faizi, R., & Afia, A. E. (2025). Sentiment Analysis in SemEval: A review of sentiment identification approaches. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2503.10457

Harish, K. B., Price, W. N., & Aphinyanaphongs, Y. (2022). Open-source clinical machine learning models: Critical appraisal of feasibility, advantages, and challenges. *JMIR Formative Research*, *6*(4), e33970.

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG: An International Journal of Obstetrics and Gynaecology*, *125*(13), 1716.

Hasan, B., Saadi, S., Rajjoub, N. S., Hegazi, M., Al-Kordi, M., Fleti, F., Farah, M., Riaz, I. B., Banerjee, I., Wang, Z., & Murad, M. H. (2024). Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evidence-Based Medicine*, *29*(6), 394–398.

Hassan, S., Rafi, M., & Shaikh, M. S. (2012). Comparing SVM and Naive Bayes classifiers for text categorization with Wikitology as knowledge enrichment. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1202.4063

Haug, J., Braun, A., Zürn, S., & Kasneci, G. (2022). Change detection for local explainability in evolving data streams. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2209.02764

Haynes, R. B., Holland, J., Cotoi, C., McKinlay, R. J., Wilczynski, N. L., Walters, L. A., Jedras, D., Parrish, R., McKibbon, K. A., Garg, A., & Walter, S. D. (2006). McMaster PLUS: a cluster randomized clinical trial of an intervention to accelerate clinical use of evidence-based information from digital libraries. *Journal of the American Medical Informatics Association: JAMIA*, *13*(6), 593–600.

He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does prompt formatting have any impact on LLM performance? In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2411.10541

He, X., Zhao, K., & Chu, X. (2019). AutoML: A survey of the state-of-the-art. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1908.00709

Hernández, A., & Amigó, J. M. (2021). Attention mechanisms and their applications to complex systems. *Entropy (Basel, Switzerland)*, *23*(3), 283.

Heßler, N., Rottmann, M., & Ziegler, A. (2020). Empirical analysis of the text structure of original research articles in medical journals. *PloS One*, *15*(10), e0240288.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1712.00409

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., Sterne, J. A. C., Cochrane Bias Methods Group, & Cochrane Statistical Methods Group. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* , *343*, d5928.

Higgins, J., & Welch, V. (2011). *Cochrane handbook for systematic reviews of interventions* (J. Higgins & S. Green, Eds.). Standards Information Network. https://training.cochrane.org/handbook

Holistic Evaluation of Language Models (HELM). (n.d.). Retrieved April 2, 2025, from https://crfm.stanford.edu/helm/

Hopewell, S., Keene, D. J., Heine, P., Marian, I. R., Dritsaki, M., Cureton, L., Dutton, S. J., Dakin, H., Carr, A., Hamilton, W., Hansen, Z., Jaggi, A., Littlewood, C., Barker, K., Gray, A., & Lamb, S. E. (2021). Progressive exercise compared with best-practice advice, with or without corticosteroid injection, for rotator cuff disorders: the GRASP factorial RCT. *Health Technology Assessment (Winchester, England)*, *25*(48), 1–158.

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., & Azim, M. A. (2022). Transfer learning: a friendly introduction. *Journal of Big Data*, *9*(1), 102.

Howarth, J. (2024, August 6). Number of parameters in GPT-4 (latest data). *Exploding Topics*. https://explodingtopics.com/blog/gpt-parameters

Hsieh, W., Bi, Z., Jiang, C., Liu, J., Peng, B., Zhang, S., Pan, X., Xu, J., Wang, J., Chen, K., Feng, P., Wen, Y., Song, X., Wang, T., Liu, M., Yang, J., Li, M., Jing, B., Ren, J., … Liang, C. X. (2024). A comprehensive guide to explainable AI: From Classical models to LLMs. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2412.00800

Irwin, A. N., & Rackham, D. (2017). Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. *Research in Social & Administrative Pharmacy: RSAP*, *13*(2), 389–393.

Issler, R. M. S., Marostica, P. J. C., & Giugliani, E. R. J. (2009). Infant sleep position: a randomized clinical trial of an educational intervention in the maternity ward in Porto Alegre, Brazil. *Birth (Berkeley, Calif.)*, *36*(2), 115–121.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: A Dataset for Biomedical Research Question Answering. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1909.06146

Justus, D., Brennan, J., Bonner, S., & McGough, A. S. (2018). Predicting the computational cost of deep learning models. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1811.11880

Kanakarajan, K. R., Kundumani, B., & Sankarasubbu, M. (2021). BioELECTRA:Pretrained Biomedical text Encoder using Discriminators. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of*

the 20th Workshop on Biomedical Language Processing (pp. 143–154). Association for Computational Linguistics.

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., & Haynes, R. B. (2009). Towards automatic recognition of scientifically rigorous clinical research evidence. *Journal of the American Medical Informatics Association: JAMIA*, *16*(1), 25–31.

Kim, H.-H., Yoo, K.-J., & Youn, Y.-N. (2023). A randomized trial of clopidogrel vs ticagrelor after off-pump coronary bypass. *The Annals of Thoracic Surgery*, *115*(5), 1127–1134.

Kim, J., Park, S., Jeong, K., Lee, S., Han, S. H., Lee, J., & Kang, P. (2023). Which is better? Exploring Prompting Strategy For LLM-based Metrics. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2311.03754

Kirman, A. P., Aumann, R. J., & Shapley, L. S. (1976). Values of non-atomic games. *Economica*, *43*(172), 445.

Kiru, G., Bicknell, C., Falaschetti, E., Powell, J., & Poulter, N. (2016). An evaluation of the effect of an angiotensin-converting enzyme inhibitor on the growth rate of small abdominal aortic aneurysms: a randomised placebo-controlled trial (AARDVARK). *Health Technology Assessment (Winchester, England)*, *20*(59), 1–180.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2205.11916

Kotelnikova, A., Paschenko, D., Bochenina, K., & Kotelnikov, E. (2021). Lexicon-based methods vs. BERT for text sentiment analysis. In *arXiv [cs.CL]*. https://doi.org/10.48550/ARXIV.2111.10097

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Basel)*, *10*(4), 150.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text classification algorithms: A survey. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1904.08067

Kumar, B., Sheetal, Badiger, V. S., & Jacintha, A. D. (2024). Sentiment Analysis for Products Review based on NLP using Lexicon-Based Approach and Roberta. *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, 1–6.

Kumar, S., Roy, P. P., Dogra, D. P., & Kim, B.-G. (2023). A comprehensive review on Sentiment Analysis: Tasks, approaches and applications. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2311.11250

Kusano, G., Akimoto, K., & Takeoka, K. (2024). Are longer prompts always better? Prompt selection in large language models for recommendation systems. In *arXiv [cs.IR]*. arXiv. http://arxiv.org/abs/2412.14454

Lai, H., Ge, L., Sun, M., Pan, B., Huang, J., Hou, L., Yang, Q., Liu, J., Liu, J., Ye, Z., Xia, D., Zhao, W., Wang, X., Liu, M., Talukdar, J. R., Tian, J., Yang, K., & Estill, J. (2024). Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Network Open*, *7*(5), e2412687.

Lange, T., Schwarzer, G., Datzmann, T., & Binder, H. (2021). Machine learning for identifying relevant publications in updates of systematic reviews of diagnostic test studies. *Research Synthesis Methods*, *12*(4), 506–515.

Le, H. T., Cerisara, C., & Denis, A. (2017). Do Convolutional Networks need to be Deep for Text Classification ? In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1707.04108

Learning to reason with LLMs. (n.d.). Retrieved April 8, 2025, from https://openai.com/index/learning-to-reason-with-llms/

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *ArXiv [Cs.CL]*. https://doi.org/10.48550/ARXIV.1901.08746

Lee, Y., Lee, Y., Lee, E., & Lee, T. (2023). Explainable artificial intelligence-based model drift detection applicable to unsupervised environments. *Computers, Materials & Continua*, *76*(2), 1701–1719.

Leong, K. C., Chen, W. S., Leong, K. W., Mastura, I., Mimi, O., Sheikh, M. A., Zailinawati, A. H., Ng, C. J., Phua, K. L., & Teng, C. L. (2006). The use of text messaging to improve attendance in primary care: a randomized controlled trial. *Family Practice*, *23*(6), 699–705.

Li, K., DeCost, B., Choudhary, K., Greenwood, M., & Hattrick-Simpers, J. (2022). A critical examination of robustness and generalizability of machine learning prediction of materials properties. In *arXiv [cond-mat.mtrl-sci]*. arXiv. http://arxiv.org/abs/2210.13597

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A survey on text classification: From shallow to deep learning. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2008.00364

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology*, *13*(2), 1–41.

Li, W., Wang, X., Li, W., & Jin, B. (2025). A survey of automatic prompt engineering: An optimization perspective. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2502.11560

Li, X., Du, M., Chen, J., Chai, Y., Lakkaraju, H., & Xiong, H. (2023). M4: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. *Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2023/file/05957c194f4c77ac9d91e1374d2def6b-Paper-Datasets_and_Benchmarks.pdf

Li, X.-L. (n.d.). *Rule-based Classi cation*. Retrieved November 13, 2024, from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8557fde1e865f0dcc209468ccaf94f12a04b7835

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2307.03172

Liu, Shicai, Tang, H., Liu, H., & Wang, J. (2021). Multi-label learning for the diagnosis of cancer and identification of novel biomarkers with High-throughput Omics. *Current Bioinformatics*, *16*(2), 261–273.

Liu, Shiwei, Ni'mah, I., Menkovski, V., Mocanu, D. C., & Pechenizkiy, M. (2021). Efficient and effective training of sparse recurrent neural networks. *Neural Computing & Applications*, *33*(15), 9625–9636.

LiveBench. (n.d.). Retrieved April 6, 2025, from https://livebench.ai/

Llama AI. (n.d.). Llamaapi. Retrieved February 24, 2025, from https://www.llama-api.com/

Lokker, C., Abdelkader, W., Bagheri, E., Parrish, R., Cotoi, C., Navarro, T., Germini, F., Linkins, L.-A., Haynes, R. B., Chu, L., Afzal, M., & Iorio, A. (2024). Boosting efficiency in a clinical literature surveillance system with LightGBM. *PLOS Digital Health*, *3*(9), e0000299.

Lokker, C., Bagheri, E., Abdelkader, W., Parrish, R., Afzal, M., Navarro, T., Cotoi, C., Germini, F., Linkins, L., Haynes, R. B., Chu, L., & Iorio, A. (2023). Deep learning to refine the identification of high-quality clinical research articles from the biomedical literature: Performance evaluation. *Journal of Biomedical Informatics*, *142*(104384), 104384.

Lokker, C., McKibbon, K. A., Afzal, M., Navarro, T., Linkins, L.-A., Haynes, R. B., & Iorio, A. (2024). The McMaster Health Information Research Unit: Over a quarter-century of health informatics supporting evidence-based medicine. *Journal of Medical Internet Research*, *26*, e58764.

López, S., & Saboya, M. (2009). On the relationship between Shapley and Owen values. *Central European Journal of Operations Research*, *17*(4), 415–423.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/bib/bbac409

Marcus, E., & Teuwen, J. (2024). Artificial intelligence and explanation: How, why, and when to explain black boxes. *European Journal of Radiology*, *173*(111393), 111393.

Markey, N., Howitt, B., El-Mansouri, I., Schwartzenberg, C., Kotova, O., & Meier, C. (2024). Clinical trials are becoming more complex: a machine learning analysis of data from over 16,000 trials. *Scientific Reports*, *14*(1), 3514.

Markov, I. L., Liu, J., & Vagner, A. (2021). Regular expressions for fast-response COVID-19 text classification. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2102.09507

Marshall, I. J., Kuiper, J., Banner, E., & Wallace, B. C. (2017-7). Automating Biomedical Evidence Synthesis: RobotReviewer. *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, *2017*, 7–12.

Marshall, I. J., Kuiper, J., & Wallace, B. C. (2015). Automating risk of bias assessment for clinical trials. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1406–1412.

Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association: JAMIA*, *23*(1), 193–201.

MCMASTER+. (n.d.). Retrieved August 13, 2024, from https://plus.mcmaster.ca/McMasterPLUSDB/

MEDLINE 2022 initiative: Transition to automated indexing. (2021). https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html

MEDLINE PubMed production statistics. (2018). https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

Melamed, O., & Caruana, R. (2023). Explaining high-dimensional text classifiers. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2311.13454

Mendlovic, J., Mimouni, F. B., Arad, I., & Heiman, E. (2022). Trends in health quality-related publications over the past three decades: Systematic review. *Interactive Journal of Medical Research*, *11*(2), e31055.

Mersha, M. A., Yigezu, M. G., & Kalita, J. (2025). Evaluating the effectiveness of XAI techniques for encoder-based language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2501.15374

Mersha, M., Lam, K., Wood, J., AlShami, A., & Kalita, J. (2024). Explainable Artificial Intelligence: A survey of needs, techniques, applications, and future direction. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2409.00265

Mete, U., & Özmen, Ö. A. (2024). Assessing the accuracy and reproducibility of ChatGPT for responding to patient inquiries about otosclerosis. *European*

Archives of Oto-Rhino-Laryngology: Official Journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): Affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery. https://doi.org/10.1007/s00405-024-09039-4

Methodological Criteria. (n.d.). Health Information Research Unit. Retrieved August 19, 2024, from https://hiruweb.mcmaster.ca/hkr/what-we-do/methodologic-criteria/

Michael, L. G., IV, Donohue, J., Davis, J. C., Lee, D., & Servant, F. (2023). Regexes are hard: Decision-making, difficulties, and risks in programming regular expressions. In *arXiv [cs.SE]*. arXiv. http://arxiv.org/abs/2303.02555

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information (Basel)*, *15*(9), 517.

Miglani, V., Kokhlikyan, N., Alsallakh, B., Martin, M., & Reblitz-Richardson, O. (2020). Investigating saturation effects in Integrated Gradients. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2010.12697

Millard, L. A. C., Flach, P. A., & Higgins, J. P. T. (2016-2). Machine learning to assist risk-of-bias assessments in systematic reviews. *International Journal of Epidemiology*, *45*(1), 266–277.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: A comprehensive review. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.03705

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A survey. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2402.06196

Moe, S. M., Zidehsarai, M. P., Chambers, M. A., Jackman, L. A., Radcliffe, J. S., Trevino, L. L., Donahue, S. E., & Asplin, J. R. (2011). Vegetarian compared with meat dietary protein source and phosphorus homeostasis in chronic kidney disease. *Clinical Journal of the American Society of Nephrology: CJASN*, *6*(2), 257–264.

Molnar, C. (2024, May 26). *9.4 Scoped Rules (Anchors)*. https://christophm.github.io/interpretable-ml-book/anchors.html

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). SHAP-Based Explanation Methods: A Review for NLP Interpretability. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4593–4603). International Committee on Computational Linguistics.

MTIX: The next-generation algorithm for automated indexing of MEDLINE. (2024). https://www.nlm.nih.gov/pubs/techbull/ma24/ma24_mtix.html

Muliono, Y., & Tanzil, F. (2018). A comparison of text classification methods k-NN, naïve Bayes, and support vector machine for news classification. *Jurnal Informatika: Jurnal Pengembangan IT*, *3*(2), 157–160.

Nassif, A. B., Darya, A. M., & Elnagar, A. (2021). Empirical evaluation of shallow and deep learning classifiers for Arabic sentiment analysis. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2112.00534

National Academies of Sciences, Engineering, Medicine, Policy, Global Affairs, Committee on Science, Engineering, Medicine, Public Policy, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Theoretical Statistics, Board on Mathematical Sciences, Analytics, Division on Earth, Life Studies, Nuclear, Radiation Studies Board, Division of Behavioral, … Replicability. (2019). *Understanding Reproducibility and Replicability*. National Academies Press.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2307.06435

Nguyen, D. (2018). Comparing automatic and human evaluation of local explanations for text classification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana. https://doi.org/10.18653/v1/n18-1097

Nielsen, D. S., Enevoldsen, K., & Schneider-Kamp, P. (2024). Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2406.13469

Noh, S.-H. (2021). Analysis of gradient vanishing of RNNs and performance comparison. *Information (Basel)*, *12*(11), 442.

Ntrougkas, M. V., Mezaris, V., & Patras, I. (2025). P-TAME: Explain any image classifier with trained perturbations. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2501.17813

Number of clinical trials by year, country, WHO region and income group (1999-2022). (n.d.). Retrieved September 16, 2024, from https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-clinical-trials-by-year-country-who-region-and-income-group

Oami, T., Okada, Y., & Nakada, T.-A. (2024). Performance of a large language model in screening citations. *JAMA Network Open*, *7*(7), e2420496.

Oleynik, M., Kugic, A., Kasáč, Z., & Kreuzthaler, M. (2019). Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association: JAMIA*, *26*(11), 1247–1254.

Open LLM Leaderboard - a Hugging Face Space by open-llm-leaderboard. (n.d.). Retrieved April 2, 2025, from https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

OpenAI platform. (n.d.). Retrieved April 7, 2025, from https://platform.openai.com/docs/guides/reasoning/advice-on-prompting?api-mode=responses

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). On the difficulty of training Recurrent Neural Networks. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1211.5063

Pasupa, K., & Sunhem, W. (2016, October). A comparison between shallow and deep architecture classifiers on small dataset. *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia. https://doi.org/10.1109/iciteed.2016.7863293

Pitre, T., Jassal, T., Talukdar, J. R., Shahab, M., Ling, M., & Zeraatkar, D. (2023). ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study. In *bioRxiv*. https://doi.org/10.1101/2023.11.19.23298727

Precios. (n.d.). Retrieved April 4, 2025, from https://openai.com/api/pricing/

Price, W. N. (2018). Big data and black-box medical algorithms. *Science Translational Medicine*, *10*(471). https://doi.org/10.1126/scitranslmed.aao5333

Pu, Q., Xi, Z., Yin, S., Zhao, Z., & Zhao, L. (2024). Advantages of transformer and its application for medical image segmentation: a survey. *Biomedical Engineering Online*, *23*(1), 14.

Qin, X., Liu, J., Wang, Y., Liu, Y., Deng, K., Ma, Y., Zou, K., Li, L., & Sun, X. (2021). Natural language processing was effective in assisting rapid title and abstract

screening when updating systematic reviews. *Journal of Clinical Epidemiology*, *133*, 121–129.

Ravi, V., Gu, Y., Gandhe, A., Rastrow, A., Liu, L., Filimonov, D., Novotney, S., & Bulyko, I. (2020). Improving accuracy of rare words for RNN-Transducer through unigram shallow fusion. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2012.00133

Rayyan. (n.d.). *Rayyan: AI-powered systematic review management platform*. Retrieved March 2, 2025, from https://www.rayyan.ai/

Rehana, H., Çam, N. B., Basmaci, M., Zheng, J., Jemiyo, C., He, Y., Özgür, A., & Hur, J. (2023). *Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text*. https://doi.org/10.48550/arXiv.2303.17728

Ribeiro, M., Malcorra, B., Mota, N. B., Wilkens, R., Villavicencio, A., Hubner, L. C., & Rennó-Costa, C. (2024). A methodology for explainable large language models with Integrated Gradients and linguistic analysis in text classification. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2410.00250

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1602.04938

Roberts, J. (2023). How powerful are decoder-only transformer neural models? In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.17026

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Sahiner, B., Chen, W., Samala, R. K., & Petrick, N. (2023). Data drift in medical machine learning: implications and potential remedies. *The British Journal of Radiology*, *96*(1150), 20220878.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2402.07927

Salih, A. M., Galazzo, I. B., Raisi-Estabragh, Z., Petersen, S. E., Menegaz, G., & Radeva, P. (2024). Characterizing the contribution of dependent features in XAI methods. *IEEE Journal of Biomedical and Health Informatics*, *28*(11), 6466–6473.

Samek, W., Binder, A., Gregoire Mon- Tavon, S., & Lapuschkin, K.-R. (2016). Evaluating the visualization of what ¨ a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*, 2660–2673.

Santos, Á. O. D., da Silva, E. S., Couto, L. M., Reis, G. V. L., & Belo, V. S. (2023). The use of artificial intelligence for automating or semi-automating biomedical literature analyses: A scoping review. *Journal of Biomedical Informatics*, *142*(104389), 104389.

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *Sn Computer Science*, *2*(6), 420.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2024). The prompt report:

A systematic survey of prompt engineering techniques. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2406.06608

Scott, S. (1999). *Feature Engineering for Text Classification." MACHINE LEARNING-INTERNATIONAL WORKSHOP …, Citeseer*.

Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons, and Fractals*, *140*(110212), 110212.

Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An independent evaluation of ChatGPT on mathematical word problems (MWP). In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2302.13814

Shanaa, A. (2021, November 8). *Rayyan – intelligent systematic review*. Rayyan; Rayyan Systems. https://www.rayyan.ai/

shap.LinearExplainer — SHAP latest documentation. (n.d.). Retrieved March 4, 2025, from https://shap.readthedocs.io/en/latest/generated/shap.LinearExplainer.html

shap.PartitionExplainer — SHAP latest documentation. (n.d.). Retrieved December 24, 2024, from https://shap.readthedocs.io/en/latest/generated/shap.PartitionExplainer.html

Shekelle, P. G., Shetty, K., Newberry, S., Maglione, M., & Motala, A. (2017). Machine Learning Versus Standard Techniques for Updating Searches for Systematic Reviews: A Diagnostic Accuracy Study. *Annals of Internal Medicine*, *167*(3), 213–215.

Shin, T., Razeghi, Y., Logan, R. L., IV, Wallace, E., & Singh, S. (2020). AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2010.15980

Sikdar, S., Bhattacharya, P., & Heese, K. (2021). Integrated directional gradients: Feature interaction attribution for neural NLP models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 865–878). Association for Computational Linguistics.

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2023). *An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing*. https://doi.org/10.48550/arXiv.2309.08008

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1706.03825

Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2020). Fast WordPiece Tokenization. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2012.15524

Song, Z., Yan, B., Liu, Y., Fang, M., Li, M., Yan, R., & Chen, X. (2025). Injecting domain-specific knowledge into large Language Models: A comprehensive survey. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2502.10708

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, *9*(26), eadh1850.

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials [Review of *RoB 2: a revised tool for assessing risk of bias in randomised trials*]. *BMJ* , *366*, l4898.

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ (Clinical Research Ed.)*, *355*, i4919.

Stryker, C., & Holdsworth, J. (2025, January 24). *What is NLP (natural language processing)?* https://www.ibm.com/think/topics/natural-language-processing

Sun, J., Mavrogenis, A. F., & Scarlat, M. M. (2021). The growth of scientific publications in 2020: a bibliometric analysis based on the number of publications, keywords, and citations in orthopaedic surgery. *International Orthopaedics*, *45*(8), 1905–1910.

Sun, Q., Akman, A., & Schuller, B. W. (2024). Explainable artificial intelligence for medical applications: A review. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2412.01829

Sunagar, P., & Kanavalli, A. (2022). A hybrid RNN based deep learning approach for text classification. *International Journal of Advanced Computer Science and Applications : IJACSA*, *13*(6). https://doi.org/10.14569/ijacsa.2022.0130636

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1703.01365

Taha, K., Yoo, P. D., Yeun, C., & Taha, A. (2024). Text classification: A review, empirical, and experimental evaluation. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2401.12982

Talebi, S., Tong, E., Li, A., Yamin, G., Zaharchuk, G., & Mofrad, M. R. K. (2024). Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment. *BMC Medical Informatics and Decision Making*, *24*(1), 40.

Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., Weng, C., & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00896-7

Tenny, S., & Varacallo, M. (2022). *Evidence Based Medicine*. StatPearls Publishing.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2007.05558

Thushari, P. D., Niazi, S., & Meena, S. (2023). Transfer learning approach to multilabel biomedical literature classification using transformer models. *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, 1–6.

van den Bulk, L. M., Bouzembrak, Y., Gavai, A., Liu, N., van den Heuvel, L. J., & Marvin, H. J. P. (2022). Automatic classification of literature in systematic reviews on food safety using machine learning. *Current Research in Food Science*, *5*, 84–95.

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. S. (2023). Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*. https://doi.org/10.21203/rs.3.rs-3483777/v1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *arXiv [cs.CL]*. https://doi.org/10.48550/ARXIV.1706.03762

Vatsal, S., & Dubey, H. (2024). A survey of prompt engineering methods in large language models for different NLP tasks. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2407.12994

Velayuthan, M., & Sarveswaran, K. (2024). Egalitarian language representation in language models: It all begins with tokenizers. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2409.11501

Verma, S., Sharan, A., & Malik, N. (2024). Efficient classification of hallmark of cancer using embedding-based support vector machine for multilabel text. *New Generation Computing*. https://doi.org/10.1007/s00354-024-00248-3

Wadden, J. J. (2021). Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics*, *48*(10), 764–768.

Waghela, H., Sen, J., & Rakshit, S. (2024). Saliency Attention and Semantic Similarity-driven adversarial perturbation. In *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/2406.19413

Wan, Z. (2023). Text classification: A perspective of deep learning methods. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2309.13761

Wang, G., Sun, Z., Gong, Z., Ye, S., Chen, Y., Zhao, Y., Liang, Q., & Hao, D. (2024). Do advanced Language Models eliminate the need for prompt engineering in software engineering? In *arXiv [cs.SE]*. arXiv. http://arxiv.org/abs/2411.02093

Wang, X., Li, C., Wang, Z., Bai, F., Luo, H., Zhang, J., Jojic, N., Xing, E. P., & Hu, Z. (2023). PromptAgent: Strategic planning with language models enables expert-level prompt optimization. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2310.16427

Wang, Y., Zhang, T., Guo, X., & Shen, Z. (2024). Gradient based feature attribution in explainable AI: A technical review. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2403.10415

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2201.11903

Wilczynski, N. L., Morgan, D., & Haynes, R. B. (2005). An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Medical Informatics and Decision Making*, *5*(1), 20.

Wilczynski, N. L., Walker, C. J., McKibbon, K. A., & Haynes, R. B. (1993). Assessment of methodologic search filters in MEDLINE. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 601–605.

Wu, X., Xiang, B., Lu, H., Li, C., Huang, X., & Huang, W. (2024). Optimizing recurrent Neural Networks: A study on gradient normalization of weights for enhanced training efficiency. *Applied Sciences (Basel, Switzerland)*, *14*(15), 6578.

Xu, Y., Zhou, Y., Sekula, P., & Ding, L. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, *6*(100045), 100045.

Yan, Y., Lu, Y., Xu, R., & Lan, Z. (2025). Do PhD-level LLMs truly grasp elementary addition? Probing rule learning vs. Memorization in Large Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2504.05262

Yanagita, Y., Yokokawa, D., Uchida, S., Tawara, J., & Ikusaka, M. (2023). Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: Evaluation study. *JMIR Formative Research*, *7*, e48023.

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *An International Journal on Information Fusion*, *77*, 29–52.

Yang, J. (2021). Fast TreeSHAP: Accelerating SHAP value computation for trees. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2109.09847

Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J., & Tang, J. (2023). GPT can solve mathematical problems without a calculator. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2309.03241

Ye, X., Wang, Z., & Wang, J. (2025). Infinite retrieval: Attention enhanced LLMs in long-context processing. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2502.12962

Yenduri, G., M, R., G, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., G, D. R., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). Generative Pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.10435

Yin, X.-C., Yang, C., Pei, W.-Y., & Hao, H.-W. (2014). Shallow classification or deep learning: An experimental study. *2014 22nd International Conference on Pattern Recognition*, 1904–1909.

Yuan, Z., Yuan, H., Tan, C., Wang, W., & Huang, S. (2023). How well do Large Language Models perform in Arithmetic tasks? In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2304.02015

Zaidan, O., & Eisner, J. (2008). Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing* (pp. 31–40).

Zeng, X. (2024). Enhancing the interpretability of SHAP values using Large Language Models. In *arXiv [cs.HC]*. arXiv. http://arxiv.org/abs/2409.00079

Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, *11*(1), 25.

Zhang, J., Wibert, M., Zhou, H., Peng, X., Chen, Q., Keloth, V. K., Hu, Y., Zhang, R., Xu, H., & Raja, K. (2024). A study of biomedical relation Extraction using GPT models. *AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science*, *2024*, 391–400.

Zhao, X., Jiang, H., Yin, J., Liu, H., Zhu, R., Mei, S., & Zhu, C.-T. (2022). Changing trends in clinical research literature on PubMed database from 1991 to 2020. *European Journal of Medical Research*, *27*(1), 95.

Zhou, F., Afzal, M., Parrish, R., Saha, A., Abdelkader, W., Brian Haynes, R., Iorio, A., & Lokkera, C. (2025). *Domain-specific Pretrained Encoder Transformers for the Identification of Methodologically Rigorous Systematic Reviews*. HEI Research Day 2025, Hamilton, Ontario, Canada.

Zhou, F., Lokker, C., Parrish, R., Brian Haynes, R., Iorio, A., Saha, A., & Afzal, M. (2025). *Fine-tuning and Benchmarking of Transformer Models for Multiclass Classification of Clinical Research Articles*.

Zhou, F., Parrish, R., Afzal, M., Saha, A., Haynes, R. B., Iorio, A., & Lokker, C. (2025). Benchmarking Domain-specific Pretrained Language Models to Identify the Best Model for Methodological Rigor in Clinical Studies. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2025.104825

Zhou, Y. (2020, April 15). A review of text classification based on deep learning. *Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis*. ICGDA 2020: 2020 3rd International Conference on Geoinformatics and Data Analysis, Marseille France. https://doi.org/10.1145/3397056.3397082

Zimmerman, J. W., Hudon, D., Cramer, K., Ruiz, A. J., Beauregard, C., Fehr, A., Fudolig, M. I., Demarest, B., Bird, Y. M., Trujillo, M. Z., Danforth, C. M., & Dodds, P. S. (2024). Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2412.10924

Zytek, A., Pido, S., Alnegheimish, S., Berti-Equille, L., & Veeramachaneni, K. (2024). Explingo: Explaining AI Predictions using Large Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2412.05145

Zytek, A., Pidò, S., & Veeramachaneni, K. (2024). LLMs for XAI: Future Directions for Explaining Explanations. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2405.06064

(N.d.-a). Retrieved April 2, 2025, from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

(N.d.-b). Retrieved March 12, 2025, from https://platform.openai.com/docs/api-reference/backward-compatibility

(N.d.-c). Retrieved February 24, 2025, from https://openai.com/index/openai-api/

(N.d.-d). Retrieved February 25, 2025, from https://platform.openai.com/docs/guides/function-calling

(N.d.-e). Retrieved February 25, 2025, from https://platform.openai.com/docs/guides/structured-outputs

(N.d.-f). Retrieved March 2, 2025, from https://platform.openai.com/docs/guides/prompt-engineering

APPENDIX A HIRU CRITERIA AND DATA

AA1. Rigour Criteria For Original Articles on Treatment, Primary Prevention, and Quality Improvement

Criterion	Description
1	In English.
2	About humans.
3	About topics that are important to the clinical practice of medicine, nursing, rehabilitation, and other health professions, other than descriptive studies of prevalence.
4	Analysis is consistent with the study question.
5	Random allocation of participants to study arms.
6	≥10 patients/participants per group completing the primary outcome assessment.
7	Primary outcome(s) assessed in ⩾80% of those randomized at the defined follow-up point.
8	At least one clinically important outcome measure.
9	If reporting subgroup analysis, it is preplanned, with groups analyzed as they were randomized and interaction between two or more subgroups reported.
Final	All 9 criteria met.

AA2. Dataset Characteristics

Dataset Name	Publication Year of the Articles	No. of Articles	Rigorous Articles (%)	Non-Rigorous Articles (%)
PLUS-2003-2023	2003 to 2023	42,575	25,561 (60.0)	17,014 (40.0)
	2003 to 2023	5,322	3,203 (60.2)	2,119 (39.8)
	2003 to 2023	5,322	3,164 (59.5)	2,158 (40.5)
PLUS-2024	2024	1,011	575 (56.9)	436 (43.1)
Clinical Hedges	2000	6,572	1,587 (24.1)	4,985 (75.9)

APPENDIX B GPT CLASSIFIER PROMPT

AB1. Prompt Dictionary

Key	Content
ROLE	You are an expert clinical researcher.
CRITERIA	Manual appraisal criteria: All of these criteria must be met to be rated as being rigorous. The article will be rated as not rigorous if any criteria are not met. The criteria are 1) in English, 2) about humans, 3) about topics that are important to the clinical practice of medicine, nursing, rehabilitation, and other health professions, other than descriptive studies of prevalence, 4) analysis of each article consistent with the study question, 5) random allocation of participants to comparison groups, 6) 10 or more patients per group completing primary outcome assessment, 7) primary outcome(s) assessed in 80% or more of those randomized at the defined follow-up point, 8) primary outcome is clinically important or ≥1 secondary outcome is clinically important, and 9) subgroup analyses must be preplanned, with groups analyzed as they were randomized; analyses must test for interaction between 2 or more subgroups.
TASK_TIAB	You are assessing whether a randomized controlled trial (RCT) meets the criteria for being considered rigorous. You will be provided the abstract of the RCT. Based on this information and the provided manual appraisal criteria, please do the following: - Reason through whether each criterion is met, unmet, or cannot be determined based on available information. - Reason through whether this RCT, based on the abstract, should be considered rigorous (True) or not (False).
TASK_TIAB_BERT	You are trying to explain the rigour classification decisions made by encoder-only transformers on a randomized controlled trial (RCT). You will be provided the input text into the model, and the model's predicted probability of a positive (rigorous)

	classification, and the rigour classification based on the default threshold of ≥0.5. Based on this information and the provided manual appraisal criteria, please do the following: - Reason through whether each criterion is met, unmet, or cannot be determined based on available information. - Reason through whether you agree with the model's prediction.
TASK_FT	You are assessing whether a randomized controlled trial (RCT) meets the criteria for being considered rigorous. You will be provided the full text of the RCT. Based on this information and the provided manual appraisal criteria, please do the following: - Reason through whether each criterion is met (True) or unmet (False) based on available information. - Reason through whether this RCT, based on the full text, should be considered rigorous (True) or not (False).
INSTANCE_TIAB_BERT	Input text: <RCT_title_abstract>. Predicted positive class (rigorous) probability: <BioLinkBERT_probability>. Classification: <Rigorous/Not rigorous>.
TASK_FT_BERT	You are trying to explain the rigour classification decisions made by encoder-only transformers on a randomized controlled trial (RCT). You will be provided the input text into the model, the full text of the article, the model's predicted probability of a positive (rigorous) classification, and the rigour classification based on the default threshold of ≥0.5. Based on this information and the provided manual appraisal criteria, please do the following: - Reason through whether each criterion is met (True) or unmet (False) based on available information. - Reason through whether you agree with the model's prediction.
INSTANCE_TIAB	RCT Abstract: <RCT_abstract>.
INSTANCE_FT	RCT full text: <RCT_full_text>.
INSTANCE_FT_BERT	Input text: <RCT_title_abstract>. RCT full text: <RCT_full_text>. Predicted positive class (rigorous) probability: <BioLinkBERT_probability>. Classification: <Rigorous/Not rigorous>.

AB2. Object Dictionary

Object Key	Content and Type
Assessment	Enum ("met", "unmet", "cannot_be_determined")
RIGOUR_ASSESSMENT_TIAB	criterion_1_justification: str criterion_1_assessment: Assessment criterion_2_justification: str criterion_2_assessment: Assessment criterion_3_justification: str criterion_3_assessment: Assessment criterion_4_justification: str criterion_4_assessment: Assessment criterion_5_justification: str criterion_5_assessment: Assessment criterion_6_justification: str criterion_6_assessment: Assessment criterion_7_justification: str criterion_7_assessment: Assessment criterion_8_justification: str criterion_8_assessment: Assessment criterion_9_justification: str criterion_9_assessment: Assessment final_justification: str final_assessment: bool
PREDICTION_EXPLAINATION_TIAB	criterion_1_justification: str criterion_1_assessment: Assessment criterion_2_justification: str criterion_2_assessment: Assessment criterion_3_justification: str criterion_3_assessment: Assessment criterion_4_justification: str criterion_4_assessment: Assessment criterion_5_justification: str criterion_5_assessment: Assessment criterion_6_justification: str criterion_6_assessment: Assessment criterion_7_justification: str criterion_7_assessment: Assessment criterion_8_justification: str criterion_8_assessment: Assessment criterion_9_justification: str criterion_9_assessment: Assessment agreement_justification: str

	agreement: bool
RIGOUR_ASSESSMENT_FT	criterion_1_justification: str criterion_1_assessment: bool criterion_2_justification: str criterion_2_assessment: bool criterion_3_justification: str criterion_3_assessment: bool criterion_4_justification: str criterion_4_assessment: bool criterion_5_justification: str criterion_5_assessment: bool criterion_6_justification: str criterion_6_assessment: bool criterion_7_justification: str criterion_7_assessment: bool criterion_8_justification: str criterion_8_assessment: bool criterion_9_justification: str criterion_9_assessment: bool final_justification: str final_assessment: bool
PREDICTION_EXPLAINATION_FT	criterion_1_justification: str criterion_1_assessment: bool criterion_2_justification: str criterion_2_assessment: bool criterion_3_justification: str criterion_3_assessment: bool criterion_4_justification: str criterion_4_assessment: bool criterion_5_justification: str criterion_5_assessment: bool criterion_6_justification: str criterion_6_assessment: bool criterion_7_justification: str criterion_7_assessment: bool criterion_8_justification: str criterion_8_assessment: bool criterion_9_justification: str criterion_9_assessment: bool agreement_justification: str agreement: bool

AB3. Prompt Chains

Sequence	Role	Key
Classification with TIAB		
1	developer/user†	{ROLE} {TASK_TIAB} {CRITERIA}
2	user	{INSTANCE_TIAB}
3	assistant	{RIGOUR_ASSESSMENT_TIAB}
Explanation with TIAB		
1	developer/user†	{ROLE} {TASK_TIAB_BERT} {CRITERIA}
2	user	{INSTANCE_TIAB_BERT}
3	assistant	{PREDICTION_EXPLAINATION_TIAB}
Classification with Full Text		
1	developer/user†	{ROLE} {TASK_FT} {CRITERIA}
2	user	{INSTANCE_FT}
3	assistant	{RIGOUR_ASSESSMENT_FT}
Explanation with Full Text		
1	developer/user†	{ROLE} {TASK_FT_BERT} {CRITERIA}
2	user	{INSTANCE_FT_BERT}
3	assistant	{PREDICTION_EXPLAINATION_FT}

†A developer prompt is used for GPT-4o. However, GPT-o3-mini does not support developer prompts, and a user prompt is used instead.

APPENDIX C SOFTWARE ENVIRONMENTS

Library	Local Version	Cloud Version
#paycheck	1.0.2	N/A
#torch	2.2.1	N/A
Cython	0.29.36	0.29.36+computecanada
GitPython	3.1.40	3.1.40+computecanada
Jinja2	3.1.2	3.1.2+computecanada
MarkupSafe	2.1.3	2.1.3+computecanada
Pillow	10.0.0	10.0.0+computecanada
PyNaCl	1.5.0	1.5.0+computecanada
PyYAML	6.0.1	6.0.1+computecanada
Pygments	2.16.1	2.16.1+computecanada
Send2Trash	1.8.2	1.8.2+computecanada
accelerate	0.27.2	0.27.2+computecanada
aiohttp	3.9.1	3.9.1+computecanada
aiosignal	1.3.1	1.3.1+computecanada
anyio	3.7.1	3.7.1+computecanada
appdirs	1.4.4	1.4.4+computecanada
arff	0.9	0.9+computecanada
argon2_cffi	23.1.0	23.1.0+computecanada
argon2_cffi_bindings	21.2.0	21.2.0+computecanada
asttokens	2.2.1	2.2.1+computecanada
async_generator	1.1	1.10+computecanada
attrs	23.1.0	23.1.0+computecanada
backcall	0.2.0	0.2.0+computecanada
backports-abc	0.5	0.5+computecanada
backports.shutil_get_terminal_size	1.0.0	1.0.0+computecanada
bcrypt	4.0.1	4.0.1+computecanada
beautifulsoup4	4.12.2	4.12.2+computecanada
bitarray	2.8.1	2.8.1+computecanada
bitstring	4.1.1	4.1.1+computecanada
bleach	6.0.0	6.0.0+computecanada
captum	N/A	0.3.0+computecanada
certifi	2023.7.22	2023.7.22+computecanada
cffi	1.15.1	1.15.1+computecanada
chardet	5.2.0	5.2.0+computecanada
charset_normalizer	3.2.0	3.2.0+computecanada
click	8.1.7	8.1.7+computecanada
comm	0.1.4	0.1.4+computecanada
contourpy	1.1.0	1.1.0+computecanada
cryptography	39.0.1	39.0.1+computecanada

cycler	0.11.0	0.11.0+computecanada
datasets	2.18.0	2.18.0+computecanada
deap	1.4.1	1.4.1+computecanada
debugpy	1.6.7.post1	1.6.7.post1+computecanada
decorator	5.1.1	5.1.1+computecanada
defusedxml	0.7.1	0.7.1+computecanada
dill	0.3.8	0.3.8+computecanada
dnspython	2.4.2	2.4.2+computecanada
docker-pycreds	0.4.0	0.4.0+computecanada
ecdsa	0.18.0	0.18.0+computecanada
entrypoints	0.4	0.4+computecanada
evaluate	0.4.1	0.4.2+computecanada
executing	1.2.0	1.2.0+computecanada
fastjsonschema	2.18.0	2.18.0+computecanada
filelock	3.13.1	3.13.1+computecanada
fonttools	4.42.1	4.42.1+computecanada
frozenlist	1.4.1	1.4.1+computecanada
fsspec	2024.2.0	2024.2.0+computecanada
funcsigs	1.0.2	1.0.2+computecanada
gitdb	4.0.11	4.0.11+computecanada
huggingface_hub	0.21.4	0.21.4+computecanada
idna	3.4	3.4+computecanada
importlib_metadata	6.8.0	6.8.0+computecanada
importlib_resources	6.0.1	6.0.1+computecanada
ipykernel	6.25.1	6.25.1+computecanada
ipython	8.15.0	8.15.0+computecanada
ipython_genutils	0.2.0	0.2.0+computecanada
jedi	0.19.0	0.19.0+computecanada
joblib	1.3.2	1.3.2+computecanada
jsonschema	4.19.0	4.19.0+computecanada
jsonschema_specifications	2023.7.1	2023.7.1+computecanada
jupyter_client	8.3.1	8.3.1+computecanada
jupyter_core	5.3.1	5.3.1+computecanada
kiwisolver	1.4.5	1.4.5+computecanada
lockfile	0.12.2	0.12.2+computecanada
matplotlib	3.7.2	3.7.2+computecanada
matplotlib_inline	0.1.6	0.1.6+computecanada
mistune	3.0.1	3.0.1+computecanada
mock	5.1.0	5.1.0+computecanada
mpmath	1.3.0	1.3.0+computecanada
multidict	6.0.5	6.0.5+computecanada
multiprocess	0.70.16	0.70.16+computecanada

nest_asyncio	1.5.7	1.5.7+computecanada
netaddr	0.8.0	0.8.0+computecanada
netifaces	0.11.0	0.11.0+computecanada
networkx	3.2.1	3.2.1+computecanada
nose	1.3.7	1.3.7+computecanada
numpy	1.25.2	1.25.2+computecanada
packaging	23.1	23.1+computecanada
pandas	2.1.0	2.1.0+computecanada
pandocfilters	1.5.0	1.5.0+computecanada
paramiko	3.3.1	3.3.1+computecanada
parso	0.8.3	0.8.3+computecanada
path	16.7.1	16.7.1+computecanada
path.py	12.5.0	12.5.0+computecanada
pathlib2	2.3.7.post1	2.3.7.post1+computecanada
paycheck	N/A	1.0.2+computecanada
pbr	5.11.1	5.11.1+computecanada
pexpect	4.8.0	4.8.0+computecanada
pickleshare	0.7.5	0.7.5+computecanada
pkgutil_resolve_name	1.3.10	1.3.10+computecanada
platformdirs	3.9.1	3.9.1+computecanada
prometheus_client	0.17.1	0.17.1+computecanada
prompt_toolkit	3.0.39	3.0.39+computecanada
protobuf	4.25.2	4.25.2+computecanada
psutil	5.9.5	5.9.5+computecanada
ptyprocess	0.7.0	0.7.0+computecanada
pure_eval	0.2.2	0.2.2+computecanada
pyarrow	15.0.1	15.0.1
pyarrow_hotfix	0.6	0.6+computecanada
pycparser	2.21	2.21+computecanada
pyparsing	3.0.9	3.0.9+computecanada
pyrsistent	0.19.3	0.19.3+computecanada
python-dateutil	2.8.2	2.8.2+computecanada
python_json_logger	2.0.7	2.0.7+computecanada
pytz	2023.3	2023.3+computecanada
pyzmq	25.1.1	25.1.1+computecanada
referencing	0.30.2	0.30.2+computecanada
regex	2023.8.8	2023.8.8+computecanada
requests	2.31.0	2.31.0+computecanada
responses	0.18.0	0.18.0+computecanada
rfc3339_validator	0.1.4	0.1.4+computecanada
rfc3986_validator	0.1.1	0.1.1+computecanada
rpds_py	0.10.0	0.10.0+computecanada

safetensors	0.4.1	0.4.1+computecanada
scikit_learn	1.3.1	1.3.1+computecanada
scipy	1.11.2	1.11.2+computecanada
sentry_sdk	1.38.0	1.38.0+computecanada
setproctitle	1.3.2	1.3.2+computecanada
shap	N/A	0.43.0+computecanada
simplegeneric	0.8.1	0.8.1+computecanada
singledispatch	4.1.0	4.1.0+computecanada
six	1.16.0	1.16.0+computecanada
sklearn	0	0.0+computecanada
smmap	5.0.1	5.0.1+computecanada
sniffio	1.3.0	1.3.0+computecanada
soupsieve	2.4.1	2.4.1+computecanada
stack_data	0.6.2	0.6.2+computecanada
sympy	1.12	1.12+computecanada
terminado	0.17.1	0.17.1+computecanada
testpath	0.6.0	0.6.0+computecanada
threadpoolctl	3.3.0	3.3.0+computecanada
tinycss2	1.2.1	1.2.1+computecanada
tokenizers	0.15.0	0.15.0+computecanada
torch	N/A	2.2.1+computecanada
tornado	6.3.3	6.3.3+computecanada
tqdm	4.66.2	4.66.2+computecanada
traitlets	5.9.0	5.9.0+computecanada
transformers	4.38.1	4.38.1+computecanada
typing_extensions	4.10.0	4.10.0+computecanada
tzdata	2023.3	2023.3+computecanada
urllib3	2.0.4	2.0.4+computecanada
wandb	0.16.0	0.16.0+computecanada
wcwidth	0.2.6	0.2.6+computecanada
webencodings	0.5.1	0.5.1+computecanada
websocket_client	1.6.2	1.6.2+computecanada
xxhash	3.2.0	3.2.0+computecanada
yarl	1.9.3	1.9.3+computecanada
zipp	3.16.2	3.16.2+computecanada

APPENDIX D GPT CLASSIFIER EXAMPLE OUTPUTS

AD1. GPT-4o - Classifier with TIAB

PMID	Label	Predicted Rigorous Probability†	Criterion	Justification	Assessment
19489805	Non-rigorous	0.608	1	The abstract is written in English.	met
			2	The study involves human participants (mothers and infants).	met
			3	The study addresses an important clinical topic related to infant sleep position and SIDS prevention.	met
			4	The analysis appears consistent with the study question, as it evaluates the effect of the intervention on the outcome.	met
			5	The study mentions random allocation of participants to intervention and control groups.	met
			6	The study includes 228 mother-infant pairs, which suggests more than 10 participants per group completed the primary outcome assessment.	met
			7	The abstract does not provide sufficient information to determine if 80% or more of those randomized were assessed for the primary outcome.	cannot_be_determined
			8	The primary outcome (infant sleep position) is clinically important for SIDS prevention.	met
			9	The abstract does not mention subgroup analyses or their preplanning.	cannot_be_determined
			Final	While most criteria are met, some cannot be determined from the abstract alone, such as the	FALSE

				percentage of participants assessed and details on subgroup analyses.	
21183586	Non-rigorous	0.288	1	The abstract is written in English.	met
			2	The study involves human participants with chronic kidney disease.	met
			3	The study addresses dietary protein sources and phosphorus homeostasis, which is relevant to clinical practice in managing chronic kidney disease.	met
			4	The analysis appears consistent with the study question, focusing on the effects of dietary protein sources on phosphorus homeostasis.	met
			5	The abstract does not specify random allocation of participants to the dietary interventions.	cannot_be_determined
			6	The study included only nine participants, which is fewer than 10 per group.	unmet
			7	The abstract does not provide information on the percentage of participants completing the primary outcome assessment.	cannot_be_determined
			8	The primary outcome, phosphorus homeostasis, is clinically important for patients with chronic kidney disease.	met
			9	The abstract does not mention subgroup analyses.	cannot_be_determined
			Final	The study does not meet the criterion of having at least 10 participants per group, which is a requirement for rigor.	FALSE
16916871	1	0.417	1	The abstract is written in English.	met
			2	The study involves human participants attending primary care clinics.	met

			3	The study addresses the clinical issue of improving attendance in primary care, which is relevant to health professions.	met
			4	The analysis appears consistent with the study question, as it evaluates attendance rates and costs.	met
			5	The study mentions random allocation to three groups, indicating randomization.	met
			6	The abstract states that 993 participants were eligible for analysis, which suggests more than 10 participants per group.	met
			7	The abstract does not specify the percentage of participants completing the primary outcome assessment, so this cannot be determined.	cannot_be_determined
			8	The primary outcome, attendance rate, is clinically important for primary care.	met
			9	The abstract does not mention subgroup analyses, so this cannot be determined.	cannot_be_determined
			Final	While most criteria are met, the abstract does not provide sufficient information to determine if criteria 7 and 9 are met. Therefore, the study cannot be confirmed as rigorous based on the abstract alone.	FALSE
36395875	1	0.943	1	The abstract is written in English.	met
			2	The study is about human patients undergoing OPCAB.	met
			3	The study addresses a clinical question relevant to medical practice, specifically the management of clopidogrel resistance after OPCAB.	met

	4	The analysis appears consistent with the study question, as it evaluates outcomes like MACEs and graft patency.	met
	5	The abstract mentions random allocation of participants to the two groups using a 1:1 ratio block table.	met
	6	Each group had 102 patients completing the primary outcome assessment.	met
	7	The abstract does not specify the percentage of randomized patients assessed for the primary outcome, but it implies that all randomized patients were included in the analysis.	cannot_be_determined
	8	The primary outcomes, such as MACEs and graft patency, are clinically important.	met
	9	The abstract does not mention any subgroup analyses.	cannot_be_determined
	Final	While most criteria are met, the information provided does not allow determination of whether all randomized patients were assessed for the primary outcome or if subgroup analyses were preplanned and appropriately conducted.	FALSE

† Predicted by the BioLinkBERT model.

AD2. GPT-o3-mini - Classifier with TIAB

PMID	Label	Predicted Rigorous Probability†	Criterion	Justification	Assessment
19489805	Non-rigorous	0.608	1	The abstract is written in English.	met
			2	The study involves human participants (motherâ€“infant pairs).	met
			3	The topic (infant sleep position related to SIDS prevention) is clinically important in public health and pediatric care, and it is an interventional study, not merely a descriptive prevalence study.	met
			4	The analysis (including logistic regression and bivariate analyses) is consistent with evaluating the effect of the educational intervention, aligning with the study question.	met
			5	The study is described as a randomized clinical trial, implying that participants were randomly allocated to the intervention or control group.	met
			6	With a total sample of 228 motherâ€“infant pairs, each group likely included well over 10 participants completing the outcome assessment.	met
			7	The abstract states that the primary outcome was assessed during a home visit at 3 months, but it does not provide information on the follow-up rate (i.e. whether 80% or more of those randomized completed the assessment).	cannot_be_determined
			8	The primary outcome (infant sleep position) is clinically relevant as it relates to SIDS	met

				prevention, which is an important clinical outcome.	
			9	The abstract does not report any subgroup analyses or discuss whether any such analyses were preplanned. Without this information, it is not possible to determine if this criterion is met.	cannot_be_determined
			Final	While the study appears to meet most of the criteria, the abstract does not provide sufficient information regarding the proportion of participants completing the primary outcome follow-up (criterion 7) and details concerning preplanned subgroup analyses (criterion 9). As per the appraisal manual, if any criterion is not clearly met, the overall study is rated as not rigorous.	FALSE
21183586	Non-rigorous	0.288	1	The abstract is written in English.	met
			2	The study involves patients with chronic kidney disease, so it is clearly about humans.	met
			3	The topic on dietary protein sources and phosphorus homeostasis is clinically relevant for managing CKD, which is important in clinical practice.	met
			4	The analysis in the abstract directly addresses the research question by comparing the effects of vegetarian versus meat diets on phosphorus homeostasis related outcomes.	met
			5	Although the study is described as a crossover trial, the abstract does not explicitly state that the order of interventions was randomized, making it unclear if random allocation was properly implemented.	cannot_be_determined

			6	The trial enrolled nine patients in total. In a crossover design, each patient serves as his/her own control; however, the criteria require 10 or more patients per group completing the primary outcome assessment. This count is not met.	unmet
			7	The studyâ€™s design included an inpatient phase with close monitoring of outcomes, and there is no indication of missing data, suggesting that primary outcomes were likely assessed in 80% or more of participants.	met
			8	The primary outcomes (e.g., serum phosphorus, FGF23 levels) are clinically important in the management of CKD and have direct implications for patient care.	met
			9	No subgroup analyses were reported. In the absence of subgroup analyses, this criterion is considered met by default.	met
			Final	While most criteria are met, the study fails to meet the criterion regarding sample size (criterion 6), and the description of random allocation (criterion 5) is ambiguous. Given that all criteria need to be fulfilled for a study to be considered rigorous, the failure to meet the sample size requirement is sufficient to rate this study as not rigorous.	FALSE
16916871	1	0.417	1	The abstract is written in English.	met
			2	The study involves primary care patients (or their caregivers) and is clearly about human subjects.	met
			3	The topic (improving attendance in primary care) is of clear importance in the clinical	met

		practice of medicine and related health fields, rather than a mere descriptive prevalence study.	
	4	The analysis directly addresses the study question by comparing attendance rates between intervention and control groups using appropriate outcome measures.	met
	5	The study is described as a multicentre threeâ€□arm randomized controlled trial, indicating random allocation of participants to comparison groups.	met
	6	With a total of 993 participants reported, it is highly likely that each group includes well over 10 patients completing the primary outcome assessment.	met
	7	The abstract reports the number of participants eligible for analysis but does not provide information on how many were randomized originally or the exact followâ€□up rate, so it cannot be determined if 80% or more of those randomized had their primary outcome assessed.	cannot_be_determined
	8	The primary outcome, attendance rate, has clinical relevance in primary care, and the study also includes a cost analysis which supports the clinical importance of the findings.	met
	9	The abstract does not mention any subgroup analyses or whether any planned subgroup analyses were performed. Thus, it cannot be determined whether subgroup analyses were preplanned and analyzed appropriately.	cannot_be_determined
	Final	While most of the key criteria are clearly met, the abstract does not provide sufficient	FALSE

				information to determine whether the primary outcome followâ€□up rate meets the 80% threshold (criterion 7) and whether any subgroup analyses, if performed, were preplanned and correctly analyzed (criterion 9). Since all criteria must be met for the study to be considered rigorous, these uncertainties lead to a final assessment of not rigorous.	
36395875	1	0.943	1	The abstract is written in English.	met
			2	The study involves human patients undergoing offâ€□pump coronary artery bypass surgery.	met
			3	The study investigates clinically relevant outcomes (graft patency and major adverse cardiovascular events) important to practice.	met
			4	The analysis directly addresses the study question by comparing ticagrelor and clopidogrel regarding clinically important outcomes using intentâ€□toâ€□treat analysis.	met
			5	Participants were randomly allocated to treatment groups using a block randomization table.	met
			6	Each group had 102 patients, which is well above the minimum of 10 per group for the primary outcome assessment.	met
			7	The abstract does not provide information on the follow-up rate or the percentage of patients who completed the primary outcome assessment, making it unclear if the 80% criterion was met.	cannot_be_determined
			8	The primary outcome (graft patency and the composite MACEs) is clinically important, thus meeting this criterion.	met

	9	There is no information provided regarding whether subgroup analyses were preplanned or if interactions were formally tested. Without this detail, it is unclear if this criterion was satisfied.	cannot_be_determined
	Final	Although the abstract meets many of the rigorous RCT criteria, the lack of details regarding the follow-up rate (criterion 7) and the preplanning/reporting of subgroup analyses (criterion 9) prevent a full appraisal based solely on the abstract. As all criteria must be met for the study to be considered rigorous, the overall assessment is that the study does not definitively meet all criteria.	FALSE

† Predicted by the BioLinkBERT model.

AD3. GPT-o3-mini - Classifier with Full Text

PMID	Label	Predicted Rigorous Probability†	Criterion	Justification	Assessment
19489805	Non-rigorous	0.608	1	The article is published in English and meets the language requirement.	TRUE
			2	The study focuses on human participants (motherâ€□infant pairs) and evaluates an educational intervention relevant to human health.	TRUE
			3	The topic addresses the prevention of SIDS through infant sleep positioning, an important clinical issue for pediatric and maternal health, rather than merely describing prevalence.	TRUE
			4	The analysis directly addresses the study question with appropriate univariate and multivariate analyses, aligning with the interventionâ€™s effect on the primary outcome.	TRUE
			5	Participants were randomly allocated to the intervention and control groups by using coin tossing in blocks, satisfying the random allocation requirement.	TRUE
			6	The sample groups at the primary follow-up (with over 90 patients in each group) exceed the minimum threshold of 10 or more per group completing the primary outcome assessment.	TRUE
			7	The primary outcome (infant sleep position) was assessed in approximately 84% of the randomized participants, which meets the 80% follow-up criterion.	TRUE
			8	The primary outcome is clinically significant for SIDS prevention, a major public health concern, thereby satisfying the clinical importance requirement.	TRUE

			9	The trial did not report any preplanned subgroup analyses with proper interaction tests as required. Although a multivariate analysis was conducted, it did not incorporate preplanned subgroup analyses with the required interaction testing.	FALSE
			Final	While the trial meets criteria 1 through 8, the absence of preplanned subgroup analyses with proper interaction tests fails to satisfy criterion 9. As all criteria must be met for a trial to be considered rigorous, the study is not considered rigorous overall.	FALSE
21183586	Non-rigorous	0.288	1	The article is written in English.	TRUE
			2	The study involves human participants with chronic kidney disease.	TRUE
			3	The topicâ€"dietary protein source affecting phosphorus homeostasis in CKDâ€"is clearly clinically relevant to medicine and related health fields.	TRUE
			4	The analyses (paired comparisons, regression analyses, interaction testing) are directly aligned with the research question regarding the effects of different diets on phosphorus and hormone levels.	TRUE
			5	The participants were randomized to the order of diet administration (vegetarian first versus meat first) in a crossover design.	TRUE
			6	Only eight patients completed both arms of the crossover. The criterion requires 10 or more patients per group completing the primary outcome, which was not met.	FALSE
			7	With 9 patients randomized and 8 finishing, over 80% of participants had the primary outcomes assessed.	TRUE
			8	The primary outcomes (serum phosphorus, FGF23 levels, etc.) are clinically important markers in CKD management.	TRUE

			9	The study preplanned subgroup analyses (e.g., testing for interaction between period and diet assignment) and analyzed groups as randomized.	TRUE
			Final	While most criteria are met, the study fails to meet criterion 6 as it had fewer than 10 patients per group completing the primary outcome assessment. Since all criteria must be met for a trial to be considered rigorous, this study does not qualify.	FALSE
16916871	1	0.417	1	The full text is written in English.	TRUE
			2	It involves human participants (patients and caregivers) in primary care settings.	TRUE
			3	The topicâ€”improving attendance in primary careâ€”is important to clinical practice, impacting appointment keeping and service efficiency.	TRUE
			4	The analysis was clearly aligned with the study question, using appropriate statistical tests and an intentionâ€□toâ€□treat approach.	TRUE
			5	Participants were randomly allocated to three groups using block randomization, as documented in the full text.	TRUE
			6	Each group had well over 10 participants (each group had over 300 participants completing the primary outcome assessment).	TRUE
			7	The primary outcome (attendance) was assessed on a large majority of those randomized (993 out of 1111 assessed, with minimal losses), meeting the 80% threshold.	TRUE
			8	Attendance is a clinically relevant outcome in the context of primary care, fulfilling the criterion for a clinically important outcome.	TRUE
			9	The report mentions subgroup analyses (e.g., by reason for followâ€□up and appointment timing), but it is unclear if	FALSE

				these were pre-planned and whether proper interaction tests were performed as required.	
			Final	While the study meets criteria 1 through 8, the requirements for preplanned subgroup analyses with appropriate interaction testing (criterion 9) are not clearly addressed. Since all criteria must be met for a rigorous rating, the RCT is not considered rigorous.	FALSE
36395875	1	0.943	1	The article is written in English.	TRUE
			2	The study involves human patients undergoing offâ€□pump coronary artery bypass surgery.	TRUE
			3	The trial addresses clinically important outcomes (graft patency and major adverse cardiovascular events) relevant to medical and surgical practice.	TRUE
			4	The analyses (comparisons of MACEs, graft patency, use of Cox regression, etc.) are consistent with the studyâ€™s primary questions.	TRUE
			5	Participants were randomly allocated to the treatment groups using a 1:1 ratio with a pre-generated randomization table.	TRUE
			6	Each group comprised 102 participants, which is well above the minimum requirement of 10 patients per group completing primary outcome assessment.	TRUE
			7	The primary outcomes (e.g., graft patency assessed via angiographic follow-up) were measured in a high percentage of participants (97.5% follow-up reported).	TRUE
			8	The primary outcomes, MACEs and graft patency, are clinically important endpoints.	TRUE
			9	There is no evidence that preplanned subgroup analyses (with interaction testing) were conducted; while multivariable analysis was performed, the requirement for	FALSE

		preplanned subgroup analyses with testing for interactions was not met.	
	Final	While the trial meets most criteria including language, relevance, randomization, adequate sample size, follow-up, and clinical importance of outcomes, it does not satisfy the criterion regarding preplanned subgroup analyses with interaction testing. Since all criteria must be met for a study to be considered rigorous, the trial is not considered rigorous.	FALSE

† Predicted by the BioLinkBERT model.

AD4. GPT-4o - Verifier with TIAB

PMID	Label	Predicted Rigorous Probability†	Criterion	Justification	Assessment
19489805	Non-rigorous	0.608	1	The text is in English.	met
			2	The study involves human participants (mothers and infants).	met
			3	The topic is relevant to clinical practice, specifically in pediatrics and maternal health.	met
			4	The analysis appears consistent with the study question, as it evaluates the effect of the intervention on infant sleep position.	met
			5	The study mentions random allocation of participants to intervention and control groups.	met
			6	The study includes 228 mother-infant pairs, which suggests more than 10 participants per group.	met
			7	The text does not specify the percentage of participants completing the primary outcome assessment.	cannot_be_determined
			8	The primary outcome (infant sleep position) is clinically important for preventing SIDS.	met
			9	The text does not mention subgroup analyses.	cannot_be_determined
			Agreement	While the study meets most criteria, the lack of information on criteria 7 and 9 makes it difficult to definitively classify it as rigorous. However, the model's classification as rigorous is reasonable given the high probability and the information provided.	TRUE
21183586		0.288	1	The text is in English.	met

	Non-rigorous		2	The study involves human participants with chronic kidney disease.	met
			3	The topic is relevant to clinical practice, specifically dietary management in chronic kidney disease.	met
			4	The analysis appears consistent with the study question, focusing on the effects of dietary protein source on phosphorus homeostasis.	met
			5	The text does not explicitly mention random allocation of participants to comparison groups.	cannot_be_determined
			6	The study involved nine participants, which is less than 10 per group.	unmet
			7	The text does not provide sufficient information to determine if the primary outcome was assessed in 80% or more of those randomized.	cannot_be_determined
			8	The primary outcome, phosphorus homeostasis, is clinically important for CKD management.	met
			9	The text does not mention subgroup analyses.	cannot_be_determined
			Agreement	The study does not meet all the criteria for being rigorous, particularly criterion 6, which is unmet. Therefore, I agree with the model's classification of not rigorous.	TRUE
16916871	1	0.417	1	The text is in English.	met
			2	The study involves human participants attending primary care clinics.	met
			3	The study addresses attendance in primary care, which is relevant to clinical practice.	met
			4	The analysis appears consistent with the study question, focusing on attendance rates and cost-effectiveness.	met

			5	The study mentions random allocation to comparison groups.	met
			6	The study reports 993 participants eligible for analysis, which suggests more than 10 per group.	met
			7	The text does not specify the percentage of participants completing the primary outcome assessment.	cannot_be_determined
			8	The primary outcome (attendance rate) is clinically important.	met
			9	The text does not mention subgroup analyses.	cannot_be_determined
			Agreement	The model's classification as not rigorous is reasonable because not all criteria can be confirmed as met.	TRUE
36395875	1	0.943	1	The text is in English.	met
			2	The study involves human participants undergoing OPCAB.	met
			3	The study is relevant to clinical practice as it evaluates treatments for cardiovascular outcomes.	met
			4	The analysis appears consistent with the study question, focusing on outcomes of the treatments.	met
			5	The study mentions random allocation of participants to groups using a 1:1 ratio block table.	met
			6	Each group had 102 participants completing the primary outcome assessment.	met

	7	The text does not specify the percentage of randomized participants assessed for the primary outcome.	cannot_be_determined
	8	The primary outcomes (graft patency and MACEs) are clinically important.	met
	9	The text does not mention subgroup analyses.	cannot_be_determined
	Agreement	While most criteria are met, some cannot be determined from the text. However, the model's high confidence suggests it inferred rigour from the provided details.	TRUE

† Predicted by the BioLinkBERT model.

AD5. GPT-4o - Verifier with Full Text

PMID	Label	Predicted Rigorous Probability†	Criterion	Justification	Assessment
19489805	Non-rigorous	0.608	1	The article is written in English, as evidenced by the text provided.	TRUE
			2	The study involves human participants, specifically mothers and their infants.	TRUE
			3	The study addresses an important clinical topic, namely the prevention of sudden infant death syndrome (SIDS) through educational interventions.	TRUE
			4	The analysis in the article appears consistent with the study question, as it evaluates the effect of the intervention on the primary outcome.	TRUE
			5	The study mentions random allocation of participants to intervention and control groups.	TRUE
			6	The study reports that 91 participants in the intervention group and 100 in the control group were analyzed at the 3-month follow-up, meeting the criterion of 10 or more patients per group.	TRUE
			7	The study reports follow-up rates and mentions that 80% or more of those randomized were assessed for the primary outcome at the defined follow-up point.	TRUE
			8	The primary outcome, infant sleep position, is clinically important as it relates to the prevention of SIDS.	TRUE
			9	The article does not mention any subgroup analyses, so this criterion is not applicable.	TRUE
			Agreement	All criteria for rigour are met based on the provided information, so I agree with the model's classification of the study as rigorous.	TRUE

21183586	Non-rigorous	0.288	1	The article is written in English, as evidenced by the text provided.	TRUE
			2	The study involves human participants, as it mentions a crossover trial conducted on patients with chronic kidney disease.	TRUE
			3	The study is relevant to clinical practice as it investigates dietary interventions for managing phosphorus homeostasis in chronic kidney disease, which is a significant clinical concern.	TRUE
			4	The analysis appears consistent with the study question, as the study design and results focus on comparing the effects of vegetarian and meat diets on phosphorus homeostasis.	TRUE
			5	The study mentions randomization of participants to diet groups, fulfilling the criterion of random allocation.	TRUE
			6	The study had 9 participants initially, but only 8 completed both arms of the crossover trial. This does not meet the criterion of having 10 or more patients per group completing the primary outcome assessment.	FALSE
			7	The study does not explicitly state whether the primary outcome was assessed in 80% or more of those randomized, but with only 8 out of 9 completing, it is likely this criterion is not met.	FALSE
			8	The primary outcome, phosphorus homeostasis, is clinically important for CKD management, so this criterion is met.	TRUE
			9	The study does not mention subgroup analyses, so this criterion is not applicable.	TRUE
			Agreement	I agree with the model's classification of the study as not rigorous because it does not meet the criteria of having 10 or more patients per group completing the primary	TRUE

				outcome assessment and assessing the primary outcome in 80% or more of those randomized.	
16916871	1	0.417	1	The article is written in English, as evidenced by the text provided.	TRUE
			2	The study involves human participants, as it is about patients attending primary care clinics.	TRUE
			3	The topic is relevant to clinical practice as it addresses attendance in primary care, which is a significant issue in healthcare management.	TRUE
			4	The analysis appears consistent with the study question, as the outcomes measured (attendance rates and cost-effectiveness) directly address the effectiveness of the interventions.	TRUE
			5	The study mentions random allocation of participants to the three groups (control, text messaging, and mobile phone reminders).	TRUE
			6	The study reports 993 participants in total, with approximately 329 participants per group, which satisfies the criterion of 10 or more patients per group.	TRUE
			7	The study does not explicitly state the percentage of participants who completed the primary outcome assessment, but it mentions that 993 participants were eligible for analysis, suggesting high follow-up rates.	TRUE
			8	The primary outcome, attendance rate, is clinically important as it impacts healthcare delivery and patient outcomes.	TRUE
			9	The study does not provide evidence that subgroup analyses were preplanned or that interaction tests were conducted between subgroups.	FALSE

			Agreement	I agree with the model's classification of the study as not rigorous because it does not meet all the manual appraisal criteria, specifically criterion 9 regarding subgroup analyses.	TRUE
36395875	1	0.943	1	The article is written in English, as evidenced by the text provided.	TRUE
			2	The study involves human participants undergoing off-pump coronary artery bypass surgery.	TRUE
			3	The study is relevant to clinical practice as it investigates the efficacy of antiplatelet therapies in a specific surgical context.	TRUE
			4	The analysis aligns with the study question, comparing outcomes between two treatment groups.	TRUE
			5	Participants were randomly allocated to the two treatment groups using a 1:1 ratio block table.	TRUE
			6	Each group had 102 participants completing the primary outcome assessment, meeting the criterion of 10 or more patients per group.	TRUE
			7	The primary outcomes were assessed in 97.5% of the randomized participants, which is above the 80% threshold.	TRUE
			8	The primary outcomes, such as major adverse cardiovascular events, are clinically important.	TRUE
			9	The study does not mention any subgroup analyses, so this criterion is not applicable.	TRUE
			Agreement	The study meets all the criteria for being classified as rigorous. The model's prediction of rigorous is correct.	TRUE

† Predicted by the BioLinkBERT model.

AD6. GPT-o3-mini - Verifier with TIAB

PMID	Label	Predicted Rigorous Probability†	Criterion	Justification	Assessment
19489805	Non-rigorous	0.608	1	The studyâ€™s title and abstract are written in English.	met
			2	The study involves human subjects (infant-mother pairs) and addresses maternal behavior regarding infant sleep positioning.	met
			3	The topic (infant sleep position to prevent SIDS) is clinically important and relevant to health professionals, rather than being merely descriptive of prevalence.	met
			4	The analysis directly addresses the study question by evaluating the effect of an educational intervention on infant sleep positioning using appropriate statistical methods.	met
			5	It is explicitly described as a randomized clinical trial with participants assigned to intervention or control groups.	met
			6	With a total sample of 228 motherâ€“infant pairs, it is reasonable to conclude that each group has more than 10 participants completing the assessment.	met
			7	The study reports outcome assessment at 3 months via home visit but does not explicitly mention the proportion of participants who completed this followâ€□up; therefore, it	cannot_be_determined

				cannot be confirmed if the threshold of 80% was achieved.	
			8	The primary outcome (infant sleep position) is a clinically significant measure given its relevance to SIDS prevention.	met
			9	There is no indication that any subgroup analyses were preplanned or that interaction tests were performed. This criterion is specifically required in the appraisal but is not met.	unmet
			Agreement	Although the model predicted the article as rigorous with a probability of 0.608, the manual appraisal indicates that not all criteria are conclusively met. In particular, the absence of documented preplanned subgroup analyses (criterion 9) and uncertain follow-up completeness (criterion 7) are critical shortcomings, which lead to a conclusion of non-rigorous study design.	FALSE
21183586	Non-rigorous	0.288	1	The text is written in English.	met
			2	The study involves patients with advanced chronic kidney disease, which means it is about humans.	met
			3	It investigates an intervention (dietary protein source) that is important to clinical practice in chronic kidney disease care, not merely a descriptive prevalence study.	met
			4	The analysis in the article aligns with the study question of comparing the effects of vegetarian versus meat diets on phosphorus homeostasis.	met

			5	While the design is noted as a crossover trial, the text does not explicitly state that participants were randomized to the order of treatment, making it unclear whether random allocation was implemented.	cannot_be_determined
			6	The study involved only nine patients, which is below the threshold of 10 or more patients per group completing the primary outcome assessment.	unmet
			7	There is evidence of intensive monitoring of outcomes during the inpatient period, indicating that primary outcomes were likely assessed in at least 80% of the participants.	met
			8	The primary outcome pertaining to phosphorus homeostasis is clinically important in the context of chronic kidney disease, and the study likely includes clinically relevant secondary outcomes.	met
			9	The abstract does not mention any preplanned subgroup analyses or interaction tests between subgroups; thus, this criterion is not satisfied.	unmet
			Agreement	Given that criteria 6 and 9 are clearly unmet (and criterion 5 remains ambiguous), the study does not meet all the criteria required for a rigorous classification. This aligns with the model's predicted classification of 'Not rigorous'.	TRUE
16916871	1	0.417	1	The text is clearly written in English, satisfying the language requirement.	met

	2	The study involves patients (or their caregivers) in a primary care setting, indicating that it is about humans.	met
	3	The study addresses an issue (attendance in primary care) that is important for clinical practice, rather than being merely a descriptive study of prevalence.	met
	4	The articleâ€™s analysis is aligned with its study question, assessing the effectiveness of text messaging reminders on appointment attendance.	met
	5	The design is described as a multicentre threeâ€arm randomized controlled trial, indicating that participants were randomly allocated to groups.	met
	6	With 993 participants analyzed across three arms, it is safe to assume that each group has more than 10 patients completing the primary outcome assessment.	met
	7	The text reports the number eligible for analysis but does not provide information on whether the primary outcome was assessed in at least 80% of those randomized, leaving this criterion unclear.	cannot_be_determined
	8	Attendance is a relevant process measure in primary care that can have implications for overall clinical care, thereby meeting the clinically important outcome criterion.	met
	9	There is no mention of preplanned subgroup analyses or any interaction tests as required, so this criterion is not met.	unmet

			Agreement	Given that one of the mandatory criteria (criterion 9) is clearly unmet and criterion 7 remains indeterminate, the overall rigour of this study cannot be affirmed. Therefore, I agree with the modelâ€™s prediction of 'not rigorous'.	TRUE
36395875	1	0.943	1	The text is clearly written in English.	met
			2	The study involves human patients undergoing offâ€pump coronary bypass surgery.	met
			3	The topic addresses clinically important treatments (antiplatelet therapy in coronary surgery) relevant to clinical practice.	met
			4	The analysis follows the study question by comparing outcomes (graft patency and MACEs) between treatment groups.	met
			5	Random allocation is explicitly described using a 1:1 ratio block table.	met
			6	Each group has 102 patients, which well exceeds the minimum requirement of 10 patients per group.	met
			7	The abstract does not specify the percentage of patients completing the primary outcome assessment at follow-up (only intent-to-treat analysis is mentioned), so the adequacy of follow-up cannot be determined from the available information.	cannot_be_determined
			8	The primary outcomes (graft patency and major adverse cardiovascular events) are clinically important endpoints.	met
			9	There is no mention of preplanned subgroup analyses or interaction tests between	cannot_be_determined

		subgroups, making it unclear if this criterion was met.	
	Agreement	While the study meets many of the rigorous design criteria (language, population, clinical relevance, randomization, adequate patient numbers, and clinically important outcomes), the abstract lacks explicit information on follow-up completeness (criterion 7) and on the conduct of preplanned subgroup analyses (criterion 9). Since all criteria must be clearly met to be rated as rigorous, the omission of these details suggests that the decision to classify the study as rigorous may be premature. Therefore, I do not fully agree with the modelâ€™s classification as rigorous.	FALSE

† Predicted by the BioLinkBERT model.

AD7. GPT-o3-mini - Verifier with Full Text

PMID	Label	Predicted Rigorous Probability†	Criterion‡	Justification	Assessment
19489805	Non-rigorous	0.608	1	The article is written entirely in English.	TRUE
			2	The study involves motherâ€□infant pairs, making it a study about human subjects.	TRUE
			3	The topicâ€”an educational intervention to improve infant sleep position for SIDS preventionâ€”is highly relevant to clinical practice in pediatrics and public health and is not merely a descriptive prevalence study.	TRUE
			4	The studyâ€™s analyses (bivariate analyses and logistic regression) directly address the study question regarding the effect of the intervention on sleep position.	TRUE
			5	Participants were randomly allocated using a coin toss and block randomization, fulfilling the requirement for random allocation to comparison groups.	TRUE
			6	Both groups had well over 10 patients (over 90 in one group and 100 in the other) completing the primary outcome assessment.	TRUE
			7	The primary outcome (infant sleep position at 3 months) was assessed in 191 of 228 participants (~84%), which exceeds the 80% minimum followâ€□up threshold.	TRUE
			8	The primary outcomeâ€”sleep positionâ€”is clinically important as it is directly linked to reduced SIDS risk, and thus has direct clinical relevance.	TRUE
			9	Although the study performed a multivariate analysis, it did not report any preplanned subgroup analyses with proper testing for interactions between subgroups, which is required by the criteria.	FALSE

			Agreement	Since all criteria must be met for a study to be rated as rigorous and criterion 9 was not met, I disagree with the modelâ€™s positive (rigorous) classification.	FALSE
21183586	Non-rigorous	0.288	1	The article is written in English.	TRUE
			2	The study involves human subjects (patients with chronic kidney disease).	TRUE
			3	The topic is clinically relevant, addressing dietary management in CKDâ€”a topic important to clinical practice.	TRUE
			4	The analysis in the study (crossover trial design with controlled diets) is consistent with the study question examining the effect of protein source on phosphorus homeostasis.	TRUE
			5	Participants were randomized to diet order, fulfilling the requirement for random allocation.	TRUE
			6	The study enrolled only 9 patients (with 8 completing both arms), which is below the threshold of 10 or more patients per group completing the primary outcome assessment.	FALSE
			7	Although only 8 patients completed the trial, this represents over 80% followâ€□up of those randomized.	TRUE
			8	The primary outcome (phosphorus homeostasis and related biomarkers) is clinically important for managing CKD.	TRUE
			9	The study did not incorporate any preplanned subgroup analyses with tests for interaction, which is a required criterion.	FALSE
			Agreement	Given that two criteria (criterion 6 on sample size per group and criterion 9 on subgroup analyses) were not	TRUE

				met, I agree with the modelâ€™s prediction that the study is not rigorous.	
16916871	1	0.417	1	The article is written in English, published in an English language journal.	TRUE
			2	The study involves human participants (patients or their caregivers) in primary care.	TRUE
			3	The research addresses an intervention (appointment reminders) that is relevant to primary care practice and affects clinical service delivery, rather than being a mere descriptive study of prevalence.	TRUE
			4	The design, execution, and reporting of the analysis are well aligned with the study question of assessing the effectiveness of text messaging reminders.	TRUE
			5	The study employed random allocation using a block randomization technique, ensuring each participant had a chance of being allocated to any arm.	TRUE
			6	Each study arm had well over the required 10 patients per group (with group sizes of 329 to 335), ensuring adequate sample size for primary outcome assessment.	TRUE
			7	The primary outcome (attendance) was assessed in nearly all randomized participants with minimal loss to followâ€□up, well exceeding the 80% threshold.	TRUE
			8	Attendance as the primary outcome is a clinically important endpoint as it impacts continuity of care and efficient healthcare delivery. Alternatively, one of the secondary outcomes (cost-effectiveness) also holds clinical relevance.	TRUE
			9	While subgroup analyses were presented, the report does not clearly state that these analyses were preplanned, nor does it detail formal interaction testing as required by the criterion.	FALSE

			Agreement	Since rigorous classification requires that all criteria are met and criterion 9 is not met, the overall evaluation of nonâ€□rigour is appropriate. The modelâ€™s predicted probability (approximately 0.417) leading to a non-rigorous classification aligns with this appraisal.	TRUE
36395875	1	0.943	1	The article is written in English, as evidenced by language and publication details.	TRUE
			2	The study involves human patients undergoing offâ€□pump coronary bypass surgery, demonstrating it is about humans.	TRUE
			3	The research topic is directly related to clinically important outcomes in cardiovascular surgery and patient care, not merely a descriptive prevalence study.	TRUE
			4	The analysis is clearly aligned with the study question (comparing ticagrelor versus clopidogrel in clopidogrel-resistant patients), with appropriate endpoints and methods described.	TRUE
			5	Random allocation is explicitly stated; patients were randomized in a 1:1 ratio using a block randomization method.	TRUE
			6	Each group comprised 102 patients, which is well above the minimum requirement of 10 patients completing the primary outcome assessment.	TRUE
			7	The study achieved a very high rate of followâ€□up (around 97.5% for angiographic evaluation), meeting the 80% threshold.	TRUE
			8	The primary outcomes (graft patency and major adverse cardiovascular events) are clinically important endpoints.	TRUE
			9	The manual criteria require that subgroup analyses must be preplanned with proper interaction testing. The article does not provide evidence of any preplanned and	FALSE

		appropriately conducted subgroup analyses. Although subgroup analysis is not always necessary if the main analysis is clear, the manual criteria list it as mandatory for a 'rigorous' rating.	
	Agreement	While all other rigorous criteria are met, the lack of evidence for preplanned subgroup analyses (and interaction testing) means that not all manual appraisal criteria are satisfied. Therefore, despite a high predicted probability and model classification as rigorous, I do not agree with the classification because criterion 9 is not met.	FALSE

† Predicted by the BioLinkBERT model.

APPENDIX E GPT PERTURBATION EXPLAINER PROMPT

AE1. Prompt Dictionary

Prompt Key	Prompt Content
ROLE	You are a machine learning model explainer.
TASK	You are tasked with explaining binary text classification encoder-only transformer model's predictions by perturbing its input tokens, similar to perturbation based XAI frameworks like LIME or SHAP.
CRITERIA	Manual appraisal criteria: All of these criteria must be met to be rated as being rigorous. The article will be rated as not rigorous if any criteria are not met. The criteria are 1) in English, 2) about humans, 3) about topics that are important to the clinical practice of medicine, nursing, rehabilitation, and other health professions, other than descriptive studies of prevalence, 4) analysis of each article consistent with the study question, 5) random allocation of participants to comparison groups, 6) 10 or more patients per group completing primary outcome assessment, 7) primary outcome(s) assessed in 80% or more of those randomized at the defined follow-up point, 8) primary outcome is clinically important or ≥1 secondary outcome is clinically important, and 9) subgroup analyses must be preplanned, with groups analyzed as they were randomized; analyses must test for interaction between 2 or more subgroups.
PROVIDED_INFO_INDEX	You will be provided the number of tokens, the logits for both positive and negative classes, and the probability for the positive class. You will NOT be provided the text.
INSTRUCTIONS_INDEX	For a given instance, determine which tokens have the greatest impact on the model's prediction by systematically masking them. You will: 1. Receive an instance with the number of tokens and model outputs without any masking.

	2. Define 'importance' for yourself. The importance should be a float. A negative and positive float should indicate that the token increases the chance of a negative and positive classification, respectively. 3. Generate a list of num_token number of lists where each list contains the index of 1 token to mask (e.g., [[0], [1], [2], [3], ...]). Tokens corresponding to the indexes will be replaced by `[MASK]` by the `mask_and_predict` function. 4. Repeat steps 4a and 4b for 10 iterations. Once all 10 iterations are complete, you will be prompted to proceed to step 5. 4a. Generate a list of lists of one or numerous index(es) to mask, based on the results of previous iterations. Start with completely random number of lists and indexes and adjust the number of lists and indexes in each list based on model outputs. DO NOT repeatedly mask the same combination of tokens. 4b. Call `mask_and_predict` with the lists of indexes to mask. The function will return a list of lists in the form of [logit_positive, logit_negative, probability_positive], in which each list corresponds with the model's output given the masked variant. 5. You will make any adjustments to your initial definition of 'importance'. 6. You will be prompted with indexes of the tokens. Please calculate the importance of each prompted token in the text. 7. For each token, output the index and the importance.
DEVELOEPR_INDEX	{ROLE} {TASK} {PROVIDED_INFO_INDEX} {INSTRUCTIONS_INDEX}
INITIAL_USER_INDEX	Number of tokens: <num_tokens>. Model output without masking: [[<logit_positive>, <logit_negative>, <probability_positive>]].
PROVIDED_INFO_TOKEN	You will be provided the number of tokens, the input tokens, the logits for both positive and negative classes, the probability for the positive

	class. You will also be provided the manual appraisal criteria (appraised using the full text).
INSTRUCTIONS_TOKEN	For a given instance, determine which tokens have the greatest impact on the model's prediction by systematically masking them. You will: 1. Receive an instance with the number of tokens, the tokens, and model outputs without any masking. 2. Define 'importance' for yourself. The importance should be a float. A negative and positive float should indicate that the token increases the chance of a negative and positive classification, respectively. 3. Generate a list of num_token number of lists where each list contains the index of 1 token to m7 Sask (e.g., [[0], [1], [2], [3], ...]). Tokens corresponding to the indexes will be replaced by `[MASK]` by the `mask_and_predict` function. 4. Repeat steps 4a and 4b for 10 iterations. Once all 10 iterations are complete, you will be prompted to proceed to step 5. 4a. Generate a list of lists of one or numerous index(es) to mask, based on which token you think is semantically important and the results of previous masking iterations. DO NOT repeatedly mask the same combination of tokens. 4b. Call `mask_and_predict` with the lists of indexes to mask. The function will return a list of lists in the form of [logit_positive, logit_negative, probability_positive], in which each list corresponds with the model's output given the masked variant. 5. You will make any adjustments to your initial definition of 'importance'. 6. You will be prompted with indexes of the tokens. Please calculate the importance of each prompted token in the text. 7. For each token, output the index and the importance.
DEVELOEPR_TOKEN	{ROLE} {TASK}

	{PROVIDED_INFO_TOKEN} {INSTRUCTIONS_TOKEN} {CRITERIA}
INITIAL_USER_TOKEN	Number of tokens: <num_tokens>. Input tokens: <token_list> Model output without masking: [[<logit_positive>, <logit_negative>, <probability_positive>]].
DEFINE_IMPORTANCE	Please define 'importance'. You will use this definition to calculate token importance when prompted later.
INDIVIDUAL_MASKING	Please generate the initial masking list and call 'mask_and_predict'. The input must be a list of <num_tokens> lists of one index each.
ITERATION_MASKING	Please proceed with iteration <iteration> of masking. Generate 10 to 30 lists. Only generate indexes from 0 to <num_tokens – 1>. Add, remove, or adjust masking based on previous maskings and model outputs. Prioritize masking tokens that seemed to be important from previous iterations AND tokens that have rarely been masked in previous iterations. DO NOT repeatedly mask the same combination of tokens. DO NOT reply with anything else other than the function call.
REDEFINE_IMPORTANCE	Based on the results of all previous masking iterations, adjust the definition of 'importance' for yourself as you see fit.
ATTRIBUTION_CALCULATION	Please calculate token importance for tokens from index <i> to <minimum(i + 20 - 1, num_token - 1)> based on previous information and your definition of 'importance'.
MASK_AND_PREDICT_OUTPUTS	[[<logit_positive_0>, <logit_negative_0>, <probability_positive_0>] …]

AE2. JSON Object Dictionary

Object Key	Object Type
TOKEN_INDEX	int
IMPORTANCE_VALUE	float
IMPORTANCE_DEFINITION	str
TOKEN_IMPORTANCE	TOKEN_INDEX: IMPORTANCE_VALUE
TOKEN_IMPORTANCE_RESULTS	[TOKEN_IMPORTANCE]
MASK_LIST	[int]
MASK_LISTS	[MASK_LIST]

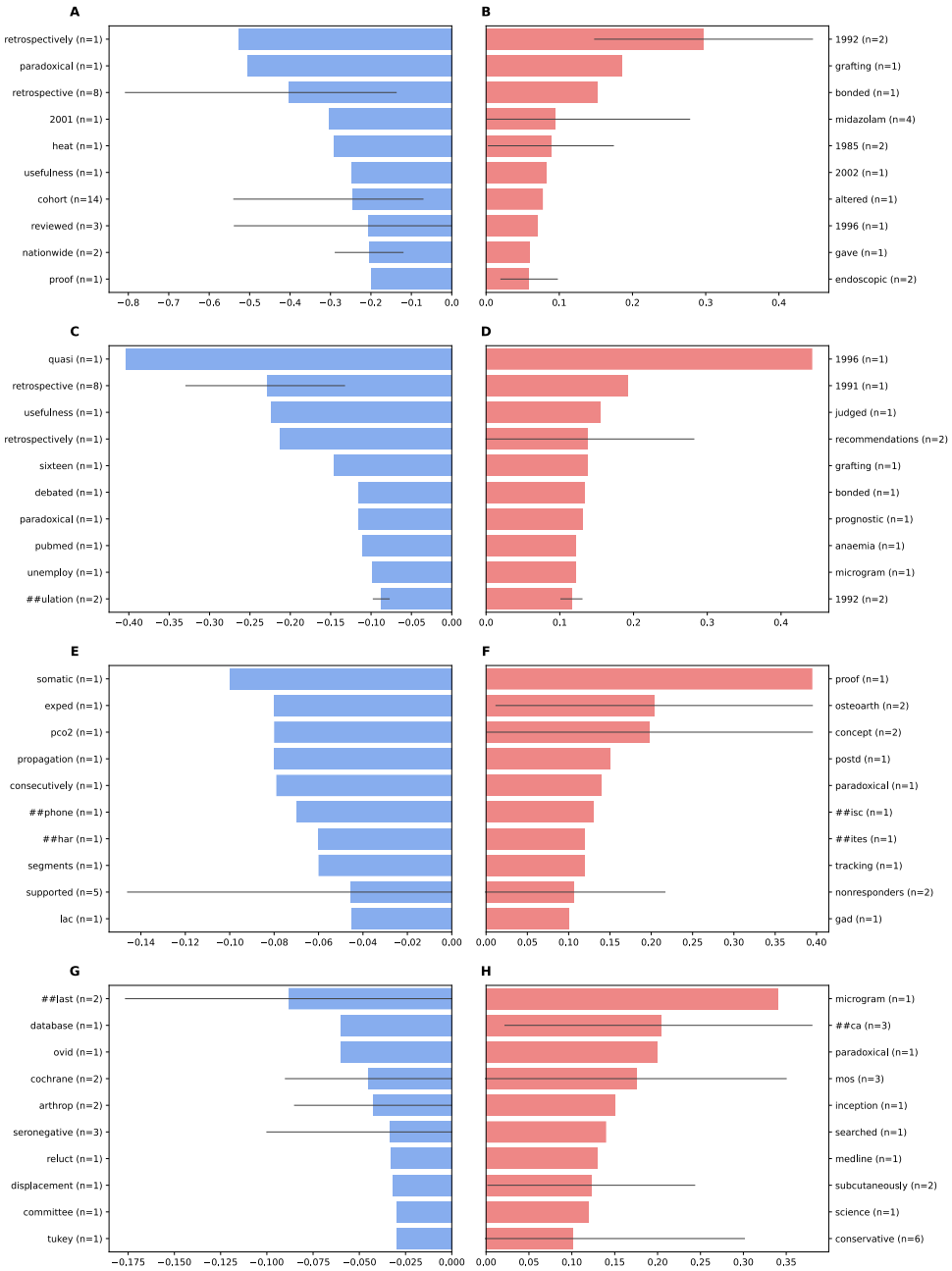
AE3. Message Chain

Sequence	Role	Key
1	developer	DEVELOPER_INDEX/DEVELOPER_TOKEN
2	user	INITIAL_USER_INDEX/INITIAL_USER_TOKEN
3	user	DEFINE_IMPORTANCE
4	assistant	IMPORTANCE_DEFINITION
5	user	INDIVIDUAL_MASKING
6	assistant	MASK_LISTS
7	tool	MASK_AND_PREDICT_OUTPUTS
8	user	ITERATION_MASKING
9	assistant	MASK_LISTS
10	tool	MASK_AND_PREDICT_OUTPUTS
11	user	ITERATION_MASKING
12	assistant	MASK_LISTS
13	tool	MASK_AND_PREDICT_OUTPUTS
14	user	ITERATION_MASKING
15	assistant	MASK_LISTS
16	tool	MASK_AND_PREDICT_OUTPUTS
17	user	ITERATION_MASKING
18	assistant	MASK_LISTS
19	tool	MASK_AND_PREDICT_OUTPUTS
20	user	ITERATION_MASKING
21	assistant	MASK_LISTS
22	tool	MASK_AND_PREDICT_OUTPUTS
23	user	ITERATION_MASKING
24	assistant	MASK_LISTS
25	tool	MASK_AND_PREDICT_OUTPUTS
26	user	ITERATION_MASKING
27	assistant	MASK_LISTS
28	tool	MASK_AND_PREDICT_OUTPUTS
29	user	ITERATION_MASKING
30	assistant	MASK_LISTS
31	tool	MASK_AND_PREDICT_OUTPUTS
32	user	ITERATION_MASKING
33	assistant	MASK_LISTS
34	tool	MASK_AND_PREDICT_OUTPUTS
35	user	ITERATION_MASKING
36	assistant	MASK_LISTS
37	tool	MASK_AND_PREDICT_OUTPUTS
38	user	REDEFINE_IMPORTANCE
39	assistant	IMPORTANCE_DEFINITION
40	user	ATTRIBUTION_CALCULATION

41	assistant	TOKEN_IMPORTANCE_RESULTS

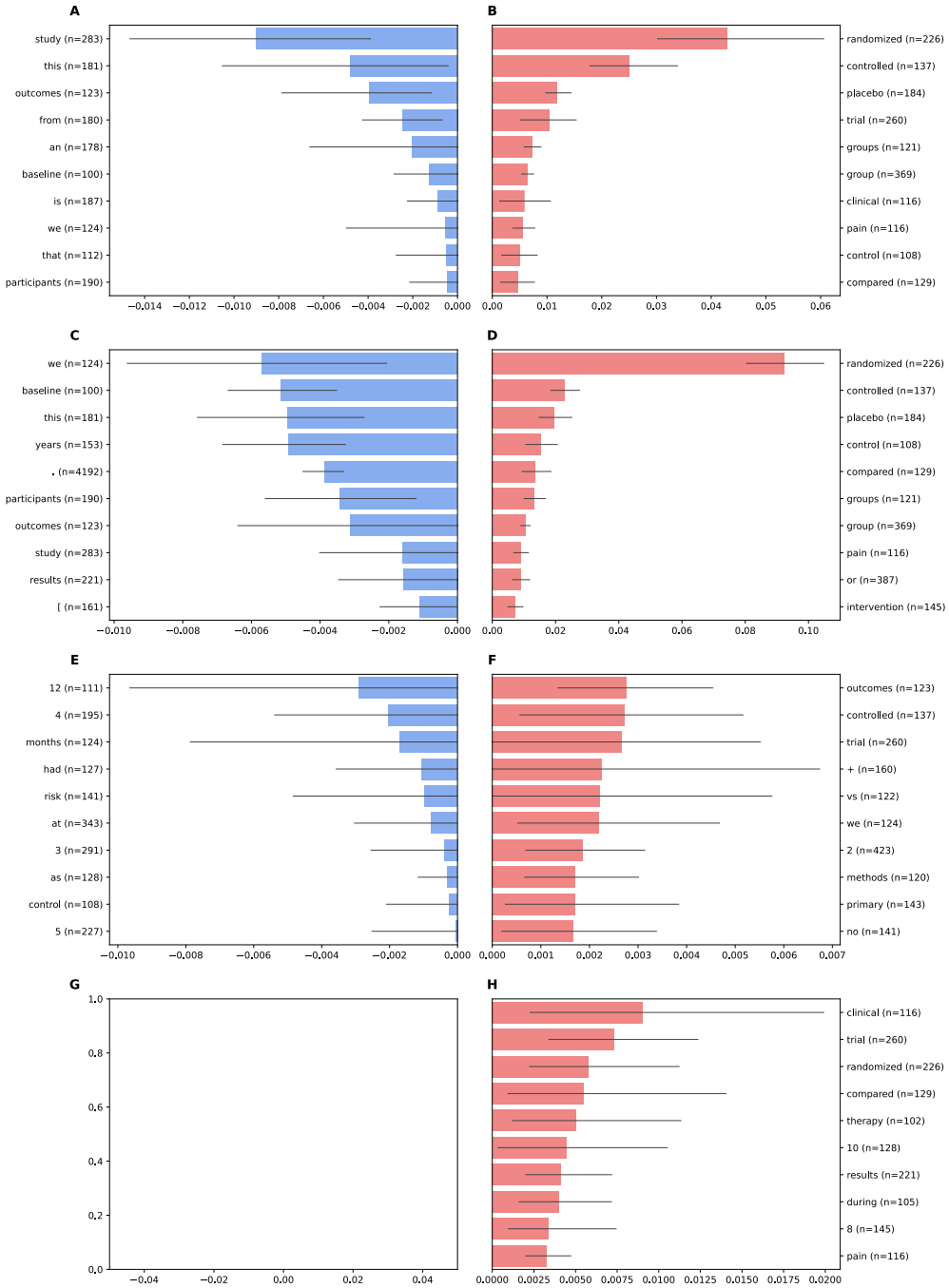
APPENDIX F GPT PERTURBATION EXPLAINER IMPORTANT TOKENS

AF1. ≥1 occurrences



A Negative tokens for SHAP; **B** Positive tokens for SHAP; **C** Negative tokens for IG; **D** Positive tokens for IG; **E** Negative tokens for GPT-index; **F** Positive tokens for GPT-index; **G** Negative tokens for GPT-token; **H** Positive tokens for GPT-token.

AF2. ≥100 occurrences



A Negative tokens for SHAP; **B** Positive tokens for SHAP; **C** Negative tokens for IG; **D** Positive tokens for IG; **E** Negative tokens for GPT-index; **F** Positive tokens for GPT-index; **G** Negative tokens for GPT-token; **H** Positive tokens for GPT-token.