# A Study on Deep Learning and Explainable & Interpretable AI: From General Domain to Healthcare

**Guangming Huang**

A thesis submitted for the degree of

**Doctor of Philosophy**

**School of Computer Science and Electronic Engineering**

**University of Essex**

April 2025

# Abstract

Deep learning has witnessed an unprecedented evolution over the past decade, transforming from theoretical concepts into practical applications that permeate numerous domains of human activity. The exponential growth in computational power, availability of large-scale data, and advancements in neural network architectures have collectively facilitated the development of increasingly sophisticated deep learning models (e.g., large language models, LLMs) with performance levels that often surpass human capabilities in specific tasks. Despite these advancements, the deployment of deep learning models for healthcare presents substantial methodological and paradigmatic challenges. The transition from general to healthcare-specific contexts requires addressing fundamental differences in (i) representation learning, (ii) domain knowledge, (iii) data characteristics and (iv) explainability and interpretability. To address these issues, this study aims to systematically investigate the methodological and paradigmatic transition of deep learning and explainable & interpretable AI from general domain to healthcare. Our contributions are in following key areas: (i) we introduce multi-label relations in multi-label supervised contrastive learning (MSCL) and propose a novel contrastive loss function, termed

Similarity-Dissimilarity Loss, which dynamically re-weights based on the computed similarity and dissimilarity factors between positive samples and anchors, applying for areas from multi-label classification to automated medical coding; (ii) We propose a Prompting Explicit and Implicit knowledge (PEI) framework for multi-hop question answering (QA) in biomedical domains, which employs CoT prompt-based learning to bridge explicit and implicit knowledge, aligning with human reading comprehension; (iii) we introduce a lexical-based imbalanced data augmentation (LIDA) for mental health moderation, which an easy-to-implement and interpretable DA method that strategically leverages sensitive lexicons by incorporating them into negative samples to transform these instances into positive examples. Through rigorous theoretical analyses and extensive experimental validation across multiple domains, this thesis contributes novel methodologies that enhance the performance, interpretability, and clinical applicability of deep learning methods from general domain to healthcare.

# Acknowledgments

The completion of this doctoral thesis represents not only years of academic research but also the culmination of countless interactions, discussions, and support from numerous individuals and institutions to whom I am profoundly indebted.

I wish to express my deepest gratitude to my principal supervisor, Dr Yunfei Long, whose unwavering guidance, incisive feedback, and intellectual rigor have fundamentally shaped this research. Your ability to challenge my thinking while providing the autonomy necessary for intellectual growth has been invaluable. I am equally grateful to my co-supervisor, Dr Cunjin Luo, whose complementary expertise and perspective broadened the scope and impact of this work.

For their emotional sustenance and unwavering belief in me, I owe an immeasurable debt to my family. To my father, whose sacrifices, encouragement, and unconditional support made this academic pursuit possible — thank you for instilling in me the value of education and perseverance. To my partner your patience, understanding, and steadfast support through the inevitable challenges of doctoral research have been my anchor. Your belief in me, particularly during moments of doubt, has been a source of immense

strength.

While this acknowledgment attempts to recognize those who have contributed to this work, it inevitably falls short of expressing the full depth of my appreciation. The journey toward this PhD has been transformative beyond the academic realm, and I am profoundly grateful to all who have been part of it.

# Publications

This thesis collects several articles and conference papers which have been published or are currently under review as first author during my PhD. As such, most of the contents of this thesis have appeared in the following publications:

- Guangming Huang, Yingya Li, Shoaib Jameel, Yunfei Long, and Giorgos Papanastasiou. "**From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?**." *Computational and Structural Biotechnology Journal*, 2024. (Chapter 1)

- Guangming Huang, Yunfei Long, and Cunjin Luo. "**Similarity-Dissimilarity Loss for Multi-label Supervised Contrastive Learning**." *Advances in Neural Information Processing Systems*, 2025. (*arXiv preprint arXiv:2410.13439*, under review) (Chapter 2)

- Guangming Huang, Yunfei Long, and Cunjin Luo. "**Improving Multi-hop Question Answering with Prompting Explicit and Implicit Knowledge Aligned on Human Reading Comprehension**." *International Journal of Machine Learning and Cybernetics*, 2025. (Chapter 3)

- Guangming Huang, Yunfei Long, Cunjin Luo, Jiaxing Shen, and Xia Sun. "**Prompting Explicit and Implicit Knowledge for Multi-hop Question Answering Based on Human Reading Process**." *In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (COLING)*, 2024. (Chapter 3)

- Guangming Huang, Yunfei Long, Cunjin Luo, and Yingya Li. "**LIDA: Lexical-Based Imbalanced Data Augmentation for Content Moderation**." *In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 2023. (Chapter 4)

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Deep Learning and Generalist Foundation Models

### 1.1.1 Deep Learning

Deep learning represents a transformative paradigm within the broader field of artificial intelligence (AI), specifically as a subset of machine learning methodologies that employ multiple layers of neural networks to progressively extract higher-level features from raw input data. Unlike traditional machine learning approaches that often require manual feature engineering, deep learning algorithms autonomously discover intricate patterns and representations within data through hierarchical learning processes [1]. This capability has fundamentally altered the landscape of computational intelligence, enabling unprecedented advances across diverse domains including computer vision, natural language processing, speech recognition, and scientific discovery.

The conceptual foundations of deep learning can be traced back to the mid-20th century with the introduction of the perceptron by Rosenblatt [2], which demonstrated the possibility of machines learning binary classification tasks. However, the limitations of single-layer neural networks, famously highlighted by Minsky and Papert [3], temporarily dampened enthusiasm for neural network research. The resurgence of interest began with backpropagation algorithms [4], allowing for efficient training of multi-layer networks, though computational constraints continued to limit practical applications. The deep learning renaissance truly commenced in the early 2010s, catalyzed by three concurrent developments: exponential growth in computational capacity through graphics processing units (GPUs), the availability of massive datasets, and algorithmic innovations that addressed previous training inefficiencies [5].

A pivotal moment in deep learning's ascendance occurred in 2012 when Krizhevsky et al. demonstrated the remarkable efficacy of convolutional neural networks (CNNs) in the ImageNet competition, reducing error rates by an unprecedented margin [6]. This watershed event, often referred to as the "ImageNet moment," precipitated a paradigm shift across computer vision and subsequently influenced numerous other fields. In the ensuing years, deep learning architectures achieved and frequently surpassed human-level performance across diverse benchmarks: from image classification [7] and object detection [8] to speech recognition [9] and natural language understanding [10].

The technical evolution of deep learning has been characterized by architectural innovations that address specific computational challenges. CNNs exploit spatially local correlation through weight sharing mechanisms, making them particularly effective for

visual data processing. Recurrent Neural Networks (RNNs), especially variants such as Long Short-Term Memory (LSTM) networks [11] and Gated Recurrent Units (GRUs) [12], capture temporal dependencies in sequential data through feedback connections. Transformative advancements emerged with attention mechanisms [13], which dynamically weight input features based on relevance, culminating in the Transformer architecture [14] that has become foundational for state-of-the-art natural language processing systems.

Critical to deep learning's success are optimization techniques that facilitate efficient training of increasingly complex models. Stochastic gradient descent variants, particularly Adam [15], have enabled stable convergence during training. Regularization strategies such as dropout [16], batch normalization [17], and weight decay mitigate overfitting and improve generalization capabilities. Additionally, architectural search methodologies [18] have automated the discovery of optimal network configurations, further enhancing performance across tasks.

The confluence of these methodological innovations with hardware advancements, which particularly specialized accelerators like TPUs [19] has enabled the training of increasingly parameter-dense models with billions or even trillions of parameters. This scaling trajectory has revealed emergent capabilities where quantitative increases in model capacity yield qualitative shifts in functional abilities, establishing the foundation for the generalist foundation models that characterize contemporary deep learning research.

### 1.1.2 Generalist Foundation Models

Generalist foundation models represent a paradigmatic evolution in AI, characterized by large-scale neural network architectures trained on diverse, expansive datasets that serve as versatile computational substrates adaptable to numerous downstream tasks with minimal task-specific training [20]. Unlike traditional machine learning systems designed for singular, narrowly-defined objectives, foundation models establish general-purpose representational frameworks that capture broad statistical patterns across multiple domains of knowledge.

The technical underpinnings of foundation models synthesize several methodological advances within deep learning. At their core, most contemporary foundation models employ Transformer architectures [14], which utilize self-attention mechanisms to dynamically model relationships between all elements in a sequence. This architectural paradigm has proven remarkably effective for capturing complex dependencies in data across modalities including text [10, 21], images [22], audio [23], and multimodal combinations thereof [24].

Large-scale pretraining constitutes the second critical component in foundation model development. This approach involves exposure to massive datasets, often encompassing hundreds of gigabytes or even petabytes of information, through self-supervised learning objectives that do not require explicit human annotation. For language models, these objectives typically involve predicting masked tokens or subsequent words in a sequence [10, 25], while vision models may employ contrastive learning between transformed

views of the same image [26]. The unprecedented scale of this pretraining enables models to internalize statistical regularities across diverse contexts, forming rich internal representations that generalize beyond specific tasks or domains.

Moreover, transfer learning represents a fundamental operational principle for foundation models, wherein knowledge acquired during pretraining is repurposed for downstream tasks through fine-tuning or adaptation techniques. Fine-tuning involves updating model parameters using task-specific data, often requiring substantially less supervision than training from scratch [27]. Parameter-efficient tuning methods such as adapter layers [28], prefix tuning [29], and Low-Rank Adaptation (LoRA) [30] reduce computational requirements by modifying only a subset of parameters while preserving general knowledge. Prompt learning has emerged as one of particularly transformative parameter-efficient tuning methods, recasting downstream tasks as variations of the original pretraining objective through carefully constructed input formulations. Initially developed as a technique to elicit knowledge from language models through "prompt engineering" [21], this approach has evolved into sophisticated methodologies including prompt tuning [31], where continuous prompt vectors are learned to optimize task performance, and chain-of-thought prompting [32], which guides models through intermediate reasoning steps. These techniques have effectively blurred the distinction between training and inference, establishing natural language as a flexible programming interface for neural computation.

The developmental trajectory of foundation models has witnessed exponential growth in both model capacity and capability. GPT (Generative Pre-trained Transformer), intro-

duced by OpenAI in 2018, demonstrated how autoregressive language modeling with modest parameter counts (117 million) could achieve strong performance across diverse language tasks [33]. Its successor, GPT-2 (1.5 billion parameters), revealed emergent text generation abilities that approximated human writing quality in certain contexts [25]. GPT-3 (175 billion parameters) represented a quantum leap, exhibiting few-shot learning capabilities where the model could adapt to novel tasks through natural language instructions alone, without parameter updates [21].

Parallel developments have occurred across other modalities. In computer vision, Vision Transformers (ViT) adapted the transformer architecture for image recognition tasks, achieving state-of-the-art performance while utilizing uniform architectural principles across modalities [22]. CLIP (Contrastive Language-Image Pre-training) demonstrated how joint training on image-text pairs could produce visual representations with remarkable zero-shot generalization capabilities to previously unseen visual concepts [24]. More recent models like DALL-E [34] and Stable Diffusion [35] extended generative capabilities to visual synthesis from textual descriptions, while models such as Flamingo [36] and GPT-4 [37] have further advanced multimodal understanding and generation capabilities.

### 1.1.3 The Generic Nature

The defining characteristic of generalist foundation models lies in their domain-agnostic versatility, namely their capacity to operate effectively across diverse contexts without

specialized architectural modifications. This "generic" nature manifests through several key properties that distinguish foundation models from their predecessors.

Foundation models exhibit cross-domain transfer capabilities, wherein knowledge acquired in one context facilitates understanding in seemingly unrelated domains through abstract pattern recognition. For instance, language models trained primarily on textual data demonstrate surprising effectiveness at tasks involving logical reasoning [32], mathematical problem-solving [38], and even protein structure prediction [39]. This transfer capability suggests the emergence of abstract representational frameworks that capture fundamental principles transcending specific domains.

The scaling laws governing foundation model performance further emphasize their generic utility. Empirical analyses indicate that model capabilities improve predictably with increases in parameter count, dataset size, and computational resources [40]. Crucially, these improvements generalize across diverse tasks without domain-specific engineering, indicating that larger models inherently develop more flexible and comprehensive internal representations. Recent work has demonstrated that this scaling trajectory often yields emergent capabilities, referring to functionalities not present in smaller models that appear spontaneously beyond certain scaling thresholds [32], such as advanced reasoning, multilingual translation, and code generation.

The adaptability of foundation models represents another dimension of their generic nature. Through techniques like in-context learning [21], prompt engineering [41], and parameter-efficient fine-tuning [30], these models can rapidly adapt to novel tasks without extensive retraining or architectural modifications. This adaptability effectively

transforms foundation models into computational substrates that can be specialized through data and instruction rather than through fundamental redesign, significantly lowering the technical barriers for deployment across diverse applications.

The advantages of generalist foundation models relative to specialized architectures are multifaceted. First, they demonstrate superior sample efficiency, requiring fewer task-specific examples to achieve high performance through transfer of generalizable knowledge. Second, they reduce the engineering overhead associated with developing domain-specific architectures, instead centralizing development efforts on foundational capabilities that benefit numerous downstream applications simultaneously. Third, they exhibit enhanced robustness to distribution shifts between training and deployment environments by learning more abstract representations that capture invariant features across contexts [42].

Perhaps most significantly, foundation models manifest emergent capabilities beyond their explicit training objectives. GPT models [33, 25, 21, 37], while trained simply to predict the next token in a sequence, demonstrate abilities ranging from mathematical reasoning to creative writing; CLIP [24], trained to align images with textual descriptions, develops visual representations that generalize to novel classification tasks without specific training. These emergent properties suggest that scale and diversity in training data naturally induce the formation of generalizable computational primitives that can be composed to address previously unseen challenges.

The economic advantages of generalist foundation models are equally compelling. By amortizing the substantial costs of pretraining across numerous applications, they en-

able more efficient resource allocation than developing specialized models for each task. Additionally, their adaptability through relatively lightweight fine-tuning reduces computational requirements for deploying AI systems in new domains, democratizing access to advanced AI capabilities for organizations lacking extensive computational resources [43].

Despite these advantages, generalist foundation models present unique challenges. Their generic nature may sacrifice performance on specialized tasks compared to tailored architectures in cases where domain-specific inductive biases provide critical constraints [44]. Moreover, their "black-box" nature and massive parameter counts complicate interpretability, raising concerns regarding trustworthiness in high-stakes applications [45]. Nevertheless, the paradigm shift toward foundation models represents a fundamental reconceptualization of AI development, moving from narrowly-specialized systems toward adaptable computational substrates that can be shaped through data and instruction rather than explicit programming.

As these generalist foundation models continue to evolve, they increasingly serve as the technological infrastructure underlying numerous AI applications, establishing a new paradigm where domain adaptation occurs through interaction with pre-trained capabilities rather than architectural engineering. This transition fundamentally alters the relationship between model development and deployment, creating new opportunities for AI applications across domains including healthcare, which will be explored in subsequent sections.

## 1.2 Foundation Models for Healthcare

The emergence of foundation models has catalyzed a paradigm shift in AI applications for healthcare, offering unprecedented opportunities to address longstanding challenges in clinical decision support, medical knowledge extraction, and health outcome prediction. These models, characterized by their massive parameter counts and broad knowledge acquisition through self-supervised pretraining, present distinct methodological pathways for healthcare adaptation [46]. Two principal strategies have emerged for leveraging foundation models in biomedical and clinical contexts: (i) the development of domain-specific foundation models through specialized pretraining or adaptation on healthcare data, exemplified by models such as BioBERT [47], PubMedBERT [48], and Med-PaLM [49]; and (ii) the utilization of prompt engineering methodologies to elicit domain knowledge from general-purpose foundation models without specialized retraining, as demonstrated by approaches such as MedPrompt [44] and clinical prompting frameworks [50, 51].

The domain-specific approach prioritizes architectural and representational specialization through exposure to biomedical literature, clinical notes, or structured health records during pretraining or fine-tuning phases. This methodology emphasizes domain adaptation at the parameter level, reconfiguring model weights to better capture the statistical patterns and semantic relationships unique to healthcare contexts [52]. In contrast, the prompting approach preserves the original parametric configuration of general-purpose foundation models while formulating inputs that strategically guide model be-

havior toward healthcare-specific reasoning pathways [44]. This methodology leverages the implicit medical knowledge embedded within models trained on diverse internet-scale corpora, which often include substantial quantities of health-related information [53].

These complementary strategies represent fundamentally different philosophical orientations toward knowledge specialization: structural adaptation versus inferential guidance, each with distinct advantages, limitations, and resource requirements that significantly impact their applicability across various healthcare domains. The optimal approach depends on numerous factors including data availability, computational constraints, task complexity, and performance requirements for specific clinical applications.

### 1.2.1 Domain-Specific Approaches

The development of domain-specific foundation models for healthcare applications represents a systematic effort to tailor general architectural frameworks to the unique linguistic, conceptual, and relational characteristics of biomedical knowledge. This approach emerged from empirical observations that general-domain language models frequently underperform when directly applied to specialized medical tasks due to distributional shifts in vocabulary, syntactic structures, and semantic relationships [52, 54].

The methodological trajectory of domain-specific models began with BioBERT [47], which extended BERT's pretraining using PubMed abstracts and PMC articles, demonstrating significant performance improvements across biomedical named entity recogni-

tion, relation extraction, and question answering tasks. This approach was further refined through PubMedBERT [48], which employed pretraining from scratch exclusively on biomedical corpora rather than continued pretraining from general-domain checkpoints, yielding additional performance gains. The progression continued with more sophisticated architectures such as BioGPT [55], which adapted autoregressive transformer models to biomedical generation tasks, and specialized clinical models like ClinicalBERT [52] and GatorTron [56], which incorporate electronic health record (EHR) data to capture clinical language patterns.

More recently, multimodal domain-specific foundation models have emerged, integrating textual, imaging, genomic, and structured clinical data. Models such as Med-PaLM Multimodal (Med-PaLM M) [57] and LLaVA-Med [58] establish unified representational frameworks across multiple healthcare data modalities, facilitating cross-modal inference tasks such as generating radiology reports from medical images or predicting clinical outcomes from multimodal inputs. These developments reflect the heterogeneous nature of healthcare data and the importance of integrating diverse information streams for comprehensive medical reasoning.

The primary advantage of domain-specific foundation models lies in their representational precision for healthcare concepts and relationships. Through specialized pretraining, these models develop nuanced semantic embeddings that capture fine-grained distinctions between medical terms, recognize domain-specific abbreviations and acronyms, and model complex biomedical relationships that may be obscured in general-domain models [48]. Empirical evaluations consistently demonstrate superior performance on

specialized tasks such as medical entity recognition, clinical relation extraction, and biomedical question answering compared to general-domain alternatives [59, 54].

Additionally, domain-specific models often exhibit enhanced sample efficiency when fine-tuned for downstream tasks, requiring fewer labeled examples to achieve competitive performance [52]. This characteristic proves particularly valuable in healthcare contexts where annotated data is frequently scarce due to privacy constraints and annotation expertise requirements. The structural adaptation of domain-specific models also facilitates integration with existing healthcare taxonomies and ontologies such as SNOMED CT, ICD-10, and UMLS, enabling more coherent alignment with established medical knowledge frameworks [60].

Despite these advantages, domain-specific foundation models face substantial limitations. The most significant constraint involves computational and data requirements for pretraining or adaptation. Specialized biomedical corpora, while extensive, represent only a fraction of the textual data available for general-domain training, potentially limiting the models' linguistic flexibility and generalization capabilities [46]. The computational resources required for pretraining large-scale models from scratch on domain-specific data often exceed the capabilities of academic or clinical research environments, creating barriers to innovation and reproducibility.

Furthermore, domain-specific models risk overfitting to particular biomedical subdomains or literature distributions, potentially compromising performance when applied to emerging medical fields or underrepresented specialties [56]. The rapid evolution of medical knowledge exacerbates this challenge, as models trained on historical literature

may struggle to incorporate recent developments without continual updating. Finally, the increased specialization may come at the cost of reduced capabilities for commonsense reasoning and general world knowledge that often prove valuable for contextualizing medical information within broader patient circumstances [53].

### 1.2.2 Prompt Engineering-Based Approaches

The prompt engineering-based paradigm for healthcare applications leverages the implicit biomedical knowledge embedded within general-domain foundation models through strategically designed input formulations. This approach operates from the premise that large-scale pretraining on internet-scale corpora inherently captures substantial medical information, which can be accessed through appropriate prompting strategies without specialized architectural modifications [50]. The methodology has gained prominence following empirical demonstrations that state-of-the-art general foundation models such as GPT-4 [37] and PaLM [61] contain considerable medical knowledge accessible through carefully constructed prompts.

The conceptual foundation for prompting based healthcare applications emerged from in-context learning capabilities demonstrated in large language models [21], where performance on specialized tasks improved dramatically through demonstration examples incorporated directly in the input context. This capability has evolved into sophisticated prompting methodologies specific to medical applications, including chain-of-thought medical reasoning [49], retrieval-augmented clinical prompting [51], and structured

medical prompting frameworks such as MedPrompt [44].

MedPrompt exemplifies the advancement of healthcare-specific prompting strategies, incorporating techniques such as knowledge retrieval from authoritative medical sources, step-by-step reasoning chains that emulate clinical diagnostic processes, and self-verification mechanisms that evaluate the model's initial responses. These approaches have demonstrated remarkable effectiveness, enabling general-domain models to perform competitively on standardized medical examinations such as USMLE and medical question-answering benchmarks without domain-specific pretraining [50, 51].

The prompt engineering-based approach offers several distinct advantages compared to domain-specific pretraining. Most significantly, it provides substantially greater flexibility and adaptability, allowing rapid iteration of prompting strategies without the computational overhead of model retraining or fine-tuning. This adaptability proves particularly valuable in healthcare contexts where requirements may vary significantly across clinical specialties, institutional settings, or patient populations [49]. Prompt engineering allows for customization of model behavior at inference time rather than training time, facilitating more agile deployment in diverse healthcare environments.

Additionally, prompt engineering-based methodologies maintain access to the broad world knowledge and commonsense reasoning capabilities inherent in general-domain foundation models, which often prove valuable for contextualizing medical information within broader patient circumstances and societal factors. The integration of medical reasoning with general knowledge supports more holistic approaches to health that consider social determinants, patient preferences, and quality-of-life factors alongside

strictly biomedical considerations [51].

From a practical implementation perspective, prompt engineering-based approaches significantly reduce the computational resources required for deployment, eliminating the need for specialized model training infrastructure while leveraging existing general-purpose foundation model capabilities. This accessibility democratizes advanced AI capabilities for healthcare organizations with limited technical resources, potentially expanding the reach of AI-assisted healthcare solutions [62].

However, prompt engineering-based approaches face several substantial limitations. The most significant concern involves the reliability and precision of medical knowledge extracted through prompting. Unlike domain-specific models that systematically incorporate comprehensive biomedical literature during pretraining, general-domain models acquire medical knowledge incidentally and potentially inconsistently through internet-scale corpora [53]. This acquisition process may introduce biases, misconceptions, or outdated information prevalent in public sources, raising concerns about factual accuracy for clinical applications.

Furthermore, general-domain models typically lack specialized representations for complex medical terminology, relationships between biomedical entities, and domain-specific reasoning patterns that may be essential for advanced clinical applications [46]. The absence of structured medical knowledge representations complicates integration with established healthcare information systems, taxonomies, and ontologies that form the backbone of clinical informatics infrastructure.

Additionally, prompt engineering-based approaches exhibit greater sensitivity to in-

put formulation variations, potentially yielding inconsistent results across similar queries with superficial linguistic differences [62]. This variability introduces concerns regarding reliability in clinical settings where consistent performance is essential for establishing practitioner trust and ensuring patient safety.

## 1.3 Explainable and Interpretable AI

Perhaps the most critical challenge for healthcare applications involves the explainability and interpretability of deep learning and foundation models. The black-box nature of deep learning, characterized by billions or trillions of parameters with complex interdependencies, fundamentally conflicts with healthcare's requirements for transparent, accountable decision processes that clinicians and patients can understand and trust. This challenge is particularly acute in high-stakes clinical contexts where model recommendations may influence diagnostic or therapeutic decisions with significant consequences for patient outcomes.

This section presents a thorough scoping background and analysis of explainable and interpretable deep learning in healthcare AI. The term "eXplainable and Interpretable Artificial Intelligence" (XIAI) is introduced to distinguish XAI from IAI. Different models are further categorized based on their functionality (model-, input-, output-based) and scope (local, global).

### 1.3.1 XAI and IAI

In healthcare settings, where trustworthy and transparent decision-making is paramount, there is a growing need for interpretable or explainable models [63]. Both explainability and interpretability pose significant challenges, as understanding how NLP embeddings translate into deep learning decision-making mechanisms remains complex [64, 65]. Fortunately, recent research in explainable [66, 67] and interpretable [68, 69] deep learning in healthcare shows promise. Furthermore, the rise of large language models (LLMs) highlights the growing importance of evaluating which explainable and interpretable methods are most beneficial for healthcare, especially as data and model complexity increase over time.

Despite extensive research in the field, a consensus regarding precise definitions and clear distinctions between interpretability and explainability remains elusive [70, 71], with numerous studies using these terms interchangeably [72]. The term "explainable artificial intelligence (XAI)" [73] has emerged as an umbrella concept encompassing both interpretability and explainability approaches. Nevertheless, this terminological ambiguity has significant implications, resulting in inconsistent model taxonomies and imprecise reporting of model outcomes across the literature.

In response to the terminological ambiguity described above, this section introduces a unified terminology of eXplainable and Interpretable Artificial Intelligence, i.e., **XIAI**, based on previous studies [70, 71, 72, 74]. To resolve inconsistencies in how these concepts are applied within deep learning research [72], we employ Rudin's statistical

definitions [74] to delineate precise boundaries: (i) **IAI** *focuses on designing inherently interpretable models*; (ii) **XAI** *aims to provide post hoc model explanations*. This fundamental distinction facilitates a systematic taxonomy of XIAI methodologies across three paradigms: model-, input-, and output-based approaches. To improve clarity and understanding, XIAI is further grouped based on their scope (local, global): **local XIAI** *yields insights derived from particular inputs, whereas* **global XIAI** *grants a wider understanding based on the entire predictive mechanism of the model* [63]. By carefully investigating XIAI in healthcare AI, our chapter aims to provide insights into the scientific and clinical impact that can potentially be important for XIAI democratization in healthcare AI.

### 1.3.2 XIAI Paradigms

We introduce three distinct XIAI categories: model-, input- and output-based methods. These definitions are conceptualized based on whether an XIAI method relies on internal/external modules (model) to perform XIAI, measures how the input features affect model decisions or explains/interprets model behaviour through analyzing prediction outcomes, respectively. Table 1.1 presents prevalent XIAI methods in healthcare, and their categories and scopes [75].

**Model-based XIAI methods** focuses on describing how deep learning functions through the use of internal or external modules, such as SHAP [76], LIME [77], and t-SNE [78], offering important advantages by providing both global and local deep learning interpretability options. These methods are designed to be transparent and accessible as

| XIAI category | XIAI methods | Explainable/Interpretable | Local/Global |
|---|---|---|---|
| Model-based | Causal graph | Interpretable | Global |
| | Logistic regression-based parametric predictor | Interpretable | Global |
| | Opinion aggregator | Interpretable | Global |
| | Case-based reasoning | Interpretable | Global |
| | Interactive classification | Interpretable | Global |
| | LIME | Explainable | Local |
| | MAXi | Explainable | Local |
| | SHAP | Explainable | Local |
| | SKET X | Explainable | Local |
| | STEP | Explainable | Local |
| | t-SNE | Explainable | Local |
| Input-based | Feature importance | Interpretable | Local |
| | Knowledge base/graph | Interpretable | Local |
| Output-based | Attention | Interpretable | Local |
| | Evidence-based | Explainable | Local |
| | Sentiment intensity score | Explainable | Local |

Table 1.1: Explainable and interpretable artificial intelligence (XIAI) methods, their category, and scope

ready-to-use entities thus, potentially being able to democratize XIAI tools for a wide audience [67]. Intuitive visualizations can to some extent elucidate complex model predictions, enhancing trust and facilitating the communication of findings. However, they also have drawbacks, such as the instability of methods like LIME, which is prone to minor data variations that can affect XAI reliability [74, 79]. Moreover, relying on separate modules for XIAI may be subject to misinterpretations, in case these modules have inherent biases or limitations. Hence, despite their ease of use, they require technical

expertise to evaluate any biases or implementation incompatibilities.

**Input-based XIAI methods** allow for deep learning model interpretation by leveraging specific input feature importance [80, 81, 82] and medical KG [83, 84, 85], to discern how different inputs influence model decisions. These methods are intuitive and accessible to computer scientists and AI researchers, due to their relative ease of being combined with deep learning architectures. In that context, the incorporation of "important features" or medical KGs into AI can enrich models with domain-specific insights, which in turn allows them to interpret even complex medical concepts that are common in clinical practice [83].

Input methods based on feature importance requires the incorporation of techniques that are not readily transparent to non-technical medical professionals (end-users). Hence, it may not always be straightforward for end-users to understand "how an important feature was derived and/or evaluated" [80, 86, 87]. The effectiveness of KG methods relies on medical professionals with expertise and capacity to develop comprehensive medical KG, posing a challenge if such expertise is scarce. Integrating medical KG into deep learning presents challenges in terms of ontology construction as well as knowledge extraction [68], which are labor-intensive techniques.

**Output-based XIAI methods**, are important to interpret deep learning outputs and can offer computational insights through mechanisms like attention [88]. Output-based methods focus on explaining/interpreting deep learning models by uncovering how internal computations within deep learning converge to output decisions, which can be mainly useful to computer scientists/ modellers. Nevertheless, attention mechanism-

based IAI faces debates [89, 90, 91, 92, 93], as high attention weights do not necessarily linearly correlate with model predictions [64]. This can lead to ambiguity while emphasizes the need for further research on IAI methods and their evaluation.

## 1.4 Research Objectives

Despite above advancements of deep learning, the deployment of foundation models from general domains to healthcare presents substantial methodological and paradigmatic challenges [94, 45, 95, 96]. The transition from general to healthcare-specific contexts requires addressing fundamental differences and challenges in (i) representation learning, (ii) domain knowledge, (iii) data characteristics and (iv) explainability and interpretability.

In general-domain foundation models, representation learning is typically optimised for broad semantic understanding across diverse, large-scale datasets such as natural language corpora or generic image collections. In healthcare, however, the nature of the input data and the semantics of the task require representations that capture clinically meaningful, fine-grained, and often subtle features [47, 81, 60]. For example, in medical imaging, disease-specific patterns may be imperceptible in the latent spaces learned from non-clinical data, while in clinical text, domain-specific terminology, abbreviations, and context-dependent meanings require tailored embeddings [97]. Bridging this gap requires developing methods to adapt or retrain representation spaces so that they capture the physiological, pathological, and procedural nuances critical for reliable healthcare

decision-making.

Foundation models from general domains typically lack embedded clinical domain knowledge, which is crucial for meaningful interpretation and decision support in healthcare. Clinical reasoning relies not only on recognising patterns but also on understanding causal relationships, disease progression, treatment effects, and medical guidelines [98]. This knowledge is often formalised in ontologies such as SNOMED CT, UMLS, or ICD-10 [99], and is implicitly embedded in clinicians' diagnostic processes [60]. Without this domain grounding, models risk making statistically plausible but clinically implausible inferences, undermining trust and utility. Incorporating domain knowledge into foundation models, whether through knowledge graphs, structured annotations, or domain-constrained pre-training, presents both methodological and computational challenges. Achieving this integration is critical for ensuring that the model's outputs align with established medical reasoning and can be validated against clinical standards.

Healthcare data are distinctive, entailing high variability, scarcity, sensitivity, and heterogeneity [100, 101]. Clinical datasets often suffer from limited size, restrictive access due to privacy regulations, and non-standard formats. In some cases, models trained on public biomedical corpora (e.g., PubMed [59], MIMICs [99]) are evaluated on small or narrowly scoped datasets, undermining generalizability. Data heterogeneity arises from differences in patient populations, imaging protocols, language variations, and care settings. For example, mental health datasets may underrepresent marginalized groups, leading to biased outcomes [102]. These challenges hinder model transferability and fairness. Addressing these issues requires novel approaches for data augmentation, bias

mitigation, and privacy-preserving learning, ensuring that adapted foundation models can perform robustly across real-world clinical settings.

These challenges necessitate novel frameworks that can effectively bridge the gap between general-domain AI advancements and the specialized requirements of healthcare applications [103, 104]. This thesis identifies and addresses above critical challenges that represent significant barrier to effective the application and methodological transition of deep learning from general to domain-specific. Firstly, the investigation confronts the methodological limitations in modeling multi-label semantic relationships within supervised contrastive learning frameworks, with particular emphasis on automated International Classification of Diseases (ICD) coding which is a multi-label classification task characterized by extreme-scale label spaces and pronounced long-tailed distributions that remain inadequately addressed by conventional methodologies [105, 106]. Secondly, complex multi-hop reasoning across the spectrum from general to biomedical domains remains an intricate challenge in knowledge fusion [44], such as how to optimally integrate prior knowledge embedded within foundation models, including domain-specific knowledge, to enhance inferential capabilities. Thirdly, data insufficiency and distributional imbalance in sensitive contexts remain persistent challenges. Data augmentation (DA) offers an effective approach; however, developing novel augmentation strategies that advance beyond generalized content moderation toward domain-specific techniques remains difficult. These techniques must be tailored to address the nuanced requirements of mental health applications, where contextual sensitivity is paramount.

Through systematic examination of these interconnected challenges, this research contributes substantive theoretical advancements and pragmatic implementation strategies that facilitate the responsible and efficacious deployment of deep learning, and explainable & interpretable AI from general domain to healthcare.

## 1.5 Contributions of This Study

This study aims to systematically investigate the methodological and paradigmatic transition of deep learning and explainable & interpretable AI from general domain to healthcare. The contributions are summarized as follows:

- We introduce multi-label relations in multi-label supervised contrastive learning (MSCL) and propose a novel contrastive loss function, termed Similarity-Dissimilarity Loss, which dynamically re-weights based on the computed similarity and dissimilarity factors between positive samples and anchors, guided by multi-label relations. Furthermore, We establish the theoretical foundations of our approach through rigorous mathematical analysis, demonstrating both the formal derivation, and the upper and lower bounds of the weighting factor. Our method is applied from multi-label classification to automatic medical coding. This work is available at [107, 108].

- We propose a Prompting Explicit and Implicit knowledge (PEI) framework, which employs CoT prompt-based learning to bridge explicit and implicit knowledge,

aligning with human reading process for multi-hop QA. PEI leverages CoT prompts to elicit implicit knowledge from LMs within the input context, while integrating question type information to boost model performance. Moreover, we propose two training paradigms to PEI, and extend our framework on biomedical domain QA to further explore the fusion and relation of explicit and implicit biomedical knowledge via employing biomedical LMs to invoke biomedical implicit knowledge and analyze the consistency of the domain knowledge fusion. This work is available at [41, 109].

- We introduce a lexical-based imbalanced data augmentation (LIDA) for content moderation, which an easy-to-implement and interpretable DA method that strategically leverages sensitive lexicons by incorporating them into negative samples to transform these instances into positive examples. Through this mechanism, LIDA facilitates the creation of balanced datasets, thus mitigating skewed distribution challenges. Furthermore, we extend the application of our method to the mental healthcare domain. This work is available at [110].

## 1.6 Thesis Outline

The sections of this thesis are organized as follows: Chapter 1 provides an overview of background, motivation, objectives and contribution in this research; Chapter 2 presents Similarity-Dissimilarity Loss for multi-label supervised contrastive learning; Chapter 3 describe PEI framework for multi-hop QA in general and biomedical domains; Chapter

4 introduce LIDA in areas from content moderation to mental health; finally, Chapter 5 give a conclusion for this thesis.

# Chapter 2

# Similarity-Dissimilarity Loss for Multi-label Supervised Contrastive Learning

Supervised contrastive learning has achieved remarkable success by leveraging label information; however, determining positive samples in multi-label scenarios remains a critical challenge. In multi-label supervised contrastive learning (MSCL), relations among multi-label samples are not yet fully defined, leading to ambiguity in identifying positive samples and formulating contrastive loss functions to construct the representation space. To address these challenges, in this chapter, we: (i) first define five distinct multi-label relations in MSCL to systematically identify positive samples, (ii) introduce a novel Similarity-Dissimilarity Loss that dynamically re-weights samples through comput-

ing the similarity and dissimilarity factors between positive samples and given anchors based on multi-label relations, and (iii) further provide theoretical grounded proof for our method through rigorous mathematical analysis that supports the formulation and effectiveness of the proposed loss function. We conduct the experiments across both image and text modalities, and extend the evaluation to automated medical coding. The results demonstrate that our method consistently outperforms baselines in a comprehensive evaluation, confirming its effectiveness and robustness. Code is available at: https://github.com/guangminghuang/similarity-dissimilarity-loss.

## 2.1 Introduction

Multi-label classification presents significant challenges due to its inherent label correlations, extreme and sparse label spaces, and long-tailed distributions. For instance, in the International Classification of Diseases (ICD) [106, 105], the presence of one label (e.g., "Pneumococcal pneumonia") may increase the probability of co-occurring labels (e.g., "fever" or "cough"). Furthermore, multi-label datasets frequently exhibit long-tailed distributions, where a small subset of labels occurs with high frequency while the majority appear rarely. This imbalance typically results in models that perform adequately on common labels but underperform on infrequent ones [111, 112]. Additionally, the number of potential label combinations increases exponentially with the number of labels, resulting in heightened computational complexity and substantial memory requirements.

Supervised contrastive learning effectively utilizes label information to yield prom-

ising results in single-label scenarios [113]. However, identifying positive samples in multi-label supervised contrastive learning (MSCL) remains a challenge. For example, consider a set of images containing cats and puppies, wherein an anchor image depicts a cat; in the single-label paradigm, positive and negative instances can be unambiguously delineated based on their corresponding taxonomic annotations. Conversely, MSCL introduces inherent classification ambiguity when determining whether an image containing both cats and puppies should be designated as a positive or negative sample in relation to the anchor.

A critical question arises: *Should a sample be considered positive when its label set partially overlaps with or exactly matches that of the anchor?* Currently, three principal strategies exist for identifying positive samples in multi-label scenarios [114]: (i) *ALL* considers only samples with an exactly matching label set as positive; (ii) *ANY* identifies samples with any overlapping class with the anchor as positive, and (iii) *MulSupCon* [114] conceptually aligns with the *ANY* approach but treats each label independently, thereby generating multiple distinct positive sets for individual anchor samples.

However, these methods have inherent limitations, since previous research has overlooked the complicated multi-label relations among samples in MSCL. As illustrated in Figure 2.1, we introduce five distinct set relations among samples to facilitate a more comprehensive identification of positive sets. The *ALL* strategy exclusively considers relation $R2$ while disregarding the potential contributions of $R3$, $R4$ and $R5$. Furthermore, long-tailed distributions, when tail samples serve as anchors, the *ALL* strategy's requirement for exact label matches significantly impedes these tail anchors from identifying ad-

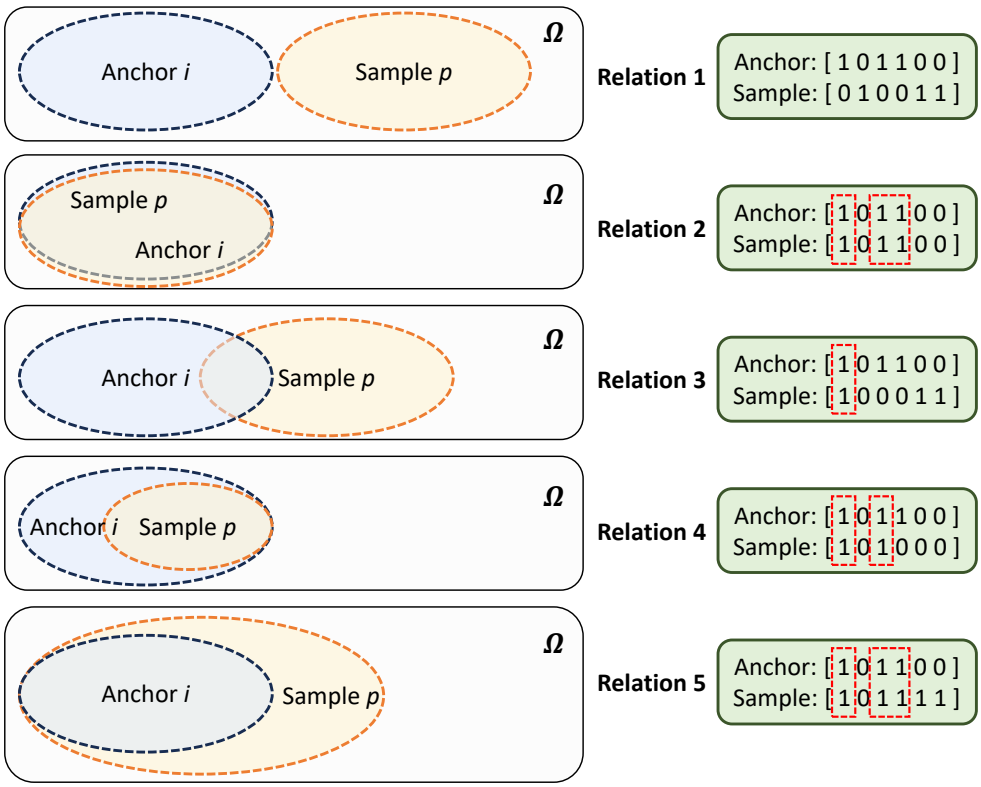


Figure 2.1: Five distinct multi-label relations between samples and a given anchor. $\Omega$ denotes a universe that contains all label entities. Here is an example with five different relations between sample $p$ and anchor $i$, where the labels are represented as one-hot vectors.

equate positive samples within a limited batch size, potentially degenerating the method to unsupervised contrastive learning in extreme scenarios [26, 115, 111]. Conversely, both *ANY* and *MulSupCon* approaches treat relations $R2$, $R3$, $R4$, and $R5$ identically with equivalent weights in contrastive loss functions, which constitutes a suboptimal approach given the inherent differences among these relations. A detailed mathematical analysis of these three methods is presented in Section 2.3.

To address these issues, we define multi-label relations and introduce a novel contrastive loss function. Our contributions are summarized as follows:

1. To the best of our knowledge, we are the first to define multi-label relations in MSCL, which facilitates the identification of complex relations in multi-label scenarios.

2. We introduce similarity and dissimilarity concepts in multi-label scenarios and propose a novel contrastive loss function, termed Similarity-Dissimilarity Loss, which dynamically re-weights based on the computed similarity and dissimilarity factors between positive samples and anchors, guided by multi-label relations.

3. We establish the theoretical foundations of our approach through rigorous mathematical analysis, demonstrating both the formal derivation, and the upper and lower bounds of the weighting factor.

4. We conduct the experiments across both image and text modalities, and extend the evaluation to medical domain. The results demonstrate that our method consistently outperforms baselines in a comprehensive evaluation, confirming its effectiveness and robustness.

## 2.2 Related Work

### 2.2.1 Multi-label Supervised Contrastive Learning

Contrastive learning aims to learn a representation of data such that similar instances are close together in the representation space, while dissimilar instances are far apart. Com-

pared to self-supervised contrastive learning, such as SimCLR [26] and MoCo [115], Khosla et al. [113] proposed supervised contrastive learning, which fully leverages class annotation information to enhance representations within the contrastive learning framework. Recent studies have extended supervised contrastive learning from single-label to multi-label scenarios by exploiting the additional information inherent in multi-label tasks. Zhang et al. [116] proposed a hierarchical multi-label representation learning framework specifically designed to utilize comprehensive label information while preserving hierarchical inter-class relationships.

In subsequent research, Zhang and Wu [114] developed Multi-Label Supervised Contrastive Learning (MulSupCon), featuring a novel contrastive objective function that expands the positive sample set based on label overlap proportions. Similarly, the Jaccard Similarity Probability Contrastive Loss (JSPCL) [117] employed the Jaccard coefficient [118] to calculate label similarity between instances, sharing conceptual foundations with MulSupCon [114] and MSC loss [119] that those approaches primarily focus on similarity only, but ignoring dissimilarity.

Despite these advancements, the intricate relationships and dependencies between multi-label samples have yet to be fully elucidated. To address this gap, we introduce multi-label relations and formalize the concepts of similarity and dissimilarity. Inspired by the idea of re-weighting of logit adjustment [120], focal loss [121] and class-balanced loss [122], we leverage the similarity and dissimilarity factors to re-weight the contrastive loss, thereby enhancing discriminative power in multi-label scenarios.

### 2.2.2 Automated Medical Coding

Medical coding refers to the process of translating free-text medical documents into predefined codes based on standardized coding systems, such as the International Classification of Diseases (ICD), which is the most widely adopted medical coding system globally [106]. The ICD system employs a tree-like hierarchical structure, also known as a medical ontology, to maintain the functional and structural integrity of the classification scheme. This standardization process enhances the accuracy and consistency of medical information, facilitating its use across various medical services and insurance claims processing [104].

However, medical coding remains a costly manual process that is prone to errors [105]. Several primary challenges hinder effective medical text processing and automated coding [104, 106, 105]: (i) the presence of noisy and lengthy clinical notes, (ii) the high dimensionality of medical codes, and (iii) the imbalanced distribution of medical codes in Electronic Health Record (EHR) systems, commonly referred to as the long-tail phenomenon.

AMC is predominantly conceptualized as a multi-label classification problem [123]. Mullenbach et al. [124] introduced the CAML model for explainable prediction of medical codes from clinical text, which has become a benchmark for ICD coding on the MIMIC datasets [100, 101]. Building on this foundation, numerous researchers [125, 126] have investigated more sophisticated methodologies that incorporate external knowledge. Yuan et al. [127] extended this approach with their MSMN, which integrates

synonym descriptions of ICD codes. Furthermore, several studies have focused on enhancing code representations through ICD relation data. Notably, Vu et al. [128] proposed the Label Attention Model (LAAT) and Nguyen et al. [129] developed the TwoStage framework, both of which predict codes hierarchically to optimize final classification outcomes.

The emergence of pre-trained language models (PLMs) has catalyzed research efforts to leverage these advanced architectures for improving ICD coding performance [125, 130]. These approaches enhance coding accuracy through the implementation of contextual prompt-based prediction techniques and hierarchical encoding methodologies. Nevertheless, despite their promising results, these PLM-based approaches continue to face significant challenges regarding computational efficiency and resource requirements.

## 2.3 Methods

In this section, we establish the preliminary notation and adhere to the conventions established in [113] to maintain consistency throughout our analysis. Subsequently, we examine the limitations of the *ALL*, *ANY*, and *MulSupCon* strategies and their corresponding loss functions. We then introduce our formulation of multi-label relations and present the Similarity-Dissimilarity Loss for MSCL. Furthermore, we provide a rigorous mathematical analysis to establish the theoretical foundations of the proposed methodology.

### 2.3.1 Preliminaries

Given a batch of $N$ randomly sample/label pairs, $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1,\dots,N}$, where $\boldsymbol{x}_i$ denotes the $i$-th sample and $\boldsymbol{y}_i$ its corresponding labels. Here, $\boldsymbol{y}_i = \{y_i^{(l)}\}_{l=1,\dots,L}$ represents the multi-labels of sample $i$, where $y_i^{(l)}$ denotes the $l$-th label of sample $i$ and $L$ is the total number of labels for sample $i$. After data augmentation, the training batch consists of $2N$ pairs, $\{\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{y}}_j\}_{j=1,\dots,2N}$, where $\tilde{\boldsymbol{x}}_{2i}$ and $\tilde{\boldsymbol{x}}_{2i-1}$ are two random augmentations of $\boldsymbol{x}_i$ ($i = 1, \dots, N$) and $\tilde{\boldsymbol{y}}_{2i-1} = \tilde{\boldsymbol{y}}_{2i} = \boldsymbol{y}_i$. For brevity, we refer to this collection of $2N$ augmented samples as a "batch" [113].

### 2.3.2 Multi-label Supervised Contrastive Loss

In MSCL, the formulation of supervised contrastive loss varies depending on the strategies employed for determining positive samples relative to a given anchor. Let $i \in \mathcal{I} = \{1, \dots, 2N\}$ denote the index of an arbitrary augmented sample. For the *ALL* strategy, the positive set is defined as follows:

$$\mathcal{P}(i) = \{p \in \mathcal{A}(i) | \forall p, \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\} \tag{2.1}$$

where $\mathcal{A}(i) \equiv I \setminus \{i\}$ [1].

Subsequently, the positive set for the *ANY* strategy is defined as follows:

$$\mathcal{P}(i) = \{p \in \mathcal{A}(i) | \forall p, \tilde{\boldsymbol{y}}_p \cap \tilde{\boldsymbol{y}}_i \neq \varnothing\} \tag{2.2}$$

[1] In contrastive learning, sample $i$ is the anchor and is supposed to be excluded out of positive sets.

In MSCL, the form of contrastive loss function for *ALL* and *ANY* is identical. For each anchor $i$, the loss function is formulated as follows:

$$\mathcal{L}_i = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \tag{2.3}$$

Here, $\tau \in \mathbb{R}^+$ represents a positive scalar temperature parameter [26], while $\boldsymbol{z}_k = Proj(Enc(\tilde{\boldsymbol{x}}_k)) \in \mathbb{R}^{D_P}$ denotes the projected encoded representation [113].

For a given batch of samples, the loss function is formulated as:

$$\mathcal{L} = \sum_{i \in I} \mathcal{L}_i \tag{2.4}$$

Zhang et la [114] propose an approach that considers each label $\tilde{y}_i^{(l)}$ independently, forming multiple positive sets for a given anchor sample $i$. For each label $\tilde{y}_i^{(l)} \in \tilde{\boldsymbol{y}}_i$, the positive set for the *MulSupCon* is defined as:

$$\mathcal{P}(i) = \{p \in \mathcal{A}(i) | \forall p, \tilde{y}_p^{(l)} \in \tilde{\boldsymbol{y}}_i\} \tag{2.5}$$

For each anchor $i$, the multi-label supervised contrastive loss for MulSupCon is represented as follows [114]:

$$\mathcal{L}_i^{\text{mul}} = \sum_{\tilde{y}_p^{(l)} \in \tilde{\boldsymbol{y}}_i} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \tag{2.6}$$

For a given batch of samples, the loss function is formulated as:

$$\mathcal{L}^{\text{mul}} = \frac{1}{\sum_i |\tilde{\boldsymbol{y}}_i|} \sum_{i \in I} \mathcal{L}_i^{\text{mul}} \tag{2.7}$$

### 2.3.3 Multi-label Relations

As illustrated in Figure 2.1, we denote each *Relation* as $R$, where, e.g., $R1$ stands for *Relation 1*. The subscripted notation $p_j$ signifies that sample $p$ corresponds to the $j$-th relation.

Let $\Omega$ denote a universal set containing all possible label entities. For any anchor $i$ and sample $p$, let $\mathcal{S}$ and $\mathcal{T}$ represent their respective label sets. The five fundamental multi-label relations are defined as follows:

$$R1 : \mathcal{S} \cap \mathcal{T} = \varnothing \tag{2.8}$$

$$R2 : \mathcal{S} = \mathcal{T} \tag{2.9}$$

$$R3 : \mathcal{S} \cap \mathcal{T} \neq \varnothing, \mathcal{S} \nsubseteq \mathcal{T}, \mathcal{T} \nsubseteq \mathcal{S} \tag{2.10}$$

$$R4 : \mathcal{S} \supsetneq \mathcal{T} \tag{2.11}$$

$$R5 : \mathcal{S} \subsetneq \mathcal{T} \tag{2.12}$$

Based on these relational definitions, we present a theoretical analysis of the limitations inherent in the *ALL*, *ANY*, and *MulSupCon* methods, illustrated via an example in Figure 2.1.

In the *ALL* method, the optimization process aims to align with the mean representation of samples sharing identical label sets [114]. As the example that is demonstrated in Figure 2.1, for a given anchor $i$, the positive set of *ALL* is:

$$\mathcal{P}(i) = \{p_2\}$$

In the *ALL* method, the sample $p_j$ in $R2$ is designated as positive sample, while those in relations $R3$, $R4$ and $R5$ are excluded from consideration. Specifically, despite their semantic similarity to anchor $i$ that those overlap labels, the feature representations of samples $p_j$ where $j \in 3, 4, 5$ are forced away from the anchor in the embedding space, as they are treated as negative examples in the contrastive learning paradigm. Consequently, the restricted size of the positive set $|\mathcal{P}(i)|$ results in a mean representation susceptible to statistical variance. Furthermore, the *ALL* method may inadvertently treat semantically related samples as negative instances in certain scenarios.

**Lemma 1.** *(Vector Similarity Under Label Equivalence). Let $i$ be an anchor and $p$ be any sample in the feature space, where $\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_p \in \mathbb{R}^d$ denote their respective label vectors. If $\tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i$, then under the contrastive learning framework [26], their corresponding projected representations $\boldsymbol{z}_i, \boldsymbol{z}_p \in \mathbb{R}^m$ satisfy $\boldsymbol{z}_i \simeq \boldsymbol{z}_p$.*

As per *ANY*'s definition, the positive set of the example in Figure 2.1 is:

$$\mathcal{P}(i) = \{p_2, p_3, p_4, p_5\}$$

By applying Lemma 1, the corresponding loss terms in Eq. (2.3) for samples in different relations exhibit approximate equality:

$$\mathcal{L}(R2) \approx \mathcal{L}(R3) \approx \mathcal{L}(R4) \approx \mathcal{L}(R5)^{2}$$

It is evident that $R2$, $R3$, $R4$ and $R5$ represent fundamentally distinct relations, each characterized by different labels and semantic information. However, the *ANY* method fails to differentiate these subtle label hierarchies, introducing substantial semantic ambiguity. Moreover, in scenarios where samples predominantly share common classes, the averaging mechanism disproportionately emphasizes these shared classes while diminishing the significance of distinctive features [114].

The *MulSupCon* method employs a positive sample identification mechanism analogous to *ANY*, samples $p_j$, where $j \in 3, 4, 5$ are designated as positive instances. However, *MulSupCon* distinguishes itself by evaluating each label individually and forming multiple positive sets for a single anchor sample. This approach aggregates positive samples based on the number of overlapping labels between the positive samples and the anchor, thereby expanding the space of positive sets:

$$\mathcal{P}(i) = \{p_2, p_2, p_2, p_3, p_4, p_4, p_5, p_5, p_5\}$$

[2]The approximation notation is used instead of equality due to vector similarity in Lemma 1 and the inherent uncertainty in deep learning's non-linear transformations.

Subsequently, the loss for $p_j$ in Eq. (2.6) are as follows by Lemma 1:

$$\mathcal{L}(R2) \approx \mathcal{L}(R5) \neq \mathcal{L}(R3) \neq \mathcal{L}(R4)$$

For this example (see Figure 2.1), the *MulSupCon* successfully discriminates $R3$ and $R4$ from $R2$ and $R5$; however, it fails to establish a distinction between $R2$ and $R5$. This limitation arises primarily because *MulSupCon* exclusively considers the overlapping regions (*Similarity* [3]) between anchor $i$ and sample $p$ (i.e., The intersection of sets $\mathcal{S}$ and $\mathcal{T}$), while disregarding the complementary non-intersecting domains (*Dissimilarity* [4]). That is to say, the similarity between positive samples and anchors is considers, but not yet dissimilarity, which is one of critical information for representation learning in MSCL.

Leveraging the proposed multi-label relations, our theoretical analysis systematically elucidates the limitations of existing methods and establishes a rigorous foundation for investigating the profound exploration of concepts of similarity and dissimilarity, and the design of contrastive loss function.

### 2.3.4 Similarity-Dissimilarity Loss

To address the aforementioned challenges, we introduce the concepts of similarity and dissimilarity based on set-theoretic relations: (i) As depicted in Figure 2.1, *Similarity* represents the intersection of sets (i.e., $\mathcal{S} \cap \mathcal{T}$), and (ii) we define *Dissimilarity* as the set difference between $\mathcal{T}$ and the intersection $\mathcal{S} \cap \mathcal{T}$ with respect to sample $p$ (i.e., $\mathcal{T} - \mathcal{S} \cap \mathcal{T}$).

[3]The definition of *Similarity* is introduced in Section 2.3.4
[4]The definition of *Dissimilarity* is introduced in Section 2.3.4

For each anchor $i$, we formulate the Similarity-Dissimilarity Loss as:

$$\mathcal{L}_i^{\text{our}} = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\mathcal{K}_{i,p}^s \mathcal{K}_{i,p}^d \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \tag{2.13}$$

Here, we define $\mathcal{K}_{i,p}^s$ and $\mathcal{K}_{i,p}^d$ that quantify the *Similarity* and *Dissimilarity* factors for a given anchor $i$ and a positive sample $p$, respectively. These factors are formally defined as follows:

$$\mathcal{K}_{i,p}^s = \frac{|\tilde{\boldsymbol{y}}_p^s|}{|\tilde{\boldsymbol{y}}_i|} = \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S}|} \tag{2.14}$$

and

$$\mathcal{K}_{i,p}^d = \frac{1}{1 + |\tilde{\boldsymbol{y}}_p^d|} = \frac{1}{1 + |\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|} \tag{2.15}$$

where we define the following set-theoretic quantities:

- $|\tilde{\boldsymbol{y}}_i| = |\mathcal{S}|$ denotes the cardinality of the label space $\tilde{\boldsymbol{y}}_i$.
- $|\tilde{\boldsymbol{y}}_p^s| = |\mathcal{S} \cap \mathcal{T}|$ measures the cardinality of the intersection of sets $\mathcal{S}$ and $\mathcal{T}$.
- $|\tilde{\boldsymbol{y}}_p^d| = |\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|$ represents the cardinality of the relative complement with respect to sample $p$.

The product of $\mathcal{K}_{i,p}^s$ and $\mathcal{K}_{i,p}^d$ is termed as *similarity-dissimilarity factor*. Moreover, the following relation holds:

$$|\tilde{\boldsymbol{y}}_p^d| = |\tilde{\boldsymbol{y}}_p| - |\tilde{\boldsymbol{y}}_p^s| \geq 0 \tag{2.16}$$

where $|\tilde{\boldsymbol{y}}_p|$ represents the cardinality of the label space associated with sample $p$.

Specifically, the Similarity-Dissimilarity Loss Loss reduces to Eq. (2.3), when the following conditions are simultaneously satisfied:

$$\begin{cases} |\tilde{\boldsymbol{y}}_i| = |\tilde{\boldsymbol{y}}_p^s| \\ |\tilde{\boldsymbol{y}}_p^d| = 0 \end{cases} \tag{2.17}$$

Accordingly, our proposed loss function constitutes a generalized form of the basic supervised contrastive loss (see Eq. (2.3)). In particular, Eq. (2.3) represents a particular case of the Similarity-Dissimilarity Loss. Moreover, our contrastive loss unifies both single-label and multi-label supervised contrastive loss functions within a comprehensive form and paradigm.

### 2.3.5 Case Analysis

Let us examine the behavior of our loss function through a detailed analysis of five distinct relational cases illustrated in Figure 2.1. Consider the following sequences of cardinalities:

$$\begin{cases} |\tilde{\boldsymbol{y}}_{p_j}^s| = \{0, 3, 1, 2, 3\}_{j=1,2,3,4,5} \\ |\tilde{\boldsymbol{y}}_{p_j}^d| = \{3, 0, 2, 0, 2\}_{j=1,2,3,4,5} \end{cases}$$

Applying these values to Eq. (2.14) and (2.15), we obtain:

$$\begin{cases} \mathcal{K}^s_{i,p} = \{0, 1, \frac{1}{3}, \frac{2}{3}, 1\} \\ \mathcal{K}^d_{i,p} = \{\frac{1}{4}, 1, \frac{1}{3}, 1, \frac{1}{3}\} \end{cases}$$

Consequently, the product of these measures yields:

$$\mathcal{K}^s_{i,p}\mathcal{K}^d_{i,p} = \{0, 1, \frac{1}{9}, \frac{2}{3}, \frac{1}{3}\}$$

When evaluating Eq. (2.13), these distinct relations ($R2$ through $R5$) generate unique loss values, establishing the following inequalities:

$$\mathcal{L}(R2) \neq \mathcal{L}(R3) \neq \mathcal{L}(R4) \neq \mathcal{L}(R5)$$

The proposed loss function effectively discriminates among the five distinct relations through a principled re-weighting mechanism, as formulated in Eq. (2.13), (2.14), and (2.15), comparing to existing methods in MSCL.

Furthermore, in contrast to *MulSupCon*, the Similarity-Dissimilarity Loss preserves the cardinality of positive sets while maintaining computational efficiency, as it requires no additional computational overhead.

### 2.3.6 Theoretical Analysis

The proposed loss function incorporates a weighting mechanism through the product of factors $\mathcal{K}^s_{i,p}$ and $\mathcal{K}^d_{i,p}$. By construction, the *similarity-dissimilarity factor* $\mathcal{K}^s_{i,p}\mathcal{K}^d_{i,p}$ is constrained to the closed interval $[0, 1]$ across all possible relational configurations. Hence, it is written as:

$$\mathcal{K}^s_{i,p}\mathcal{K}^d_{i,p} \in [0, 1] \tag{2.18}$$

For notational conciseness, let us denote the product of Similarity and Dissimilarity factors across the five relations as $\{\mathcal{K}^s_m\mathcal{K}^d_m\}_{m=1,2,3,4,5}$.

**Theorem 1.** *Let $\mathcal{K}^s_m$ and $\mathcal{K}^d_m$ be the Similarity and Dissimilarity operators, respectively, as defined in Eq. (2.14) and (2.15). For the case $m = 1$, their product vanishes:*

$$\mathcal{K}^s_m\mathcal{K}^d_m = 0, \quad \textit{when } m = 1 \tag{2.19}$$

*Proof.* Consider the case where $m = 1$. By definition, we have $\mathcal{S} \cap \mathcal{T} = \varnothing$. This implies:

$$|\tilde{\boldsymbol{y}}^s_p| = |\mathcal{S} \cap \mathcal{T}| = |\varnothing| = 0$$
$$\therefore \mathcal{K}^s_1 = \frac{|\tilde{\boldsymbol{y}}^s_p|}{|\tilde{\boldsymbol{y}}_i|} = \frac{0}{|\tilde{\boldsymbol{y}}_i|} = 0$$

Since $\mathcal{K}_1^s = 0$ and $\mathcal{K}_1^d$ is finite by construction, we conclude:

$$\mathcal{K}_1^s \mathcal{K}_1^d = 0 \cdot \mathcal{K}_1^d = 0 \tag{2.20}$$

□

**Theorem 2.** *Consider the Similarity operator $\mathcal{K}_m^s$ and Dissimilarity operator $\mathcal{K}_m^d$ as defined in Eq. (2.14) and (2.15). For the case $m = 2$, their product equals unity:*

$$\mathcal{K}_m^s \mathcal{K}_m^d = 1, \quad \textit{when } m = 2 \tag{2.21}$$

*Proof.* Consider the case where $m = 2$. By hypothesis, we have $\mathcal{S} = \mathcal{T}$. This equality implies:

$$\mathcal{K}_2^s = \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S}|} = \frac{|\mathcal{S}|}{|\mathcal{S}|} = 1$$
$$\mathcal{K}_2^d = \frac{1}{1 + |\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|} = \frac{1}{1 + |\varnothing|} = 1$$

where we have used the fact that $\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T}) = \varnothing$ when $\mathcal{S} = \mathcal{T}$. Thus, we conclude:

$$\mathcal{K}_2^s \mathcal{K}_2^d = 1 \cdot 1 = 1 \tag{2.22}$$

□

**Theorem 3.** *Let $\mathcal{K}_m^s$ and $\mathcal{K}_m^d$ be the Similarity and Dissimilarity operators as defined in Eq.*

*(2.14) and (2.15), respectively. For $m \in \{3, 4, 5\}$, their product is strictly bounded between 0 and 1:*

$$0 < \mathcal{K}_m^s \mathcal{K}_m^d < 1 \tag{2.23}$$

*Proof.* Consider $m \in \{3, 4, 5\}$. Under these cases, we have:

$$\mathcal{S} \cap \mathcal{T} \neq \varnothing \tag{2.24}$$

$$\mathcal{S} \neq \mathcal{T} \tag{2.25}$$

We first establish the strict positivity. Given $|\mathcal{S}| > 0$ and conditions (2.24)-(2.25), we have:

$$\mathcal{K}_m^s = \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S}|} > 0$$
$$\mathcal{K}_m^d = \frac{1}{1 + |\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|} > 0$$

For the upper bound, we consider three cases:

Case 1 ($m = 3$): By Eq. (2.10), we have three conditions: $\mathcal{S} \cap \mathcal{T} \neq \varnothing$, $\mathcal{S} \nsubseteq \mathcal{T}$, and $\mathcal{T} \nsubseteq \mathcal{S}$. These conditions lead to:

$$|\mathcal{S} \cap \mathcal{T}| < |\mathcal{S}| \implies \mathcal{K}_3^s < 1$$
$$|\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})| > 0 \implies \mathcal{K}_3^d < 1$$

Therefore, $\mathcal{K}_3^s \mathcal{K}_3^d < 1$.

Case 2 ($m = 4$): When $m = 4$, by Eq. (2.11), we have $\mathcal{S} \supseteq \mathcal{T}$. This subset relation implies:

$$\mathcal{K}_4^s = \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S}|} = \frac{|\mathcal{T}|}{|\mathcal{S}|} < 1$$
$$\mathcal{K}_4^d = \frac{1}{1 + |\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|} = \frac{1}{1 + |\varnothing|} = 1$$

where the strict inequality $\mathcal{K}_4^s < 1$ follows from $|\mathcal{T}| < |\mathcal{S}|$ (since $\mathcal{S} \supsetneq \mathcal{T}$), and $\mathcal{K}_4^d = 1$ is a consequence of $\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T}) = \varnothing$ when $\mathcal{S} \supsetneq \mathcal{T}$. Therefore:

$$\mathcal{K}_4^s \mathcal{K}_4^d = \mathcal{K}_4^s \cdot 1 = \mathcal{K}_4^s < 1$$

Case 3 ($m = 5$): When $m = 5$, by Eq. (2.12), we have $\mathcal{S} \subsetneq \mathcal{T}$. This subset relation implies:

$$\mathcal{K}_5^s = \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S}|} = \frac{|\mathcal{S}|}{|\mathcal{S}|} = 1$$
$$\mathcal{K}_5^d = \frac{1}{1 + |\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|} = \frac{1}{1 + |\mathcal{T} \setminus \mathcal{S}|} < 1$$

where $\mathcal{K}_5^s = 1$ follows from the fact that $\mathcal{S} \cap \mathcal{T} = \mathcal{S}$ when $\mathcal{S} \subsetneq \mathcal{T}$. The strict inequality

$\mathcal{K}_5^d < 1$ holds because:

$$\begin{aligned} \mathcal{S} \subsetneq \mathcal{T} &\implies |\mathcal{T} \setminus \mathcal{S}| > 0 \\ &\implies 1 + |\mathcal{T} \setminus \mathcal{S}| > 1 \\ &\implies \frac{1}{1 + |\mathcal{T} \setminus \mathcal{S}|} < 1 \end{aligned}$$

Therefore, we can conclude:

$$\mathcal{K}_5^s \mathcal{K}_5^d = 1 \cdot \mathcal{K}_5^d = \mathcal{K}_5^d < 1$$

Combining the results with Propositions 1 and 2, we obtain complete ordering for all $m \in \{1, 2, 3, 4, 5\}$. The products $\mathcal{K}_m^s \mathcal{K}_m^d$ satisfy:

$$0 = \mathcal{K}_1^s \mathcal{K}_1^d < \mathcal{K}_m^s \mathcal{K}_m^d < \mathcal{K}_2^s \mathcal{K}_2^d = 1, \quad m \in \{3, 4, 5\} \tag{2.26}$$

□

Based on Theorem 1, 2, and 3, the product of weighting factors $\mathcal{K}_{i,p}^s$ and $\mathcal{K}_{i,p}^d$ is bounded within the interval $[0, 1]$, which aligns with fundamental principles of loss functions and set-theoretic relations. The non-negative lower bound adheres to the essential property of loss functions being strictly positive [1]. Given that our proposed loss function generalizes the supervised contrastive loss [113] and incorporates multi-label relation definitions, the upper bound naturally equals 1. Furthermore, this mathematical

framework demonstrates that our proposed contrastive loss can dynamically adjust the weighting factor within $[0, 1]$, effectively differentiating sample features with rigorous mathematical justification for both the formulation and efficacy of the loss function.

**Theorem 4.** *Let $i \in \mathcal{I}$ be a fixed anchor sample, and let $p_3, p_4 \in \mathcal{P}(i)$ be positive samples corresponding to relations $R_3$ and $R_4$, respectively. Suppose their label spaces satisfy the cardinality constraint:*

$$|\tilde{\boldsymbol{y}}_{p_3}| = |\tilde{\boldsymbol{y}}_{p_4}| \tag{2.27}$$

*Then, the product of similarity and dissimilarity operators satisfies the strict inequality:*

$$\mathcal{K}_4^s \mathcal{K}_4^d > \mathcal{K}_3^s \mathcal{K}_3^d \tag{2.28}$$

*Proof.* Let us establish the strict inequality $\mathcal{K}_4^s \mathcal{K}_4^d > \mathcal{K}_3^s \mathcal{K}_3^d$ through direct comparison. From definitions (2.14) and (2.15), we have:

$$\mathcal{K}_4^s \mathcal{K}_4^d = \frac{|\tilde{\boldsymbol{y}}_{p_4}|}{|\tilde{\boldsymbol{y}}_i|} > \mathcal{K}_3^s \mathcal{K}_3^d = \frac{|\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}_{p_3}^d|}{|\tilde{\boldsymbol{y}}_i|} \cdot \frac{1}{1 + |\tilde{\boldsymbol{y}}_{p_3}^d|}$$

$\Rightarrow$

$$\frac{|\tilde{\boldsymbol{y}}_{p_4}|(1 + |\tilde{\boldsymbol{y}}_{p_3}^d|)}{|\tilde{\boldsymbol{y}}_i|(1 + |\tilde{\boldsymbol{y}}_{p_3}^d|)} > \frac{|\tilde{\boldsymbol{y}}_{p_4} - \tilde{\boldsymbol{y}}_{p_3}^d|}{|\tilde{\boldsymbol{y}}_i|(1 + |\tilde{\boldsymbol{y}}_{p_3}^d|)}$$

$\Rightarrow$

$$|\tilde{\boldsymbol{y}}_{p_4}|(1 + |\tilde{\boldsymbol{y}}_{p_3}^d|) > |\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}_{p_3}^d|$$

By the cardinality constraint (2.27) in the theorem:

$$|\tilde{\boldsymbol{y}}_{p_3}|(1+|\tilde{\boldsymbol{y}}^d_{p_3}|) > |\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}^d_{p_3}|$$

where the strict inequality follows from the fact that for any positive real numbers $a, b > 0$:

$$a(1+b) > a - b$$

This inequality holds trivially, thereby establishing the original claim $\mathcal{K}^s_4\mathcal{K}^d_4 > \mathcal{K}^s_3\mathcal{K}^d_3$. □

**Theorem 5.** *Let $i \in \mathcal{I}$ be a fixed anchor sample, and let $p_3, p_5 \in \mathcal{P}(i)$ be positive samples corresponding to relations $R_3$ and $R_5$, respectively. Suppose:*

$$|\tilde{\boldsymbol{y}}^d_{p_5}| \leq |\tilde{\boldsymbol{y}}^d_{p_3}| \tag{2.29}$$

*Then, the product of Similarity and Dissimilarity operators satisfies the strict inequality:*

$$\mathcal{K}^s_5\mathcal{K}^d_5 > \mathcal{K}^s_3\mathcal{K}^d_3 \tag{2.30}$$

*Proof.* From definitions (2.14) and (2.15), we have:

$$\mathcal{K}^s_3\mathcal{K}^d_3 = \frac{|\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}^d_{p_3}|}{|\tilde{\boldsymbol{y}}_i|} \cdot \frac{1}{1+|\tilde{\boldsymbol{y}}^d_{p_3}|}$$

$$\mathcal{K}^s_5\mathcal{K}^d_5 = \frac{1}{1+|\tilde{\boldsymbol{y}}^d_{p_5}|}$$

Taking the ratio:

$$\frac{\mathcal{K}_5^s \mathcal{K}_5^d}{\mathcal{K}_3^s \mathcal{K}_3^d} = \frac{|\tilde{\boldsymbol{y}}_i|(1+|\tilde{\boldsymbol{y}}_{p_3}^d|)}{|\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}_{p_3}^d|(1+|\tilde{\boldsymbol{y}}_{p_5}^d|)}$$

By the properties of cardinality and set difference:

$$|\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}_{p_3}^d| \leq |\tilde{\boldsymbol{y}}_i|$$

Given the constraint (2.29), $|\tilde{\boldsymbol{y}}_{p_5}^d| \leq |\tilde{\boldsymbol{y}}_{p_3}^d|$, we have:

$$\frac{|\tilde{\boldsymbol{y}}_i|(1+|\tilde{\boldsymbol{y}}_{p_3}^d|)}{|\tilde{\boldsymbol{y}}_{p_3} - \tilde{\boldsymbol{y}}_{p_3}^d|(1+|\tilde{\boldsymbol{y}}_{p_5}^d|)} > 1$$

Therefore, $\mathcal{K}_5^s \mathcal{K}_5^d > \mathcal{K}_3^s \mathcal{K}_3^d$. □

Theorem 4 and 5 establish strict dominance relations between relation types $R3$, $R4$, and $R5$, demonstrating that $\mathcal{K}_4^s \mathcal{K}_4^d > \mathcal{K}_3^s \mathcal{K}_3^d$ when $|\tilde{\boldsymbol{y}}_{p_3}| = |\tilde{\boldsymbol{y}}_{p_4}|$ and $\mathcal{K}_5^s \mathcal{K}_5^d > \mathcal{K}_3^s \mathcal{K}_3^d$ when $|\tilde{\boldsymbol{y}}_{p_5}^d| \leq |\tilde{\boldsymbol{y}}_{p_3}^d|$. These inequalities, proved through careful mathematical derivation using set cardinality properties and fundamental principles of real analysis, reveal a well-defined hierarchical structure in the weighting factors. This hierarchical relations ensures that our loss function appropriately modulates the contribution of different relation types during the learning process, providing theoretical guarantees for the effectiveness of our proposed approach in capturing complex relations within the data.

Our theoretical analysis establishes a comprehensive mathematical foundation for the proposed loss function through five key theorems. These theoretical guarantees, derived through rigorous set-theoretic analysis, demonstrate that our loss function effectively

modulates the contribution of different relation types while maintaining proper mathematical bounds, thereby providing a solid theoretical foundation for its application in multi-label contrastive learning.

## 2.4 Experiments

The previous theoretical analysis establishes a rigorous mathematical foundation for our method, validating both the formulation and efficacy of the proposed loss function. In our experimental evaluation, we focus on assessing the effectiveness and robustness of Similarity-Dissimilarity Loss in the MSCL framework. Rather than comparing with other multi-label classification approaches, we emphasize that Similarity-Dissimilarity Loss primarily aims to enable models to learn generalizable and transferable features that enhance performance across diverse downstream tasks (classification, detection, and clustering) instead of optimizing for any specific task. We conduct the experiments to compare Similarity-Dissimilarity Loss with current contrastive loss functions (*ALL*, *ANY*, and *MulSupCon*) in a comprehensive evaluation, considering: (i) Data modality: image and text data; (ii) Domain-specific: general text data (AAPD) and medical domain (MIMIC III and IV); (iii) Data distribution: full setting (extreme long-tailed distribution) and top-50 frequent labels setting; (iv) ICD code versions: ICD-9 and ICD-10, and (v) Models: ResNet-50, RoBERTa-based, Llama-3.1-8B, and PLM-ICD.

| Dataset | Train | Val | Test | Total # labels | Avg # labels |
|---|---|---|---|---|---|
| MS-COCO | 82.0k | 20.2k | 20.2k | 80 | 2.9 |
| PASCAL | 5.0k | 2.5k | 2.5k | 20 | 1.5 |
| NUS-WIDE | 125.4k | 41.9k | 41.9k | 81 | 2.4 |
| AAPD | 37.8k | 6.7k | 11.3k | 54 | 2.4 |
| MIMIC-III-Full | 47,723 | 1,631 | 3,372 | 8,692 | 15.7 |
| MIMIC-III-50 | 8,066 | 1,573 | 1,729 | 50 | 5.7 |
| MIMIC-IV-ICD9-Full | 188,533 | 7,110 | 13,709 | 11,145 | 13.4 |
| MIMIC-IV-ICD9-50 | 170,664 | 6,406 | 12,405 | 50 | 4.7 |
| MIMIC-IV-ICD10-Full | 110,442 | 4,017 | 7,851 | 25,230 | 16.1 |
| MIMIC-IV-ICD10-50 | 104,077 | 3,805 | 7,368 | 50 | 5.4 |

Table 2.1: Statistics of datasets.

### 2.4.1 Datasets and Metrics

To rigorously evaluate the efficacy of our proposed loss function, we conducted comprehensive experiments across three distinct data modalities: visual data, textual data, and specialized medical corpus data (MIMIC datasets). The MIMIC datasets are particularly noteworthy for their exceptionally large label space and pronounced long-tailed distributions [75]. This long-tailed characteristic, which is especially prevalent in multi-label classification scenarios, facilitates a robust assessment of the performance of our loss function across heterogeneous data distributions. Comprehensive statistical analyses of all experimental datasets are presented in Table 2.1.

- **MS-COCO** (Microsoft Common Objects in Context) [131] consists of over 330,000 images annotated across 80 object categories, providing rich semantic information for object detection, segmentation, and captioning tasks that has significantly ad-

vanced computer vision research since its introduction by Microsoft.

- **PASCAL VOC** [132] contains 9,963 natural images with standardized annotations spanning 20 object categories, enabling rigorous evaluation of classification, detection, and segmentation algorithms in computer vision.

- **NUS-WIDE** [133] is a large-scale web image collection comprising approximately 269,000 Flickr images annotated with 81 concept categories and user tags, widely used as a benchmark for multi-label image classification.

- **AAPD** (Arxiv Academic Paper Dataset) [134] is a text corpus containing 55,840 scientific paper abstracts from arXiv with multi-label annotations across various subject categories, designed specifically for benchmarking multi-label text classification and document categorization algorithms.

- **MIMIC-III** [5] [100] includes records labeled with expert-annotated ICD-9 codes, which identify diagnoses and procedures. We adhere to the same splits as in previous works [124], employing two settings: MIMIC-III-Full, which includes all ICD-9 codes, and MIMIC-III-50, which includes only the 50 most frequent codes.

- **MIMIC-IV** [6] [101] contains records annotated with both ICD-9 and ICD-10 codes, where each code is subdivided into sub-codes that often capture specific circumstantial details. we follow prior studies [99] and utilize four settings: MIMIC-IV-ICD9-Full, MIMIC-IV-ICD9-50, MIMIC-IV-ICD10-Full, and MIMIC-IV-ICD10-50.

[5] We have been granted access to MIMIC-III Clinical Database (v1.4)
[6] We have been granted access to MIMIC-IV (v2.2)

**Metrics**. Consistent with prior research [124, 99], we report macro/micro-AUC, macro/micro-F1, and precision at K (P@$\mathcal{K}$) metrics on MIMIC datasets, where $\mathcal{K} = \{5, 8\}$ for different settings. Moreover, micro/macro-F1 and mAP are used for image datasets following [115, 114, 119].

### 2.4.2 Baseline Loss Functions and Encoders

This study evaluates the proposed Similarity-Dissimilarity Loss in comparison with three established baseline loss functions: (i) *ALL*, (ii) *ANY*, and (iii) *MulSupCon* [114], all implemented within the MSCL framework.

For experimental evaluation, we employ modality-specific encoder architectures tailored to each data type. For image data, ResNet-50 [7] serves as the encoder architecture, consistent with established methodologies [115, 26, 114]. For textual data, we utilize pre-trained large language models (LLMs), specifically RoBERTa-base [135] and Llama-3.1-8B [136] with Low-Rank Adaptation (LoRA) [30]. Additionally, for the specialized task of ICD coding on MIMIC datasets, we implement PLM-ICD [137], a model specifically designed for ICD coding using LLMs.

### 2.4.3 Implementation Details

Within the MSCL framework, we implement a two-phase training method as established by Khosla [113]: (i) encoder training, wherein the model learns to generate vector representations that maximize similarity between instances of the same class while distin-

guishing them from other classes; and (ii) classifier training, which utilizes the trained encoder and freeze it to train the classifier.

In the representation training, we use a standard cosine learning rate scheduler with a $0.05$ warm-up period and set the temperature $\tau = 0.07$. The projection head comprises two MLP layers with ReLU activation function and employs contrastive loss function for the training, where the projected representation $\boldsymbol{z}_k = Proj(Enc(\tilde{\boldsymbol{x}}_k)) \in \mathbb{R}^{D_P}$. Here $h = Enc(\tilde{\boldsymbol{x}}_k)$ denotes the encoded feature vectors and the projection dimension $D_P = 256$. For subsequent classifier training, the projection head is removed, a linear layer is appended to the frozen encoder, and binary cross-entropy (BCE) loss is utilized for optimization.

For image data, we employ ResNet-50 using stochastic gradient descent (SGD) with momentum. The input images are set up at a resolution of $224 \times 224$ pixels. For text data, RoBERTa-base and Llama-3.1-8B serve as backbone encoders implemented via Huggingface platform [138]. RoBERTa configures with a dropout rate of $0.1$ and AdamW optimizer with a weight decay of $0.01$, exempting bias and LayerNorm from weight decay. Compared with full-parameter fine-tuning, we employ LoRA [30] to efficiently fine-tune large model Llama. LoRA configures with the low-rank dimension $r = 16$, scaling factor $\alpha = 32$ and dropout as $0.1$. There is no KV cache to save memory during training. To enhance computational efficiency, BFloat16 precision is used for the training. The hyperparameters and detailed configuration are shown our code [7].

[7] https://github.com/guangminghuang/similarity-dissimilarity-loss

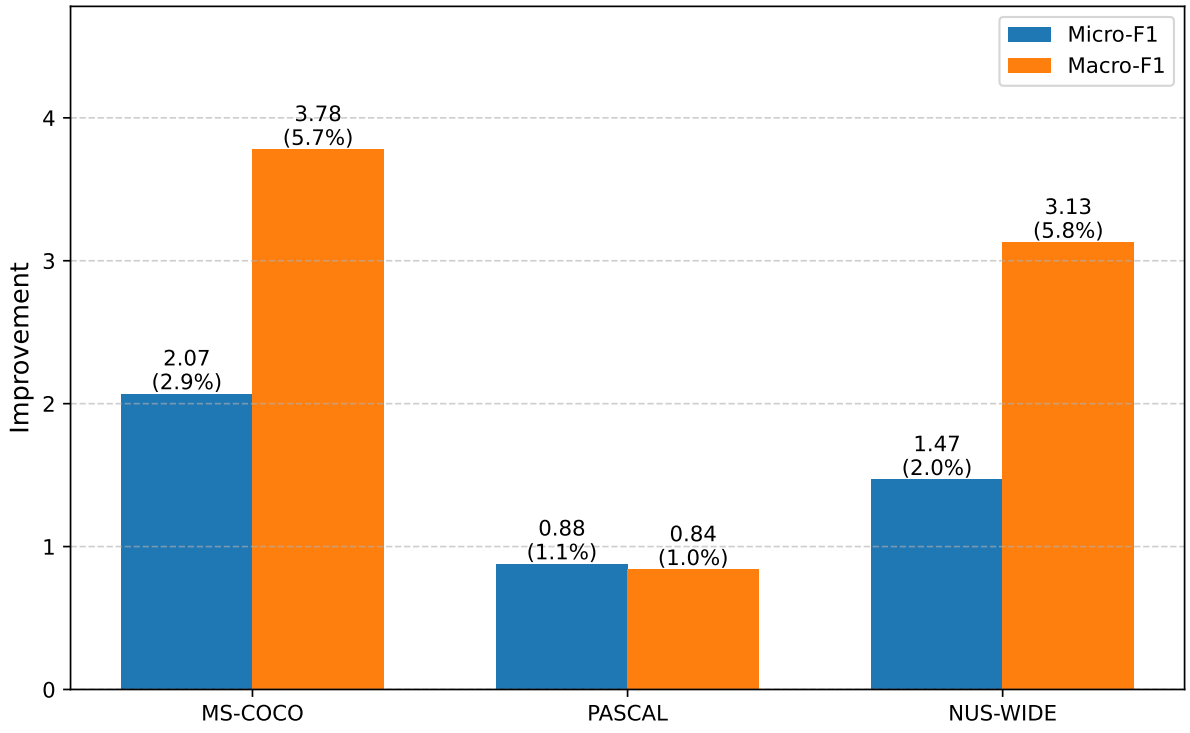


Figure 2.2: Comparison of performance improvements between Similarity-Dissimilarity Loss and *MulSupCon*.

## 2.5 Results and Analysis

### 2.5.1 Evaluation on Image

The experimental results in Table 2.2 demonstrate that our proposed loss function outperforms baselines across all metrics, including micro-F1, macro-F1, and mAP, on all image datasets (MS-COCO, PASCAL, and NUS-WIDE). Compared to MulSulCon, Similarity-Dissimilarity Loss achieves significant improvements of $2.07/3.78/1.51$ in Micro-F1, Macro-F1, and mAP on MS-COCO and $1.47/3.13/4.27$ on NUS-WIDE.

Figure 2.2 illustrates the comparison between Similarity-Dissimilarity Loss and *MulSup-*

| Method | MS-COCO | | | PASCAL | | | NUS-WIDE | | |
|---|---|---|---|---|---|---|---|---|---|
| | Micro-F1 | Macro-F1 | mAP | Micro-F1 | Macro-F1 | mAP | Micro-F1 | Macro-F1 | mAP |
| ALL | 68.93 | 63.32 | 64.11 | 82.53 | 79.87 | 79.32 | 70.25 | 52.84 | 51.35 |
| ANY | 64.80 | 57.37 | 56.90 | 82.31 | 79.65 | 79.15 | 68.42 | 50.65 | 49.28 |
| MulSupCon | 71.33 | 66.25 | 67.69 | 82.75 | 80.26 | 79.58 | 71.88 | 54.36 | 52.47 |
| Ours | **73.40** | **70.03** | **69.20** | **83.63** | **81.10** | **79.75** | **73.35** | **57.49** | **56.74** |

Table 2.2: Results on image datasets. We compare our method with baselines (*ALL*, *ANY* and *MulSupCon*) on MS-COCO, PASCAL and NUS-WIDE. The mAP standards for mean Average Precision.

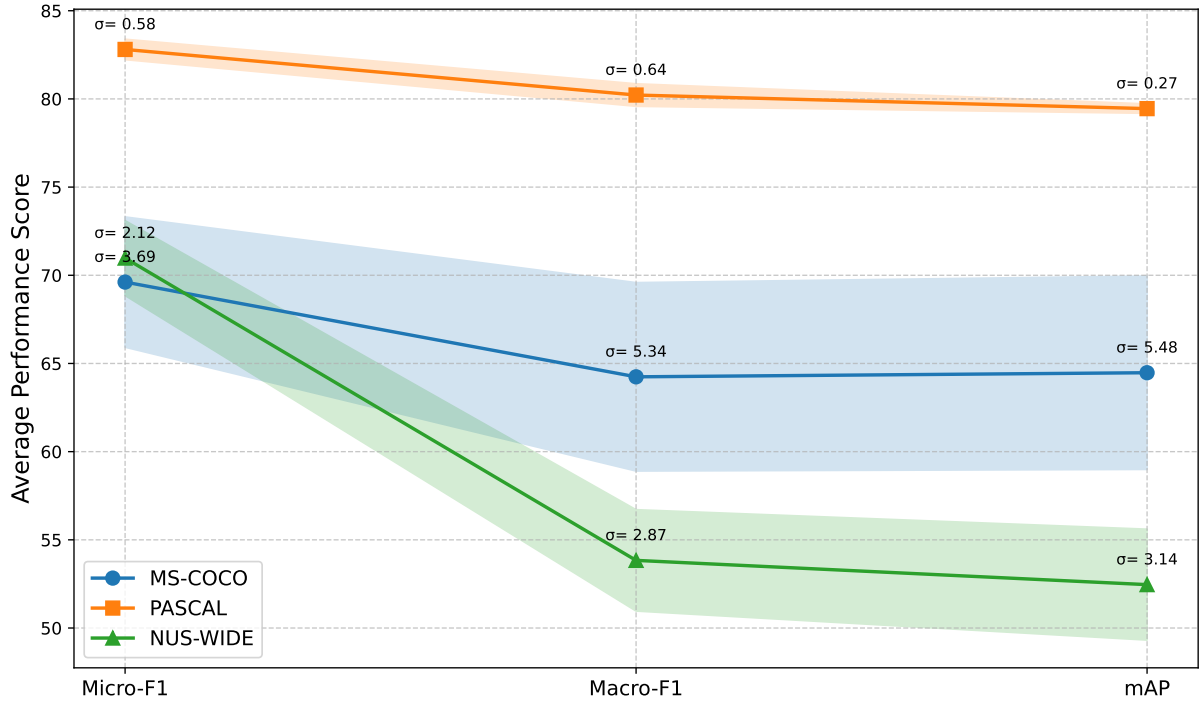


Figure 2.3: Comparison standard deviation of image datasets on micro-F1, macro-F1 and mAP metrics.

*Con* as measured by micro- and macro-F1 metrics. The results indicate that our method yields substantially greater improvements in macro-F1 compared to micro-F1 across all image datasets. Specifically, macro-F1 increases by $5.7\%$ on MS-COCO and $5.8\%$ on NUS-WIDE, whereas micro-F1 exhibits more modest improvements of $2.9\%$ and $2.0\%$, respectively. Macro-F1 assigns equal importance to each class regardless of its frequency, rendering it particularly appropriate for evaluating performance on imbalanced datasets where minority class prediction accuracy is critical [1, 111]. In contrast, micro-F1 places more considerable weight on classes with more samples, making it more appropriate when larger classes should have a more potent influence on the overall score

[1, 121]. Multi-label classification inherently faces more pronounced challenges with long-tailed distributions than single-label classification due to exponential output space complexity, intricate label co-occurrence patterns, and high annotation costs [111]. The observed superior improvement in macro-F1 metrics provides compelling evidence that our method demonstrates exceptional efficacy in addressing long-tailed distribution challenges, a capability particularly crucial in multi-label scenarios.

However, on the PASCAL dataset, our method demonstrates mere marginal improvements, with gains of $0.88/0.84/0.17$ in micro/macro-F1/mAP, respectively. This limited enhancement can be attributed to the structural characteristics of PASCAL, wherein the average number of labels per instance is approximately $1.5$ (as detailed in Table 2.1), causing the task to approximate single-label classification, particularly when the batch size is limited [113]. Consequently, loss functions specifically designed for multi-label scenarios exert minimal influence on model performance under these conditions. As Audibert et al. [119] have demonstrated, the cardinality of the label space constitutes a significant determinant of model efficacy within MSCL .

Furthermore, Figure 2.3 reveals that the standard deviation across four methods for PASCAL equals $0.58/0.64/0.27$ in micro/macro-F1/mAP, which are considerably lower than the corresponding standard deviations observed for the MS-COCO and NUS-WIDE. This statistical finding suggests that the efficacy of specialized multi-label loss functions diminishes significantly when the average label cardinality per instance approaches $1$ in MSCL. This finding further corroborates our theoretical analysis and hypothesis in the Section 2.3, wherein Similarity-Dissimilarity Loss degenerates to single-label scenarios

(see Eq. (2.17)).

### 2.5.2 Evaluation on Text

We further evaluate our method on general text data, and the results demonstrate that our proposed loss function consistently surpasses baseline methods for both RoBERTa and Llama models across all metrics on the AAPD dataset (See Table 2.3). In contrast to the significant performance gains observed on image data, Similarity-Dissimilarity Loss achieves more modest enhancements of $0.90/1.79$ in micro/macro-F1 scores on RoBERTa, and $0.89/1.84$ on Llama. This attenuated performance differential can be attributed to the extensive knowledge already encoded within LLMs through their comprehensive pre-training paradigms [139].

Moreover, as illustrated in Figure 2.4, performance variations of contrastive loss functions for MSCL on both RoBERTa and Llama models are relatively minimal. Specifically, the standard deviations in micro-F1 are $0.80$ and $0.79$ on RoBERTa and Llama, respectively, while the corresponding standard deviations for macro-F1 metrics are $1.41$ and $1.42$. Unlike image classification in MSCL paradigm, performance improvements in text classification are predominantly attributable to the intrinsic representational capabilities of model architecture of LLMs. Consequently, while fine-tuning the pre-trained weights of LLMs during the contrastive learning phase can yield marginal performance improvements, this methodological approach demonstrates substantially greater efficacy for visual classification tasks compared to textual classification.

| Method | RoBERTa | | Llama | |
|---|---|---|---|---|
| | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 |
| ALL | 73.23 | 59.41 | 74.32 | 60.47 |
| ANY | 72.31 | 58.55 | 73.41 | 59.63 |
| MulSupCon | 73.64 | 60.52 | 74.72 | 61.58 |
| Ours | **74.54** | **62.31** | **75.61** | **63.42** |

Table 2.3: Results on AAPD Dataset. We compare our proposed Similarity-Dissimilarity Loss with baselines on general text data using RoBERTa-based and Llama-3.1-8B models.

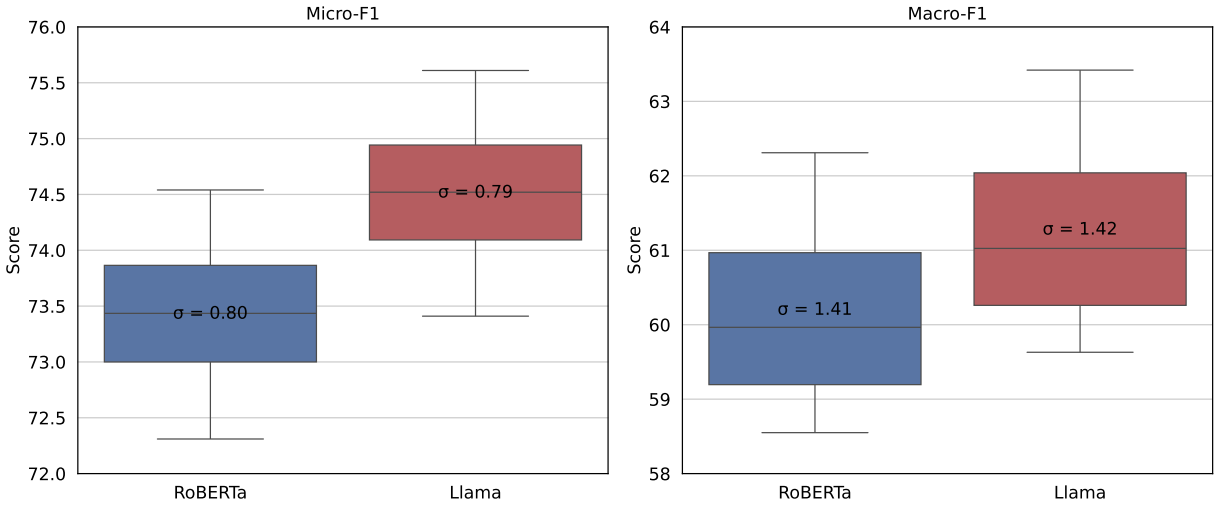


Figure 2.4: Comparison of RoBERTa and Llama across micro-F1 and macro-F1 on AAPD Dataset.

### 2.5.3 Evaluation on Medical Domain

We extend and evaluate our method on the medical domain, specifically for ICD coding. The results in Tables 2.4 and 2.4 demonstrate that our proposed loss function consistently surpasses baselines across all metrics in a comprehensive evaluation, considering: (i) Diverse data distribution: full setting (long-tailed distribution) and top-50 frequent labels setting; (ii) Model architectures: RoBERTa, LLaMA, and domain-specialized PLM-

ICD; and (iii) ICD code versions: ICD-9 and ICD-10. The consistent performance improvements observed across these multidimensional evaluation criteria provide substantial empirical evidence for the efficacy and generalizability of our proposed approach.

In the full setting, macro-F1 performance exhibits considerably lower compared to micro-F1, whereas the top-50 setting achieves approximately equal macro and micro-F1 scores. This disparity indicates that extreme long-tailed distributions remain challenging for both the MSCL framework and our method, despite the improvements achieved.

Table 2.4 reports that our method achieves superior results on MIMIC-IV-ICD9-Full compared to MIMIC-III-Full, despite both datasets employing identical ICD-9 coding standards. This marked performance differential can be attributed primarily to the more extensive training corpus available in MIMIC-IV-ICD9-Full (see in Table 2.1). While MIMIC-IV-ICD10-Full similarly comprises a substantial volume of clinical data, its considerably expanded label taxonomy introduces increased representational sparsity and presents additional computational and methodological challenges [99]. Moreover, the MIMIC-IV-ICD10-50 dataset demonstrates consistent performance metrics in this restricted setting, providing empirical evidence that label space dimensionality constitutes a critical determinant of model training efficacy.

Comparative analysis of model performance reveals that Llama significantly outperforms RoBERTa across evaluation metrics, a finding attributable to scaling laws of LLMs and the extensive knowledge and training corpus during the pre-training phase [40, 140]. Although LLMs demonstrate considerable efficacy in domain-specific applications [44], our results indicate that PLM-ICD consistently surpasses both RoBERTa

| Method | MIMIC-III-Full | | | | | MIMIC-IV-ICD9-Full | | | | | MIMIC-IV-ICD10-Full | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | F1 | | P@8 | AUC | | F1 | | P@8 | AUC | | F1 | | P@8 |
| | Macro | Micro | Macro | Micro | | Macro | Micro | Macro | Micro | | Macro | Micro | Macro | Micro | |
| **RoBERTa** | | | | | | | | | | | | | | | |
| ALL | 89.87 | 95.83 | 7.94 | 53.08 | 71.06 | 93.04 | 98.57 | 11.76 | 57.73 | 64.75 | 89.46 | 98.19 | 4.23 | 53.82 | 64.17 |
| ANY | 88.15 | 94.18 | 7.13 | 51.35 | 68.92 | 92.86 | 98.14 | 11.17 | 57.42 | 64.48 | 89.09 | 98.07 | 4.02 | 52.28 | 62.65 |
| MulSupCon | 90.37 | 96.38 | 8.64 | 54.16 | 71.24 | 93.87 | 99.34 | 12.83 | 58.67 | 65.89 | 90.53 | 98.74 | 4.56 | 54.09 | 65.46 |
| Ours | **90.78** | **96.67** | **9.19** | **54.63** | **71.38** | **94.13** | **99.36** | **13.08** | **58.85** | **66.29** | **90.68** | **98.86** | **4.72** | **54.89** | **66.07** |
| **Llama** | | | | | | | | | | | | | | | |
| ALL | 91.27 | 96.94 | 8.38 | 54.75 | 72.63 | 94.52 | 98.93 | 12.34 | 58.97 | 66.35 | 90.78 | 98.57 | 4.53 | 54.98 | 65.31 |
| ANY | 90.64 | 96.38 | 7.82 | 53.97 | 71.85 | 94.19 | 98.74 | 11.93 | 58.68 | 65.92 | 90.36 | 98.32 | 4.37 | 54.24 | 64.78 |
| MulSupCon | 91.68 | 97.23 | 8.79 | 55.36 | 72.94 | 94.87 | 99.42 | 12.96 | 59.35 | 66.73 | 91.15 | 98.97 | 4.72 | 55.29 | 66.16 |
| Ours | **91.93** | **97.57** | **9.26** | **55.87** | **73.28** | **95.14** | **99.58** | **13.37** | **59.69** | **67.14** | **91.38** | **99.15** | **4.94** | **55.67** | **66.59** |
| **PLM-ICD** | | | | | | | | | | | | | | | |
| ALL | 92.58 | 98.69 | 10.73 | 60.06 | 76.84 | 96.95 | 99.28 | 14.18 | 62.83 | 70.53 | 91.87 | 98.79 | 4.83 | 57.36 | 69.29 |
| ANY | 91.09 | 97.36 | 9.24 | 58.87 | 75.38 | 95.85 | 98.17 | 12.64 | 61.82 | 69.58 | 90.54 | 97.72 | 4.54 | 55.86 | 68.17 |
| MulSupCon | 93.46 | 99.13 | 11.68 | 61.42 | 77.65 | 97.86 | 99.32 | 14.47 | 64.23 | 71.97 | 92.83 | 99.38 | 5.43 | 58.15 | 70.19 |
| Ours | **94.47** | **99.43** | **12.46** | **62.34** | **78.42** | **98.47** | **99.59** | **15.04** | **64.95** | **72.95** | **93.75** | **99.57** | **5.74** | **58.76** | **70.79** |

Table 2.4: Results on MIMIC-III-Full, MIMIC-IV-ICD9-Full and MIMIC-IV-ICD10-Full test sets. RoBERTa, Llama and PLM-ICD are used as backbone encoder models.

| Method | MIMIC-III-50 | | | | | MIMIC-IV-ICD9-50 | | | | | MIMIC-IV-ICD10-50 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | | F1 | | P@5 | AUC | | F1 | | P@5 | AUC | | F1 | | P@5 |
| | Macro | Micro | Macro | Micro | | Macro | Micro | Macro | Micro | | Macro | Micro | Macro | Micro | |
| **RoBERTa** | | | | | | | | | | | | | | | |
| ALL | 87.73 | 90.57 | 57.38 | 61.84 | 61.29 | 93.84 | 94.46 | 67.63 | 72.24 | 60.92 | 91.43 | 93.52 | 64.86 | 67.65 | 60.07 |
| ANY | 87.36 | 89.42 | 56.25 | 60.83 | 60.32 | 93.37 | 93.73 | 67.39 | 71.97 | 60.24 | 90.06 | 92.03 | 64.09 | 66.54 | 58.08 |
| MulSupCon | 88.02 | 91.24 | 57.83 | 62.26 | 61.53 | 94.73 | 95.28 | 68.63 | 73.32 | 61.98 | 92.09 | 93.95 | 65.43 | 68.54 | 61.36 |
| Ours | **88.86** | **93.14** | **60.03** | **62.43** | **62.06** | **94.92** | **95.43** | **69.05** | **73.54** | **62.23** | **92.43** | **94.34** | **66.07** | **70.24** | **62.09** |
| **Llama** | | | | | | | | | | | | | | | |
| ALL | 88.93 | 91.67 | 60.32 | 64.58 | 62.87 | 94.32 | 95.28 | 69.18 | 73.72 | 61.42 | 92.35 | 94.57 | 66.38 | 69.83 | 61.75 |
| ANY | 88.57 | 91.09 | 59.72 | 64.03 | 62.19 | 94.05 | 94.85 | 68.79 | 73.19 | 61.07 | 91.89 | 93.97 | 65.82 | 69.09 | 61.12 |
| MulSupCon | 89.21 | 92.13 | 60.85 | 65.12 | 63.23 | 94.74 | 95.83 | 69.76 | 74.46 | 61.95 | 92.73 | 95.12 | 66.85 | 70.57 | 62.43 |
| Ours | **89.54** | **92.49** | **61.32** | **65.67** | **63.69** | **94.97** | **96.07** | **70.21** | **74.87** | **62.32** | **93.07** | **95.53** | **67.23** | **71.23** | **62.81** |
| **PLM-ICD** | | | | | | | | | | | | | | | |
| ALL | 90.13 | 93.02 | 65.18 | 69.43 | 65.26 | 95.18 | 96.42 | 71.31 | 75.83 | 62.45 | 93.53 | 95.97 | 68.96 | 73.14 | 64.52 |
| ANY | 89.03 | 92.07 | 63.73 | 68.14 | 63.84 | 93.73 | 95.34 | 70.23 | 74.43 | 61.42 | 92.27 | 94.42 | 67.95 | 71.83 | 63.17 |
| MulSupCon | 91.23 | 94.04 | 66.17 | 70.32 | 66.42 | 96.32 | 97.63 | 72.64 | 76.93 | 63.83 | 94.43 | 97.32 | 70.15 | 74.23 | 65.63 |
| Ours | **91.82** | **94.63** | **67.15** | **71.07** | **67.32** | **97.28** | **98.32** | **73.52** | **77.84** | **64.82** | **94.93** | **97.85** | **70.62** | **75.14** | **66.23** |

Table 2.5: Results on MIMIC-III-50, MIMIC-IV-ICD9-50 and MIMIC-IV-ICD10-50 test sets. RoBERTa, Llama and PLM-ICD are used as backbone encoder models.

and Llama across all experimental configurations. This hierarchical performance pattern aligns with theoretical expectations, as PLM-ICD incorporates architecture and training paradigms specifically optimized for automated ICD coding tasks [137]. Despite the increasing generalization capabilities of foundation models in diverse applications, significant questions persist regarding their capacity to achieve state-of-the-art performance on highly specialized tasks, particularly within the medical domain, without substantial domain-specific training or parameter-efficient adaptation techniques [141]. Contemporary research on foundation model applications in biomedical domain has predominantly relied on specialized adaptation methods tailored to specific domain requirements. The comparative advantages of domain-specific pre-training becomes particularly evident following the development of initial foundation model architectures, as exemplified by widely implemented medical models such as Med-PaLM [53] and Med-Gemini[141].

Therefore, compared with the enhancements via the contrastive training phase, the intrinsic knowledge within LLMs contributes substantially more to ICD coding efficacy. In particular, domain-specific knowledge representations emerge as critical factors of LLMs performance in medical applications.

## 2.6 Conclusion

Multi-label classification poses a compelling challenge in applying contrastive learning due to the diverse ways of defining relations between multi-label samples. In this chapter, we introduce multi-label relations and formalize the concepts of similarity and dissim-

ilarity. Then, we propose a Similarity-Dissimilarity Loss for MSCL, which dynamically re-weights the loss by the combination of similarity and dissimilarity factors. We provide theoretical grounded proof for our method through rigorous mathematical analysis that supports the formulation and effectiveness of the proposed loss function. Then, We conduct a comprehensive experiments, considering data modality, domain-specific, data distribution and backbone models to further evaluation our method. The results show that our proposed loss outperforms the baselines (*ALL*, *ANY* and MulSupCon) across all the configurations and confirm the effectiveness and robustness of our method in image, text and medical domain.

# Chapter 3

# Prompting Explicit and Implicit Knowledge for Multi-hop QA

Language models (LMs) utilize chain-of-thought (CoT) to imitate human reasoning and inference processes, achieving notable success in multi-hop question answering (QA). Despite this, a disparity remains between the reasoning capabilities of LMs and humans when addressing complex challenges. Psychological research highlights the crucial interplay between explicit content in texts and prior human knowledge during reading. However, current studies have inadequately addressed the relationship between input texts and the pre-training-derived knowledge of LMs from the standpoint of human cognition. In this chapter, we propose a **P**rompting **E**xplicit and **I**mplicit knowledge (PEI) framework, which employs CoT prompt-based learning to bridge explicit and implicit knowledge, aligning with human reading comprehension for multi-hop QA. PEI leverages CoT prompts to elicit implicit knowledge from LMs within the input context, while

integrating question type information to boost model performance. Moreover, we propose two training paradigms to PEI, and extend our framework on biomedical domain QA to further explore the fusion and relation of explicit and implicit biomedical knowledge via employing biomedical LMs in the Knowledge Prompter to invoke biomedical implicit knowledge and analyze the consistency of the domain knowledge fusion. The experimental results indicate that our proposed PEI performs comparably to the state-of-the-art on HotpotQA, and surpasses baselines on 2WikiMultihopQA and MuSiQue. Additionally, our method achieves significant improvement compared to baselines on MEDHOP. Ablation studies further validate the efficacy of PEI framework in bridging and integrating explicit and implicit knowledge.

## 3.1 Introduction

Multi-hop question answering (QA) poses a significant challenge, requiring sophisticated reasoning and inference across multiple sources to derive a coherent and accurate answer [142]. Chain-of-thought (CoT) mimics human reasoning by generating a series of intermediate natural language steps that guide the model toward the final answer for complex reasoning tasks. Recent studies utilizing CoT prompt-based learning on language models (LMs) have shown considerable effectiveness in tackling multi-hop QA [143, 144, 145].

Despite the advancements in LMs, there remains a significant gap between their reasoning abilities and human cognitive processes in addressing intricate problems. Current

research has yet to adequately investigate the interplay between input texts and the pre-training-derived knowledge of LMs, particularly through the lens of cognitive science.

In studies of human reading comprehension, Smith [146] suggests that information is often reiterated during reading, resulting in redundancies at various linguistic levels, including letter-to-letter, word-to-word, sentence-to-sentence, and text-to-text. As a result, readers are able to reduce their dependence on explicit information details within the text by integrating external sources of information, such as world knowledge [147]. According to the findings of Clarke and Silberstein [148], readers engage in reading comprehension and question-answering process while reading, drawing upon both the explicit information conveyed in the text and their pre-existing language knowledge, background knowledge, and world knowledge derived from that explicit information. Certain studies have pointed out that a critical factor in reading ability is what the reader brings to the text, or what is generally referred to as prior knowledge [149, 150, 151]. Related experimental findings further reveal a significant positive correlation between human reading comprehension and prior knowledge [151].

For instance, as depicted in Figure 3.1, consider the question "*Was Morris Lee born in the capital of the Democratic Republic of the Congo?*". A human reader would retrieve relevant information from the provided passages and, based on the auxiliary verb "*was*" in the yes-no question, infer the answer "*yes*" or "*no*" drawing upon linguistic knowledge (as part of implicit knowledge), even in the absence of information regarding the capital of Congo.

Therefore, an inherent and inseparable connection prevails between the explicit in-

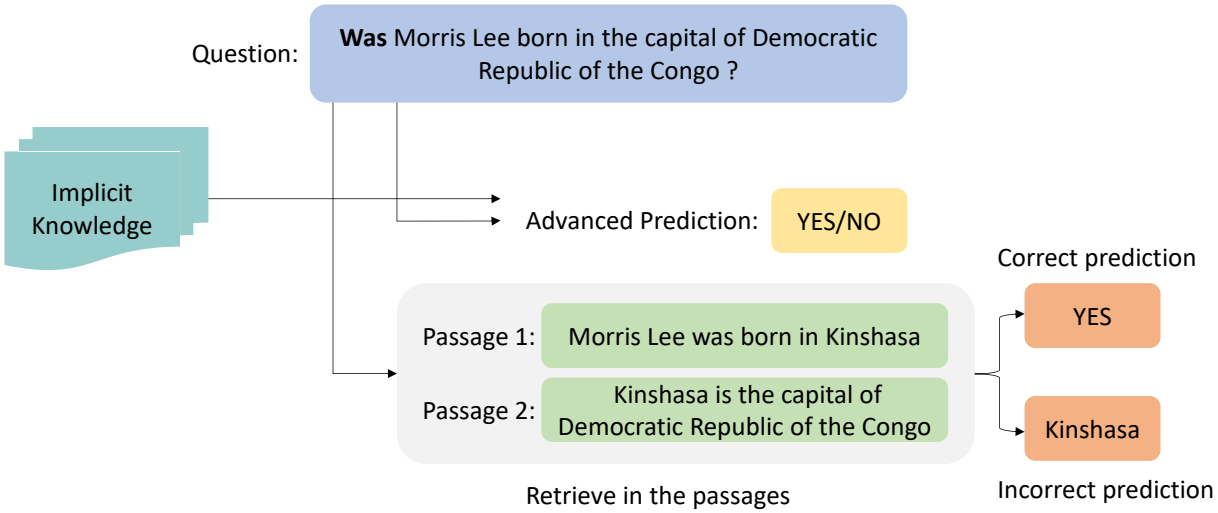


Figure 3.1: An example of the significance of implicit knowledge in reading comprehension.

formation within text context and pre-existing prior knowledge of human being. The prior knowledge lessens the dependence on explicit details, thus reducing the necessities for redundant information during inference and reasoning. In addition, the harmonious fusion of explicit information and prior knowledge improve the effectiveness of reading process, contributing to enhanced comprehension and deeper engagement.

Building on insights from the theories of human cognition mention above, we introduce a novel framework, referred to as **P**rompting **E**xplicit and **I**mplicit knowledge (PEI), to address the challenges multi-hop QA. In this framework, readers are analogized to LMs, where their prior knowledge represents implicit knowledge gained through pre-training, and the explicit information within passages serves as the input context conveying explicit knowledge. While acknowledging the inherent differences between LMs and and human beings, and recognizing the limitations in directly considering readers

as LMs, Jin and Rinard [152] argue that LMs surpass beyond mere "stochastic parrots" [153], as they possess the capacity to acquire meaningful semantic information during pre-training. Complex question answering encompasses high-complexity, non-factoid inquiries that require multi-step decomposition and integration of multiple information sources. The decomposition process is fundamental to this problem-solving paradigm, as it transforms initially intractable complex questions into manageable subproblems. Furthermore, as evidenced in recent chain-of-thought approaches and prompt-based methodologies for LLMs, these explicit reasoning steps not only enhance the models' problem-solving capabilities but also render the solutions auditable, verifiable, and interpretable when errors occur [154, 155].

To make use of these knowledge sources, we utilize CoT prompting to capture explicit knowledge and activate implicit knowledge. Intuitively, this approach effectively bridges these knowledge types, thereby enhancing the reasoning performance of PEI framework for multi-hop QA. Additionally, PEI reduces dependency on the explicit information details contained within input passages by enabling the selective removal of irrelevant or "redundant" information unrelated to the questions, aligning with Smith [146]'s theory. To further demonstrate the significant role of implicit knowledge in boosting the proposed framework's performance, we conduct ablation studies, which corroborates our hypothesis (refer to Section 3.5.4).

As illustrated in Figure 3.2, our proposed PEI framework consists of three main components: (i) **The Type Prompter** is designed to identify and learn the weights of reasoning types for given questions; (ii) **The Knowledge Prompter** acquires implicit know-

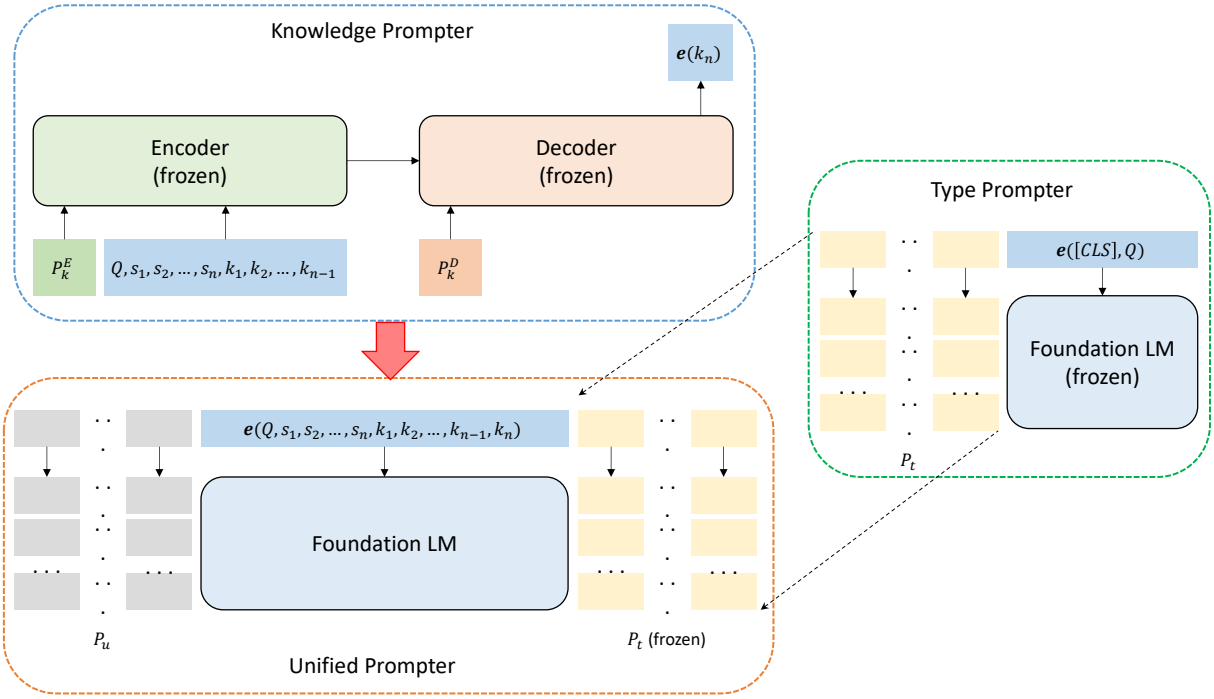


Figure 3.2: The overview of our proposed PEI framework for multi-hop QA. The right green dashed block is the Type Prompter; the top blue dashed block refers to the Knowledge Prompter; and the bottom orange dashed block is the Unified Prompter.

ledge by leveraging explicit knowledge, which employs an encoder-decoder foundation language module, mirroring human reading comprehension based on psychological cognition theories; (iii) **The Unified Prompter** fuses explicit, implicit knowledge and question types for multi-hop QA, which uses the same foundation model as Type Prompter. Moreover, our proposed framework offers flexibly replace the foundation models to adapt to different requirements, such as those specific to the biomedical domain or constraints related to computational cost. Hence, We propose two training paradigms to PEI framework, which employs various sizes of foundation LMs (i.e., Llama 3.1-8B and ELECTRA).

Tackling questions in the biomedical field frequently requires multi-step reasoning.

For example, a clinician might inquire, "Which tests are required for patients exhibiting [specific symptoms]?" To answer, models must: (i) deduce the possible diseases involved and (ii) determine the appropriate tests for differential diagnosis. To address these challenges, we extend our proposed PEI framework to biomedical domain for further exploring the fusion and connection of explicit and implicit biomedical knowledge in addition to general domain. Biomedical QA significantly deviates from general QA in content, scope, and methodology because of the complex nature of biomedical information [156, 75]. Zweigenbaum [157] was the first to highlight the distinctive characteristics of Biomedical QA compared to general domain QA. Biomedical QA deals with specific technical terms (e.g., "pharmacokinetics," "tyrosine kinase inhibitors" and "monoclonal antibodies") and entails the comprehension of detailed, evidence-based information, such as the interpretation of clinical trials and understanding action mechanisms. Additionally, biomedical information is frequently incomplete, ambiguous, or subject to constant change. Therefore, it is essential to comprehensively evaluate our proposed framework within the biomedical domain to ensure its effectiveness and reliability. We utilize a biomedical encoder-decoder foundation model within the Knowledge Prompter to harness specialized implicit knowledge, which is then integrated into the Unified Prompter with explicit knowledge derived from the provided texts. Moreover, we examine the consistency of the domain knowledge integration.

Our contributions are summarized as follows:

- We introduce the PEI framework that offers a proficient method for multi-hop QA,

based on the human reading process, by modeling the input passages or context as explicit knowledge and invoking the pre-trained knowledge of LMs as implicit knowledge that mirroring with human prior knowledge.

- We propose two training paradigms to PEI framework, which employs various sizes of foundation LMs (i.e., Llama 3.1-8B and ELECTRA). The experiment results show that the performance of Llama-based PEI with prompt tuning is slightly lower that of the standard PEI, but significantly reduces the number of trainable parameters while maintaining comparable reasoning performance.

- Our PEI framework demonstrates performance on par with state-of-the-art baseline evaluating on the benchmark HotpotQA dataset. Furthermore, PEI shows consistent effectiveness and robustness on single-hop sub-questions and additional multi-hop datasets (2WikiMultiHopQA and MuSiQue).

- We further evaluate PEI framework on biomedical domain to comprehensive explore the explicit and implicit knowledge fusion in specify domain. The experimental results shows that the proposed PEI framework significantly outperforms the baselines on MEDHOP dataset.

- The ablation studies corroborate that implicit knowledge improves the reasoning abilities of PEI framework, thereby supporting our hypothesis regarding PEI, which grounded in the human cognition theories for reading comprehension.

## 3.2 Related Work

### 3.2.1 Chain-of-Thought Prompting

Prompt tuning [1] has been acknowledged as a potent method to tune LMs to harness pertinent knowledge for targeted downstream tasks [158]. CoT prompting, which is a prompt-driven strategy, has surfaced as a technique to extract implicit knowledge from large language models (LLMs) for intricate reasoning tasks. It mirrors the sequential and coherent thought processes of humans by creating intermediate reasoning steps in natural language that culminate in the final result [32, 159]. Manual-CoT [32] aimed to extract CoT reasoning capability via manual demonstrations. Subsequently, Kojima et al. [160] showed that LLMs can effectively act as zero-shot reasoners, producing rationales that inherently contain CoT reasoning by using the phrase "*Let's think step by step*" to encourage a detailed thought process before deriving answers. AutoCoT, an automatic CoT prompting approach [159], employed various question sampling and reasoning chain construction to form demonstrations, thereby reducing the need for human input. Recent research have delved into CoT prompt learning for multi-hop QA [161, 144]. Building on aforementioned studies, our study investigates the application of CoT prompting to extract implicit knowledge from LMs. Unlike CoT, which produces intermediate steps in natural language, our approach produces continuous embeddings to represent implicit knowledge.

[1] The term "prompt tuning" refers to a broad array of methods rather than a specific approach.

### 3.2.2 Prompt-based Learning for Multi-hop QA

Significant advancements in recent research have been made by incorporating prompts for multi-hop QA [162, 163, 41]. For instance, PromptRank [164] developed an instruction-based prompt, integrating a candidate document pathway to calculate the relevance between a given question and the documented path. This relevance is evaluated via the conditional likelihood of the question in relation to the path prompt, as assessed by a language model. In contrast, IRCoT [144] implemented a system that alternated between CoT generation and knowledge retrieval steps, leveraging CoT prompting to direct the retrieval process. Wang et al. [161] proposed an iterative CoT prompting method that progressively extracts knowledge from LMs using a sequence-to-sequence BART-large model, thereby recalling natural language sequences for multi-hop QA. Each triplet in the evidence path is transformed into a natural language statement through a straightforward template, cumulatively generating the final statement. Building on this concept, our method employs a similar encoder-decoder foundation LM (i.e., BART-large) for recalling implicit knowledge by an approach of iterative prompting.

Comparing to the aforementioned studies, our proposed method distinguishes in the following three aspects: (i) PEI eliminates the need to transform triple evidence paths into natural language statements; (ii) we utilizes input passages to explicitly draw upon implicit knowledge from LMs, which previous methods have yet explored; (iii) PEI represents the recalled implicit knowledge as continuous embeddings, as opposed to using natural language statements or lexical knowledge [110]. Consequently, our framework

is not dependent on natural language statements originating from evidence paths.

### 3.2.3 Biomedical Multi-hop QA

Biomedical multi-hop QA is a specialized domain of multi-hop QA that involves reasoning over multiple interconnected biomedical facts to answer complex queries [156]. Recent studies utilize a combination of LMs and structured knowledge representations such as biomedical knowledge graphs [165, 166], or external medical knowledge [167]. Du et al. [166] proposed Adversarial Entity Graph Convolutional Networks (AEGCN), constructing an enriched entity graph with innovative edge relationships derived from supporting text while leveraging adversarial entities during training to enhance the model's resistance to interference. MedKGQA [167] addressed drug-drug interaction (DDI) prediction and medical reasoning by combining external medical knowledge bases with "drug-protein" triplets and graph neural networks (GNNs) to navigate and extract answers from biomedical pathways effectively.

Comparing with the previous methods introduced in related works, our proposed PEI framework employs CoT prompting to elicit implicit biomedical knowledge from LMs, reducing the cost of constructing of knowledge bases. Meanwhile, PEI framework is flexible to utilize various foundation LMs to adapt specialty knowledge requirements.

## 3.3 Methodology

### 3.3.1 Problem Statement

Multi-hop QA represents a challenging natural language processing task wherein a system must generate responses by retrieving and reasoning across multiple evidence fragments from disparate textual sources. In contrast to single-hop QA, which extracts answers from individual passages, multi-hop QA necessitates complex inferential reasoning to synthesize information across disconnected documents. Formally, given a complex query $Q$ and a collection of supporting sentences $S_n = [s_1, s_2, ..., s_i, ..., s_n]$, the objective is to derive the correct answer $A$ through sequential inferential processes that traverse the relevant textual evidence, and the corresponding supporting sentences $S_k$, where $S_k \subseteq S_n$.

### 3.3.2 Framework Overview

As illustrated in Figure 3.2, our proposed PEI framework comprises three primary components: (i) Type Prompter identifies reasoning types for given questions and learns their respective weights; (ii) Knowledge Prompter acquires implicit knowledge by leveraging explicit knowledge through an encoder-decoder foundation language model, reflecting the human reading process as informed by psychological cognition theories; and (iii) Unified Prompter, which integrates explicit knowledge, implicit knowledge, and question types for multi-hop QA using the identical foundation backbone model as Type Prompter.

PEI framework facilitates flexible substitution of foundation models to accommodate diverse requirements, such as domain-specific adaptation for biomedical applications or optimization of computational resource allocation.

**Pre-training on single-hop QA.** To analyze the capabilities of QA models throughout each step of reasoning processes for multi-hop QA, our study utilizes a foundational LM, namely ELECTRA [2] [168], trained on the single-hop QA dataset SQuAD [169]. Following this, we employ the pre-trained ELECTRA model as the foundation LM for our Type Prompter component. By deploying the ELECTRA model trained on single-hop tasks, we endeavor to investigate the interplay between the model's behavior and reasoning processes across multi-hop reasoning stages.

Note that the LLMs (i.e. Llama 3.1 [170]) also be employed as the foundation model, however, we do not fine-tune Llama on SQuAD since the computational cost of full parameter fine-tuning for LLMs is costly. Additionally, SQuAD dataset is a sub-task of the LLM benchmark.

### 3.3.3 Type Prompter

The Type Prompter is designed to enhance the training of acquired weights for soft prompts, allowing them to adeptly learn the unique features of different question types. As illustrated in Figure 3.2, the yellow blocks $P_t$ denote trainable prompt embeddings, whereas the blue blocks represent the input embeddings and frozen foundation LM.

[2]The foundational LMs could be displace with more advanced models. Consistent with previous studies [143] on prompt-based learning, we selected ELECTRA.

The P-tuning v2 approach [171] is utilized to trainable soft prompts $P_t$, learning the weights and capturing specific-type information of the given queries. Initially, the foundation LM remains frozen while the trainable soft prompt $P_t$ is optimized. Upon training, the updated $P_t$ is linked to the Unified Prompter module, while preserving its fixed nature throughout subsequent operations. The input sequence for the model includes both the trainable prompt embeddings and the token embeddings of the given question $Q$:

$$\mathbf{H}_{in} = [\mathbf{e}(P_t); \mathbf{e}(\texttt{[CLS]}); \mathbf{e}(Q)] \tag{3.1}$$

where $\mathbf{e}(P_t)$ is the trainable prefix embeddings (prompt tokens added before the input text), $\mathbf{e}(Q)$ denotes the token embeddings of the input question $Q$, and the specific token `[CLS]` is used as classification.

The extended input $\mathbf{H}_{in}$ is passed through the LM to compute hidden representations:

$$\mathbf{H}_{out} = \text{Model}(\mathbf{H}_{in}) \tag{3.2}$$

here, $\mathbf{H}_{out} \in \mathbb{R}^{(d+l)\times m}$, where We denote $d$ as the embedding dimension of the foundation LM, $l$ denotes the length of trainable prompt $P_t$ and $m$ is the hidden dimension of the LM.

In this module, the total number of trainable parameters can be calculated as $\Theta(d{\cdot}h{\cdot}l)$, where $h$ as the number of layers within the LM.

Comparing to full-parameter fine-tuning, Type Prompter module that employing p-

tuning v2 decreases the number of training parameters while effectively capturing type-specific information. Additionally, it allows for the transfer updated weights of $P_t$, encompassing type-specific information, to the Unified Prompter module. Furthermore, by utilizing p-tuning v2, a broader feature spectrum can be efficiently captured and learned compared to the prompt-tuning [31].

### 3.3.4 Knowledge Prompter

The Knowledge Prompter leverages textual input to activate and integrate LMs' innate prior knowledge, thereby enhancing the fusion of explicit and implicit information for effective reading comprehension.

Figure 3.2 illustrates how the Knowledge Prompter employs an iterative encoder-decoder LM to retrieve implicit knowledge through prefix tuning [3] [29]. Trainable prompt embeddings, labeled as $P_k^E$ for the encoder and $P_k^D$ for the decoder, are integrated within each LM layer. This method facilitates the efficient retrieval and application of explicit knowledge during the iterative phases of encoding and decoding.

Given a multi-hop query $Q$ and a serious of supporting sentences $S_n = [s_1, s_2, ..., s_i, ..., s_n]$, we aim to retrieve and extract a sequence of knowledge $K_n = [k_1, k_2, .., k_i, ..., k_n]$ that provides sufficient information for determining the response to both $Q$ and $S_n$, where $n$ represents the number of supporting sentences. Our focus lies in the development of prompt-based learning method, where we intend to construct trainble prompts $P_k^E$ and

[3] We employ prefix tuning method for the Knowledge Prompter inspired by the context-aware prompter design [161].

$P_k^D$ to lead the encoder-decoder LM in recalling the desired knowledge $K_n$. Notably, we maintain fixed parameters for the encoder-decoder LM, thereby allowing us to direct its retrieval process via trainable prompts.

Motivated by the sequential nature observed in multi-step reasoning tasks [161], we adopt an iterative approach as below:

$$P(k_j|Q, S_j, K_{j-1}) = \prod_{j=1}^{n} P(k_j|Q, s_1, ...,s_j, k_1, ..., k_{j-1}) \tag{3.3}$$

$$\text{decoder}(k_j) = \text{encoder}(Q, S_j, K_{j-1}) \tag{3.4}$$

where at each step $j$, LM recalls the next piece of knowledge $k_j$ conditioned on the query $Q$ and supporting sentences $s_1, ..., s_j$ and gathered knowledge $k_1, ..., k_{j-1}$.

More specially, when $j = 1$, it is written as following based on Equation (3.3) and (3.4):

$$\text{decoder}(k_1) = \text{encoder}(Q, s_1) \tag{3.5}$$

### 3.3.5 Unified Prompter

As illustrated in Figure 3.2, we suture the Unified Prompter module with $P_t$, integrating weights tailored for specific reasoning types. Additionally, the implicit knowledge $K_n$ derived from the Knowledge Prompter module serves as supplementary input. This fusion

of information intuitively boost reasoning capabilities of PEI framework based on human reading process.

In Unified Prompter module, the trainable prompt embeddings are denoted as $P_u$, where the updated prompt embeddings $P_t$ are frozen. To preserve the learned weights of $P_t$ derived from the Type Prompter, we adopt the identical architecture of the foundation LM, allowing for seamless concatenation of $P_t$ with the Unified Prompter.

The input sequence for the Unified Prompter includes the trainable prompt embeddings $P_u$ and the frozen prompt embeddings $P_t$:

$$\mathbf{H} = [\mathbf{e}(P_u); \mathbf{e}(Q, S_n, K_n); \mathbf{e}(P_t)] \tag{3.6}$$

Subsequently, we perform two training settings depending on various foundation LM. For ELECTRA, we employ joint p-tuning and full parameter fine-tuning to this module. For Llama, we use p-tuning to optimize the trainable $P_u$ and freeze the foundation LM.

**Prediction Module**[4]. After encoding in Unified Prompter, we design a prediction module to jointly perform answer and supporting evidence prediction, followed by [172, 143]. To determine the answer span, two linear layers are utilized on the context representation to ascertain the start and end positions of the response. Meanwhile, a binary linear layer is deployed for predicting supporting evidence by assigning a binary relevance label at the beginning of each supporting sentence `[SE]`.

[4] https://github.com/Tswings/PCL

## 3.4 Experiments

### 3.4.1 Dataset and Metrics

**HotpotQA** [142] comprises a dataset of 113,000 question-answer pairs sourced from Wikipedia. Furthermore, HotpotQA includes sentence-level supporting facts critical for reasoning, thereby enabling QA systems to carry out inference with strong supervision and articulate their predictions.

**2WikiMultiHopQA** [173], comprises more than 192,000 entries, distributed across 167,000 for training, 12,500 for evaluation, and 12,500 for testing. While its structure is largely aligned with HotpotQA [142], this dataset introduces improvements by offering a wider spectrum of reasoning categories for questions and detailed annotations of the evidence trajectories linked to each question.

**MuSiQue** [174] consists of 25,000 examples featuring questions that require 2 to 4 reasoning steps. This collection is curated through a structured method of selecting compatible single-hop question pairs that manifest logical links, thereby crafting a comprehensive set of multi-hop inquiries.

**Sub-question QA dataset** [175] was developed to support the examination of multi-hop QA models' reasoning abilities at each phase of the reasoning process. To assess the models' effectiveness, the authors assembled a dedicated dataset composed of single-hop sub-questions. This collection encompasses 1k samples, manually validated from the HotpotQA development set, thereby providing a high-quality evaluation resource for

the research.

To achieve uniformity and comparability among the datasets employed in our experimental analyses, we classify the question types from the three datasets being examined into the overarching categories of comparison and bridge. This classification aids in establishing a standardized methodology for managing various question structures across the datasets subjected to our evaluation.

**MEDHOP**[176] is a benchmark resource designed for multi-hop reasoning in the biomedical domain. It belongs to the QAngaroo collection[5], which emphasizes the integration of data from multiple documents or textual units to resolve intricate queries. The dataset comprises 1,620 training instances, 342 validation instances, and 546 test instances, culminating in a total of 2,508 instances. However, to the best of our knowledge, only MEDHOP evaluates multi-hop reasoning abilities, while almost all other Biomedical QA datasets focus on single-hop reasoning[156].

**Metrics.** Consistent with prior research [142], We report Exact Match ($EM$) and Partial Match ($F1$) to evaluate the efficacy and performance of our proposed framework concerning both answer and supporting facts prediction. Furthermore, the joint $EM$ and $F1$ are used to assess the overall performance. Specially, we use accuracy to evaluate the performance of our PEI and baselines on MEDHOP following by Welbl et al. [176].

[5]http://qangaroo.cs.ucl.ac.uk

### 3.4.2 Selected Baselines

To comprehensive assess the performance of PEI, we compare with a series of selected and state-of-art baselines, including 1) general domain methods on HotpotQA, 2WikiMultiHopQA, MuSiQue and Sub-question QA, and 2) biomedical domain methods that is conducted on MEDHOP dataset.

These baselines for general multi-hop QA as follows:

- Baseline Model [142] serves as the initial baseline for HotpotQA.
- DecompRC [177] transforms complex queries into easier sub-questions, enabling resolution via existing single-hop reading comprehension frameworks.
- OUNS [178] represents an algorithm designed for One-to-N Unsupervised Sequence transduction. It converts intricate, multi-step queries into a range of straightforward, single-step questions to enhance the QA process by decomposing complexities.
- QFE [179] model incrementally identifies evidence sentences through an RNN with attention focused on the query, drawing inspiration from models used in extractive summarization.
- Longformer [180] employs an attention mechanism that increases linearly with the sequence length, facilitating the examination of extended texts in a multi-hop QA context.

- Beam Retrieval [145] integrates the retrieval operation in an end-to-end manner by synchronously optimizing an encoder and two classification outputs throughout all steps.

The following baselines apply GNN-based methods for general multi-hop QA:

- DFGN [181] dynamically constructs an entity graph from the text and incrementally identifies supporting entities relevant to the query within the provided documents.
- SAE-large [182] utilizes a GNN where contextual sentence embeddings serve as nodes, bypassing the use of entities as nodes, thereby directly predicting supporting sentences alongside the answer.
- C2F Reader [183] integrates task-specific prior knowledge via graph structures and adjacency matrices, employing graph attention as a variant of self-attention.
- HGN [172] introduces a hierarchical graph with nodes representing varying granularities, such as questions, paragraphs, sentences, and entities, leveraging pre-trained contextual encoders for initialization.
- AMGN [184] incorporates GNN-based methodologies to asynchronously update multi-grained nodes by modeling relationships across different levels, reflecting the logical progression of multi-hop reasoning.
- S2G [185] implements a select-to-guide (S2G) strategy to retrieve evidence paragraphs in a coarse-to-fine manner, augmented by two novel attention mechanisms,

which effectively align with the inherent nature of multi-hop reasoning.

These baselines utilize CoT prompting methods for general multi-hop QA as follows:

- iCAP [161] adopts an iterative prompting framework designed to incrementally extract pertinent knowledge from pre-trained language models (PLMs), enabling step-by-step inference.

- PCL [143] introduces a Prompt-based Conservation Learning (PCL) framework, wherein soft prompts are optimized to guide sub-networks in executing type-specific reasoning tasks.

The biomedical domain baselines are as follows:

- FastQA [186] employs a single bi-directional recurrent neural network followed by an answer prediction layer that independently identifies the start and end points of the answer span.

- BiDAF [187] adopts a hierarchical approach, representing contextual information at varying levels of granularity and leveraging a bidirectional attention flow mechanism to construct a query-aware context representation without premature summarization.

- Document-cue [176] emphasizes the model's ability to leverage document-answer co-occurrence patterns to identify relevant information.

- BAG [188] introduces a bidirectional attention entity graph convolutional network that captures relationships between graph nodes and employs attention mechanisms to link the query with the entity graph.

- EPAr [189] (Explore-Propose-Assemble reader) mimics human-like reading strategies by adopting a coarse-to-fine approach for reasoning and answering QA tasks.

- NLProlog [190] integrates symbolic reasoning and rule learning with distributed representations of sentences and entities to perform rule-based multi-hop reasoning over natural language inputs.

- DrKIT [191] simulates traversals in a knowledge base (KB) constructed over a text corpus, providing ability to follow relations in the "virtual" KB over text for multi-hop questions.

- DILR-BERT [192] combines transformer-based model with inductive logic reasoning, first extracting query-relevant data and then applying rule induction to conduct logical reasoning across the filtered information.

- ClueReader [165] employs a heterogeneous graph attention network inspired by the grandmother cell concept, aggregating semantic features at multiple levels and dynamically focusing or suppressing information for reasoning.

- MedKGQA [167] integrates external biomedical knowledge bases, including "drug-protein" triplets, with graph neural networks (GNNs) to facilitate effective navigation and answer retrieval within biomedical pathways, particularly for drug-drug

interaction (DDI) prediction and medical reasoning.

- AEGCN [166] enhances entity graph representations with enriched edge relationships derived from supporting texts, employing adversarial training to improve resistance and interference.

### 3.4.3 Implementation Details

ELECTRA-large [168] and Llama 3.1 8B [170] serve as the foundation LM for both Type Prompter and Unified Prompter modules. In Type Prompter, p-tuning v2 [171] is used for the prompt tuning to acquiring the weights of specific-type information of the given questions. In Unified Prompter, we conduct two training settings depends on various foundation LM: 1) for ELECTRA, we employ joint p-tuning and full parameter fine-tuning to the whole module, and 2) for Llama, we use p-tuning to optimize the trainable Pu and freeze the Llama. Inspired by the studies of Wang et al. [161], we adopt BART-large [193] as the foundation LM in the Knowledge Prompter module. Specially, BioELECTRA [194] and BioBART [195] are employed as the foundation LMs for biomedical domain extension.

Our implementation is built upon the Huggingface platform [138]. For model optimization, we employ the AdamW optimizer [196] along with a linear learning rate scheduler with a warmup ratio of $0.05$.

In terms of hyperparameters, we conduct a search for the optimal batch size. We explored batch sizes of $\{4, 8, 12, 16, 32\}$ respectively. Additionally, we performed a tuning

process for the learning rate, considering values from $\{2e-5, 4e-5, 8e-5, 2e-4, 4e-4, 8e-4, 2e-3, 4e-3, 8e-3, 2e-2, 4e-2, 8e-2\}$. Moreover, we conducted tuning experiments for the length of the encoder/decoder prompts $P_k$ type prompts $P_t$ and the unified prompt $P_u$, exploring values from $\{20, 40, 60, 80, 100, 120, 150\}$.

## 3.5 Results and Analysis

### 3.5.1 Evaluation on HotpotQA

Initially, we evaluate PEI framework on the test set of HotpotQA in the distractor setting comparing with peer-reviewed baselines, including the baseline model of HotpotQA [142], Beam Retrieval [145] that is the state-of-art model on the leaderboard, iCAP [161] and PCL [143] which we inspired by, and other baselines. For a reminder, we will refer to the PEI(ELECTRA$_{PT+FT}$) framework that employs ELECTRA with full parameter fine-tuning and prompt tuning as PEI to be more brief.

As depicted in Table 3.1, PEI framework outperforms all baseline models across the evaluated metrics, with the sole exception of Beam Retrieval. Notably, PEI achieves performance comparable to that of Beam Retrieval [145] on the HotpotQA, demontrating the substantial advancements facilitated by PEI in addressing multi-hop QA.

More specifically, comparing with Beam Retrieval, PEI achieves an improvement of $0.20/0.28/0.30$ in answer EM, answer F1 score and join F1 score, respectively. In contrast, Beam Retrieval exhibits superior results with an improvement of $1.22$ in supporting EM,

| Models | Ans | | Sup | | Joint | |
|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 |
| Baseline Model [142] | 45.60 | 59.02 | 20.32 | 64.49 | 10.83 | 40.16 |
| DecompRC [177] | 55.20 | 69.63 | - | - | - | - |
| OUNS [178] | 66.33 | 79.34 | - | - | - | - |
| QFE [179] | 53.86 | 68.06 | 57.75 | 84.49 | 34.63 | 59.61 |
| DFGN [181] | 56.31 | 69.69 | 51.50 | 81.62 | 33.62 | 59.82 |
| SAE-large [182] | 66.92 | 66.92 | 61.53 | 86.86 | 45.36 | 71.45 |
| C2F Reader [183] | 67.98 | 81.24 | 60.81 | 87.63 | 44.67 | 72.73 |
| Longformer [180] | 68.00 | 81.25 | 63.09 | 88.34 | 45.91 | 73.16 |
| HGN [172] | 69.22 | 82.19 | 62.76 | 88.47 | 47.11 | 74.21 |
| AMGN [184] | 70.53 | 83.37 | 63.57 | 88.83 | 47.77 | 75.24 |
| S2G [185] | 70.72 | 83.53 | 64.30 | 88.72 | 48.60 | 75.45 |
| iCAP † [161] | 68.61 | 81.82 | 62.80 | 88.51 | 47.02 | 74.11 |
| PCL [143] | 71.76 | 84.39 | 64.61 | 89.20 | 49.27 | 76.56 |
| Beam Retrieval [145] | 72.69 | 85.04 | **66.25** | **90.09** | **50.53** | 77.54 |
| PEI(ELECTRA$_{PT+FT}$) | **72.89** | **85.32** | 65.03 | 89.81 | 49.91 | **77.84** |
| PEI(Llama$_{PT}$) | 71.45 | 85.05 | 64.00 | 88.97 | 48.89 | 76.41 |

Table 3.1: Results on the blind test set of HotpotQA in the distractor setting. "-" denotes the case where no results are available. † denotes that we implement the code. "Ans" represents the metrics for answer; "Sup" denotes the metrics for supporting facts; "Joint" is the joint metrics that combine the evaluation of answer spans and supporting facts. "FT" refers to full fine-tuning and "PT" represents prompt tuning.

$0.28$ in supporting F1 score, and $0.62$ in joint EM compared to PEI.

The difference in performance between PEI and Beam Retrieval in answer prediction versus supporting prediction could be attributed to the distinct methodologies employed by the two approaches. Beam Retrieval preserves multiple partial hypotheses of relevant passages at each step, expanding the search space (albeit at the expense of an exponentially complex retrieval process) and reducing the risk of missing relevant passages.

Consequently, it excels in supporting prediction. On the other hand, PEI draws inspiration from human reading processes by integrating implicit knowledge and type-specific information, which enhances its accuracy in answer prediction. However, this design may limit its efficacy in supporting prediction compared to Beam Retrieval, owing to the differences in their retrieval strategies.

Though PCL and PEI adopts the same backbone LM (i.e., ELECTRA) and prediction module, PEI framework demonstrates a significant improvement of $0.64/1.28$ in the Joint EM/F1 score compare to PCL. Compare with the question classification that PCL trains a PLM to acquired the reasoning type, PEI leverages the prompt tuning to learn the type-specific knowledge and transfers the trained weight to the Unifier Prompter, which effectively reduces computational cost.

Moreover, the proposed PEI framework demonstrates a notable improvement over iCAP, achieving a $2.89/3.73$ increase in joint EM/F1 scores, despite both models utilizing the same encoder-decoder architecture (BART) as their foundational LMs. When compared to the graph-based AMGN model, PEI achieves even greater gains, with a $2.14/2.6$ enhancement in joint EM/F1 scores.

**Comparison on LM Architecture and Training Paradigms**

In Table 3.1, the performance of PEI(Llama$_{PT}$) framework based on Llama with prompt tuning is slightly lower that that of the standard PEI in both answer and supporting facts prediction. Because Unified Prompter actually acts as a context encoder, the performance of PEI depends on the architecture of foundation LMs and training paradigms.

Extractive multi-hop QA (e.g., HotpotQA) often requires token-level attention over the input representation to have an edge in span-based extraction, which is a native strength of encoder-only architectures such as ELECTRA [197]. In contrast, decoder-only models (such as Llama) are not inherently designed to output spans of text, instead, they excel at generative QA tasks, where their ability to synthesize and produce text is advantageous [198].

Compared to standard PEI, which employs full parameter fine-tuning, the Llama-based PEI utilizing prompt tuning significantly reduces the number of trainable parameters while maintaining comparable reasoning performance. However, it still incurs higher computational cost and longer inference time due to the large-scale parameters of Llama and the associated memory requirements.

**Comparison with LLMs under Zero-Shot**

Table 3.2 presents the performance of open-source LLMs in a zero-shot setting for multi-hop QA. The results clearly indicate that zero-shot LLMs perform significantly inferior than PEI and other baseline models (see Tables 3.1 and 3.3) across the HotpotQA, 2WikiMultihopQA, and MuSiQue datasets. This demonstrates that extractive multi-hop QA still remains a challenge to LLMs in zero-shot setting. Notably, Llama-based PEI surpass all LLM baselines, including Llama 3.1-8B, across diverse datasets. Compare to zero-shot Llama 3.1-8B, the PEI (Llama 3.1-8B$_{PT}$) attains a notable performance boost with EM and F1 improvements of $32.0/39.1$ on HotpotQA and $12.8/36.0$ on 2WikiMultihopQA, respectively. On MuSiQue, while the performance gains are less pronounced (an

| Models | HoptpotQA | | 2WikiMultihopQA | | MuSiQue | |
|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 |
| Mistral-7B | 30.6 | 37.2 | 27.4 | 29.8 | 25.2 | 28.9 |
| Qwen 2-7B | 36.2 | 43.5 | 31.7 | 35.8 | 28.2 | 31.2 |
| Llama 2-7B | 34.5 | 41.3 | 30.6 | 34.7 | 31.7 | 35.6 |
| Llama 3.1-8B † | 39.4 | 45.9 | 33.1 | 37.5 | 32.9 | 36.5 |
| PEI (Llama 3.1-8B$_{PT}$) | 71.4 | 85.0 | 45.9 | 73.5 | 40.5 | 67.2 |

Table 3.2: Performance of Llama-based PEI compared to open source LLMs with zero-shot settings on HotpotQA, 2WikiMultihopQA and MuSiQue. † denotes that is our implementation.

improvement of $7.6/30.7$ in EM/F1 compared to Llama 3.1-8B), PEI still leads with an EM of 40.5 and F1 of 67.2, demonstrating consistent performance across diverse datasets. These findings demonstrate the effectiveness of leveraging task-specific prompt tuning to enhance the reasoning capabilities of LLMs. By enabling task-specific fine-tuning, PEI enhances the model's ability to fuse explicit and implicit knowledge effectively for complex reasoning tasks.

### 3.5.2 Evaluation of Robustness

To further evaluate the robustness of proposed PEI framework, we conduct three aspects experiments: (i) assessing PEI on other multi-hop QA datasets; (ii) evaluation on sub-question dataset in composing answers from solved sub-questions; and (iii) effect of foundation LMs in the same training paradigm.

| Models | 2WikiMultihopQA | | MuSiQue | |
|---|---|---|---|---|
| | EM | F1 | EM | F1 |
| iCAP | 42.80 | 47.90 | - | - |
| HGN | 38.74 | 68.69 | 39.42 | 65.12 |
| PCL | 46.03 | 73.42 | 41.28 | 67.34 |
| PEI (Ours) | **47.32** | **74.56** | **41.97** | **67.85** |

Table 3.3: Results of our proposed PEI compared to PCL, HGN and iCAP on 2WikiMultihopQA and MuSiQue multi-hop QA test set. "-" denotes the case where no results are available. PEI refers to version of PEI(ELECTRA$_{PT+FT}$).

**Evaluation on Other Multi-hop Datasets**

To evaluate generalization, we validate the PEI framework on the 2WikiMultihopQA and MuSiQue datasets. As presented in Table 3.3, PEI consistently outperforms all baseline models across both EM and F1 metrics. Notably, although both PEI and iCAP utilize the same encoder-decoder architecture (BART), PEI achieves a significant improvement of $4.52/26.66$ in answer EM/F1 scores on the 2WikiMultihopQA dataset. Additionally, PEI demonstrates superior performance over PCL, with gains of $1.29/1.14$ and $0.69/0.51$ in answer EM/F1 scores on the 2WikiMultihopQA and MuSiQue, respectively.

**Evaluation on Sub-question Dataset**

To assess the efficacy of the PEI model in multi-hop reasoning, particularly in synthesizing answers from resolved sub-questions, we conduct an evaluation on the sub-question QA dataset [175]. Each parent question, denoted as $q$, is associated with two corresponding sub-questions, $q_{sub1}$ and $q_{sub2}$. As presented in Table 3.4, the PEI model achieves a

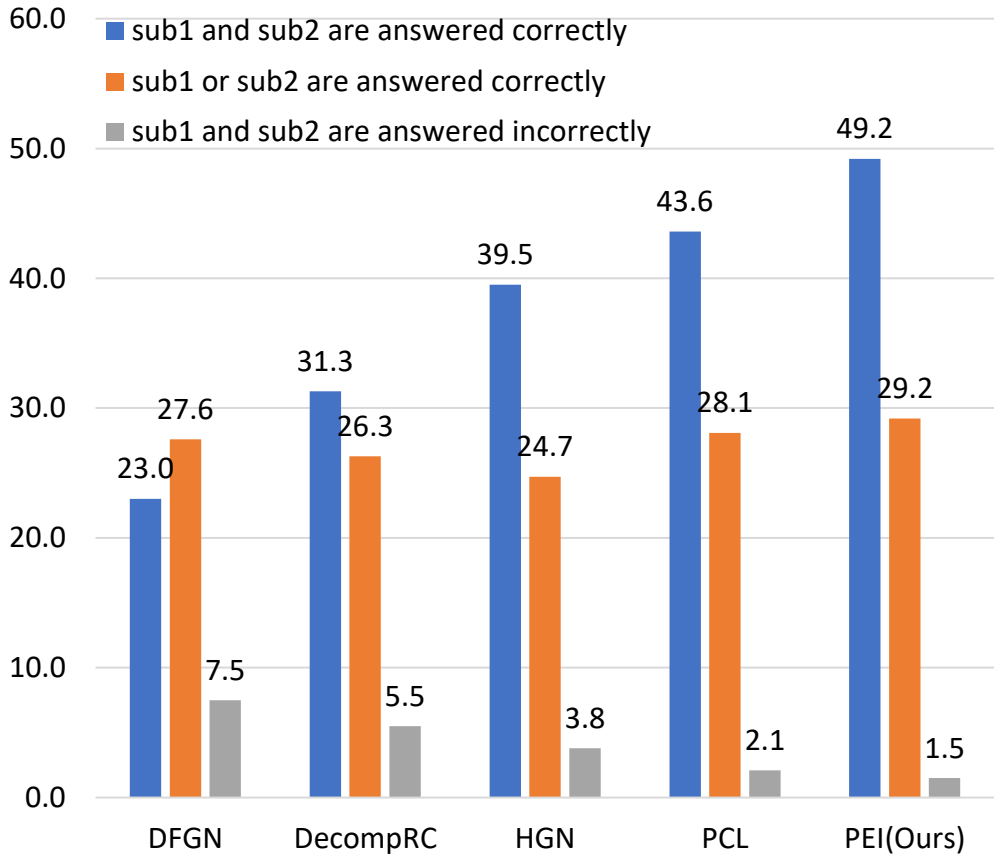


Figure 3.3: The success rate (%) of five multi-hop QA models. $sub1$ denotes the first sub-question and $sub2$ is the second sub-question of corresponding question $q$.

success rate of $97.62\%$ in correctly answering the parent multi-hop question $q$ while both sub-questions $q_{sub1}$ and $q_{sub2}$ are answered correctly [6]. This highlights the proficiency of PEI in retaining acquired knowledge through the integration of explicit and implicit knowledge, surpassing other baseline models. Interestingly, PEI also demonstrates a notable success rate of $36.55\%$ in correctly answering the parent multi-hop question even when only one of the sub-questions is answered correctly [7]. Figure 3.3 further illustrates the sub-question-dependent success rates for various multi-hop QA models, showing that

[6]The calculation process: $49.2/(49.2 + 1.2) = 97.62\%$
[7]The calculation process: $(7.1 + 22.1)/(49.2 + 7.1 + 22.1 + 1.5) = 36.55\%$

| $q$ | $q_{sub1}$ | $q_{sub2}$ | DFGN | DecompRC | HGN | PCL | PEI(Ours) |
|---|---|---|---|---|---|---|---|
| c | c | c | 23.0 | 31.3 | 39.5 | 43.6 | 49.2 |
| c | c | w | 9.7 | 7.2 | 5.1 | 6.8 | 7.1 |
| c | w | c | 17.9 | 19.1 | 19.6 | 21.3 | 22.1 |
| c | w | w | 7.5 | 5.5 | 3.8 | 2.1 | 1.5 |
| w | c | c | 4.9 | 3.0 | 2.8 | 1.7 | 1.2 |
| w | c | w | 17.0 | 18.6 | 16.7 | 16.3 | 13.4 |
| w | w | c | 3.5 | 3.4 | 2.6 | 1.1 | 1.0 |
| w | w | w | 16.5 | 11.9 | 9.9 | 7.1 | 4.5 |

Table 3.4: Results on sub-question dataset. The notation $c/w$ indicates that a question has been answered correctly or wrongly, respectively. For each primary question $q$, we denote its constituent components as $sub1$ and $sub2$, representing the first and second sub-questions, respectively.

these models frequently predict the correct parent question answer even when only one sub-question is answered correctly. This observation highlights a persistent challenge in multi-hop QA: models often exploit unreliable reasoning shortcuts for answer prediction, a phenomenon that deviates from the expected logical reasoning process [143].

**Effect of Foundation LMs**

To evaluate the effects of foundation LMs, we conduct a comparative analysis of PEI against PCL and HGN under identical experimental conditions, including the same datasets and foundation models. As shown in Table 3.5, PEI consistently exceeds both PCL and HGN across all evaluation metrics, demonstrating its effectiveness and robustness across various foundation LMs. Furthermore, PEI that implemented with ALBERT achieves an improvement of $0.62/0.23$ in Answer/Support F1 scores compared to its implementation with RoBERTa. This aligns with prior findings that ALBERT surpasses

| Model | Ans F1 | Sup F1 | Joint F1 |
|---|---|---|---|
| HGN (RoBERTa) | 82.22 | 88.58 | 74.37 |
| HGN (ELECTRA) | 82.24 | 88.63 | 74.51 |
| HGN (ALBERT) | 83.46 | 89.20 | 75.79 |
| PCL (RoBERTa) | 84.33 | 90.75 | 77.12 |
| PCL (ELECTRA) | 84.42 | 91.15 | 77.76 |
| PCL (ALBERT) | 85.47 | 91.28 | 78.76 |
| PEI (RoBERTa) | 85.61 | 92.02 | 78.95 |
| PEI (ELECTRA) | 85.68 | 92.11 | 79.02 |
| PEI (ALBERT) | **86.23** | **92.25** | **79.11** |
| PEI (Llama$_{PT}$) | 85.05 | 88.97 | 76.41 |

Table 3.5: Results with different LMs on the development set of HotpotQA.

RoBERTa on the GLUE benchmark under a single-model configuration [8]. These results affirm that incorporating a more advanced foundation LM can significantly enhance the performance of the PEI framework.

### 3.5.3 Evaluation on Biomedical Inference

#### Results and Discussion

To comprehensively evaluate PEI framework on biomedical domain, we compare PEI with diverse baselines including biomedical knowledge-based and non-biomedical knowledge-based models. As shown in Table 3.6, the vanilla PEI, which employs BART and ELECTRA foundation models without biomedical pretraining, exceeds all the baselines on MEDHOP dataset and outperforms the state-of-art model, AEGCN [166] with $0.75\%$ in

[8] https://github.com/google-research/albert

| Model | Accuracy |
|---|---|
| FastQA [186] | 23.10 |
| BiDAF [187] | 47.80 |
| Document-cue [176] | 44.90 |
| BAG [188] | 64.50 |
| EPAr [189] | 64.90 |
| NLProlog [190] | 65.78 |
| DrKIT [191] | 67.25 |
| DILR-BERT [192] | 71.35 |
| ClueReader [165] | 46.00 |
| MedKGQA [167] | 64.80 |
| AEGCN [166] | 72.28 |
| PEI (Vanilla) | 73.03 $\uparrow_{0.75}$ |
| PEI (Biomedical) | **75.62** $\uparrow_{3.34}$ |

Table 3.6: Results on test set of MEDHOP. PEI (vanilla) denotes that PEI model employs BART and ELECTRA as foundation models without biomedical pre-trained knowledge. PEI (biomedical) denotes that bioBART and bioELECTRA serve as foundation models for PEI, which integrate pre-trained biomedical knowledge

accuracy. This shows that even without using a biomedical LMs, the PEI framework exhibits significant reasoning ability in biomedical multi-hop QA.

Although general capabilities of foundation models have become increasingly apparent, there remain open questions regarding whether exceptional performance can be achieved on specialized tasks, such as those in the medical field, without extensive domain-specific training or fine-tuning of these general models [44, 107]. Much of the research on the application of foundation models in biomedicine has been heavily reliant on fine-tuning tailored to specific domains and tasks. With the advent of first-generation foundation models, the benefits of domain-specific pretraining became evid-

ent, as demonstrated by widely used models in the biomedical domain, such as PubMedBERT [48] and BioGPT [55]. In Table 3.6, the experimental results show that the biomedical PEI, which employs both BioBART and BioELECTRA as fundamental models to make fully use of pretrained biomedical domain knowledge, achieve a significant $3.34\%$ improvement in accuracy compared with AEGCN. Compared with the $0.75\%$ accuracy improvement of vanilla PEI, it is obvious that biomedical PEI has better performance and stronger reasoning ability in the biomedical domain.

**Effect of Biomedical Knowledge**

To further analyze the effect of specialty knowledge of PEI framework for biomedical domain, we explore the various settings of foundation models (see Table 3.7). As illustrated in Table 3.7, the experimental results show that biomedical knowledge boost performance of PEI. The PEI with bioBART shows an improvement of $0.96\%$ in accuracy compared with the vanilla PEI, demonstrating the Knowledge Prompter module is able to elicit the implicit specialty knowledge via CoT prompting. Moreover, the Unified Prompter makes fully use of the implicit knowledge $K_n = [k_1, k_2, .., k_i, ..., k_n]$ and fuse with explicit knowledge $S_n = [s_1, s_2, ..., s_i, ..., s_n]$ to enhance the inference ability. These results provide evidence that implicit knowledge plays a crucial role in improving the model's reasoning capabilities, thereby comfirming the hypothesis that underpins our proposed PEI framework, which is inspired by human reading process.

Additionally, the PEI with BioELECTRA outperforms the vanilla PEI with an improvement of $1.27\%$ in accuracy, further emphasizing the reliance of most investigations into

| Components | | Accuracy |
|---|---|---|
| BioBART | BioELECTRA | |
| ✗ | ✗ | 73.03 |
| ✓ | ✗ | 73.99 $\uparrow_{0.96}$ |
| ✗ | ✓ | 74.30 $\uparrow_{1.27}$ |
| ✓ | ✓ | **75.62** $\uparrow_{2.59}$ |

Table 3.7: Ablation study of comparison on vanilla PEI and biomedical PEI with different components.

the efficacy of foundation models in biomedical tasks on extensive domain- and task-specific fine-tuning.

### 3.5.4 Ablation Studies

To investigate the contributions of each components within PEI framework, we perform a series of ablation studies on the validation set of HotpotQA.

#### Effect of Implicit Knowledge

To validate the hypothesis that implicit knowledge enhances reasoning capabilities in multi-hop QA, we conduct an ablation study comparing the ELECTRA model with and without implicit knowledge integration (specifically referring to the implicit knowledge $K_n$ derived from the Knowledge Prompter). Table 3.8 demonstrates that incorporating implicit knowledge yielded significant improvements of $3.10/2.05/2.30$ in Answer F1, Supporting F1, and Joint F1 scores, respectively, compared to the baseline model without implicit knowledge. These findings empirically confirm that implicit knowledge

substantially augments models' reasoning abilities, aligning with the cognitive theories underpinning the PEI framework, which draws inspiration from the human reading process.

**Effect of Type Prompts**

To assess the influence of type-specific prompts and the model's capacity for type-driven reasoning in multi-hop question answering, we perform a comparative analysis of the ELECTRA language model, both with and without the Type Prompter module. As presented in Table 3.8, integrating the Type Prompter with the language model results in significant improvements of $3.02/1.17/2.18$ in Answer F1, Supporting F1, and Joint F1 scores, respectively, over the model without Type Prompter. These results highlight the effectiveness of incorporating question type information through the Type Prompter in enhancing the model's overall performance and facilitating type-specific reasoning. Furthermore, the findings support the alignment of PEI framework architecture with cognitive process observed in human reasoning, as type information can be regarded as a form of implicit knowledge.

**Effect of Pre-training on Single-hop**

We initially trained an ELECTRA-based QA model on the single-hop QA dataset, SQuAD [169], and subsequently fine-tuned it on the HotpotQA dataset. While conservation learning [143] is not utilized in our approach, we assess the model's performance both with and without pre-training to evaluate its impact on single-hop QA. As indicated in

| Model | Ans F1 | Sup F1 | Joint F1 |
|---|---|---|---|
| ELECTRA | 78.12 | 88.20 | 73.50 |
| - Type Prompter | 81.14 $\uparrow_{3.02}$ | 89.37 $\uparrow_{1.17}$ | 75.68 $\uparrow_{2.18}$ |
| - Pre-trained | 78.82 $\uparrow_{0.70}$ | 88.82 $\uparrow_{0.62}$ | 74.54 $\uparrow_{1.04}$ |
| - Implicit knowledge | 81.22 $\uparrow_{3.10}$ | 90.25 $\uparrow_{2.05}$ | 75.80 $\uparrow_{2.30}$ |
| PEI | **85.68** $\uparrow_{7.56}$ | **92.11** $\uparrow_{3.91}$ | **79.02** $\uparrow_{5.52}$ |

Table 3.8: Ablation Study of PEI on the development set of HotpotQA. Ans F1 stands for answer F1; Sup F1 is supporting F1.

Table 3.8, incorporating pre-training resulted in improvements of $0.70/0.62/1.04$ in Answer F1, Supporting F1, and Joint F1 scores, respectively, compared to the model without pre-training. These results suggest that pre-training in the single-hop QA task helps the model capture valuable information, thereby boosting its performance. However, it is important to note that the observed improvements are relatively modest in the absence of conservation learning.

## 3.6 Conclusion

In this chapter, we present a novel framework inspired by human cognitive theories, which utilizes prompts to connect explicit and implicit knowledge. Our approach incorporates chain-of-thought prompts to extract implicit knowledge from LMs within the given input context, while also integrating question type information to improve the overall performance of the model. Moreover, we propose two training paradigms to PEI framework, and extend PEI on biomedical domain QA to further explore the fusion and relation of explicit and implicit biomedical knowledge and analyze the consistency of the

domain knowledge fusion. Experimental results show that PEI performs comparably to the state-of-the-art on HotpotQA, and excels all baselines on MEDHOP.

Additionally, ablation studies affirm the efficacy and resilience of PEI framework in mirroring human reading comprehension. Moving forward, we intend to expand and apply cognitive theories of human reading to a wider range of reasoning tasks, with the goal of fostering more advanced and intricate reasoning capabilities.

# Chapter 4

# Lexical-based Imbalanced Data Augmentation for Mental Healthcare Moderation

Data augmentation (DA) has garnered significant attention as an alternative method for expanding datasets without requiring additional human annotation efforts, particularly in low-resource, sensitive, and class-imbalanced tasks. However, the majority of contemporary approaches are designed for general domains with relatively balanced data distributions, whereas specific applications such as content moderation frequently exhibit highly skewed distributions. This challenge is further exacerbated by data sensitivity concerns, which render additional human annotations either prohibitively expensive or infeasible. To address this research gap, we introduce a lexical-based imbalanced data

augmentation (LIDA) for content moderation, which an easy-to-implement and interpretable DA method that strategically leverages sensitive lexicons by incorporating them into negative samples to transform these instances into positive examples. Through this mechanism, LIDA facilitates the creation of balanced datasets, thus mitigating skewed distribution challenges. We evaluate our approach on Wiki-TOX and Wiki-ATT, demonstrating the superior performance of our proposed algorithm compared to rule-based baselines, with statistical significance confirmed through comprehensive $p$-value analyses. Furthermore, we extend the application of our method to the mental healthcare domain, validating its efficacy on the Kooth Mental Health dataset. The results substantiate that LIDA is effectively transferable to the mental healthcare domain.

## 4.1 Introduction

The proliferation of cyberbullying and harassment constitutes a significant societal concern due to their deleterious effects on users exposed to inappropriate user-generated content, including violent, disturbing, depressive, or fraudulent materials. Such exposure frequently results in adverse mental health outcomes [199, 200]. Consequently, content moderation represents a domain of substantial business and research significance for online mental health and social communities [201].

Online moderation conducted by service providers or community members represents a human-centered process that enhances the safety, engagement, and efficacy of online mental health community (OHMC) conversations [102]. Moderation strategies

can be classified as either preventive or interventive based on their temporal relationship to user behavior [202]. While the moderation industry has established extensive human-reliant systems, implementing these systems exclusively presents several significant challenges. First, content moderators frequently encounter disturbing images, videos, and text during their work, which may compromise their psychological wellbeing and potentially lead to post-traumatic stress disorder (PTSD). For instance, Facebook-contracted moderators in Arizona have developed mental health conditions attributed to persistent exposure to violent content [1]. Second, the escalating volume of community participants and comment traffic renders the moderation of user-generated content increasingly demanding and resource-intensive. Prolonged periods of continuous moderation activity demonstrably reduce moderator accuracy and effectiveness. Third, extended response times during peak usage periods negatively impact user satisfaction and community engagement. To address these limitations, automated content moderation systems leverage machine learning models trained on extensive textual corpora [203].

As online content continues to expand exponentially, automated moderation techniques have emerged as viable solutions [204], which can be considered as specialized text classification task. However, the domain-specific nature of content moderation presents notable challenges in developing comprehensive gold-standard datasets, which typically require extensive domain expertise and considerable resources. To address these limitations, data augmentation (DA) has been explored to enhance training

[1] https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviewstrauma-working-conditions-arizona

data diversity and quantity without necessitating additional data collection or human annotation. This approach demonstrates significant potential for improving the performance and generalizability of content moderation systems [205].

However, current DA methods predominantly focus on general text classification domain [206, 207, 208, 209], while research in content moderation remains notably deficient, particularly regarding the moderation of toxic and abusive information [210]. Unlike general domain, content moderation typically presents challenges of imbalanced data distribution. For example, previous research demonstrates that across seven Twitter dataset samples, the proportion of negative samples (content adherent to Twitter community guidelines, without hate speech, racist, or gender-discriminatory information) exceeds $80\%$ on average [211].

To address these issues, in this chapter, we propose a lexical-based imbalanced DA (LIDA) method, which is an easy-to-implement yet effective DA methods for automatic content moderation. Compared with prior approaches that heavily depend on extensive required sensitive lexicons [212], LIDA randomly selects sensitive words from a 104-word lexicon collected from Wiktionary [2] and Hatebase [3], and then strategically inserts the sensitive words into negative samples to convert them into positives. Through the utilization of structured lexical knowledge and linguistic features, we generate a more balanced dataset without relying on probabilistic soft labeling techniques [213, 214].

For example, as illustrated in Figure 4.1, we randomly select two sensitive words,

[2] https://en.wiktionary.org/wiki/Category:English_swear_words
[3] https://hatebase.org/

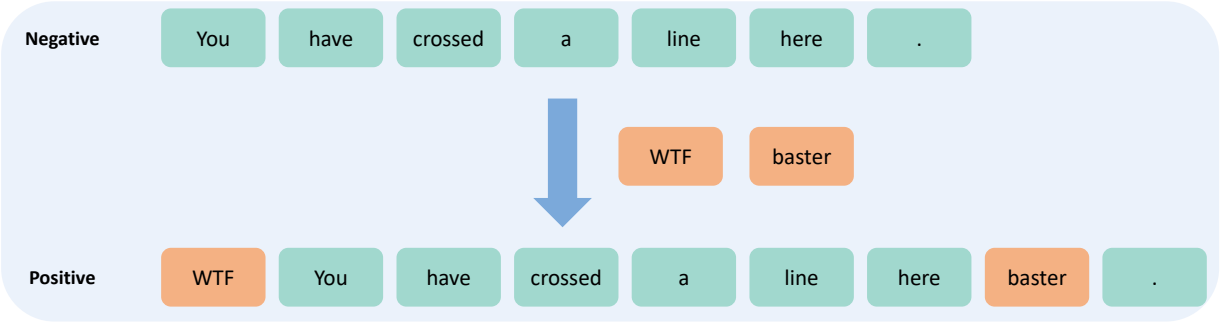


Figure 4.1: Example of the LIDA Method. Sensitive words, *WTF* and *baster* (highlighted as orange blocks), are inserted into a negative sample to transform it into a positive sample. Here, negative samples refer to those that pass content moderation, whereas positive samples denote those that fail moderation.

*WTF* and *baster*, from our wordlist, and subsequently insert them into a negative sample "*You have crossed a line here*". This lexical augmentation transforms the previously benign utterance into positive (content classified as inappropriate). Notably, the positional placement of these sensitive terms within the original text is inconsequential; any resulting augmented text containing such sensitive words would necessarily be categorized as a positive sample warranting moderation due to the inherent inappropriateness of the inserted lexical elements.

Furthermore, we extend the application of our method to mental healthcare moderation contexts. Research indicates that in England, approximately two-thirds of individuals requiring mental health support neither access nor seek assistance through traditional face-to-face services [215]. Young people, who simultaneously exhibit the highest risk for mental health disorders [199] and demonstrate the greatest internet usage rates, represent a critical demographic for intervention. Connecting these young individuals with peer-supervised internet communities, which are facilitated by those with lived experi-

ence, offers potential solutions to the escalating public health challenge of increasing demand. E-mental health services, designed specifically to address treatment barriers, have proliferated globally. OHMC constitute a particularly significant digital health platform category. While these forums demonstrate promising outcomes in supporting youth populations, they may concurrently expose vulnerable users to inappropriate user-generated content, including violent, disturbing, depressing, or fraudulent materials, potentially compromising their psychological wellbeing [216]. Effective mental health promotion, similar to physical health initiatives, necessitates balancing regulatory oversight with engaging user experiences that encourage sustained participation.

Our contributions are summarized as follows:

- We introduce a lexical-based imbalanced data augmentation (LIDA) for content moderation, which an easy-to-implement and interpretable DA method that strategically leverages sensitive lexicons by incorporating them into negative samples to transform these instances into positive examples. Through this mechanism, LIDA facilitates the creation of balanced datasets, thus mitigating skewed distribution challenges.

- We evaluate our approach on Wiki-TOX and Wiki-ATT, demonstrating the superior performance of our proposed algorithm compared to rule-based baselines, with statistical significance confirmed through comprehensive $p$-value analyses.

- Furthermore, we extend the application of our method to the mental healthcare domain, validating its efficacy on the Kooth Mental Health dataset. The results

substantiate that LIDA is effectively transferable to the mental healthcare domain.

**Content Warning.** This article contains examples of hateful and abusive language. All examples are taken from Wikitionary [2] and Hatebase [3] to illustrate its composition.

## 4.2 Related Work

Current DA methods could be roughly classified into two categories: rule-based data augmentation, and generative model-based data augmentation.

### 4.2.1 Rule-Based DA

Rule-based DA approaches typically manipulate words, phrases, or sequences in original textual data through various operations, including swapping, deletion, insertion, and replacement. Wei and Zou [217] introduced the Easy Data Augmentation (EDA) method, which implements four operations on a given sentence: (i) randomly selecting $n$ words to be replaced by their synonyms; (ii) inserting synonyms of random words at random positions; (iii) swapping positions of two words; and (iv) randomly deleting words with probability $p$. While EDA employs word-level operations that potentially alter the semantic class of augmented data, Karimi et al. [206] proposed An Easier Data Augmentation (AEDA), which solely inserts punctuation marks into the original text, thereby preserving class label invariance. Furthermore, researchers have developed learnable compositional paradigms for DA, exemplified by Text AutoAugment (TAA) [207], which represents another significant rule-based mixed DA technique. Moreover, Xiang et al.

[208] developed a part-of-speech (POS) focused lexical substitution approach for DA (PLSDA). This method leverages POS information to identify candidate words for replacement and implements various augmentation strategies to generate semantically related substitutions based on WordNet synonyms.

However, the aforementioned rule-based DA methods demonstrate significant limitations when handling imbalanced datasets, as they fail to rectify skewed distributions post-augmentation. To address this critical limitation, our proposed LIDA method achieves balanced data distribution by strategically incorporating lexical features into negative instances, thereby transforming them into positive instances. Additionally, unlike alternative approaches, our method operates independently of soft label predictions.

### 4.2.2 Generative-Based DA

Generative-based DA approaches predominantly utilize large language models (LLMs) to synthesize novel augmented samples derived from original textual data [218, 209, 219, 220, 221]. Anaby-Tavor et al. [218] developed a DA pipeline leveraging generative pre-training (GPT) [33] with limited labeled data, subsequently filtering the augmented corpus using a classifier trained on the original dataset. Similarly, Yoo et al. [209] employed a GPT-based model to synthesize realistic text samples through GPT-3 [21], while integrating textual perturbations and knowledge distillation from pre-trained transformer-based language models for soft-label prediction. Building upon the enhanced language comprehension capabilities demonstrated by ChatGPT, Dai et al. [219] introduced Aug-

GPT, a text DA approach that reformulates each training sentence into multiple variants that maintain conceptual similarity while exhibiting semantic diversity. These augmented samples subsequently facilitate improved performance in downstream model training tasks.

Despite their capacity to generate linguistically diverse and fluent augmented samples, generative DA methods encounter significant computational constraints related to pre-training and inference processes. Furthermore, these approaches predominantly depend on predicted soft labels for effective data augmentation, introducing additional methodological complexities.

DA strategies range from rule-based manipulations to generative models, with an optimal strategy balancing implementation simplicity against performance enhancement capabilities. This balance is critical given that DA's primary function is to serve as an efficient alternative to collecting additional training data. Extant literature indicates that researchers frequently navigate trade-offs between implementation complexity and performance gains when developing augmentation methods [205].

### 4.2.3 Mental Health Moderation

The effective utilization of online platforms to facilitate connection among individuals with mental health conditions represents one of the ten principal unresolved issues in digital mental health [222]. Research indicates a potential dichotomy in outcomes for OHMC users [223, 224]. The pursuit of social connection constitutes a critical juncture in

the recovery process for clinically isolated and vulnerable individuals [225]. Significant concerns exist among mental health professionals, researchers, and platform administrators regarding potential harm resulting from deliberate provocative communications or unintended adverse effects [226]. While community moderators frequently utilize content removal as an intervention strategy [227], such actions within OHMCs may be perceived as censorship that inhibits the therapeutic self-disclosure these platforms aim to support [228, 224].

The "protective cloak" of anonymity [229] and disinhibition of online communication [230] promote self disclosure; however, content deemed acceptable on mainstream social media platforms [231] may be misinterpreted within OHMCs due to heightened sensitivity to rejection [232] and stigmatizing attitudes [233]. Furthermore, even in the absence of intentional provocation, self-centered communications can negatively impact users, particularly those experiencing elevated depressive symptoms [234]. Emotionally vulnerable individuals attempting interpersonal engagement may develop maladaptive comparative processes and experience vicarious psychological distress [235].

Moderators fulfill diverse functions within OHMCs [236], including providing quasi-therapeutic guidance [102] and maintaining constructive, relevant discourse. Harding and Chung [237] assert that moderation approach and intensity serve as primary distinguishing characteristics between online communities, a finding substantiated in research across various mental health conditions [238].

## 4.3 Methodology

### 4.3.1 Lexical Features

We systematically collect English profanity terms from Wikitionary [2] and hate words from Hatebase [3], the latter being a collaborative and regionalized repository of multilingual hate speech developed through a partnership between the Dark Data Project [4] and The Sentinel Project [5]. This preliminary corpus consisted of 140 lexical items. Subsequently, the compilation underwent critical evaluation and refinement by two doctoral candidates who are native English speakers, as lexical ambiguity in sensitive terms has been identified as a potential confounding factor for model performance (e.g., **northern monkey**). We further comprehensively discuss and analyze the lexical ambiguity and insertion strategy in Section 4.5.5. The final refined corpus comprises 104 sensitive items (the complete wordlist in Appendix A.2).

### 4.3.2 LIDA Algorithm

We present LIDA algorithm in Algorithm 1. Let a training sentence be denoted as $s = [w_1, w_2, ..., w_i, ..., w_l]$, where $w_i$ represents the $i^{th}$ token and $l$ denotes the sentence length. Given a batch of $M$ training samples $\mathcal{D}_{\text{orig}}$, comprising $N$ negative sample $\mathcal{D}_N$ and $P$ positive samples $\mathcal{D}_P$, we introduce an augmentation ratio $t$ as a hyperparameter to control the ratio of data augmentation since excessive perturbations may compromise

[4] https://darkdataproject.org/
[5] https://thesentinelproject.org/

**Algorithm 1:** LIDA Algorithm

**Input:** Given a batch of $M$ training samples $\mathcal{D}_{\text{orig}}$, comprising $N$ negative sample $\mathcal{D}_N$ and $P$ positive samples $\mathcal{D}_P$

**Output:** augmented positive sample $s'$. $\mathcal{D}_{\text{combined}} = \mathcal{D}_{\text{aug}} \cup \mathcal{D}_{\text{orig}}$, where $N$ negative samples and $P + \lfloor N \times t \rfloor$ positive samples

**Initialize:** Augmentation ratio $t \in [0, 1]$

**for** $i = 1$ **to** $\lfloor N \times t \rfloor$ **do**

- Select augmentation depth $d \in \{1, 2, 3\}$ ; `/* number of lexical features */`
- Select lexical feature(s) [LF]$^d$ based on depth $d$ ;
- Generate augmented sample $s'_i = \text{Insert}([\text{LF}]^d, s_i)$, where $s_i$ is randomly selected from negative samples;
- Add $s'_i$ to the augmented dataset $\mathcal{D}_{\text{aug}}$;

**end**

**return** $\mathcal{D}_{combined}$ *with* $N$ *negative samples and* $P + \lfloor N \times t \rfloor$ *positive samples*

model performance. While data noise can be effective for DA, operations like insertion may disrupt the sentence structure, potentially leading to information loss, noise introduction, and even label changes [239]. For example, Karimi et al. [206] propose a method that involved either replacing words selected from the unigram frequency distribution or inserting underscore characters as placeholders for DA. However, excessive noise can mislead the model and degrade performance. Therefore, the augmentation ratio $t$ serves as a crucial hyperparameter in our proposed LIDA algorithm.

For each iteration, we randomly set $d = \{1, 2, 3\}$ as the number of lexical features $[LF]$ to be selected from the lexicon. This operation is written as:

$$[LF]^d = \text{select}(d), \text{ where } d \in 1, 2, 3. \tag{4.1}$$

Parameter $d$ plays a significant role in LIDA (as $d$ in Section 4.5.5), mitigating word ambiguities. Subsequently, we insert $[LF]^d$ into a negative sentence to convert it to be positive instance $s'$:

$$s' = \text{insert}([LF]^d, s), \text{ where } d \in 1, 2, 3. \tag{4.2}$$

The augmented instance $s'$ is combined with the original $s$ to form our new training batch. Since the lexical features consist of sensitive words that violate content moderation policies, intuitively, it is reasonable to assume that negative samples would be converted to positive ones after adding these features.

### 4.3.3 Content Moderation Pipeline

Following the data augmentation procedure, we acquire a balanced corpus with equalized class distributions, thus addressing the inherent imbalance in content moderation datasets. Subsequently, both the augmented dataset $\mathcal{D}_{\text{aug}}$ and the original dataset $\mathcal{D}_{\text{orig}}$ can be utilized for downstream model training purposes, with the combined dataset $\mathcal{D}_{\text{combined}} = \mathcal{D}_{\text{aug}} \cup \mathcal{D}_{\text{orig}}$ serving as the foundation for robust model development.

In this section, we delineate the comprehensive pipeline for content moderation utilizing BERT in conjunction with our proposed LIDA framework. The pipeline consists of several sequential phases, each critical to the overall efficacy of the system. Initially, we leverage LIDA for data augmentation to obtain a balanced training corpus, where negative samples $\mathcal{D}_N$ are transformed into positive samples $\mathcal{D}'_P$ through the strategic insertion of lexical features, as previously formalized. This augmentation process is governed by

the hyperparameter $t$, which controls the ratio of negative samples subjected to transformation.

The content moderation procedure employing BERT encompasses multiple stages of data processing and model development. First, the textual data undergoes preprocessing, where each input text $s_i$ is tokenized using BERT's WordPiece tokenizer, resulting in a token sequence $T(s_i) = [t_1, t_2, \ldots, t_k]$, where $k \leq 512$ due to BERT's architectural constraints. Special tokens, namely `[CLS]` and `[SEP]`, are inserted at the beginning and end of each sequence, respectively. The tokenized sequences are subsequently converted into numerical representations through token embeddings, positional embeddings, and segment embeddings, which are additively combined to form the input representation matrix $\mathbf{X} \in \mathbb{R}^{k \times d}$, where $d = 768$ for $\text{BERT}_{\text{BASE}}$ and $d = 1024$ for $\text{BERT}_{\text{LARGE}}$.

The architecture of our model is constructed by augmenting the pre-trained BERT model with task-specific layers. Specifically, we extract the contextualized representation $\mathbf{h}_{\text{[CLS]}} \in \mathbb{R}^d$ from the final hidden layer corresponding to the `[CLS]` token, which captures the holistic semantic content of the input text. This representation is then processed through a pooling layer to obtain a fixed-dimensional feature vector. Subsequently, we implement a series of fully connected layers with decreasing dimensionality:

$$\mathbf{z}_1 = \sigma(\mathbf{W}_1 \mathbf{h}_{\text{[CLS]}} + \mathbf{b}_1) \tag{4.3}$$

$$\mathbf{z}_2 = \sigma(\mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2) \tag{4.4}$$

$$\hat{y} = \text{softmax}(\mathbf{W}_3 \mathbf{z}_2 + \mathbf{b}_3) \tag{4.5}$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ represent the weight matrices and bias vectors, respectively, and $\sigma(\cdot)$ denotes the activation function, specifically ReLU in our implementation.

The model is fine-tuned using the augmented dataset $\mathcal{D}_{\text{combined}}$, optimizing the cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}_{\text{combined}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\text{combined}}} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \tag{4.6}$$

where $C$ represents the number of classes, $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability for class $c$. We employ the Adam optimizer with a learning rate schedule incorporating warm-up and linear decay phases, formulated as:

$$\eta_t = \begin{cases} \eta_{\text{base}} \cdot \frac{t}{t_{\text{warmup}}} & \text{if } t \leq t_{\text{warmup}} \\ \eta_{\text{base}} \cdot \max\left(0, \frac{t_{\text{total}} - t}{t_{\text{total}} - t_{\text{warmup}}}\right) & \text{if } t > t_{\text{warmup}} \end{cases} \tag{4.7}$$

where $\eta_t$ represents the learning rate at step $t$, $\eta_{\text{base}}$ is the base learning rate, $t_{\text{warmup}}$ is the number of warm-up steps, and $t_{\text{total}}$ is the total number of training steps.

For inference on new data, the process involves preprocessing the input text following the same tokenization and embedding procedures outlined above, followed by forward propagation through the fine-tuned model to obtain class probabilities. The final prediction is determined by selecting the class with the highest probability:

$$\hat{c} = \arg\max_{c \in \{1,2,\ldots,C\}} \hat{y}_c \tag{4.8}$$

It is noteworthy that our proposed LIDA method is a "plug-in" approach, which exhibits considerable adaptability and can be seamlessly integrated with various classification models beyond BERT.

## 4.4 Experiments

### 4.4.1 Datasets

We conduct experiments on two public datasets from the Wikipedia Talk corpus [240]: Wikipedia Toxic dataset[6] and Wikipedia Personal Attack dataset[7]. Both datasets contain human annotations for toxic and personal attack behaviors, respectively. We pre-process these datasets by converting the multiple classification labels into binary classification format for content moderation purposes, referring to them as Wiki-TOX and Wiki-ATT. Unlike the balanced datasets used in prior work [206, 217], our datasets exhibit class imbalance, with positive samples comprising only approximately $10.1\%$ and $7.5\%$ of Wiki-TOX and Wiki-ATT, respectively. Moreover, we evaluate our method on a private mental health moderation dataset provided by Kooth [8], hereafter referred to as the Kooth Mental Health dataset. The statistics of datasets show in Table 4.1.

**Wiki-ATT**. The Wikipedia Personal Attacks dataset [240] is a subset of the Wikipedia Comment Corpus, containing 63M comments from English discussion pages and articles

[6] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

[7] https://www.kaggle.com/datasets/jigsaw-team/wikipedia-talk-labels-personal-attacks

[8] https://www.kooth.com/

| | Data | | | Train | | | Val | | | Test | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Total** | **N** | **P** | **Total** | **N** | **P** | **Total** | **N** | **P** | **Total** | **N** | **P** |
| Wiki-TOX | 159,495 | 143,346 | 16,149 | 139,495 | 125,413 | 14,082 | 10,000 | 8,966 | 1,034 | 10,000 | 8,967 | 1,033 |
| WIKI-ATT | 115,864 | 107,190 | 8,674 | 95,864 | 88,762 | 7,102 | 10,000 | 9,218 | 782 | 10,000 | 9,210 | 790 |
| Kooth | 201,848 | 198,629 | 3,219 | 181,848 | 178,956 | 2,892 | 10,000 | 9,836 | 164 | 10,000 | 9,837 | 163 |

Table 4.1: The statistics of Wiki-TOX and Wiki-ATT. Train, Val and Test represent the training set, validation set and test set respectively. $N$ and $P$ refer to the number of negative and positive samples, respectively.

during 2004-2015. Each comment is annotated and identified as personal attacks by at least 10 workers. The dataset comprises five classes: quoting attack, recipient attack, third-party attack, other attack, and no attack. For our binary classification purposes, we aggregate all four attack categories (quoting, recipient, third-party, and other) as positive samples, while designating the no-attack category as negative samples.

**Wiki-TOX**. Wulczyn et al. [240] introduce the toxic comment dataset, which has been extensively utilized for toxic content detection research [241]. The dataset classifies comments into six types of toxicity: toxic, severe toxic, obscene, threat, insult, identity hate. For our binary classification approach, we designate comments exhibiting any form of toxicity across these six categories as positive samples, while comments without any toxic attributes are classified as negative samples.

**Kooth Mental Health**. Mental health moderation research lacks standardized datasets for comparative evaluation. This study presents a novel corpus derived from Kooth, a leading provider of online mental health services for children, adolescents, and adults in the United Kingdom. This dataset, designated "Kooth Mental Health," comprises moderation decisions regarding user-generated content. The platform employs pre-moderation protocols wherein moderators evaluate all user comments for compliance with established community guidelines prior to publication. Comments violating these standards are withheld from public display. This research corpus has been developed with explicit user consent and organizational authorization for academic investigation. To ensure robust privacy protection, all personally identifiable information has been systematically removed, including but not limited to: user identifiers, identifiable location, identifiable

name, and institutional affiliations.

### 4.4.2 Selected Baselines

We systematically evaluate our proposed LIDA algorithm against established DA methods, including two rule-based approaches: EDA [217] and AEDA [206]; and a generative-based approach, GPT3Mix [209]. The evaluation is conducted across three neural network architectures: (i) a Convolutional Neural Network (CNN) [242] utilizing pre-trained GloVe [243] word embeddings; (ii) a Long Short-Term Memory (LSTM) network [244] comprising a single layer with $128$ hidden units and randomly initialized embedding weights; and (iii) BERT (base, cased version) [10].

- **Easy Data Augmentation (EDA)** [217]. EDA comprises four stochastic operations applied to a given training sample: (i) synonym replacement, where randomly selected $n$ words are substituted with their synonyms; (ii) synonym insertion, where synonyms of randomly chosen words are inserted at random positions, repeating it $n$ times; (iii) random swap, where the positions of two random words are exchanged $n$ times; and 4) random deletion, where each word is removed with probability $p$. Following the authors' recommendations, we implement EDA with a word modification rate of $0.05$[9].

- **An Easier Data Augmentation (AEDA)** [206] presents a minimalist approach to text DA through the random insertion of punctuation marks into the source text.

[9] https://github.com/jasonwei20/eda_nlp

Following previous research for rule-based augmentation techniques [207], we implement AEDA with a punctuation insertion ratio of $0.3$, adhering to the parameters specified in the original implementation[10].

- **GPT3Mix** [209] employs a generative approach wherein it combines real instances to synthesize realistic training examples utilizing GPT-3 [21]. The method incorporates textual perturbation techniques and knowledge distillation processes from pre-trained transformer-based language models to generate probabilistic soft labels, enhancing model robustness.

### 4.4.3 Experimental Settings

We adopt the Adam optimizer [15] along with a linear learning rate scheduler with a warm-up ratio of $0.05$. All experiments are conducted on NVIDIA RTX 6000 GPU and GeForce RTX 3090 GPU, each with 24GB of VRAM. To ensure statistical robustness, each model are executed $10$ times per experimental task, with performance metrics reported as mean values. Statistical significance is assessed through appropriate $p$-value analysis to validate the reliability and stability of our experimental findings.

## 4.5 Results and Analysis

In this section, we present a comprehensive evaluation of our proposed LIDA method through comparative analyses against vanilla deep neural networks (DNNs) and baseline

[10] https://github.com/akkarimi/aeda_nlp/blob/master/code/aeda.py

DA methods, with additional validation in the mental health domain. In section 4.5.5, our ablation study comprises two experimental sets designed to investigate the impact of augmentation ratio and the effectiveness of various insertion strategies within our LIDA method.

### 4.5.1 Compare to Vanilla DNNs

Table 4.2 demonstrates that LIDA consistently enhances performance across all model architectures on both Wiki-TOX and Wiki-ATT datasets compared to non-augmented models. Specifically, LIDA yields average improvements of $4.62/2.55/6.13$ on F1-score and $4.15/3.87/2.39$ on AUC, respectively. Statistical analysis confirms the significance of these improvements, with $p$-values below the $0.05$ threshold across all models, establishing that LIDA-augmented training produces statistically significant performance gains over non-augmented models.

### 4.5.2 Compare to Rule-Based Methods

As demonstrated in Table 4.3, our proposed LIDA method consistently outperforms both rule-based DA baselines, EDA and AEDA, across all model architectures and evaluation metrics on the Wiki-TOX and Wiki-ATT datasets. On the Wiki-TOX dataset, LIDA achieves F1/AUC improvements of $4.30/4.07$ on CNN, $1.48/4.66$ on LSTM, and $4.43/1.95$ on BERT when compared to EDA. Similarly, when compared to AEDA, LIDA demonstrates performance enhancements of $1.26/2.45$ (F1/AUC) on CNN, $0.66/4.23$ on LSTM, and $1.60/1.29$ on

| Datasets | Models | No Aug | | LIDA | | Improvement | | P-Value | |
|---|---|---|---|---|---|---|---|---|---|
| | | F1 | AUC | F1 | AUC | F1 | AUC | F1 | AUC |
| Wiki-TOX | CNN | 65.67 | 75.99 | 72.38 | 82.11 | 6.71 | 6.12 | 0.0000 | 0.0000 |
| | LSTM | 36.84 | 61.26 | 38.52 | 66.06 | 1.68 | 4.80 | 0.0000 | 0.0000 |
| | BERT | 75.59 | 88.95 | 83.65 | 91.57 | 8.06 | 2.62 | 0.0000 | 0.0082 |
| Wiki-ATT | CNN | 69.02 | 79.26 | 71.54 | 81.44 | 2.52 | 2.18 | 0.0066 | 0.0175 |
| | LSTM | 43.30 | 64.16 | 46.51 | 67.10 | 3.21 | 2.94 | 0.0000 | 0.0000 |
| | BERT | 76.82 | 89.04 | 81.02 | 91.20 | 4.20 | 2.16 | 0.0001 | 0.0051 |
| Average | CNN | 67.34 | 77.62 | 71.96 | 81.77 | **4.62** | **4.15** | - | - |
| | LSTM | 40.07 | 62.71 | 42.52 | 66.58 | **2.45** | **3.87** | - | - |
| | BERT | 76.20 | 88.99 | 82.33 | 91.38 | **6.13** | **2.39** | - | - |

Table 4.2: Compare LIDA to models without augmentation. Each experiment have been conducted $10$ times, with reported results representing the mean values. "-" denotes the case where no results are available.

BERT. Comparable performance advantages for LIDA are also observed across all model architectures and metrics on the Wiki-ATT dataset. Statistical validation presented in Table 4.4 confirms the significance of these improvements, with all $p$-values falling below the $0.05$ threshold across all three model architectures when comparing LIDA against both EDA and AEDA. These results hold particular significance for content moderation applications, where high sensitivity in detecting harmful content is critical to minimize false negatives.

We observe that the performance enhancement afforded from our algorithm is less pronounced for LSTM architectures compared to CNN and BERT models. This differential effect can be attributed to the architectural simplicity of the LSTM implementation employed in our study (detailed in Section 4.4.2), particularly its absence of pre-trained

| Models | Wiki-TOX | | Wiki-ATT | |
|---|---|---|---|---|
| | **F1** | **AUC** | **F1** | **AUC** |
| CNN | 65.67 | 75.99 | 69.02 | 79.26 |
| +EDA | 68.08 | 78.04 | 69.17 | 79.71 |
| +AEDA | 71.12 | 79.66 | 69.41 | 79.89 |
| +LIDA | **72.38** | **82.11** | **71.54** | **81.44** |
| LSTM | 36.84 | 61.26 | 43.30 | 64.16 |
| +EDA | 37.04 | 61.40 | 44.38 | 65.45 |
| +AEDA | 37.86 | 61.83 | 45.24 | 65.91 |
| +LIDA | **38.52** | **66.06** | **46.51** | **67.10** |
| BERT | 75.59 | 88.95 | 76.82 | 89.04 |
| +EDA | 79.22 | 89.62 | 77.51 | 90.39 |
| +AEDA | 82.05 | 90.28 | 77.97 | 90.64 |
| +LIDA | **83.65** | **91.57** | **81.02** | **91.20** |

Table 4.3: Compare to rule-based baselines. overall performance is measured by F1 and AUC. F1: F1-score, AUC: Area Under the Receiver Operating Characteristics (AUC-ROC). Each experiment have been conducted 10 times, with reported results representing the mean values.

word embeddings necessary for effectively capitalizing on lexical information. Given the specific configuration of our LSTM architecture, the LIDA method yields comparatively modest performance improvements relative to those observed with CNN and BERT implementations. This finding indicates that the efficacy of DA methods is substantially modulated by model architecture and complexity, which is an observation further substantiated by the considerable influence of pre-trained word embeddings on overall system performance.

| | | Wiki-TOX | | Wiki-ATT | |
|---|---|---|---|---|---|
| | | **F1** | **AUC** | **F1** | **AUC** |
| CNN | EDA | 0.0000 | 0.0000 | 0.0127 | 0.0713 |
| | AEDA | 0.0398 | 0.0055 | 0.0455 | 0.0376 |
| LSTM | EDA | 0.0004 | 0.0000 | 0.0031 | 0.0022 |
| | AEDA | 0.0331 | 0.0000 | 0.0277 | 0.0075 |
| BERT | EDA | 0.0000 | 0.0017 | 0.0024 | 0.0087 |
| | AEDA | 0.0018 | 0.0192 | 0.0013 | 0.0206 |

Table 4.4: Statistical significance as measured by $p$-values is presented for comparisons between LIDA and rule-based baselines (EDA and AEDA).

### 4.5.3 Compare to Generative-Based Methods

To evaluate the comparative efficacy of our proposed rule-based DA algorithm, we conduct systematic analyses against both rule- and generative-based baselines, such as GPT3Mix [209], a state-of-the-art generative-based DA method. The results presented in Table 4.5 demonstrate that GPT3Mix achieves superior performance across CNN, LSTM, and BERT architectures relative to LIDA, which is an expected outcome given the capacity of GPT3Mix to generate linguistically diverse and contextually coherent augmented samples through its utilization of large-scale pre-trained language models, as previously discussed in Section 4.2.

Nevertheless, GPT3Mix entails substantial computational overhead for both pre-training and fine-tuning processes. GPT-based architectures necessitate significant computational infrastructure and exhibit prolonged training durations for task-specific adaptation [245, 139]. The foundation model, GPT-3, comprises 175 billion parameters and requires train-

| Datasets | Models | GPT3Mix | | LIDA | | Improvement | |
|---|---|---|---|---|---|---|---|
| | | F1 | AUC | F1 | AUC | F1 | AUC |
| Wiki-TOX | CNN | 74.48 | 83.98 | 72.38 | 82.11 | -2.10 | -1.87 |
| | LSTM | 38.94 | 66.39 | 38.52 | 66.06 | -0.42 | -0.33 |
| | BERT | 86.32 | 94.53 | 83.65 | 91.57 | -2.67 | -2.96 |
| Wiki-ATT | CNN | 72.94 | 83.30 | 71.54 | 81.44 | -1.40 | -1.86 |
| | LSTM | 47.03 | 67.73 | 46.51 | 67.10 | -0.52 | -0.63 |
| | BERT | 84.28 | 93.11 | 81.02 | 91.20 | -3.26 | -1.91 |
| Average | CNN | 73.71 | 83.64 | 71.96 | 81.77 | **-1.75** | **-1.87** |
| | LSTM | 42.99 | 67.06 | 42.52 | 66.58 | **-0.47** | **-0.48** |
| | BERT | 85.30 | 93.82 | 82.33 | 91.38 | **-2.97** | **-2.44** |

Table 4.5: Comparison between the performance of LIDA and GPT3Mix on Wiki-TOX and Wiki-ATT.

ing on a $45$ TB corpus [21]. Fine-tuning GPT3Mix demands considerably greater computational resources and processing time compared to rule-based alternatives such as EDA, AEDA, and our proposed LIDA method. Consequently, the computational efficiency-performance tradeoff presents a compelling rationale for implementing lightweight augmentation methods that offer reasonable performance while maintaining minimal resource requirements.

### 4.5.4 Evaluation on Mental Health Domain

A notable observation from Table 4.6 is the consistent pattern of zero values ($0.00$) across all evaluation metrics (e.i., precision, recall, and F1-score) for vanilla CNN, LSTM, and BERT models, without data augmentation. The models fail to predict any positive class instances ($TP = 0$), which indicates that it classified all samples as negative. Although

| Model | Vanilla | | | LIDA | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| CNN | 0.00 | 0.00 | 0.00 | 30.23 | 26.71 | 28.31 |
| LSTM | 0.00 | 0.00 | 0.00 | 22.73 | 10.27 | 14.16 |
| BERT | 0.00 | 0.00 | 0.00 | 54.17 | 35.37 | 42.77 |

Table 4.6: Comparison the performance of LIDA with vanilla models which are without data augmentation on the Kooth Mental Health dataset.

the true negative count was substantial ($9,837$), reflecting that the majority of samples were indeed negative class instances, the models nevertheless fail to identify all $163$ positive cases ($FN = 163$). This represents a classic case of "accuracy paradox," where the overall accuracy metric appears satisfactory, but the model demonstrates complete inefficacy in detecting positive class instances, which are often of primary interest in classification tasks, specially in medical domain.

In contrast, our proposed LIDA method demonstrates substantial performance improvements across all architectures. The BERT model with LIDA achieves the highest performance metrics $54.17/35.37/42.77$ in precision/recall/F1, followed by CNN and LSTM. This dramatic improvement from complete non-performance to functional classification capability underscores the critical importance of DA method when applying deep learning models with LIDA to mental health moderation.

These findings suggest that standard implementation approaches may be inadequate for the complexities inherent in mental health data, where subtle linguistic patterns and context-dependent features play crucial roles. The LIDA method appears to provide a fair balance dataset that enable these models to capture relevant patterns that remain

entirely undetected in the vanilla setting.

### 4.5.5 Ablation Studies

**Effect of Augmentation ratio.** We investigate the influence of augmentation ratio on model performance within the Wiki-TOX dataset for CNN and LSTM models. Our analysis indicates that augmentation ratio constitutes a crucial hyperparameter that significantly affects model performance, even when insertion strategies and additional parameters remain constant. As illustrated in Figure 4.2, empirical results demonstrate that optimal performance is consistently achieved when the augmentation ratio falls within the $[50, 60]$ percentage interval. Although establishing a universally optimal augmentation ratio range requires further investigation across diverse datasets and model architectures, our findings substantiate the critical importance of augmentation ratio calibration in maximizing the efficacy of DA techniques.

**Effect of Insertion Strategy.** We conduct a comparison of three insertion strategies, specifically, $d = \{1\}$, $d = \{2\}$, and $d = \{1, 2, 3\}$ to determine their relative efficacy in DA. Results presented in Table 4.7 demonstrate that these strategies significantly influence model performance. The $d = \{1\}$ strategy consistently yielded the least favorable outcomes across all evaluated models. This performance deficit can be attributed to lexical ambiguity phenomena, exemplified by phrases such as "**northern monkey**". This expression functions as a pejorative term in southern England, connoting perceived intellectual inferiority and cultural unsophistication of northern residents—despite its reclamation

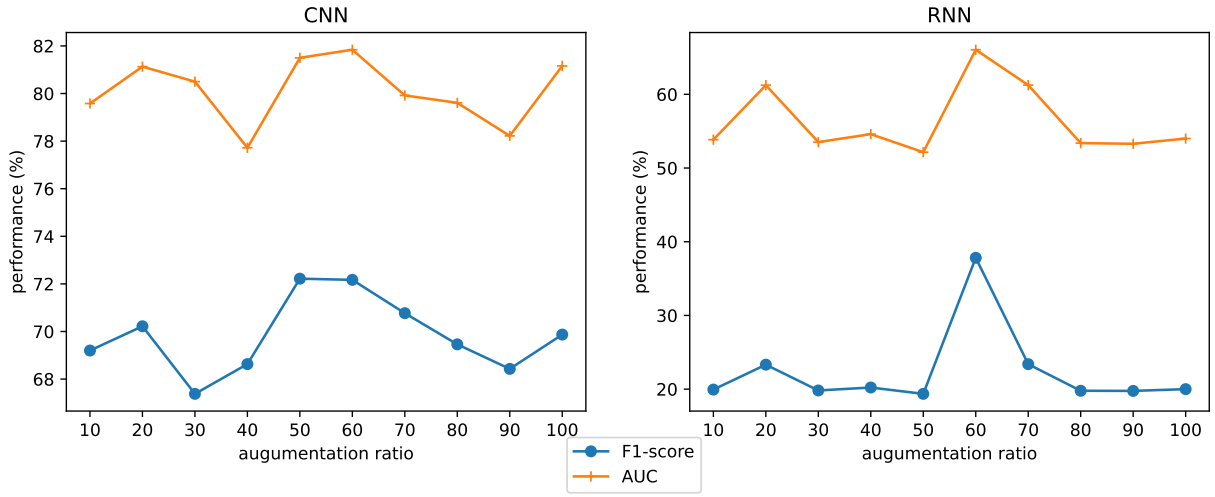


Figure 4.2: Effect of augmentation ratio on Wiki-TOX dataset.

| | CNN | | LSTM | | BERT | |
|---|---|---|---|---|---|---|
| | **F1** | **AUC** | **F1** | **AUC** | **F1** | **AUC** |
| d=1 | 65.79 | 79.79 | 37.49 | 65.62 | 80.81 | 86.22 |
| d=2 | 70.81 | 81.22 | 37.36 | **66.09** | 81.89 | 88.33 |
| d=1, 2, 3 | **72.38** | **82.11** | **38.52** | 66.06 | **83.65** | **91.57** |

Table 4.7: Results of different insertion strategies on Wiki-TOX. $d$ denotes the number of lexical features that are inserted into the original sample.

in some northern contexts, as evidenced by a Leeds establishment named "The Northern Monkey." When employed antagonistically toward northerners, the phrase constitutes hate speech and can be inserted into neutral text to alter sentiment classification from negative to positive. However, in alternative contexts, the phrase may simply denote simian species inhabiting northern regions. To address such contextual ambiguities, we implemented a randomized combinatorial insertion strategy that mitigates phrase-level semantic ambiguity through multi-level contextual embedding.

## 4.6 Conclusion

In this chapter, we introduce a simple yet effective data augmentation (DA) method termed lexical-based imbalanced data augmentation (LIDA) in areas from content moderation to mental health. LIDA utilizes lexical features to transform negative samples into positives, thereby achieving balanced datasets without requiring soft labels or human annotation. Our experimental results demonstrate that LIDA significantly enhances model generalization capabilities while reducing the burden of manual annotation. We evaluate the efficacy of our approach across multiple datasets. The results indicate that our algorithm outperforms other rule-based baselines, with statistical analysis of $p$-values confirming the effectiveness and stability of the LIDA method. Consequently, our approach presents a competitive alternative to traditional augmentation techniques for imbalanced data. Although LIDA exhibits lower performance compared to generative DA methods based on LLMs, it remains valuable in automated content moderation contexts. This value stems from its limited computational requirements, robust performance metrics, enhanced explainability, reduced privacy concerns, and capacity to incorporate human moderation expertise.

# Chapter 5

# Conclusion

This thesis investigates the methodological and paradigmatic transition of deep learning and explainable & interpretable AI from general domain to healthcare. Through a series of interconnected studies, we have addressed critical challenges in representation learning, knowledge integration, data augmentation, and explainability & interpretability. Our contributions advance both theoretical frameworks and practical implementations that facilitate the responsible deployment of AI in healthcare contexts.

## 5.1 Limitations and Future Directions

Despite the advances presented in this thesis, several limitations warrant acknowledgment and point toward future research directions. The Similarity-Dissimilarity Loss, while effective across multimodal data in MSCL, may benefit from further optimization for extreme-scale label spaces, more sophisticated modeling of label dependencies,

and more advanced weighting factors. Future work could explore non-linear similarity-dissimilarity weighting factors and learnable hyperparameter for contrastive loss function to capture more complex semantic relations and long-tailed distribution.

The PEI framework, though promising, currently relies on LMs that may not fully capture specialized domain knowledge. Future research should investigate how to efficiently integrate structured domain knowledge (e.g., medical ontologies) with the implicit knowledge in LMs to enhance reasoning capabilities in specialized domains.

For LIDA, while our approach offers advantages in interpretability and efficiency, generative-based augmentation methods demonstrated superior performance in some experiments. Hybrid approaches that combine the interpretability of lexical augmentation with the diversity of generative methods represent a promising direction for future work.

In the broader context of XIAI in healthcare, significant challenges remain in balancing model performance with transparency, addressing potential biases in explanations, and ensuring that interpretability methods are accessible to healthcare practitioners. Developing standardized evaluation frameworks for XIAI methods in healthcare contexts will be crucial for benchmarking progress and facilitating clinical integration.

### 5.1.1 Challenges and Opportunities for Clinical Translation

We further discuss the key opportunities and challenges surrounding the clinical translation of XIAI. XIAI presents significant opportunities such as leveraging attention mech-

anisms to combine and interpret multi-modal information (text, images, genetic data, clinical history) for personalized medicine and combining deep learning with causal modelling towards further enhancing inherent interpretability. However, our findings also suggest that all these methods require strong in-house technical expertise to infer XIAI. Other key challenges include the lack of established best practices for XIAI selection based on data and problem type, as well as the unmet need for systematic evaluation methods and high-quality benchmarks. In the following paragraphs, we expand on these challenges and opportunities, identified from the perspective of clinical translation.

Considering their early development phase, more extensive studies are needed in the future to develop XIAI methods that will allow "global" (end-to-end) interpretations that go beyond visualization maps, devise more transparent XIAI that will require less technical oversight and design XIAI evaluation metrics that are critical to establish best practices for XIAI selection. The following points can therefore guide future work towards strengthening and democratizing XIAI in healthcare AI.

**Challenges:** The majority of research focus on local XIAI, with limited studies involving a global XIAI approach in healthcare based on our findings [41]. Moving beyond local XIAI (which provides relatively limited insights based on specific inputs or features) towards "globally" enhancing the transparency of the entire modeling process might be necessary to reduce the requirement of in-house technical expertise and to develop readily translational XIAI methods for end-users.

While attention-based IAI holds promise, it primarily relies on visualizing important attention weights as heat maps [246, 247]. Heat maps may reveal patterns in the

way the model prioritizes specific types of words or grammatical structures, providing some insight into which parts of the processed text are most influential. However, heat maps highlight what the model "looks at," but not necessarily how it interprets that information. Moreover, attention weights do not linearly correlate with model outcomes [64, 79]. These two main limitations currently compromise attention-based IAI. The adoption of current XIAI in healthcare and industry systems is therefore challenging, given the limited access to global XIAI techniques and the absence of robust XIAI metrics, ground truths and benchmark data/ studies, which form an important barrier to their systematization.

XIAI systematization and deployment necessitate further thorough work. A possible practical mitigation to accelerate the systematization of robust and transparent XIAI developments for AI in healthcare is to bring "humans into the deep learning loop" [248, 249]: domain experts, end-users, policymakers and patients will be able to contribute to the XIAI method design, development and evaluation. This collective approach can potentially lead to the emergence of robust and ready-to-use XIAI methods across different AI and medical tasks.

**Opportunities:** Although Transformers are the dominant deep learning backbones recently, individual attention mechanisms are still one of the most frequently and diversified used IAI technique in healthcare [75]. Attention and self-attention mechanisms have also recently been used in combination with CNN models for medical image analysis, as lighter alternatives of Transformers [250]. This demonstrates their diverse applicability across data domains and tasks. A major opportunity identified is the versatility of at-

tention mechanisms to be combined with multiple deep learning structures (e.g., CNN [251], RNN [252], BERT [253] and full Transformers) [79]. Another translational opportunity identified via the use of attention mechanisms is to simultaneously model and interpret information from variable multi-modal data (e.g., text, images, genetics, clinical history) [250]. Based on the previous, developing multi-modal XIAI can potentially support personalized medicine which benefits from combining patient-level information from multiple sources [254]. Despite the challenges associated with attention-based heat maps described above, combining information from multiple sources e.g., images and text, can potentially enhance interpretability. A characteristic example are deep learning models developed for MIMIC data analysis, in which features extracted from X-ray images and radiology reports are combined to create a unified representation that fuses information from both modalities [68, 69, 255, 256, 257]. Attention heat maps can highlight which parts of the image and text were most influential in the model decision. Combining image and text data can offer insights into a model's decision-making that go beyond the limitations of single-modal (text-only) attention heat maps.

While our research [75] highlights the benefits of using IAI against XAI, their combination can be useful as an auxiliary assessment to cross-evaluate IAI and XAI outcomes. For instance, the SHAP method that can provide both global and local explanations [76], has been effectively combined with a Transformer encoder for suicide risk prediction [258]. Although this work focused on the SHAP method to perform XAI, future work could aim to also co-extract IAI information from specific attention mechanisms within Transformers.

Causality is an emerging topic in deep learning which aims to improve model interpretability, fairness and generalization [259, 260]. The fundamental aim of causal deep learning is to unravel causal relationships between variables which determine the model's decision-making process [259]. Understanding how the data are causally related can help us design better deep learning models. This leads to more reliable predictions and a deeper understanding of how our models work [261]. Next to identifying causal paths between data, central to causality is also to understand the relationships between cause and effect [260]. Understanding cause and effect is crucial for many important decisions. In clinical trials for example, doctors need to know if a new drug actually improves patient outcomes (not just whether there is a statistical correlation). In our study, a Bayesian Network-derived causal graph has been fed into a TabNet, showing promising IAI results [262]. Further work in this field can significantly enhance IAI in deep learning for healthcare. We endeavour to inspire and guide relevant benchmark studies to thoroughly examine XIAI in terms of strengthening AI applications in healthcare. For example, leveraging causal mechanisms to better understand how foundation models such as Transformers interact with data or prompts, may be a critical path forward.

### 5.1.2 Inspired by the Future

**Go Large:** Recently, LLMs have attracted significant attention in AI [263, 21, 245]. The intersection of LLMs and healthcare creates unique opportunities towards designing future studies, from drug discovery to personalized diagnosis and treatment [264, 57, 53].

Healthcare data analysis is one of the high gain-high risk domains for LLMs [265]. One of the limitations of LLMs is that they sometimes tend to "hallucinate" results [263, 21, 245, 263, 21, 245]. This can add considerable barriers to their utilization in healthcare settings [264, 57, 53]. As discussed in subsection 5.1.1, causal inference can be an effective solution towards enhancing IAI [259]. Although designing causal logic inside LLMs is challenging due to their architectural complexity and the fact that models are already "trained" on a causal-agnostic mode, there have been recent attempts which aim to develop causal reasoning between prompts and responses [266, 267].

**Go Small:** An ongoing discussion in the community is whether LLMs or domain-specific smaller models can be more robust solutions for healthcare data [268]. Recently, smaller-parameter domain-specific LMs have outperformed larger LMs when examined on clinical notes from large public databases (MIMIC) [269]. This approach has several possible benefits: (i) small-parameter LMs can be trained using in-house computational capabilities, which (ii) minimizes the risks associated with transmitting sensitive patient data to cloud-based or external servers (thus, adhering to privacy regulations). Moreover, (iii) causal deep learning techniques can be more optimally designed and validated since they can be trained on domain-specific data from scratch.

## 5.2 Broader Impact and Concluding Remarks

The methodological contributions of this thesis collectively advance the state-of-the-art in applying deep learning and XIAI techniques from general domains to healthcare-specific

applications. Our work demonstrates that specialized approaches addressing the unique challenges of healthcare data, such as extreme label spaces, complex reasoning requirements, sensitive information, and strict transparency demands, which can significantly improve model performance and usability in clinical contexts.

In conclusion, this thesis contributes to bridging the gap between general-domain AI and healthcare applications through methodological innovations in representation learning, knowledge integration, data augmentation, and explainability & interpretability. While challenges remain, the frameworks and methods presented provide a foundation for developing AI systems that are not only effective but also transparent, fair, and aligned with the unique requirements of healthcare applications. As AI continues to transform healthcare, such responsible approaches will be essential for realizing the technology's full potential to improve patient outcomes and healthcare delivery.

# Appendix A

# Appendix

## A.1 Computational Cost Analysis of Similarity-Dissimilarity Loss

The key computation in all supervised contrastive loss functions, including our proposed Similarity-Dissimilarity Loss, is the similarity between representations and the size of the positive set $\mathcal{P}(i)$ for each anchor $i$. All methods share the same core structure:

$$\mathcal{L}_i = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \tag{A.1}$$

Our method introduces a reweighting factor $\mathcal{K}^s_{i,p}\mathcal{K}^d_{i,p} \in [0, 1]$ only in the numerator of

the positive terms:

$$\mathcal{L}_i^{\text{our}} = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\mathcal{K}_{i,p}^s \mathcal{K}_{i,p}^d \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau)} \tag{A.2}$$

Importantly, the denominator (negative set) remains unchanged across all methods, including *ALL*, *ANY*, *MulSupCon*, and our proposed approach. The positive set $\mathcal{P}(i)$ in our method is structurally identical to that of *ANY*, and while it is typically larger than the set used in *ALL*, it is generally smaller than or equal to that in *MulSupCon*, which constructs multiple positive sets. The computation of $\mathcal{K}_{i,p}^s = \frac{|\mathcal{S} \cap \mathcal{T}|}{|\mathcal{S}|}$ and $\mathcal{K}_{i,p}^d = \frac{1}{1+|\mathcal{T} \setminus (\mathcal{S} \cap \mathcal{T})|}$ relies only on simple set operations applied to one-hot label vectors, making it both efficient and linear in the number of labels. Furthermore, our method does not require any additional encoder forward passes or the introduction of extra model parameters.

Memory and Runtime Considerations. The proposed Similarity-Dissimilarity Loss introduces negligible overhead in terms of both memory and runtime. Specifically, it does not increase memory consumption, as the encoder architecture remains unchanged and the feature dimensionality is preserved. The additional runtime cost is also minimal, arising only from lightweight scalar computations required for each positive pair (i.e., computing $\mathcal{K}_{i,p}^s \mathcal{K}_{i,p}^d$).

In contrast, methods like *MulSupCon* may require repeated label-wise positive sets and multiple forward loss computations per label, depending on implementation. Thus, our method maintains comparable or better efficiency while offering a more principled weighting mechanism.

## A.2 Sensitive Wordlist

**ID Sources Words**

001 Wikitionary *arse*

002 Wikitionary *ass*

003 Wikitionary *asshole*

004 Wikitionary *bastard*

005 Wikitionary *bitch*

006 Wikitionary *bollocks*

007 Wikitionary *brotherfucker*

008 Wikitionary *bugger*

009 Wikitionary *bullshit*

010 Wikitionary *child-fucker*

011 Wikitionary *Christ on a bike*

012 Wikitionary *Christ on a cracker*

013 Wikitionary *cocksucker*

014 Wikitionary *crap*

015 Wikitionary *cunt*

016 Wikitionary *damn*

017 Wikitionary *effing*

018 Wikitionary *fatherfucker*

019 Wikitionary *frigger*

020 Wikitionary *fuck*

021 Wikitionary *goddamn*

022 Wikitionary *godsdamn*

023 Wikitionary *hell*

024 Wikitionary *holy shit*

025 Wikitionary *horseshit*

026 Wikitionary *in shit*

027 Wikitionary *Jesus Christ*

028 Wikitionary *Jesus fuck*

029 Wikitionary *Jesus H. Christ*

030 Wikitionary *Jesus Harold Christ*

031 Wikitionary *Jesus wept*

032 Wikitionary *Jesus, Mary and Joseph*

033 Wikitionary *Judas Priest*

034 Wikitionary *motherfucker*

035 Wikitionary *nigga*

036 Wikitionary *piss*

037 Wikitionary *prick*

038 Wikitionary *shit*

039 Wikitionary *shit ass*

040 Wikitionary *sisterfucker*

041 Wikitionary *slut*

042 Wikitionary *son of a bitch*

043 Wikitionary *son of a whore*

044 Wikitionary *sweet Jesus*

045 Wikitionary *twat*

046 Hatebase *buttfucker*

047 Hatebase *assplay*

048 Hatebase *sucker*

049 Hatebase *homophobic slurs*

050 Hatebase *nerdiness*

051 Hatebase *putz*

052 Hatebase *ass-rape*

053 Hatebase *ponce*

054 Hatebase *narcism*

055 Hatebase *muthafucker*

056 Hatebase *dastardliness*

057 Hatebase *african-negros*

058 Hatebase *virgin*

059 Hatebase *arsehole*

060 Hatebase *crook*

061 Hatebase *self-destruction*

062 Hatebase *self-annihilation*

063 Hatebase *vestal*

064 Hatebase *pervert*

065 Hatebase *self harm*

066 Hatebase *slay*

067 Hatebase *felon*

068 Hatebase *virgo the virgin*

069 Hatebase *outrage*

070 Hatebase *self injury*

071 Hatebase *shoot down*

072 Hatebase *whoreson*

073 Hatebase *ill-treat*

074 Hatebase *terrorist*

075 Hatebase *bastard*

076 Hatebase *blackguard*

077 Hatebase *maltreat*

078 Hatebase *ill-usage*

079 Hatebase *mistreat*

080 Hatebase *suicide*

081 Hatebase *dickhead*

082 Hatebase *maltreatment*

083 Hatebase *virginal*

084 Hatebase *prick*

085 Hatebase *shit*

086 Hatebase *ravish*

087 Hatebase *rape*

088 Hatebase *ill-use*

089 Hatebase *slaying*

090 Hatebase *sexually assault*

091 Hatebase *violate*

092 Hatebase *cocksucker*

093 Hatebase *wtf*

094 Hatebase *self loathe*

095 Hatebase *gay*

096 Hatebase *lesbian*

097 Hatebase *terrorist*

098 Hatebase *murder*

099 Hatebase *assault*

100 Hatebase *kill*

101 Hatebase *robbery*

102 Hatebase *dumbcunt*

103 Hatebase *topless*

104 Hatebase *dickdipper*

# Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[3] Marvin Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, 1969.

[4] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[5] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[9] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations

using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR)*, 2015. arXiv preprint arXiv:1409.0473.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2014.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[18] Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017.

[19] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*, pages 1–12, 2017.

[20] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[23] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[27] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 328–339. Association for Computational Linguistics, 2018.

[28] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-

efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[29] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *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)*, pages 4582–4597, 2021.

[30] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[31] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[33] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[36] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[37] OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[38] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

[39] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[40] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[41] Guangming Huang, Yunfei Long, Cunjin Luo, Jiaxing Shen, and Xia Sun. Prompting explicit and implicit knowledge for multi-hop question answering based on human reading process. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13179–13189, 2024.

[42] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2744–2751. Association for Computational Linguistics, 2020.

[43] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. Artificial intelligence index report 2025, 2025.

[44] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo

Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[45] MD Imran Hossain, Ghada Zamzmi, Peter R Mouton, MD Sirajus Salekin, Yu Sun, and Dmitry Goldgof. Explainable ai for medical data: current methods, limitations, and future directions. *ACM Computing Surveys*, 57(6):1–46, 2025.

[46] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[47] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[48] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[49] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[50] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[51] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

[52] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[53] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[54] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[55] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation

and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[56] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

[57] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[58] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[59] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[60] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of*

*the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, 2021.

[61] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[62] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[63] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, 2020.

[64] Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*, 2021.

[65] Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. Explainable artificial intel-

ligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185, 2020.

[66] Ozan Ozyegen, Devika Kabe, and Mucahit Cevik. Word-level text highlighting of medical texts for telehealth services. *Artificial Intelligence in Medicine*, 127:102284, 2022.

[67] Hans-Christian Thorsen-Meyer, Davide Placido, Benjamin Skov Kaas-Hansen, Anna P Nielsen, Theis Lange, Annelaura B Nielsen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, et al. Discrete-time survival analysis in the critically ill: a deep learning approach using heterogeneous data. *NPJ digital medicine*, 5(1):142, 2022.

[68] Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. Explainable prediction of medical codes with knowledge graphs. *Frontiers in bioengineering and biotechnology*, 8:867, 2020.

[69] Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, 116:103728, 2021.

[70] Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[71] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[72] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[73] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[74] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[75] Guangming Huang, Yingya Li, Shoaib Jameel, Yunfei Long, and Giorgos Papanastasiou. From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Computational and Structural Biotechnology Journal*, 2024.

[76] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[77] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[78] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[79] Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945, 2021.

[80] Hali Lindsay, Johannes Tröger, and Alexandra König. Language impairment in alzheimer's disease—robust and explainable evidence for ad-related deterioration of spontaneous speech through multilingual machine learning. *Frontiers in aging neuroscience*, page 228, 2021.

[81] Diego Garcia-Olano, Yasumasa Onoe, Ioana Baldini, Joydeep Ghosh, Byron C Wallace, and Kush Varshney. Biomedical interpretable entity representations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3547–3561, 2021.

[82] Charlene Jennifer Ong, Agni Orfanoudaki, Rebecca Zhang, Francois Pierre M Caprasse, Meghan Hutch, Liang Ma, Darian Fard, Oluwafemi Balogun, Matthew I Miller, Margaret Minnig, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PloS one*, 15(6):e0234908, 2020.

[83] Giacomo Frisoni and Gianluca Moro. Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge. In *Data Management Technologies and Applications: 9th International Conference, DATA 2020, Virtual Event, July 7–9, 2020, Revised Selected Papers 9*, pages 293–318. Springer, 2021.

[84] Alexios Mandalios, Alexandros Chortaras, Giorgos Stamou, and Michalis Vazirgiannis. Enriching graph representations of text: Application to medical text classification. In *Complex Networks & Their Applications IX: Volume 2, Proceedings of the Ninth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2020*, pages 92–103. Springer, 2021.

[85] Ya Zhang, Shuai Yang, Hao Wen, Pei Xie, and Yuanqing Wu. Unified framework for ner and re tasks with strong interpretability on chinese medicine instruction parsing. In *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pages 460–466. IEEE, 2022.

[86] Eben Holderness, Nicholas Miller, Philip Cawkwell, Kirsten Bolton, Marie Meteer, James Pustejovsky, and Mei-Hua Hall. Analysis of risk factor domains in psychosis patient health records. *Journal of Biomedical Semantics*, 10:1–10, 2019.

[87] Elena Álvarez Mellado, Eben Holderness, Nicholas Miller, Fyonn Dhang, Philip Cawkwell, Kirsten Bolton, James Pustejovsky, and Mei Hua-Hall. Assessing the efficacy of clinical sentiment analysis and topic extraction in psychiatric readmis-

sion risk prediction. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 81–86, 2019.

[88] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4942–4950, 2018.

[89] Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

[90] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, 2020.

[91] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019.

[92] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*, 2019.

[93] Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations*, 2019.

[94] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[95] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[96] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[97] Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. Rethinking masked image modelling for medical image representation. *Medical Image Analysis*, 98:103304, 2024.

[98] Jianguo Mao, Jiyuan Zhang, Zengfeng Zeng, Weihua Peng, Wenbin Jiang, Xiangdong Wang, Hong Liu, and Yajuan Lyu. Hierarchical representation-based dynamic reasoning network for biomedical question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1480–1489, 2022.

[99] Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. Mimic-iv-icd: A new benchmark for extreme multilabel classification. *arXiv preprint arXiv:2304.13998*, 2023.

[100] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[101] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55, 2020.

[102] David Wadden, Tal August, Qisheng Li, and Tim Althoff. The effect of moderation on online mental health conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 751–763, 2021.

[103] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[104] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159, 2022.

[105] Shaoxiong Ji, Xiaobo Li, Wei Sun, Hang Dong, Ara Taalas, Yijia Zhang, Honghan Wu, Esa Pitkänen, and Pekka Marttinen. A unified review of deep learning for automated medical coding. *ACM Computing Surveys*, 56(12):1–41, 2024.

[106] Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. Automated medical coding on mimic-iii and mimic-iv: a critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2572–2582, 2023.

[107] Guangming Huang, Yunfei Long, and Cunjin Luo. Similarity-dissimilarity loss for multi-label supervised contrastive learning, 2025.

[108] Guangming Huang, Yunfei Long, and Cunjin Luo. Similarity-dissimilarity loss for multi-label supervised contrastive learning. *arXiv preprint arXiv:2410.13439*, 2024.

[109] Guangming Huang, Yunfei Long, and Cunjin Luo. Improving multi-hop question answering with prompting explicit and implicit knowledge aligned human reading comprehension. *International Journal of Machine Learning and Cybernetics*, pages 1–16, 2025.

[110] Guangming Huang, Yunfei Long, Cunjin Luo, and Yingya Li. Lida: Lexical-based imbalanced data augmentation for content moderation. In *Proceedings of the 37th*

*Pacific Asia Conference on Language, Information and Computation*, pages 59–69, 2023.

[111] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

[112] Haohui Wang, Weijie Guan, Chen Jianpeng, Zi Wang, and Dawei Zhou. Towards heterogeneous long-tailed learning: Benchmarking, metrics, and toolbox. In *Advances in Neural Information Processing Systems*, volume 37, pages 73098–73123, 2024.

[113] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[114] Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16786–16793, 2024.

[115] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[116] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022.

[117] Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, 2023.

[118] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

[119] Alexandre Audibert, Aurélien Gauffre, and Massih-Reza Amini. Exploring contrastive learning for long-tailed multi-label text classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 245–261. Springer, 2024.

[120] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[121] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[122] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[123] Fei Teng, Yiming Liu, Tianrui Li, Yi Zhang, Shuangqing Li, and Yue Zhao. A review on deep neural networks for icd coding. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4357–4375, 2022.

[124] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, 2018.

[125] Boon Liang Clarence Ng, Diogo Santos, and Marek Rei. Modelling temporal document sequences for clinical icd coding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1640–1649, 2023.

[126] Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. Corelation: Boosting automatic icd coding through contextualized code relation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3997–4007, 2024.

[127] Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 808–814, 2022.

[128] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341, 2021.

[129] Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Ramesh Kashyap, and Stefan Winkler. A two-stage decoder for efficient icd coding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4658–4665, 2023.

[130] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6, 2024.

[131] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[132] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[133] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[134] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, 2018.

[135] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[136] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[137] Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. Plm-icd: Automatic icd coding with pretrained language models. In *Proceedings of the 4th Clinical Natural*

*Language Processing Workshop*, pages 10–20, 2022.

[138] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[139] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024.

[140] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

[141] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[142] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, 2018.

[143] Zhenyun Deng, Yonghua Zhu, Yang Chen, Qianqian Qi, Michael J Witbrock, and Patricia Riddle. Prompt-based conservation learning for multi-hop question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1791–1800, 2022.

[144] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, 2023.

[145] Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. End-to-end beam retrieval for multi-hop question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1718–1731, 2024.

[146] Frank Smith. *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Holt, Rinehart and Winston, New York, 1971.

[147] Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. Integration of word meaning and world knowledge in language comprehension. *science*, 304(5669):438–441, 2004.

[148] Mark A Clarke and Sandra Silberstein. Toward a realization of psycholinguistic principles in the esl reading class 1. *Language learning*, 27(1):135–154, 1977.

[149] Koh Moy Yin. The role of prior knowledge in reading comprehension. 1985.

[150] R Scott Baldwin, Ziva Peleg-Bruckner, and Ann H McClintock. Effects of topic interest and prior knowledge on reading comprehension. *Reading research quarterly*, pages 497–504, 1985.

[151] Noureldin Mohamed Abdelaal and Amal Saleh Sase. Relationship between prior knowledge and reading comprehension. *Advances in Language and Literary Studies*, 5(6):125–131, 2014.

[152] Charles Jin and Martin Rinard. Evidence of meaning in language models trained on programs. *arXiv preprint arXiv:2305.11169*, 2023.

[153] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[154] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[155] Giovanni Maria Biancofiore, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Fedelucio Narducci. Interactive question answering systems: Literature review. *ACM Computing Surveys*, 56(9):1–38, 2024.

[156] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36, 2022.

[157] Pierre Zweigenbaum. Question answering in biomedicine. In *Proceedings Workshop on Natural Language Processing for Question Answering, EACL*, volume 2005, pages 1–4. Citeseer, 2003.

[158] Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[159] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[160] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

[161] Boshi Wang, Xiang Deng, and Huan Sun. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730, 2022.

[162] Zefeng Lin, Weidong Chen, Yan Song, and Yongdong Zhang. Prompting few-shot multi-hop question generation via comprehending type-aware semantics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3730–3740, 2024.

[163] Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, 2024.

[164] Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. Few-shot reranking for multi-hop qa via language model prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15882–15897, 2023.

[165] Peng Gao, Feng Gao, Peng Wang, Jian-Cheng Ni, Fei Wang, and Hamido Fujita. Cluereader: Heterogeneous graph attention network for multi-hop machine reading comprehension. *Electronics*, 12(14):3183, 2023.

[166] Yongping Du, Rui Yan, Ying Hou, Yu Pei, and Honggui Han. Adversarial entity graph convolutional networks for multi-hop inference question answering. *Expert Systems with Applications*, 258:125098, 2024.

[167] Peng Gao, Feng Gao, Jian-Cheng Ni, Yu Wang, Fei Wang, and Qiquan Zhang. Medical knowledge graph question answering for drug-drug interaction predic-

tion based on multi-hop machine reading comprehension. *CAAI Transactions on Intelligence Technology*, 2024.

[168] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[169] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[170] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[171] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[172] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, 2020.

[173] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, 2020.

[174] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

[175] Yixuan Tang, Hwee Tou Ng, and Anthony Tung. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, 2021.

[176] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

[177] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109, 2019.

[178] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised question decomposition for question answering. In *Proceedings of the*

*2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880, 2020.

[179] Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2335–2345, 2019.

[180] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[181] Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150, 2019.

[182] Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9073–9080, 2020.

[183] Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu. Is graph structure necessary for multi-hop question answering? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7187–7192, 2020.

[184] Ronghan Li, Lifang Wang, Shengli Wang, and Zejun Jiang. Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension. In *IJCAI*, pages 3857–3863, 2021.

[185] Bohong Wu, Zhuosheng Zhang, and Hai Zhao. Graph-free multi-hop reading comprehension: A select-to-guide strategy. *arXiv preprint arXiv:2107.11823*, 2021.

[186] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Making neural qa as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, 2017.

[187] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*, 2017.

[188] Yu Cao, Meng Fang, and Dacheng Tao. Bag: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 357–362, 2019.

[189] Yichen Jiang, Nitish Joshi, Yen-Chun Chen, and Mohit Bansal. Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2714–2725, 2019.

[190] Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. Nlprolog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6151–6161, 2019.

[191] Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W. Cohen. Differentiable reasoning over a virtual knowledge base. In *International Conference on Learning Representations*, 2020.

[192] Wenya Wang and Sinno Pan. Deep inductive logic reasoning for multi-hop reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4999–5009, 2022.

[193] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[194] Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th workshop on biomedical language processing*, pages 143–154, 2021.

[195] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model.

In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, 2022.

[196] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[197] Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16339–16347, 2024.

[198] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[199] Vikram Patel, Alan J Flisher, Sarah Hetrick, and Patrick McGorry. Mental health of young people: a global public-health challenge. *The Lancet*, 369(9569):1302–1313, 2007.

[200] Rosemary Sedgwick, Sophie Epstein, Rina Dutta, and Dennis Ougrin. Social media, internet use and suicide attempts in adolescents. *Current opinion in psychiatry*, 32(6):534, 2019.

[201] Sally McManus, Paul E Bebbington, Rachel Jenkins, and Terry Brugha. *Mental health and wellbeing in England: the adult psychiatric morbidity survey 2014*. NHS digital, 2016.

[202] Tanner Skousen, Hani Safadi, Colleen Young, Elena Karahanna, Sami Safadi, and Fouad Chebib. Successful moderation in online patient communities: inductive case study. *Journal of medical Internet research*, 22(3):e15983, 2020.

[203] Pablo Navarro, Matthew Bambling, Jeanie Sheffield, Sisira Edirippulige, et al. Exploring young people's perceptions of the effectiveness of text-based online counseling: Mixed methods pilot study. *JMIR Mental Health*, 6(7):e13152, 2019.

[204] Ariadna Matamoros-Fernández and Johan Farkas. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224, 2021.

[205] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, 2021.

[206] Akbar Karimi, Leonardo Rossi, and Andrea Prati. Aeda: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, 2021.

[207] Shuhuai Ren, Jinchao Zhang, Lei Li, Xu Sun, and Jie Zhou. Text autoaugment: Learning compositional augmentation policy for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9029–9043, 2021.

[208] Rong Xiang, Emmanuele Chersoni, Qin Lu, Chu-Ren Huang, Wenjie Li, and Yunfei Long. Lexical data augmentation for sentiment analysis. *Journal of the Association for Information Science and Technology*, 72(11):1432–1447, 2021.

[209] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, 2021.

[210] Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 875–878. IEEE, 2018.

[211] Ziqi Zhang and Lei Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945, 2019.

[212] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, 2020.

[213] Soonki Kwon and Younghoon Lee. Explainability-based mix-up approach for text data augmentation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2022.

[214] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34, 2021.

[215] S McManus, P Bebbington, R Jenkins, and T Brugha. A survey carried out for nhs digital by natcen social research and the department of health sciences. 2014.

[216] Rosemary Sedgwick, Sophie Epstein, Rina Dutta, and Dennis Ougrin. Social media, internet use and suicide attempts in adolescents. *Current opinion in psychiatry*, 32(6):534–541, 2019.

[217] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, 2019.

[218] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390, 2020.

[219] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023.

[220] Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. *International journal of machine learning and cybernetics*, 14(1):135–150, 2023.

[221] Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. Data noising as smoothing in neural network language models. In *In 5th International Conference on Learning Representations, ICLR*, 2017.

[222] Chris Hollis, Stephanie Sampson, Lucy Simons, E Bethan Davies, Rachel Churchill, Victoria Betton, Debbie Butler, Kathy Chapman, Katherine Easton, Toto Anne Gronlund, et al. Identifying research priorities for digital technology in mental health care: results of the james lind alliance priority setting partnership. *The Lancet Psychiatry*, 5(10):845–854, 2018.

[223] Sumaira H Malik and Neil S Coulson. Computer-mediated infertility support groups: an exploratory study of online experiences. *Patient education and counseling*, 73(1):105–113, 2008.

[224] Katherine Easton, Jacob Diggle, Mabel Ruethi-Davis, Megan Holmes, Darian Byron-Parker, Jessica Nuttall, Chris Blackmore, et al. Qualitative exploration of the potential for adverse events when using an online peer support network for mental health: cross-sectional survey. *JMIR mental health*, 4(4):e8168, 2017.

[225] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122, 2016.

[226] Katja Machmutow, Sonja Perren, Fabio Sticca, and Françoise D Alsaker. Peer victimisation and depressive symptoms: can specific coping strategies buffer the negative impact of cybervictimisation? *Emotional and Behavioural Difficulties*, 17(3-4):403–420, 2012.

[227] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21, 2019.

[228] Mario Alvarez-Jimenez, JF Gleeson, S Rice, Cesar Gonzalez-Blanch, and Sarah Bendall. Online peer-to-peer support in youth mental health: seizing the opportunity. *Epidemiology and psychiatric sciences*, 25(2):123–126, 2016.

[229] Russell Spears and Martin Lea. Panacea or panopticon? the hidden power in computer-mediated communication. *Communication research*, 21(4):427–459, 1994.

[230] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.

[231] Mohammadreza Rezvan, Saeedeh Shekarpour, Faisal Alshargi, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. Analyzing and learning the language for different types of harassment. *Plos one*, 15(3):e0227330, 2020.

[232] Anna Ehnvall, Philip B Mitchell, Dusan Hadzi-Pavlovic, Gordon Parker, Andrew Frankland, Colleen Loo, Michael Breakspear, Adam Wright, Gloria Roberts, Phoebe Lau, et al. Rejection sensitivity and pain in bipolar versus unipolar depression. *Bipolar disorders*, 16(2):190–198, 2014.

[233] Brandon A Gaudiano, Casey A Schofield, Carter Davis, and Lara S Rifkin. Psychological inflexibility as a mediator of the relationship between depressive symptom severity and public stigma in depression. *Journal of Contextual Behavioral Science*, 6(2):159–165, 2017.

[234] Yoshimitsu Takahashi, Chiyoko Uchida, Koichi Miyaki, Michi Sakai, Takuro Shimbo, Takeo Nakayama, et al. Potential benefits and harms of a peer support social network service on the internet for people with depressive tendencies: qualitative content analysis and social network analysis. *Journal of medical Internet research*, 11(3):e1142, 2009.

[235] Brian A Feinstein, Rachel Hershenberg, Vickie Bhatia, Jessica A Latack, Nathalie Meuwly, and Joanne Davila. Negative social comparison on facebook and depressive symptoms: Rumination as a mechanism. *Psychology of popular media culture*, 2(3):161, 2013.

[236] Richard M Smedley and Neil S Coulson. A thematic analysis of messages posted by moderators within health-related asynchronous online support forums. *Patient Education and Counseling*, 100(9):1688–1693, 2017.

[237] Claire Harding and Henry Chung. Behavioral health support and online peer communities: international experiences. *Mhealth*, 2:43, 2016.

[238] Bruno Biagianti, Sophia H Quraishi, and Danielle A Schlosser. Potential benefits of incorporating peer-to-peer interactions into digital interventions for psychotic disorders: a systematic review. *Psychiatric services*, 69(4):377–388, 2018.

[239] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, 2020.

[240] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[241] Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, and Yaser Al-Onaizan. Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145.

[242] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[243] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[244] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879, 2016.

[245] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[246] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[247] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, 2016.

[248] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[249] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

[250] Giorgos Papanastasiou, Nikolaos Dikaios, Jiahao Huang, Chengjia Wang, and Guang Yang. Is attention all you need in medical image analysis? a review. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2023.

[251] Owen Trigueros, Alberto Blanco, Nuria Lebeña, Arantza Casillas, and Alicia Pérez. Explainable icd multi-label classification of ehrs in spanish with convolutional attention. *International Journal of Medical Informatics*, 157:104615, 2022.

[252] Jinghe Zhang, Kamran Kowsari, James H Harrison, Jennifer M Lobo, and Laura E Barnes. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

[253] Pei-Fu Chen, Tai-Liang He, Sheng-Che Lin, Yuan-Chia Chu, Chen-Tsung Kuo, Feipei Lai, Ssu-Ming Wang, Wan-Xuan Zhu, Kuan-Chih Chen, Lu-Cheng Kuo, et al. Training a deep contextualized language model for international classification of diseases, 10th revision classification via federated learning: Model development and validation study. *JMIR Medical Informatics*, 10(11):e41342, 2022.

[254] Massimo Salvi, Hui Wen Loh, Silvia Seoni, Prabal Datta Barua, Salvador García, Filippo Molinari, and U Rajendra Acharya. Multi-modality approaches for medical

support systems: A systematic review of the last decade. *Information Fusion*, page 102134, 2023.

[255] Elvira Amador-Domínguez, Emilio Serrano, Daniel Manrique, and Javier Bajo. A case-based reasoning model powered by deep learning for radiology report recommendation. 2021.

[256] Adam Gabriel Dobrakowski, Agnieszka Mykowiecka, Małgorzata Marciniak, Wojciech Jaworski, and Przemysław Biecek. Interpretable segmentation of medical free-text records based on word embeddings. *Journal of Intelligent Information Systems*, 57:447–465, 2021.

[257] Joshua R Minot, Nicholas Cheney, Marc Maier, Danne C Elbers, Christopher M Danforth, and Peter Sheridan Dodds. Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance. *ACM Transactions on Computing for Healthcare*, 3(4):1–41, 2022.

[258] Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G Dunn. Hybrid text representation for explainable suicide risk identification on social media. *IEEE transactions on computational social systems*, 2022.

[259] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[260] Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2), 2010.

[261] Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

[262] Dehua Chen, Hongjin Zhao, Jianrong He, Qiao Pan, and Weiliang Zhao. An causal xai diagnostic model for breast cancer based on mammography reports. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3341–3349. IEEE, 2021.

[263] OpenAI. Chatgpt: Optimizing language models for dialogue. 2022.

[264] Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

[265] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[266] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[267] Marius Hobbhahn, Tom Lieberum, and David Seiler. Investigating causal understanding in llms. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

[268] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, pages 804–823. PMLR, 2023.

[269] Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR, 2023.