UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

LLM Fine-Tuning With Biomedical Open-Source Data

Christopher Anaya

Mestrado em Ciências de Dados

Dissertação orientada por:
Prof. Doutor Francisco José Moreira Couto
Doutora Maria Isabel Mou Sequeira Fernandes

2025

Acknowledgments

I would like to thank everyone whom I've met during my time here in Lisbon. Coming to a new country to study has been a difficult but gratifying experience. To my family, I love you. And to my advisors, your guidance and unwavering patience have been invaluable.

Dedicatória.

Resumo

A emergência de desafios com a sobrecarga de literatura biomédica, com milhões de artigos publicados anualmente, demonstra a necessidade de soluções automatizadas para a eficaz procura e síntese de informação no campo biomédico. Os métodos tradicionais, centrados na recuperação de documentos com base em palavras-chave, tornaram-se insuficientes, pois não respondem de maneira direta a perguntas complexas ou compostas. O surgimento dos Modelos de Linguagem de Grande Escala (LLMs, do inglês *Large Language Models*) especializados veio alterar este panorama. Modelos como o BioBERT, PubMedBERT, BioGPT ou BioMistral demonstraram que a adaptação a corpora biomédicos melhora significativamente a adequação terminológica e a precisão das respostas. Contudo, problemas como a tendência para gerar respostas factualmente incorretas, conhecidas como alucinações, a dificuldade em lidar com formatos de resposta estruturados exigidos por competições como o BioASQ e as limitações das métricas tradicionais de avaliação continuam a constituir barreiras ao uso destes sistemas em cenários clínicos reais, onde a veracidade e a interpretabilidade são cruciais.

Por outro lado, os LLMs têm revelado, nos últimos anos, capacidades notáveis em tarefas de compreensão e geração de linguagem natural. A possibilidade de treinar estes modelos em enormes volumes de dados textuais permitiu avanços consideráveis em aplicações que vão desde a tradução automática até à redação assistida, mas uma das áreas em que o impacto se tem mostrado mais promissor é a dos sistemas de pergunta-resposta. Em domínios especializados como o biomédico, a necessidade de ferramentas capazes de sintetizar informação com precisão, clareza e veracidade é particularmente urgente, dado que decisões clínicas, investigações científicas e práticas académicas dependem diretamente da qualidade das respostas fornecidas. A presente dissertação investiga o potencial de LLMs otimizados para o domínio biomédico, com ênfase na utilização de técnicas de afinação eficiente de parâmetros, nomeadamente o QLoRA, aplicadas ao modelo Mistral-7B. O objetivo principal é desenvolver um sistema de QA biomédico que responda de forma robusta a diferentes tipos de perguntas—de sim/não, factuais, de lista e de resposta ideal ou sumário—garantindo qualidade, consistência e fidelidade às fontes científicas.

A metodologia seguida neste trabalho foi delineada para enfrentar estes desafios de forma sistemática. A primeira etapa consistiu na seleção e harmonização dos dados, recorrendo ao conjunto de treino oficial do BioASQ como base de referência e complementando-o com dados provenientes de fontes como Gene Ontology, DrugBank e BiQA. A integração de múltiplas fontes exigiu processos de normalização e harmonização, sobretudo no caso do BiQA, cujos formatos de anotação

divergiam dos restantes. Esta diversidade permitiu treinar o modelo em diferentes tipos de perguntas e domínios semânticos, aumentando a sua capacidade de generalização e robustez em cenários variados. A segunda etapa centrou-se na configuração do modelo, escolhendo-se o Mistral-7B, um dos modelos abertos mais avançados e recentemente propostos, ajustado com QLoRA em modo 4-bit (NF4). Esta técnica de afinação eficiente de parâmetros tornou possível treinar e ajustar o modelo em GPUs de gama média, como a Tesla P4, reduzindo drasticamente os requisitos de memória sem perdas significativas de desempenho. Desta forma, o trabalho contribui também para a democratização do uso de LLMs biomédicos de grande porte.

Para garantir que o sistema funcionasse adequadamente, desenvolveram-se pipelines de pré e pós-processamento específicos para cada tipo de pergunta. As perguntas factuais, por exemplo, exigiram heurísticas de normalização e formatação que assegurassem consistência entre diferentes respostas, enquanto as perguntas de resposta ideal foram processadas com técnicas de sumarização e avaliadas com métricas como ROUGE-L e BLEU. As perguntas de sim/não beneficiaram da utilização de formatos de saída estruturados em JSON, o que permitiu uma avaliação automática mais confiável e evitou ambiguidades. Já as perguntas de lista constituíram um desafio adicional, devido à necessidade de remover redundâncias e garantir a padronização das respostas múltiplas. A formulação dos *prompts* constituiu outra dimensão crucial, sendo ajustada de acordo com o tipo de pergunta de forma a melhorar a precisão e reduzir alucinações. Todo o processo experimental foi documentado com controlo rigoroso de hiperparâmetros, *random seeds* e versões de dependências, de modo a garantir reprodutibilidade.

Os resultados obtidos evidenciam desempenhos sólidos nos conjuntos de teste do BioASQ, embora com diferenças claras entre os tipos de perguntas. As de sim/não apresentaram resultados particularmente robustos em termos de *macro F1*, atingindo valores próximos de 0.76, e a utilização de *datasets* curados como Gene Ontology e DrugBank mostrou ganhos consistentes, confirmando a importância da qualidade e adequação das fontes de conhecimento. As perguntas factuais, por sua vez, revelaram maior sensibilidade à formulação dos *prompts* e à cobertura terminológica das bases de dados utilizadas, resultando em variações significativas de desempenho. As perguntas de lista mostraram-se mais difíceis, uma vez que exigiram maior rigor na normalização do formato de saída e no tratamento de redundâncias, sendo um dos pontos de maior espaço para melhoria. Já as perguntas de resposta ideal foram avaliadas com métricas de sumarização, e apesar de apresentarem respostas informativas e fluentemente estruturadas, verificou-se a necessidade de melhorias na coesão, concisão e fundamentação factual, destacando os limites atuais dos modelos em tarefas abertas de síntese textual. De modo geral, o modelo demonstrou competitividade mesmo em ambientes de hardware limitado, e a utilização do QLoRA revelou-se fundamental para tornar possível a afinação de modelos de grande escala com custos computacionais reduzidos.

A dissertação inclui ainda uma análise crítica das métricas de avaliação utilizadas. Verificou-se que métricas tradicionais como F1, MRR, ROUGE e BLEU, embora úteis para comparações quantitativas, nem sempre refletem a factualidade ou a utilidade clínica das respostas. Foram discutidas alternativas, incluindo métricas semânticas baseadas em julgamentos humanos ou em

avaliadores automáticos baseados em modelos, que capturam melhor critérios como confiança, clareza e consistência. Além disso, procedeu-se a uma análise qualitativa de erros, de modo a identificar padrões recorrentes de falhas, como respostas parcialmente corretas, redundantes ou especulativas, fornecendo pistas para melhorias futuras.

As perspetivas futuras identificadas a partir deste estudo são várias. A integração de técnicas de *Retrieval-Augmented Generation* surge como um caminho promissor para enriquecer as respostas com evidências científicas recuperadas em tempo real, reduzindo o risco de alucinações. A infusão de conhecimento estruturado proveniente de ontologias e grafos semânticos biomédicos, como UMLS e MeSH, poderá aumentar a factualidade, a interpretabilidade e a explicabilidade das respostas. A adaptação a contextos multilingues constitui outra direção relevante, considerando a diversidade linguística da produção científica na área biomédica. Finalmente, defende-se a adoção de métodos de avaliação mais sofisticados, baseados em modelos julgadores como GEval ou QuestEval, capazes de medir dimensões como factualidade e confiabilidade para além da simples sobreposição lexical.

Em conclusão, a dissertação demonstra que é possível desenvolver sistemas de QA biomédico eficazes utilizando modelos modernos como o Mistral-7B, mesmo em ambientes de hardware limitado. O recurso ao QLoRA permitiu afinar o modelo de forma eficiente e acessível, democratizando a utilização de LLMs de grande porte em contextos académicos e de investigação aplicada. A integração de dados biomédicos estruturados, a engenharia de *prompts* e a implementação de pipelines de processamento adequados contribuíram para aumentar a qualidade, a consistência e a confiabilidade das respostas. Para além de avanços técnicos, o trabalho destaca também a importância da reprodutibilidade, da ciência aberta e da reflexão crítica sobre métricas de avaliação. Assim, este estudo não só reforça o papel dos LLMs no avanço da área de QA biomédico, como também estabelece bases sólidas para o desenvolvimento de sistemas mais robustos, transparentes e aplicáveis em cenários clínicos, científicos e industriais no futuro.

Palavras-chave: Respostas a perguntas biomédicas, Modelos de Linguagem de Grande Escala, Afinação Eficiente de Parâmetros, BioASQ, Métricas de Avaliação

Abstract

Biomedical question answering (QA) systems aim to support researchers and clinicians by providing accurate, context-aware answers to complex information needs. Recent advances in large language models (LLMs) have significantly improved QA performance across domains, yet challenges remain in the biomedical domain due to terminology complexity, limited data availability, and the risk of generating hallucinated content. This thesis investigates the application of parameter-efficient fine-tuning techniques to adapt LLMs for biomedical QA, focusing on the BioASQ challenge Task B Phase B, which includes yes/no, factoid, list, and ideal questions.

A comprehensive review of biomedical QA datasets and LLM adaptations highlights the evolving landscape of knowledge-infused models. The thesis presents a fine-tuning pipeline based on QLoRA, a memory-efficient method for adapting the Mistral-7B-Instruct-v0.1 model using quantized weights. Domain-specific prompt templates were designed for each question type to optimize answer formatting and reduce hallucinations. The experimental setup included training on a curated dataset comprising the training dataset provided by BioASQ, Gene Ontology, DrugBank, and BiQA-derived examples.

Results show that the proposed system achieves competitive performance across question types, particularly for yes/no questions, attaining F1 scores of 0.76, where structured JSON outputs enabled reliable automatic evaluation. For ideal (free-text) questions, the system demonstrated fluent but occasionally speculative responses, highlighting the trade-offs between informativeness and factual grounding. Evaluation metrics such as F1, MRR, and ROUGE were complemented by qualitative error analysis to assess system robustness.

The study concludes that combining domain-adapted prompts with QLoRA fine-tuning offers a promising approach for deploying efficient and effective biomedical QA systems. Future work should explore retrieval-augmented generation, deeper integration of biomedical ontologies, and improved evaluation frameworks tailored to the nuances of clinical and research settings.

Keywords: Biomedical Question Answering, Large Language Models, Parameter-Efficient Fine-Tuning, BioASQ, Evaluation Metrics

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the biomedical domain has experienced a remarkable increase in the volume of scientific publications, clinical guidelines, and research outputs. With millions of new publications being published annually in databases such as PubMed and PMC [51], biomedical professionals and researchers face an ever-growing challenge: how to efficiently and accurately filter through this immense publication output to find and read new information relevant to their research. The ability to navigate, interpret, and apply knowledge from this vast corpus is crucial not only for accelerating scientific discovery but also for improving evidence-based medical practice and informed decision-making in clinical settings[23]. This in turn can lead to better patient outcomes and more effective treatments, and key to improving health outcomes across society[74].

Manual review and exploration of biomedical literature at the scale that it is generated has become increasingly impossible[34]. As a result, highly specialized researchers often struggle to stay current with the literature in their subfields, let alone beyond. Early attemps to create Biomedical QA systems focused on modular pipelines that included a question classifier, an information retrieval component, and a final answer extraction step [14]. However, these systems often struggled with the complexity and variability of natural language questions, and relied heavily on well-defined rules and structured data sources. These systems were sufficiently limited in performance to not be widely adopted in practice.

In this context, systems that can automatically process natural language questions and provide relevant, structured, and accurate answers represent a valuable technological frontier. Biomedical Question Answering (QA) systems, powered by advances in natural language processing (NLP) and large-scale language models (LLMs), offer a promising solution to this problem [32]. These systems aim to provide direct answers to user queries expressed in natural language, using curated knowledge sources, scientific literature, or learned representations.

The emergence of instruction-tuned LLMs, such as ChatGPT [56], BioGPT [52], and more recently open-source models like Mistral-7B [1] and its biomedical variants, has significantly transformed the landscape. These models can be fine-tuned or prompted to generate coherent and informative responses across a wide range of biomedical topics. Nevertheless, adapting such models to highly specialized domains remains a non-trivial task, requiring access to domain-specific data, robust fine-tuning techniques, and careful evaluation protocols [21, 13, 48].

The evolution of biomedical QA systems has also seen the emergence of benchmarks that can quantify and contextualize their improvements; one such example is BioASQ [73], a recurring competition designed to assess performance of biomedical QA systems. BioASQ Task B, in particular, offers a demanding yet practical set of requirements. Systems must answer a variety of question types—yes/no, factoid, list, and ideal—using biomedical abstracts and domain-specific knowledge as their reference. This diversity makes BioASQ a rigorous and well-established evaluation framework, one that reflects many of the challenges faced by real-world biomedical QA applications. Participation also enables comparison under standardized conditions and contributes to a shared research community.

1.1 Scope and Research Context

This thesis aims to investigate the feasibility and effectiveness of adapting a powerful open-source language model to create a Biomedical Question Answering model using parameter-efficient fine-tuning. Specifically, this work employs QLoRA [16], a recent technique that reduces computational and memory overheads during training and allows for the use of consumer grade, off-the-shelf hardware. The goal is to demonstrate that, with the right methodology and careful engineering, a biomedical QA system with competitive performance to peer systems can be built with commonly available computing resources.

Unlike some systems that incorporate large-scale document retrieval components or rely on proprietary datasets, the approach presented here focuses on generative question answering within a purely open-source setting. The model is trained on publicly available datasets, including BioASQ [73], BiQA [39], DrugBank [76], and Gene Ontology [4], and evaluated using the standard BioASQ metrics. This approach aligns with principles of reproducibility and scientific transparency [61]. While prior work has adapted large models for biomedical question answering, this thesis seeks to demonstrate that a competitive system can be developed exclusively using open-source tools and consumer-grade hardware, thereby promoting accessibility and lowering barriers to entry for the broader research community.

1.2 Objectives

The overarching objective of this thesis is to use external knowledge to improve the performance of a generative model on biomedical question answering tasks. This is achieved through the following aims:

The first aim of this thesis is to fine-tune a generative model, Mistral-7B-Instruct-v0.1, using the QLoRA method on curated biomedical QA datasets. This includes harmonizing input formats, designing type-specific prompts, and applying lightweight, resource-conscious training strategies, providing an ample training set with which to fine tune the model. The second aim is to implement a complete question answering pipeline that supports input preprocessing, inference, answer formatting, and post-processing.

In addition to system development, the third aim is to evaluate the model's performance on the set of metrics using both official leaderboard from BioASQ[73] and supplementary internal evaluation scripts. This enables a more granular understanding of system behavior and supports systematic error analysis.

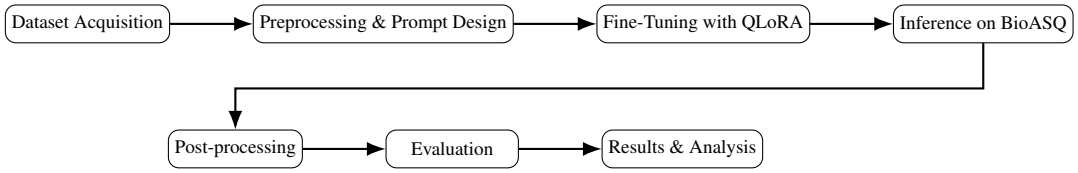

Figure 1.1: Overview of the biomedical QA pipeline presented in this thesis.

1.3 Methodology Overview

To meet these objectives, the thesis follows a modular and reproducible methodology. The pipeline begins with dataset acquisition and preprocessing, including filtering, cleaning, and formatting QA pairs from multiple biomedical sources. These are transformed into a unified structure and encoded using Hugging Face tokenization tools [77].

Next, the model is fine-tuned using QLoRA. This involves inserting low-rank adapters into the model's attention layers and using 4-bit quantized base weights to reduce memory usage. Training is conducted on a set of 8 Tesla M10 GPUs, with gradient accumulation and early stopping to manage batch size and optimize convergence. Each dataset targets specific question types, and custom prompts steer the model's generative behavior.

Inference applies the fine-tuned model to unseen BioASQ Task 12b questions, generating raw outputs that are subsequently normalized to match BioASQ's submission schema. Post-processing is essential, particularly for factoid and list questions, where formatting inconsistencies can lead to invalid or penalized outputs.

Evaluation is conducted using BioASQ's official metrics—accuracy, macro-F1, MRR, and ROUGE—as well as internal analysis scripts for examining consistency, common error patterns, and the effects of post-processing.

The diagram in Figure 1.1 illustrates the overall pipeline architecture, which includes dataset acquisition, preprocessing, fine-tuning, inference, post-processing, evaluation, and results analysis. Each step is designed to ensure that the model can effectively handle the diverse question types present in the BioASQ Task 12b dataset.

1.4 Contributions

This thesis makes three main contributions to biomedical QA and resource-efficient NLP model development. First, it implements and publishes a fine-tuned Large Language Model (LLM) based on Mistral-7B-Instruct-v0.1, using QLoRA to achieve efficient parameter tuning on consumer-grade hardware. This model is specifically adapted for the BioASQ Task 12b dataset, demonstrat-

ing that high-quality biomedical QA systems can be built without access to large-scale computational resources. This QA pipeline is designed to handle multiple question types, including yes/no, factoid, list, and ideal questions, using only open datasets and tooling.

This work also carries out a comprehensive evaluation of the model's performance across these question types, using both official BioASQ metrics and internal evaluation scripts. The results reveal important trends in model behavior, such as the impact of dataset curation on performance and the challenges of generating coherent answers for complex question types.

The code base is available on GitHub, at `https://github.com/chranama/biollm-finetune`, allowing for reproducibility and further experimentation. The model, datasets, and evaluation scripts are all documented to facilitate future research in this area.

Finally, a paper was accepted at the BioASQ 2024 workshop proceedings[3], which presents the findings of our work and discusses the implications for future research in biomedical question answering. The paper is available at `https://ceur-ws.org/Vol-3740/`.

1.5 Structure of the Document

The remainder of this document is structured as follows. Chapter 2 provides a review of related work and theoretical background, including biomedical QA architectures, dataset resources, fine-tuning techniques, and evaluation strategies. Chapter 3 presents the methodology in detail, describing the model, data pipeline, training strategy, and inference setup. Chapter 4 discusses the results, analyzing the system's performance across question types and identifying key success factors and failure modes. Chapter 5 concludes with a summary of contributions and outlines several avenues for future research, particularly in the areas of knowledge infusion, multilingual adaptation, and hybrid architectures.

Chapter 2

Related Work

Biomedical Question Answering (QA) occupies a unique intersection between natural language processing, biomedical informatics, and knowledge representation. The development of QA systems for the biomedical domain has closely followed broader advances in neural architectures, pretraining strategies, and evaluation methodologies. This chapter reviews key research directions that have informed the present work, beginning with the foundations of large language models and Transformer architectures, and proceeding through the evolution of biomedical QA systems, domain adaptation strategies, model architectures, evaluation metrics, and parameter-efficient fine-tuning methods.

2.1 Large Language Models and Transformers

Large Language Models (LLMs) are advanced neural architectures designed to process and generate human language, typically based on the Transformer framework [75]. These models are trained on large-scale text corpora using language modeling objectives that predict the next token in a sequence. Decoder-only Transformer architectures, such as those used in the GPT family [62, 11], have gained particular prominence due to their effectiveness in open-ended text generation tasks, including question answering, summarization, and dialogue.

Text processing begins with tokenization, where input is segmented into subword units using algorithms like Byte Pair Encoding (BPE) [68] or SentencePiece [37]. This is especially valuable in biomedical NLP, where rare and technical terms are common [52]. Tokens are then mapped to dense vector embeddings that reflect semantic relationships, while positional encodings convey sequence order.

Transformers compute token representations through stacked layers of self-attention and feed-forward neural networks. Self-attention enables each token to contextualize itself relative to all others in the sequence, capturing both local and long-range dependencies [75]. Multi-head attention enhances modeling capacity by allowing the model to focus on different aspects of the input simultaneously.

In decoder-only models, masked self-attention prevents access to future tokens during training, ensuring autoregressive generation. Each token is predicted iteratively, conditioned on the

preceding sequence. Decoding strategies such as greedy decoding, nucleus sampling [26], and beam search [35] influence the diversity and coherence of generated text.

Recent work has adapted Transformer-based LLMs to specialized domains, including biomedicine. Domain-specific models like BioGPT [52] retain the core Transformer architecture but are pretrained on biomedical corpora, demonstrating the versatility of LLMs across both general and specialized tasks.

Having established the foundational architectures underlying modern NLP, the next sections examine how these models have been adapted to biomedical QA specifically.

2.2 Evolution of Biomedical QA

The field of biomedical question answering (QA) has undergone several distinct phases, each reflecting broader technological shifts in natural language processing and biomedical informatics. Early biomedical QA systems were rooted in information retrieval and symbolic reasoning. These systems typically relied on keyword matching, Boolean queries, and domain-specific query expansion using ontologies such as the Unified Medical Language System (UMLS)[8] or SNOMED-CT[12]. For example, questions were mapped to biomedical concepts using MeSH terms or UMLS Concept Unique Identifier, and answer candidates were extracted from retrieved documents using hand-crafted rules or regular expressions [5]. These approaches benefited from interpretability and modularity but suffered from brittleness, low recall, and inability to handle linguistic variation or complex inference[32].

The advent of neural networks and distributed representations marked a major shift in QA methodologies. Early distributed, vector space word representations like word2vec[54] and GloVe[59] introduced the idea of word embeddings, which captured statistical semantics from large corpora. However, these were static and failed to represent polysemy, where one word has multiple meanings. The introduction of contextualized embeddings with BERT[17] fundamentally changed the landscape. BERT's transformer-based architecture allowed for bidirectional encoding of context, enabling it to model the subtle semantics of biomedical language more effectively than earlier approaches.

Domain-specific adaptations of BERT soon followed. BioBERT [41] was pretrained on millions of PubMed abstracts and PMC articles, significantly improving performance on named entity recognition (NER), relation extraction (RE), and factoid QA. SciBERT [6], trained on a multidisciplinary scientific corpus, provided a more general scientific language model, while PubMedBERT [20] went further by training from scratch exclusively on biomedical text. These models enabled end-to-end fine-tuning for QA tasks using relatively modest annotated datasets and quickly became foundational in biomedical NLP.

More recently, there has been a transition from extractive encoder-based systems to generative models. While the proximate factor was massive investment in generative models, this was in turn spurred by the versatility that generative models demonstrate in a wide swath of NLP tasks. And inspired by developments in open-domain QA, models such as BioGPT [52], PMC-LLaMA[72],

and BioMistral[38] adopt decoder-only architectures, allowing for the generation of fluent, multi-sentence answers from prompts. These systems can address yes/no and ideal questions in BioASQ with greater flexibility, though they also introduce challenges related to factuality, hallucination, and evaluation.

This historical trajectory—from rule-based systems to transformer-based encoders to generative LLMs—has been shaped by advances in model architecture, access to large-scale biomedical corpora, and the standardization of evaluation through shared tasks such as BioASQ and TREC. A timeline of key system milestones is provided in Figure 2.1.

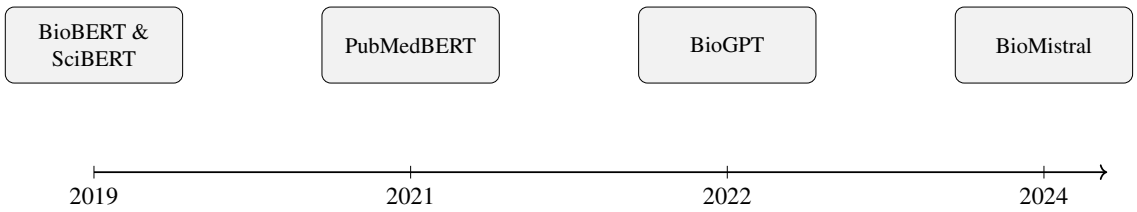

Figure 2.1: Timeline of key pretrained language models for biomedical question answering

2.3 Adaptation and Biomedical Pretraining

Domain adaptation is a critical concern in biomedical natural language processing (NLP), where general-purpose large language models (LLMs) frequently fall short in understanding domain-specific terminology, abbreviations, and linguistic conventions. Most foundation models, such as GPT-3[11] and LLaMA[72], are pretrained on open web corpora that include news articles, Wikipedia, and Common Crawl, as well as other open-source text data like books and academic papers. While these sources offer broad linguistic coverage, they lack the depth and precision required for biomedical applications, where accurate interpretation of gene symbols, clinical trial terminology, and drug interactions is essential.

To address these shortcomings from general-purpose pretraining, the biomedical NLP community has developed a range of domain-adapted models, each trained or fine-tuned on large biomedical text corpora. These include PubMed abstracts, PMC full-text articles, and electronic health records such as those found in the MIMIC-III dataset[33]. The benefits of domain-specific pretraining have been consistently demonstrated across tasks such as named entity recognition (NER), relation extraction (RE), medical entity linking (MEL), and question answering (QA). For instance, BioBERT [41] and PubMedBERT [20] have achieved strong performance gains over the original version of BERT on biomedical benchmark datasets.

Several recent models represent a further step toward fully biomedical LLMs at scale, reflecting the broader transition that has been evidenced from extractive to generative models. BioMedLM[10], released by MosaicML, is a decoder-only model trained on PubMed abstracts and clinical trial texts. As well, Galactica-Med is a domain-specific checkpoint of Meta's Galactica model, adapted for scientific and biomedical literature generation [71]. And PubMedGPT [20] was trained from

scratch using the GPT-2 architecture on biomedical abstracts only, producing superior factual consistency for downstream tasks. These models are summarized in Table 2.1.

Model	Architecture	Pretraining Data	Notes
BioMedLM[10]	Decoder-only	PubMed, ClinicalTrials	Instruction-tuned; released by MosaicML
Galactica-Med[71]	Decoder-only	Scientific literature	Biomedical specialization of Galactica
PubMedGPT[20]	GPT-2 variant	PubMed abstracts	Trained from scratch; improved factuality
BioBERT[41]	BERT	PubMed, PMC	Continued pretraining from BERT base

Table 2.1: Comparison of biomedical language models for domain adaptation

However, domain adaptation is not solely about the source of pretraining data. Prompt design also plays essential roles in adapting models to biomedical downstream tasks. Instruction-tuned models—those trained to follow structured natural language prompts—are particularly well suited for question answering. This is because they internalize patterns for task-switching, answer formatting, and conversational flow. The Mistral-7B-Instruct-v0.1 model, used as the base for this thesis, is an open-access, instruction-tuned variant of Mistral-7B. While it was not originally pretrained on biomedical text, its strong instruction-following behavior makes it an effective foundation for lightweight adaptation through parameter-efficient methods.

2.4 Question Answering Architectures

Question answering (QA) systems have evolved significantly in recent years, driven by changes in underlying model architectures and the increasing availability of domain-specific datasets. Traditionally, QA systems were designed as multi-stage pipelines consisting of question classification, information retrieval, passage ranking, and answer extraction. These systems often relied on keyword matching, query expansion using biomedical ontologies, and syntactic parsing to identify answer spans within retrieved documents. While modular and interpretable, pipeline architectures struggled in the biomedical domain, where terminology is highly variable, and correct answers may require complex reasoning or background knowledge [5].

The advent of neural embedding models and transformer-based encoders introduced a new paradigm: extractive QA. In this approach, models such as BERT [17] and BioBERT [41] receive a concatenated input of the question and context, and are trained to predict start and end positions of the answer span within the context. These models have shown strong performance in tasks like SQuAD and BioASQ factoid questions, where answers are relatively short and can be localized in a single passage. However, extractive models are limited when it comes to multi-hop reasoning, summarization, or answering questions that require combining knowledge from multiple sources; this is due to the fact that their general objective is to predict a span of text from a fixed context window, which may not contain all relevant information [32].

More recently, there has been a shift toward generative architectures for QA. Decoder-only models, such as GPT-3 [11], BioGPT [52], and instruction-tuned variants like Mistral-7B-Instruct[30], generate answers token-by-token in free-form text. These models are particularly effective for generating ideal answers in BioASQ or for supporting conversational agents that require natural language fluency. Unlike extractive models, generative LLMs can synthesize information from prompts without a fixed context window, and are capable of handling open-ended questions.

However, generative systems come with their own challenges. Without access to explicit context, these models may ”hallucinate” plausible-sounding but factually incorrect information [29]. This issue is especially problematic in biomedical settings, where factual correctness is critical. Additionally, generative models often require post-processing to align outputs with task-specific formats (e.g., for factoid or list questions), and may produce inconsistent or overly verbose answers. To address these issues, some systems incorporate retrieval-augmented generation (RAG) techniques [45], which ground generation in relevant documents retrieved from biomedical literature.

Figure 2.2 summarizes key differences between extractive and generative QA systems in the biomedical domain.

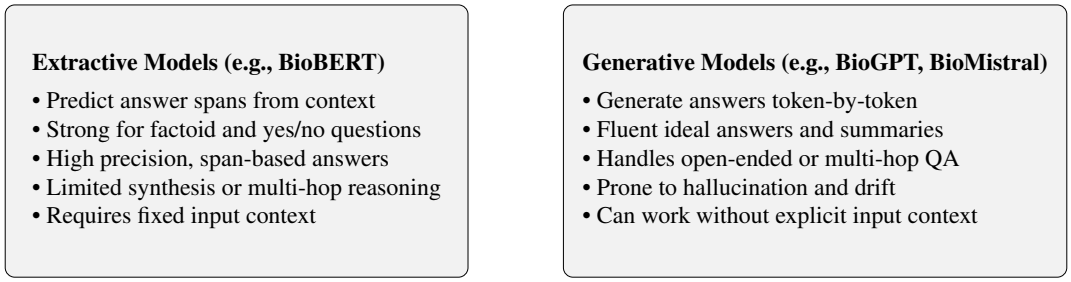

Figure 2.2: Comparison of extractive and generative QA architectures in biomedical NLP

2.5 Prompt Engineering in Biomedical QA

Prompt engineering has emerged as a central technique for controlling the behavior of instruction-tuned large language models (LLMs), particularly in few-shot or zero-shot scenarios, referring to either a limited context or no context in a prompt[58]. This would contrast with a traditional machine learning task where extensive labeled data is available for training. Rather than fine-tuning model parameters, prompt engineering relies on carefully phrased natural language instructions to steer the model toward the desired output. The structure, specificity, and tone of a prompt can dramatically influence the quality, factuality, and format of the generated answer [57]. This makes prompt design especially important in biomedical question answering, where answers must conform to both domain-specific semantics and strict evaluation formats.

In biomedical QA, ambiguous or poorly framed prompts can lead to vague, speculative, or overly verbose responses, in addition to completely incorrect answers[32]. Furthermore, biomedical terms often have multiple valid interpretations—“RA” may refer to “rheumatoid arthritis” or

"right atrium"—and models are prone to hallucinate plausible-sounding but incorrect information when not grounded in precise instructions [83]. Prompt design must therefore account not only for linguistic clarity but also for medical disambiguation and output format compliance. Instruction-tuned models like Mistral-7B-Instruct are particularly sensitive to task framing, as they have been trained to follow instructions structured as natural language templates.

2.6 Challenges in Biomedical QA

Biomedical question answering (QA) presents further challenges that go well beyond those encountered in general-domain QA. These challenges stem from the nature of biomedical language, the interpretive variability of medical knowledge, and the high stakes associated with factual correctness. Addressing these challenges requires more than generic language understanding—it demands domain-specific reasoning, careful ambiguity resolution, and robust handling of semantic nuance.

One of the most pervasive issues is the complexity of biomedical terminology. Biomedical texts are rich with abbreviations, synonyms, and polysemous terms that differ subtly in meaning depending on context. For example, the abbreviation "RA" may refer to "rheumatoid arthritis," "right atrium," or "renal artery" depending on the specialty and sentence. Similarly, entity names like "ACE" might denote an enzyme (angiotensin-converting enzyme), a drug class (ACE inhibitors), or even a protein-coding gene. These linguistic challenges complicate both answer generation and evaluation, as models must distinguish between highly similar but clinically distinct concepts [70].

Equally important is the demand for semantic precision. Biomedical QA systems must carefully differentiate between concepts such as correlation and causation, efficacy and effectiveness, or risk reduction and prevention. A model that incorrectly states that "aspirin prevents heart attacks" instead of "aspirin reduces the risk of heart attacks" introduces a serious factual inaccuracy. The latter reflects a probabilistic relationship supported by studies; the former implies certainty. This level of precision is difficult to achieve with generative models that prioritize fluency over factual grounding [29].

Another challenge lies in the inherent ambiguity and disagreement that exists within biomedical knowledge itself. Clinical evidence evolves rapidly, and even domain experts may disagree on what constitutes the "correct" answer to a question. Differences in interpretation between guidelines can lead to multiple plausible answers. These ambiguities make annotation difficult, as human annotators may interpret the same question differently, leading to lower inter-annotator agreement [64]. This, in turn, complicates the use of supervised learning techniques that depend on high-quality, consistent labels.

The final and perhaps most significant concern is ethical. Biomedical QA systems, especially those built on large language models, may produce outputs that appear fluent and authoritative but are factually incorrect or misleading. In high-risk domains like medicine, these hallucinations can have serious consequences, especially if such systems are used in patient-facing applications

or clinical decision support tools. As noted by recent work on clinical LLMs, models must be designed and deployed with caution, transparency, and robust fail-safes to prevent misuse [29]. Over-reliance on these systems without verification could lead to the propagation of misinformation, inappropriate treatment recommendations, or erosion of trust in digital health technologies.

These challenges are summarized in Table 2.2, which categorizes the most pressing concerns in biomedical QA and provides illustrative examples.

Challenge Category	Example / Implication
Terminological complexity	"RA" could mean rheumatoid arthritis, right atrium, or renal artery
Factual precision	"Aspirin prevents MI" implies certainty; "Aspirin reduces MI risk" is probabilistic
Expert disagreement	Guidelines for statin use differ between American and European cardiology societies
Ethical risk	Hallucinated dosage information could lead to treatment errors

Table 2.2: Categories of challenges unique to biomedical question answering

In summary, biomedical QA requires systems that are not only linguistically competent, but also medically cautious, interpretively flexible, and ethically grounded. These demands elevate the importance of robust evaluation methods, domain-specific training data, and transparent system design—principles that guided the development of the QA system in this thesis.

2.7 Evaluation Metrics

Evaluating biomedical QA systems remains a significant challenge, especially for generative models. Unlike classification or span-selection tasks, QA involves open-ended language generation where multiple valid answers may differ lexically but be semantically equivalent. This complexity has led to the adoption of a variety of evaluation metrics, each with distinct strengths and weaknesses. In this section, we divide evaluation metrics into three broad categories: surface-level string metrics, embedding-based semantic metrics, and factuality-aware or hybrid metrics.

Surface-Level Metrics

Traditional metrics such as accuracy, macro-F1, and mean reciprocal rank (MRR) are commonly used for yes/no and factoid questions. These metrics are straightforward to compute and easy to interpret but are often brittle. For example, yes/no performance in BioASQ is typically evaluated with both accuracy and macro-F1 to account for class imbalance. Factoid questions are scored using strict match accuracy and MRR, rewarding systems that rank correct answers higher but penalizing slight formatting variations or explanatory phrasing. A correct answer like "alpha-synuclein" may go unrecognized if embedded in a sentence like "Alpha-synuclein is a key protein in Parkinson's disease."

For ideal answers, the BioASQ challenge uses ROUGE-2 and ROUGE-SU4, which rely on n-gram overlap. ROUGE-SU4, in particular, incorporates skip-bigram and unigram matches while accounting for stopwords [46]. However, these metrics often fail to capture semantic similarity between paraphrased answers. For instance, "reduces blood glucose levels" and "improves glycemic control" express the same idea but yield low ROUGE scores.

Embedding-Based Metrics

To address the limitations of surface-level comparisons, several semantic similarity metrics have been proposed. BERTScore [82] measures alignment between tokens in the generated and reference answers using contextualized embeddings. BLEURT [67] uses a fine-tuned BERT model trained on human-annotated quality scores to predict how similar two answers are from a human judgment perspective.

These metrics improve evaluation for paraphrased outputs but still struggle with factual accuracy. A fluent but factually incorrect answer—such as "Aspirin cures stroke" instead of "Aspirin reduces stroke risk"—may receive a high BLEURT score due to linguistic fluency and vocabulary similarity, despite conveying a dangerous misinterpretation.

Factuality-Aware and Hybrid Metrics

Recent work has sought to develop evaluation metrics that better reflect factuality, faithfulness, and grounding. QuestEval [66] uses a question generation and answering framework to assess how much information from the reference is preserved in the generated text. More recently, LLM-based evaluators such as G-Eval have emerged, where large language models serve as automatic judges by assessing answer quality along dimensions such as coherence, correctness, and relevance [50].

While promising, these newer methods are not yet widely adopted in shared tasks due to reproducibility concerns, computational cost, and the lack of standardized scoring scripts. Human evaluation remains the most reliable method, particularly in biomedical QA where domain knowledge is essential for assessing answer plausibility. However, manual scoring is resource-intensive and subject to annotator disagreement, especially when multiple answers are technically correct but framed differently.

Metric Type	Examples	Strengths / Limitations
Surface-level	Accuracy, F1, MRR, ROUGE	Easy to compute; sensitive to formatting; ignores semantics
Embedding-based	BERTScore, BLEURT	Captures paraphrasing; fails on factuality; high correlation with fluency
Factuality-aware / Hybrid	QuestEval, PARENT, LLM-as-judge	Evaluate meaning and grounding; costly; limited standardization

Table 2.3: Comparison of evaluation metric types for biomedical QA

2.8 Parameter-Efficient Fine-Tuning for QA

As large language models (LLMs) continue to grow in size and complexity, the cost of full fine-tuning has become increasingly prohibitive for many research environments. Fine-tuning a 7B or 13B parameter model typically requires GPUs with 40–80 GB of memory, extended compute time, and considerable energy consumption. These requirements are particularly challenging for biomedical QA, where datasets are relatively small and many academic or clinical institutions lack access to high-end infrastructure.

Parameter-efficient fine-tuning (PEFT)[22] offers an attractive alternative. Rather than updating the entire model, PEFT methods introduce small, trainable modules—such as adapters or low-rank projection layers—into a frozen base model. This dramatically reduces the number of trainable parameters, leading to faster convergence, reduced memory usage, and easier model reuse across tasks. In biomedical applications, where task-specific adaptation is often required across multiple subdomains (e.g., oncology, pharmacology, genetics), PEFT allows for modular tuning without catastrophic forgetting.

Among PEFT methods, LoRA (Low-Rank Adaptation) [28] has become particularly popular. LoRA freezes the pretrained weights and injects trainable low-rank matrices into the attention layers of the model. This enables adaptation with fewer than 1% of the total parameters and no change to inference behavior. Building upon LoRA, QLoRA [16] further reduces memory usage by quantizing the frozen model weights to 4-bit precision using NormalFloat4 (NF4), while retaining 16-bit precision for the adapter layers. The result is a fine-tuning scheme that enables high-quality adaptation of models like Mistral-7B on consumer-grade GPUs. Table 2.4 compares full fine-tuning, LoRA, and QLoRA in terms of trainable parameters and memory footprint.

The convergence of transformer-based architectures, domain-adapted pretraining, generative modeling, and parameter-efficient fine-tuning has redefined the state of biomedical question answering. Yet, successfully adapting these techniques into a practical system requires careful integration of datasets, prompt engineering, post-processing, and evaluation strategies. The following chapter describes the specific methodological decisions taken in this thesis to implement and optimize a biomedical QA system leveraging these advances.

Method	**Trainable Params**	**Memory Footprint**	**Suitable For**
Full Fine-Tuning	100%	40–80 GB	High-end clusters
LoRA	$< 1\%$	24–32 GB	Research-grade consumer GPUs
QLoRA	$< 1\%$	12–16 GB	Budget GPUs (e.g., Tesla P4, T4)

Table 2.4: Comparison of fine-tuning strategies for LLMs

Chapter 3

Methods

This chapter presents the methodology developed to build and evaluate a biomedical question answering (QA) system based on the Mistral-7B-Instruct model. Unlike many high-resource pipelines in NLP, this system was designed with constraints in mind: limited hardware, full reproducibility, and reliance on only open datasets. The design process was iterative, shaped by both technical requirements and the structured goals of the BioASQ shared task. What follows is a detailed description of each component in the system pipeline, including data selection and harmonization, prompt engineering, fine-tuning with QLoRA, inference and output formatting, and the internal evaluation framework that supported the model development lifecycle.

3.1 System Design Rationale

This thesis developed a system built on top of `Mistral-7B-Instruct-v0.1`, an instruction-tuned, decoder-only large language model (LLM). This model was selected based on a combination of architectural, practical, and reproducibility considerations. As an publicly available model with open-source licensing, it provides an ideal foundation for an academic system that must be freely distributable and easy to adapt. Mistral-7B-Instruct is compact relative to models like GPT-3[11] or Falcon-40B[2], yet large enough to encode rich linguistic and task-following behavior.

Decoder-only transformer architectures are particularly well suited to generative question answering tasks. Unlike encoder-only models such as BERT or BioBERT, which are typically used for extractive QA, decoder-only models like Mistral can generate full-sentence or multi-paragraph answers from scratch, enabling support for open-ended or "ideal" question types. Furthermore, instruction tuning makes the model highly responsive to prompt structure, which is essential for aligning responses with BioASQ's task formats[81].

Reproducibility was another critical factor. Many high-performing biomedical LLMs are either closed-source or require licensing barriers that complicate academic use. Mistral-7B-Instruct is openly available through HuggingFace, supported by a well-documented ecosystem that includes the PEFT and Transformers libraries[77].

To adapt the model for biomedical QA, this thesis employs QLoRA, a parameter-efficient fine-tuning method that allows large models to be trained on modest hardware. QLoRA works

by freezing the model weights and quantizing them to 4-bit precision, while injecting low-rank adapter modules into attention layers [16]. Only these adapters are updated during training, reducing GPU memory requirements and compute load. This approach enabled full fine-tuning of the 7B parameter model on a set of eight NVIDIA Tesla M10 GPUs with just 8 GB of VRAM—a configuration that would be infeasible using traditional full fine-tuning.

The scope of this project was further refined by focusing on three question types from the BioASQ challenge[73]: yes/no, factoid, and ideal. These represent a broad spectrum of QA complexity, from binary classification to entity recognition and abstractive summarization. List-type questions, which often require predicting multiple entities and matching structured JSON schemas, were excluded due to their formatting complexity, evaluation sensitivity, and limited prompt compatibility in early tests. However, our internal model evaluation included list questions. The inclusion rationale for each question type is summarized in Table 3.1.

Type	QA Behavior	Rationale for Inclusion / Exclusion
Yes/No	Binary answer generation	Simple structure; tests classification and domain reasoning
Factoid	Entity identification	Requires precise entity grounding and short-form generation
Ideal	Abstractive summarization	Evaluates fluency, synthesis, and biomedical context comprehension
List	Multi-entity aggregation	Excluded due to output formatting complexity and evaluation brittleness

Table 3.1: BioASQ question types and rationale for inclusion in this thesis

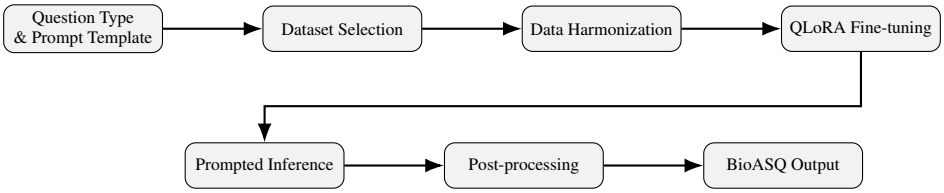

Figure 3.1: System architecture overview for biomedical QA pipeline. Top row: data preparation and model training. Bottom row: inference and output processing for BioASQ task.

3.2 Dataset Selection and Harmonization

The quality, diversity, and structure of the training data play a pivotal role in the success of any supervised NLP task, particularly in specialized domains like biomedical question answering (QA). This thesis draws on four complementary sources—BioASQ [73], BiQA [39], Gene Ontology (GO) [4] definitions, and DrugBank [76] profiles—to construct a harmonized training dataset. These sources were selected to span a range of biomedical subdomains, answer types, and linguistic registers, ensuring that the fine-tuned QA model would be exposed to both expert-curated content and more informal, user-driven queries.

BioASQ served as the original training dataset supplied by the competition. It contains over a decade of manually curated biomedical questions derived from PubMed articles, complete with gold-standard answers and references. Only Task B QA pairs were used, and entries were grouped by question type (yes/no, factoid, list, and ideal). Questions with incomplete metadata or ambiguous type labels were discarded. As the training data was already structured in a well-formed question-answer format with clearly delineated types and clean annotations, no further formatting or answer reconstruction was required beyond basic filtering. BioASQ was chosen for its structured format, high-quality annotations, and widespread use as a biomedical QA benchmark.

BiQA, by contrast, required substantial preprocessing and was the primary focus of the data harmonization workflow illustrated in Figure 3.2. The dataset consisted of biomedical questions posed by users on public forums such as Reddit and Medium, paired with PubMed article abstracts that were crowdsourced as potentially relevant. However, these abstracts were not synthesized into final answers. To prepare this data, I manually reviewed every question and excluded those that were ambiguous, low-quality, or outside the scope of structured biomedical QA. For each retained question, I examined the associated abstracts. If a text span in the abstract directly answered the question, I used it as the answer; otherwise, I wrote a single-sentence answer that reflected the most relevant or consensus information across the snippets. The resulting QA pairs were then formatted into a consistent JSONL structure. Table 3.2 shows three representative examples from this process, including a discarded input, a case where the answer was directly sourced from the abstract, and one requiring abstraction beyond the retrieved evidence.

GO Terms were reformulated into question-answer pairs by converting ontology entries into natural language prompts. Specifically, each term definition was used to construct a question of the form "What is *X*?", with *X* being the name of the GO term, and the original ontology definition serving as the answer. This approach allowed for the generation of many short, well-defined summary questions from a structured vocabulary.

DrugBank entries were processed using the same "What is *X*?" strategy. Each entry was parsed to extract the name, description, mechanism of action, and therapeutic indications of the drug. Depending on the content of the extracted fields, these entries were converted into summary questions. To ensure brevity and focus, overly long answers were truncated using simple sentence-count heuristics.

To enable consistent multi-dataset training, a harmonization pipeline was implemented. All samples were normalized into a shared JSONL structure with fields for question, answer, source, type, token length, and formatting instructions. Output answers were stripped of HTML tags, normalized to plain UTF-8 text, and deduplicated where necessary. Type mappings were collapsed to three categories—yes/no, factoid, and ideal—aligning with the scope defined in Section 3.1.

3.3 Prompt Engineering and Templates

Instruction-tuned language models are highly sensitive to prompt phrasing, particularly when applied to new domains like biomedical QA. Their behavior is shaped by prior instruction-tuning

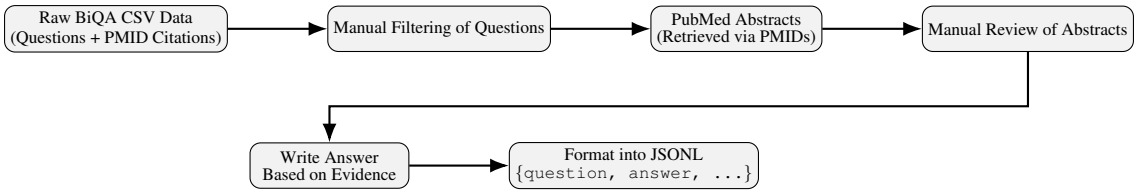

Figure 3.2: BiQA data harmonization pipeline. Top row: data ingestion and abstract enrichment. Bottom row: manual annotation and formatting into JSONL.

Question	Ideal Answer (Excerpt)	Sample Snippet (Excerpt)
How many times did endosymbiosis occur?		*Almost all aphid species (Homoptera, Insecta) have 60–80 huge cells called bacteriocytes, within which are round-shaped bacteria....Buchnera is completely symbiotic and viable only in its limited niche, the bacteriocyte.*
Under what conditions do dendritic spines form?	Spine growth precedes synapse formation and that new synapses form preferentially onto existing boutons.	***Our data show that spine growth precedes synapse formation and that new synapses form preferentially onto existing boutons.***
How long does antibiotic-dosed LB maintain good selection?	No loss of bioactivity was demonstrated after 30 days of storage in plates containing various antibiotics.	***No loss of bioactivity was demonstrated after 30 days of storage in plates containing methicillin, erythromycin, cephalothin...***

Table 3.2: Examples from cleaned BiQA data illustrating harmonization cases: unusable input, abstract-sourced answer, and answer not strictly found in abstracts.

datasets[81], where the format, tone, and specificity of prompts correlate closely with output quality. Accordingly, prompt engineering was a critical component of this project, as it served to align model outputs with BioASQ's task structure and evaluation metrics.

Prompt templates were designed for each supported question type: yes/no, factoid, and ideal. The design goals varied depending on the output expectations. Yes/no questions required strict binary answers, where even slight elaboration ("Yes, because...") could lead to evaluation penalties. Factoid questions demanded entity-style responses without full-sentence generation. Ideal questions, in contrast, benefited from multi-sentence summaries grounded in biomedical context. Early experiments revealed that generic or weak prompts—such as "Answer the following question"—resulted in verbose, unfocused, or hallucinated outputs. The final templates were developed through iterative testing to mitigate these issues and standardize output structures.

Yes/No:

```
Instruction: This is an example of yes/no question and the
respective answer in the intended format. The answer can
only use JSON format. The format must be {"exact_answer":"",
"ideal_answer":""} where exact_answer should be 'yes' or 'no',
and ideal_answer is a short conversational response starting
with yes/no then follow on the explanation. An example
of a Yes/No question: Is the protein Papilin secreted? An
example of a Yes/No answer: {"exact_answer":"yes",
```

```
"ideal_answer":"Yes,  papilin is a secreted protein"}
Question: Can insulin resistance lead to diabetes?
```

Factoid:

```
Instruction: This is an example of factoid question and
the respective answer in the intended format. The answer
can only use JSON format. The format must be {"exact_answer":
[], "ideal_answer":""} where exact_answer is a list of
precise key entities to answer the question, and ideal_answer
is a short conversational response containing an explanation.
Question: What is the most common anticoagulant used during
surgery?
```

List:

```
Instruction: This is an example of list question and
the respective answer in the intended format. The answer
can only use JSON format. The format must be {"exact_answer":
[], "ideal_answer":""} where exact_answer is a list of
precise key entities to answer the question, and ideal_answer
is a short conversational response containing an explanation.
Question: Which genes are associated with Parkinson's disease?
```

Ideal:

```
Instruction: Provide a concise and informative answer to
the biomedical question.
Question: How does aspirin reduce the risk of cardiovascular
events?
```

Table 3.3 summarizes the prompt designs and their intended behavioral constraints. These templates were also used at inference time during evaluation and analysis to ensure that model responses remained aligned with BioASQ scoring expectations.

These prompts were applied to the data during the runtime of the fine-tuning scripts, which I describe next.

3.4 Fine-Tuning with QLoRA

To adapt the base model (Mistral-7B-Instruct-v0.1) for biomedical question answering, this thesis employed QLoRA, a parameter-efficient fine-tuning method that enables low-resource adaptation

Type	Prompt Template	Design Goal
Yes/No	"Answer the following biomedical question with 'yes' or 'no'."	Enforce binary output; avoid explanatory drift
Factoid	"Answer...with a short phrase or name."	Promote concise, entity-style answers
Ideal	"Provide a concise and informative answer..."	Encourage multi-sentence, informative summaries
List	"List all relevant entities that answer the following biomedical question."	Elicit comprehensive, enumerated responses in a structured format

Table 3.3: Prompt templates and design goals per question type

of large models. QLoRA combines 4-bit quantization of the pretrained model weights with trainable low-rank adapters (LoRA) inserted into specific transformer layers [16]. This setup drastically reduces the GPU memory footprint while maintaining model performance close to full fine-tuning.

Specifically, the model was quantized using the NormalFloat4 (NF4) quantization scheme, supported by the `bitsandbytes` library [15]. LoRA adapters were injected into the query and value projection matrices of the attention mechanism, enabling the model to learn task-specific transformations while leaving the quantized base model untouched. This architecture was implemented using Hugging Face's PEFT and Transformers libraries [77], which provided streamlined support for adapter-based training, 4-bit inference, and gradient accumulation.

The training was conducted on a set of eight NVIDIA Tesla M10 GPUs (8 GB VRAM each). Due to hardware constraints, the batch size was set to 2, with gradient accumulation steps set to 4 to simulate larger batches. The optimizer used was AdamW, with a fixed learning rate of 2.5×10^{-4}. Early stopping was enabled with a patience threshold of 10 steps. Each dataset was fine-tuned independently for 100–500 steps depending on its size, as summarized in Table 3.4. No additional warm-up or learning rate decay schedule was applied.

Parameter	Value / Setting
Model base	Mistral-7B-Instruct-v0.1
PEFT method	QLoRA (4-bit quantized, LoRA adapters)
Quantization scheme	NF4 (NormalFloat4), double quantization
Trainable modules	Attention projection matrices
GPU hardware	Tesla M10 (8 GB VRAM)
Batch size	2 (gradient accumulation steps = 4)
Optimizer	AdamW, learning rate 2.5×10^{-4}
Training steps	100–500 per dataset
Early stopping	Enabled (patience = 10)

Table 3.4: Training configuration for QLoRA fine-tuning

To ensure reproducibility, random seeds were fixed across runs, and all model checkpoints were saved at regular intervals. Training logs included step-level loss metrics and evaluation

scores, timestamped for traceability. The environment was containerized with fixed package versions using a `Dockerfile`. All scripts, configuration files, and command-line invocations are documented in the project's GitHub repository.

This setup demonstrates that with modern PEFT methods, domain adaptation of large generative models is feasible even under stringent computational constraints—offering a reproducible pathway for other researchers working in biomedical NLP with limited hardware.

3.5 Inference and Post-Processing Pipeline

Once fine-tuning was complete, inference was conducted on the official test questions from BioASQ Task 12b. The model operated in zero-shot mode, that is, without training examples entered into the prompt, using the templates described in Section 3.3. Responses were generated using greedy decoding with a maximum output length of 256 tokens. No sampling was applied (i.e., temperature was set to 0), and other generation parameters such as repetition penalty and top-k remained at their default values.

Following inference, a post-processing pipeline was applied to transform raw text outputs into evaluation-ready JSON files compliant with the BioASQ submission schema. This pipeline was question-type specific and evolved through iterative testing to handle common formatting issues. The transformation logic per question type is summarized in Table 3.5.

Yes/No answers were stripped of surrounding text, punctuation, and capital letters. Many initial outputs included expansions like "Yes, it can." or "No, this is unlikely," which were trimmed to the canonical "yes" or "no" strings. If no valid token could be extracted, the fallback rule inserted "unknown" to avoid null entries.

Factoid answers were normalized to extract a concise entity mention. Regular expressions were used to isolate the first capitalized noun phrase or medically-relevant term. Punctuation and trailing verbs or articles were removed to avoid format mismatches. In some cases, redundant prefixes (e.g., "The name of the drug is…") were stripped prior to extraction.

List answers were parsed to extract individual entities, typically separated by commas, newlines, or bullet points. Each item was stripped of punctuation, lowercased, and deduplicated. Responses that lacked a clear list structure were split heuristically using conjunctions or separators. If no valid items were found, "unknown" was returned to preserve completeness.

Ideal answers were subject to length and quality constraints. Any output exceeding 200 words was truncated by token count. A lightweight sentence filter removed common hedging or filler phrases, such as "This is a complex issue," "Many factors may contribute," or "It is important to note." These patterns were identified through manual review of outputs that scored poorly under ROUGE evaluation.

After post-processing, the outputs were assembled into a JSON object containing required BioASQ fields: question ID, type, system name, and answer content. A schema validation script ensured that all entries conformed to expected formats. Early submission attempts revealed issues such as missing fields and non-string values, which were resolved by adding fallback values and

Question Type	Post-Processing Rules	Rationale
Yes/No	Strip whitespace and extra tokens; convert to lowercase	Enforce exact "yes"/"no" match for scoring
Factoid	Regex extraction of key phrase; remove articles/punctuation	Conform to entity-style output; avoid sentence fragments
Ideal	Truncate to 200 words; filter out generic/hedging phrases	Enhance answer informativeness; align with ROUGE scoring
List	Split into items using commas or newlines; lowercase and deduplicate	Produce clean, structured lists for evaluation alignment

Table 3.5: Post-processing rules applied per question type

assert checks. Once the pipeline was finalized, submission failures were reduced to zero.

This multi-stage inference and post-processing setup helped ensure that model outputs were both high-quality and formally compliant—two essential factors for achieving reliable evaluation in the BioASQ framework.

With the fine-tuned model deployed and a structured inference pipeline established, the system was capable of generating evaluation-compliant answers across the diverse question types posed in the BioASQ challenge. The careful integration of prompt engineering, deterministic decoding, and post-processing ensured that outputs adhered to strict formatting constraints while preserving answer quality. The following chapter presents the empirical evaluation of this system, analyzing its performance across BioASQ batches and exploring its strengths, limitations, and broader implications for biomedical question answering.

Chapter 4

Results and Discussion

This chapter presents the results of the biomedical QA system built around the Mistral-7B-Instruct model and fine-tuned using QLoRA. The evaluation is structured by question type, with results broken down across the four batches from the BioASQ Task 12b challenge. In addition to reporting standard evaluation metrics, this chapter interprets the system's behavior, highlights strengths and limitations, and explores failure cases in depth. Special attention is given to post-processing effectiveness, error typologies, metric limitations, and broader implications for biomedical NLP.

4.1 Evaluation Framework and Performance Metrics

The system was evaluated using the official BioASQ Task 12b leaderboard, which provides a standardized benchmark for biomedical question answering systems. Evaluating on the official platform ensures that system performance is comparable to prior work and adheres to the same task definitions and scoring criteria. External validation is critical, particularly in biomedical QA, where internal-only testing risks misrepresenting generalization or robustness [55].

Each system submission was evaluated independently across four sequential test batches, with questions spanning yes/no, factoid, list, and ideal types. Following the project scope defined in Section 3.1, only yes/no, factoid, and ideal answers were generated and submitted. The evaluation metrics applied by BioASQ varied by question type, as summarized in Table 4.1.

Question Type	**Metric**	**Evaluation Focus**
Yes/No	Accuracy, Macro-F1	Classification correctness and class balance
Factoid	Strict Accuracy, MRR	Precise entity prediction and ranking quality
List	Precision, Recall, F1	Completeness and correctness of multiple entity answers
Ideal	ROUGE-2, ROUGE-SU4	Summarization quality and content overlap

Table 4.1: Evaluation metrics used for each BioASQ question type

Official leaderboard results are summarized in Tables 4.2 and 4.3. The best yes/no perfor-

mance was achieved in Batch 1 with an accuracy of 64% and a Macro-F1 score of 0.5322, peaking at 0.5486 in Batch 2. For factoid questions, strict accuracy remained low, reaching a maximum of 15.8% in Batch 4, though MRR scores indicated that correct entities often appeared among the top-ranked candidates. Summary (ideal) questions showed the highest ROUGE performance in Batch 3, with ROUGE-2 F1 reaching 0.0543 and ROUGE-SU4 F1 peaking at 0.0719.

Table 4.2: Evaluation results from BioASQ Task 12b – yes/no and factoid questions

Batch	**Yes/No**				**Factoid**		
	Accuracy	F1 Yes	F1 No	Macro F1	Strict Acc.	Lenient Acc.	MRR
Batch 1	0.6400	0.7568	0.3077	0.5322	–	–	–
Batch 2	0.6154	0.7222	0.3750	0.5486	–	–	–
Batch 3	0.5000	0.2500	0.6250	0.4375	0.0769	0.0769	0.0769
Batch 4	0.4074	0.5294	0.2000	0.3647	0.1579	0.1579	0.1579

Table 4.3: Evaluation results from BioASQ Task 12b – summary (ideal) questions

Batch	R-2 (Recall)	R-2 (F1)	R-SU4 (Recall)	R-SU4 (F1)
Batch 1	0.0516	0.0331	0.0808	0.0432
Batch 2	0.0416	0.0227	0.0783	0.0404
Batch 3	0.1122	0.0474	0.1436	0.0607
Batch 4	0.0845	0.0543	0.1172	0.0719

To complement the leaderboard results, internal evaluations were also conducted on held-out development sets across four datasets: BioASQ, GO Terms, DrugBank, and BiQA. Macro-averaged F1 scores for yes/no classification peaked on the GO Terms dataset (0.6217), followed closely by DrugBank (0.6120). Factoid MRR remained at 0.0000 across all datasets, as shown in Table 4.5, indicating persistent challenges in exact entity localization. List-type questions similarly produced near-zero scores, while ROUGE scores were low across question types but slightly higher for summary-type prompts.

Table 4.4: Macro-Averaged F1 Scores for Yes/No Questions

Dataset	**F1 (Skipping 'n/a')**	**F1 ('n/a' as Wrong)**
BioASQ	0.5407	0.5414
GO Terms	0.6217	0.6067
DrugBank	0.6120	0.5888
BiQA	0.5018	0.5018

Inference was conducted using greedy decoding with prompt templates tailored to each question type, as described in Section 3.3. Outputs were passed through the post-processing pipeline detailed in Section 3.5, ensuring normalization and compliance with BioASQ submission specifications. Without these steps, significant portions of the system's output would have been invalidated or unfairly penalized during evaluation.

Table 4.5: Mean Reciprocal Rank (MRR) for Factoid Questions

Dataset	MRR
BioASQ	0.0000
GO Terms	0.0000
DrugBank	0.0000
BiQA	0.0000

Table 4.6: Mean Average F1 Scores for List-Type Questions

Dataset	Mean Average F1
BioASQ	0.0000
GO Terms	0.0000
DrugBank	0.0000
BiQA	0.0000

Table 4.7: ROUGE Scores (F1) Across Question Types and Datasets

Question Type	Dataset	ROUGE-2	ROUGE-SU4
Yes/No	BioASQ	0.0000	0.0164
	GO Terms	0.0001	0.0145
	DrugBank	0.0000	0.0162
	BiQA	0.0000	0.0170
List	BioASQ	0.0001	0.0120
	GO Terms	0.0000	0.0141
	DrugBank	0.0000	0.0168
	BiQA	0.0001	0.0140
Summary	BioASQ	0.0000	0.0222
	GO Terms	0.0002	0.0226
	DrugBank	0.0000	0.0250
	BiQA	0.0000	0.0210
Factoid	BioASQ	0.0000	0.0174
	GO Terms	0.0000	0.0157
	DrugBank	0.0002	0.0161
	BiQA	0.0001	0.0183

In summary, the evaluation framework combined external benchmarking against the BioASQ gold standard with internal quality assurance processes, enabling a rigorous and interpretable assessment of system performance.

4.2 Yes/No Question Performance

Among the supported question types, yes/no classification yielded the strongest performance. According to the official BioASQ Task 12b leaderboard (Table 4.2), accuracy scores ranged from 40.7% in Batch 4 to a peak of 64.0% in Batch 1. Corresponding macro-F1 scores spanned from 0.3647 to 0.5486, with the highest value achieved in Batch 2. These results confirm that the system maintained moderate classification capability across batches, though performance showed variability between evaluation rounds.

A mild affirmative bias was observed in both leaderboard and internal evaluations. When uncertain, the model tended to answer "yes" rather than "no." Table 4.2 shows a noticeable imbalance in F1 scores across classes, with higher F1 for "yes" (e.g., 0.7568 in Batch 1) than for "no" (e.g., 0.3077 in the same batch), supporting this observation.

Answer formatting consistency was high across batches. Post-processing modules enforced exact lowercase responses ("yes" or "no"), preventing format-related evaluation penalties. Internal evaluations (Table 4.4) confirmed consistent macro-F1 scores across datasets, with GO Terms and DrugBank reaching 0.6217 and 0.6120 respectively. These results suggest that while factual grounding remains a challenge, the system demonstrates robust binary classification capability, with potential for further improvement through calibration and bias mitigation.

4.3 Factoid Question Performance

Factoid questions presented significant challenges for the system. According to the official leaderboard results (Table 4.2), strict accuracy scores were consistently low, peaking at 15.8% in Batch 4 and dropping to 7.7% in Batch 3. Mean Reciprocal Rank (MRR) scores reflected similar limitations, with a maximum of 0.1579 in Batch 4. These results indicate that the system rarely returned the correct entity in the required strict format and struggled to rank correct entities consistently.

Internal evaluation results (Table 4.5) further confirmed poor performance on factoid questions, with MRR scores of 0.0000 across all internal datasets. This suggests the system consistently failed to provide short-form entity answers as required, across both external and internal evaluations.

Qualitative inspection indicates that the model often generated full sentences containing the correct entity instead of returning the isolated entity itself. While this reflects underlying domain knowledge, it failed strict evaluation protocols based on exact matching. Additionally, entity ambiguity contributed to errors, with the model returning higher-level concepts (e.g., drug classes) instead of specific entities required by the task.

Overall, these findings suggest that while the model possesses latent biomedical knowledge,

it lacks mechanisms to constrain output to strict entity formats, significantly limiting its factoid question answering performance.

4.4 List Question Performance

List-type questions presented the weakest performance of all question types. As shown in Table 4.6, the system achieved a mean average F1 of 0.0000 across all four internal datasets, indicating a complete failure to produce list answers in the required format. Similarly, ROUGE-2 and ROUGE-SU4 F1 scores for list questions were negligible across all datasets (Table 4.7).

Qualitative analysis revealed that the model typically returned free-text responses containing a small number of relevant entities, but failed to output the structured list format expected by the evaluation framework. For example:

> **Q:** Which cytokines are involved in the inflammatory response of rheumatoid arthritis?
> **A:** "TNF-alpha and IL-6 are key cytokines implicated."

While such responses occasionally included correct items, they omitted others and did not conform to the required list structure, leading to zero evaluation scores.

These results suggest that the system is not capable of handling list-type questions in its current form. The model was neither trained nor prompted to produce multiple-entity outputs, and its decoder-only architecture with instruction tuning appears biased toward single-sentence or single-entity generation.

In summary, despite possessing some latent knowledge of relevant biomedical entities, the system lacks the necessary output control mechanisms to generate structured lists, resulting in systematically poor performance on list-style questions.

4.5 Ideal Answer Performance

Ideal answers, requiring concise free-form summaries, were evaluated using ROUGE-2 and ROUGE-SU4 metrics. As shown in Table 4.3, system performance was consistently low across all batches, with ROUGE-2 F1 scores ranging from 0.0227 to 0.0543 and ROUGE-SU4 F1 scores from 0.0404 to 0.0719. While the highest scores occurred in Batch 4, overall metric values remained modest.

Internal evaluation results were even lower. Table 4.7 shows ROUGE-2 F1 scores near zero for all internal datasets, reinforcing that the model's generated summaries had limited lexical overlap with reference answers.

Qualitatively, generated answers were generally fluent and coherent but tended toward brevity and omission of detailed content, likely contributing to the low ROUGE scores. For example:

> **Q:** How does metformin improve insulin sensitivity?
> **A:** "Metformin activates AMPK and reduces hepatic glucose production, improving insulin sensitivity."

While correct, this response was incomplete relative to reference summaries, omitting alternative mechanisms or supporting details.

Overall, the system demonstrated an ability to produce fluent, domain-relevant summaries but struggled with factual coverage and recall, as reflected by both leaderboard and internal metrics. These findings suggest that the model, in its current configuration, prioritizes conciseness over completeness and lacks mechanisms to ensure full content coverage required for higher ROUGE performance.

4.6 Error Analysis and Failure Modes

While leaderboard metrics provide a broad view of system performance, they often fail to reveal the underlying reasons behind incorrect or suboptimal answers. In biomedical question answering, where factual correctness and answer type precision are critical, a deeper understanding of model behavior is essential. To this end, a qualitative error taxonomy was developed to classify recurring issues observed across BioASQ Task 12b batches. These categories illuminate the model's limitations and suggest avenues for improvement in future iterations.

Formatting Errors were especially prevalent in factoid questions. In many cases, the model produced full sentences or descriptive phrases when concise entities were required. This behavior contributed to low strict accuracy scores for factoid questions (e.g., 0.0769 in Batch 3 and 0.1579 in Batch 4, Table 4.2), as entity detection within longer responses was not supported by the evaluation script. Internal MRR scores of 0.0000 across datasets (Table 4.5) further highlight this issue.

Content Errors included both incorrect facts and incomplete responses. In several instances, responses lacked the necessary specificity, providing broad classes of answers where specific entities were required. These types of errors negatively affected both factoid and list question performance, where precise entity extraction is critical.

Factual Drift was observed in responses that were fluent but incorrect. This likely contributed to low ROUGE scores across question types (Table 4.7), as the model occasionally produced answers that, while syntactically correct, deviated from factual content.

Overconfidence was evident in yes/no classification tasks. The model frequently returned definitive "yes" or "no" answers without apparent uncertainty, even on questions where evidence might be inconclusive. This behavior may reflect biases introduced during pretraining and instruction tuning, leading to categorical predictions in the absence of strong supporting information.

A summary of representative error types is presented in Table 4.8.

Understanding these failure modes is essential not only for improving performance but also for informing model design. Many of the issues identified could potentially be mitigated through entity-aware prompting, retrieval-augmented generation, uncertainty-aware calibration, or post-hoc fact-checking. As biomedical QA systems evolve toward real-world deployment, ensuring answer correctness, interpretability, and humility will be as important as maximizing leaderboard scores.

Error Type	Description
Incorrect Fact	Response included an incorrect biomedical assertion
Correct Entity in Sentence	Returned correct entity embedded in a full sentence
Overly Broad Response	Returned a general category instead of a specific entity
Negation Confusion	Misinterpreted negation, leading to reversed answers

Table 4.8: Representative error types observed across batches.

4.7 Comparative Reflections and Broader Implications

Compared to earlier extractive systems based on BioBERT or SciBERT, the current model offers clear advantages in fluency and generative coherence, particularly for ideal answers. However, it continues to lag behind retrieval-enhanced architectures in factoid precision and evidence-grounded factuality. This is reflected in low strict accuracy scores (e.g., 0.0769 in Batch 3) and zero MRR across all internal datasets (Table 4.5). These shortcomings illustrate the limitations of closed-book generation when high-fidelity entity grounding is required.

Architecturally, the system's strengths lie in its simplicity and accessibility. It was trained on a series of 8 GB GPUs, required no retrieval infrastructure, and relied exclusively on open-source tools and datasets (where licensing allowed). The use of QLoRA reduced both memory and computational demands without degrading model expressivity. This lightweight setup significantly lowers the barrier to entry for academic groups, non-profit researchers, and students lacking access to enterprise-scale hardware. It aligns with broader goals of democratizing machine learning research and minimizing the environmental cost of experimentation [24].

However, the results also underscore the limitations of purely generative biomedical QA. The system struggled with factual grounding, entity disambiguation, and list-style responses—all of which demand reasoning over structured knowledge or external context. As shown in Table 4.6, performance on list questions remained near zero across all datasets, despite the model occasionally recalling relevant entities. These findings point toward the need for future biomedical QA architectures to incorporate knowledge graphs, entity linkers, or biomedical ontologies that can support evidence verification and reduce hallucination. Hybrid models that blend symbolic inference with neural generation are already being explored in the QA community [78].

Additionally, these reflections highlight the importance of evaluation frameworks that move beyond surface-level string matching. As discussed in Section 4.1, metrics such as ROUGE and strict accuracy often penalize semantically correct but paraphrased responses. Complementary evaluation protocols—including fuzzy entity matching, semantic similarity metrics, and human-in-the-loop assessment—may better capture the full range of model capabilities, especially for free-form generative tasks.

Chapter 5

Conclusion and Future Work

The research presented in this thesis contributes to the field of biomedical question answering (QA) through the design, implementation, and evaluation of a generative QA system based on the Mistral-7B-Instruct model, fine-tuned using the QLoRA framework [16]. This study has demonstrated that it is possible to develop domain-specific QA systems within constrained computational environments by leveraging modern parameter-efficient fine-tuning (PEFT) techniques alongside curated, open-source biomedical datasets.

5.1 Summary of Contributions

This thesis fulfills the objectives outlined in the introduction by advancing system development, empirical evaluation, and conceptual understanding in biomedical question answering (QA).

To achieve the first objective—fine-tuning a generative model on curated biomedical datasets—Mistral-7B-Instruct-v0.1 was fine-tuned using the QLoRA framework on harmonized data drawn from BioASQ, BiQA, DrugBank, and Gene Ontology. Prompt templates specific to each question type were designed, and resource-conscious training procedures were implemented. The QLoRA approach enabled efficient adaptation of a large language model on consumer-grade GPUs, demonstrating that high-quality fine-tuning is feasible in low-resource settings.

For the second objective, a complete and modular biomedical QA pipeline was developed. This system includes components for dataset harmonization, prompt generation, structured inference, post-processing tailored to question type, and output formatting for BioASQ evaluation. Each component was designed to be reproducible, extensible, and independent, enabling future researchers to reuse or adapt the system. The entire codebase has been made publicly available at `https://github.com/chranama/biollm-finetune`.

To address the third objective, the system was evaluated using both the official BioASQ Task 12b leaderboard and internal evaluation scripts. Performance was assessed across all question types—yes/no, factoid, list, and ideal—and type-specific metrics such as accuracy, MRR, and ROUGE were applied. The system performed best on yes/no questions, achieving strong macro-F1 scores, while factoid and list questions exposed persistent issues in formatting and entity disambiguation. Ideal answers demonstrated fluent generation but lacked comprehensive factual

coverage. These findings led to the development of a taxonomy of common failure modes and informed several proposed directions for future system improvements.

Beyond the specific aims, this thesis contributes to the broader conversation about instruction-tuned, decoder-only LLMs in biomedical contexts. The results demonstrate that open-source generative models, when fine-tuned using parameter-efficient methods, can deliver meaningful performance without proprietary datasets or enterprise-scale infrastructure. The findings reinforce concerns in the literature about grounding and factual accuracy in generative biomedical NLP models [42, 29] and point toward promising hybrid strategies that combine generation with structured knowledge.

Through these contributions, the core objectives of the thesis have been met, laying the groundwork for more accessible, robust, and semantically grounded biomedical QA systems.

5.2 Reflections on Methodology

Beyond model architecture, this research underscores the critical role of system-level design choices in determining downstream performance. Elements such as prompt structure, dataset formatting, and the implementation of post-processing routines were shown to exert significant influence over evaluation outcomes. In particular, tasks reliant on strict string matching—such as factoid and list question evaluation—were highly sensitive to formatting discrepancies, highlighting the need for meticulous attention to surface-level outputs when working with generative models.

The adoption of QLoRA for fine-tuning proved effective, although not without practical complexities. While quantization facilitated the training of a large-scale model within the constraints of limited memory, it introduced additional considerations related to memory management, checkpointing, and debugging. Nonetheless, QLoRA confirmed its value as a pragmatic approach for resource-efficient model adaptation, enabling experimentation that would otherwise have been infeasible in constrained environments.

The evaluation process itself provided further insights into the limitations of standard benchmarking practices. Variability in system performance across BioASQ batches emphasized the influence of question phrasing, domain specificity, and batch composition, suggesting that aggregate metrics alone are insufficient for fully characterizing model capabilities. Consequently, this work advocates for more granular, question-level, or typological analyses when evaluating biomedical QA systems.

5.3 Limitations

Despite the contributions outlined above, several limitations characterize the current system. Foremost among these is its reliance on a closed-book generative architecture, which prevents access to external biomedical knowledge during inference. This design choice constrains the system's factual grounding and limits its ability to retrieve or verify evidence when answering complex

questions.

Additionally, challenges relating to entity disambiguation and precise answer formatting persisted, particularly in the context of factoid and list-type questions. The reliance on strict string-matching metrics further exacerbated these weaknesses, as the generative outputs frequently failed to meet the exacting syntactic requirements imposed by the BioASQ evaluation framework.

The evaluation methodology itself represents another limitation. This thesis depended exclusively on automated metrics such as accuracy, macro-F1, ROUGE, and mean reciprocal rank (MRR), which offer only partial insights into model performance. These metrics often fail to capture semantic correctness, clinical relevance, or factual accuracy, especially in generative tasks where multiple valid phrasings may convey the same factual content.

Finally, the system was developed and evaluated solely in the English language. As a result, its applicability to multilingual biomedical contexts remains unexplored, limiting its potential usefulness in real-world, global healthcare environments.

5.4 Directions for Future Work

Several research directions arise from the limitations and findings of this thesis, offering clear opportunities for improving the system's capabilities and practical applicability.

Foremost among these is the incorporation of retrieval-augmented generation (RAG) architectures. The system's closed-book configuration currently limits its ability to ground answers in external evidence, contributing to hallucinations and factual inaccuracies. Integrating a retrieval component—allowing the model to dynamically access biomedical literature such as PubMed during inference—would address this limitation directly. RAG techniques, combining dense passage retrieval with generative decoding, have demonstrated considerable promise in related natural language tasks [45], and their application in biomedical QA represents a particularly promising avenue for improving factual grounding and answer trustworthiness.

In parallel, leveraging structured biomedical knowledge sources should be prioritized. Ontologies such as UMLS, MeSH, and SNOMED-CT encode curated entity relationships and semantic hierarchies that could support more accurate entity disambiguation, enhance concept grounding, and enable systematic post-hoc validation of generated answers. Approaches that incorporate these resources, whether through entity-aware prompting, retrieval filtering, or post-processing validation layers, hold potential to address many of the entity-level errors and ambiguity challenges observed in this study [9, 79].

A further direction involves extending the system's capabilities beyond English. Biomedical knowledge and clinical expertise are inherently multilingual, and enabling question answering in languages such as Portuguese and Spanish would substantially broaden the system's applicability. Recent advances in cross-lingual transfer learning and the growing availability of multilingual biomedical ontologies suggest that such expansion is both technically feasible and societally valuable [40].

From a methodological perspective, alternative parameter-efficient fine-tuning strategies war-

rant exploration. While QLoRA proved effective within this project, techniques such as adapter-based architectures, prompt-tuning methods, and curriculum learning may offer advantages in terms of generalization, multi-task learning, and training efficiency [27, 43]. Evaluating these methods within the biomedical QA context could yield insights into their relative strengths and trade-offs.

Finally, improving evaluation methodology emerges as a necessary step. While automated metrics such as ROUGE and BERTScore provide quantitative benchmarks, they fail to capture critical dimensions of answer quality, including factual accuracy and clinical relevance. Incorporating human-in-the-loop evaluation—particularly involving biomedical experts—would provide richer and more reliable assessments of system outputs, especially for generative tasks where nuance and domain knowledge are essential [66].

Collectively, these future directions outline a clear trajectory for advancing biomedical question answering systems toward greater accuracy, reliability, and real-world utility. Addressing the identified limitations through retrieval integration, structured knowledge incorporation, multilingual expansion, methodological refinement, and expert-driven evaluation will be essential in progressing toward robust and clinically relevant QA solutions.

Bibliography

[1] Mistral AI. Mistral-7b-instruct-v0.1 model card. `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1`, 2023. Accessed 2025-04-29.

[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[3] Christopher Anaya, Maria Fernandes, and Francisco M. Couto. Llm fine-tuning with biomedical open-source data. In *CLEF 2024 Working Notes*, pages 68–77. CEUR-WS, 2024.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, J Michael Davis, Kara Dolinski, Suzanne S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[5] Sofoklis Athenikos and Hyoil Han. Biomedical question answering: a survey. *Computer methods and programs in biomedicine*, 99(1):1–24, 2010.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *EMNLP*, 2019.

[7] Anya Belz and Craig Thomson et al. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp, 2023.

[8] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[9] Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 2004.

[10] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, G irish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[12] Eunsuk Chang, Hyeoun-Ae Kim, Yoon Kim, Hyeoun-Ae Kim, and Yoon Kim. The use of snomed ct, 2013–2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026, 2021.

[13] Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. Med42 – evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches, 2024.

[14] Dina Demner-Fushman and Jimmy Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[15] Tim Dettmers. Bitsandbytes: 8-bit optimizers and quantization routines for pytorch. *GitHub repository*, 2022.

[16] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient fine-tuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[18] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Robustness evaluation of qa models in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[19] Adam Fisch, Matt Gardner, Andrew Yates, Minjoon Seo, Bo Lin Tseng, and Luke Zettlemoyer. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering (MRQA)*, 2019.

[20] Yu Gu, Robert Tinn, Hao Cheng, Matt 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, 2022.

[21] Yu Gu and Robert et. al Tinn. Domain-specific language model pretraining for biomedical natural language processing. 3(1), October 2021.

[22] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[23] R. Brian Haynes and Andrew Haines. Barriers and bridges to evidence based clinical practice. *BMJ*, 317(7153):273–276, 1998.

[24] Brent Hecht, Yoshua Bengio, et al. The carbon footprint of machine learning training will plateau, then shrink. In *Communications of the ACM*, 2021. Special issue on climate and computing.

[25] William Hersh, Aaron Cohen, Jianji Yang, Ritu Bhupatiraju, Kirk Roberts, and Marti Hearst. Trec 2006 genomics track overview. In *TREC*, 2007.

[26] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

[27] Neil Houlsby, Andrei Giurgiu, et al. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

[28] Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[29] Zhewei Ji, Nayeon Lee, Jason Fries, and Danqi Yu. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38, 2023.

[30] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[31] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of EMNLP*, 2019.

[32] Qiao et. al Jin. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, 55(2):1–36, January 2022.

[33] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Leo H. Lehman, Mengling Feng, Marzyeh Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[34] Irma Klerings, Alexander S Weinhandl, and Kylie J Thaler. Information overload in healthcare: too much of a good thing? *ZEFQ - Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 109(4-5):285–290, 2015.

[35] Philipp Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. *Machine Translation: From Real Users to Research*, pages 115–124, 2004.

[36] Kalpesh Krishna, Abhinav Panda, and Mohit Iyyer. The parent metric for controllable text generation. In *Proceedings of ACL*, 2021.

[37] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

[38] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand, 2024.

[39] Andre Lamurias, Diana Sousa, and Francisco M. Couto. Generating biomedical question answering corpora from q&a forums. *IEEE Access*, 8:161042–161051, 2020.

[40] Anne Lauscher, Goran Glavaš, and Simone Ponzetto. Specializing unsupervised pretraining for cross-lingual transfer. In *Findings of EMNLP*, 2020.

[41] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon 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.

[42] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C Wallace. What lies beneath the mask: Understanding masked factual knowledge in biomedical language models. In *EMNLP*, 2021.

[43] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

[44] Patrick Lewis, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, Oren Etzioni, and Peter Clark. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of ACL*, 2021.

[45] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kulkarni, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[46] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[47] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *Proceedings of ACL*, 2022.

[48] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2024.

[49] Fangyu Liu, Ivan Vulić, and Anna Korhonen. 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.

[50] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023.

[51] Zhiyong Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011.

[52] Renqian Luo, Xiaofei Sun, Qingyu Xia, Bin Qin, Ting Liu, and Heng Xu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

[53] Chen Ma, Ming Zhang, Zhen Liu, Lemao Liu, Mian Tan, and Yue Zhang. Entity-aware instruction tuning for open-domain question answering. In *Proceedings of ACL*, 2023.

[54] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[55] Diego Molla. Evaluation of biomedical question answering systems: A review of methodologies and metrics. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, 2020.

[56] OpenAI. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`, 2022. Accessed: 2025-06-12.

[57] Long Ouyang, Jeff 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.

[58] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, volume 22, pages 1410–1418, 2009.

[59] 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.

[60] Jonas Pfeiffer, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Peft: A library for parameter-efficient fine-tuning of transformers, 2022.

[61] Joelle Pineau, Philippe Vincent-Lamarre, Jean-Francois Fortin, Xavier Bouthillier, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

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

[63] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of ACL*, 2020.

[64] Kirk Roberts and Dina Demner-Fushman. Annotating logical forms for ehr questions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1378–1383, 2016.

[65] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. The (un)reliability of nlp research: A meta-analysis of recently published papers. *Proceedings of ACL*, 2020.

[66] Thomas Scialom, Paul-Antoine Dray, Gabriel Staerman, Patrick Gallinari, and Jan Szafran. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of EMNLP*, 2021.

[67] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*, 2020.

[68] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725, 2016.

[69] Linlin Shen, Yue Zhang, Weidong Chen, Xiang Wei, Chuanqi Tan, Dong Yu, and Xiaodong Jin. K-bert: Enabling language representation with knowledge graph. In *Proceedings of AAAI*, 2021.

[70] Sungrim Sohn, Donald C Comeau, Won Kim, and W John Wilbur. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9(1):402, 2008.

[71] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Alex Hartshorn, Elvis Saravia, Yujia Xu, Xin Wang, Robert Stojnic, and Joelle Pineau. Galactica: A large language model for science. `https://galactica.org`, 2022.

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[73] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Apostolos Pyrgelis, Eric Gaussier, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. In *BMC bioinformatics*, volume 16, page 138. Springer, 2015.

[74] United Nations. Sustainable Development Goal 3: Good Health and Well-Being. `https://sdgs.un.org/goals/goal3`, 2015. "Ensure healthy lives and promote well-being for all at all ages.".

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

[76] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jon Grant, Tanvir Sajed, Daniel Johnson, Cecilia Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[77] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[78] Qinyuan Yao, Yuxiang Sun, and Xiang Ren. Kat: A knowledge-aware toolkit for hybrid question answering. In *Proceedings of NAACL*, 2022.

[79] Lei Zhang, David Perez, David Campos, et al. Triplet-based contrastive learning for medical entity linking with umls. In *Proceedings of ACL*, 2023.

[80] Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Knowledge-rich self-supervision for biomedical entity linking. *arXiv preprint arXiv:2112.07887*, 2021.

[81] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[82] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[83] Zhengxuan Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

Appendix A

Dataset Sizes

The following table reports the dataset sizes referenced in the thesis.

Table A.1: Dataset sizes used for fine-tuning and evaluation

Dataset	Entries
BioASQ training data	4,719
BiQA	769
DrugBank	9,377
GO Terms	47,735