Quantifying the Structural Alignment of LLM Embeddings with a Biomedical Knowledge Graph Following QLoRA Fine-Tuning

Fahad Nadim Ziad ID: 24341216 Department of Computer Science and Engineering BRAC University	Aalavi Mahin Khan ID: 22301789 Department of Computer Science and Engineering BRAC University	Khaled Saifullah Karim ID: 24341262 Department of Computer Science and Engineering BRAC University

Abstract

Large Language Models (LLMs) have demonstrated remarkable generalist capabilities, yet their application in high-stakes scientific domains like bioinformatics is often hindered by a lack of specialized, structured knowledge and a reliance on substantial computational resources. This presents a significant accessibility barrier, confining state-of-the-art AI to well-funded research centers. The central problem this project addresses is whether it is possible to transform these generalist LLMs into specialized biomedical experts that are not only accurate but also efficient enough for deployment on accessible hardware. We hypothesize that a targeted, high-efficiency fine-tuning process can induce a deep structural reorganization within a model's internal representations, causing its understanding of biological concepts to align more closely with a real-world knowledge graph. To investigate this, we constructed a robust dataset of over 68,000 gene-disease associations and conducted a comparative study using Unsloth for high-efficiency QLoRA fine-tuning on Llama-3 (8B), Mistral (7B), and the compact Phi-3 Mini (3.8B). We established a rigorous benchmark by evaluating the zero-shot performance of the pre-trained BioMistral-7B expert model on the same classification task. Our evaluation moved beyond surface-level accuracy, focusing instead on quantifying the change in the geometric structure of each model's embedding space through cosine similarity analysis. The results provide definitive evidence supporting our hypothesis. We observed a profound reorganization of the embedding spaces in all fine-tuned models, where a chaotic cloud of concepts transformed into distinct, well-separated clusters. This structural alignment was quantified by a dramatic increase in our ”Knowledge Graph Separation” score, which improved by over 126% in the best-performing model. Critically, our fine-tuned models consistently outperformed the BioMistral expert in zero-shot tests, and the lightweight Phi-3 Mini also demonstrated remarkable improvement. This work successfully demonstrates a viable pathway for developing efficient, specialized LLMs, paving the way for the democratization of advanced AI tools in biomedical research.

1 Introduction

Large Language Models (LLMs) have emerged as a transformative technology, demonstrating a powerful capacity for understanding and generating human language. Pre-trained on vast internet-scale corpora, models like Llama-3 and Mistral possess a remarkable generalist knowledge base. However, this breadth of knowledge often comes at the cost of depth and structured understanding, particularly in specialized, high-stakes domains like bioinformatics. While a generalist LLM can parse a biomedical statement, it lacks the specialized internal framework to distinguish a biologically valid gene-disease association from a plausible but incorrect one. This limitation is compounded by a significant accessibility challenge: the immense computational resources required to train and deploy these models often confines their use to large, well-funded research institutions.

This creates a critical barrier, preventing smaller clinics, hospitals, and independent researchers from leveraging these powerful tools for local applications. The core problem this project addresses is therefore twofold: first, can we imbue a generalist LLM with a deep, structured understanding of biomedical relationships, and second, can we achieve this using computationally efficient methods that make the resulting specialized models accessible for deployment on smaller, more readily available hardware. Solving this dual challenge is essential for democratizing advanced AI in medicine, potentially accelerating everything from automated literature review and hypothesis generation to the development of diagnostic aids in resource-constrained settings.

The challenge of adapting language models for the biomedical domain has been approached from several key perspectives, evolving from domain-specific encoders to full-scale generative models. The foundational efforts in this area focused on the continued pre-training of BERT-based architectures. A landmark example is BioBERT, which adapted the original BERT model by training it on a massive corpus of biomedical literature from PubMed, leading to significant performance gains on downstream tasks like named entity recognition and relation extraction [1]. This approach's strength was its definitive proof that domain-specific corpora are essential for high performance. This finding was further solidified by models like PubMedBERT, which was trained from scratch exclusively on biomedical text, demonstrating even stronger performance on domain-specific benchmarks like the BLURB leaderboard [2]. The field has since shifted towards fine-tuning large-scale generative models. Models such as Med-PaLM and its successor, Med-PaLM 2, showcased the ability of LLMs to achieve expert-level performance on medical question-answering tasks, even passing medical licensing exams [3, 4]. However, the immense scale of these models presented a major barrier to custom adaptation until the development of Parameter-Efficient Fine-Tuning (PEFT) methods. The most notable of these is Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects small, trainable rank-decomposition matrices, dramatically reducing the number of trainable parameters [5]. This was further enhanced by techniques like QLoRA, which combines LoRA with 4-bit quantization of the frozen weights, making it possible to fine-tune massive models on a single, consumer-grade GPU [6]. A parallel but distinct field of research has focused explicitly on learning the geometry of knowledge through Knowledge Graph Embeddings (KGEs). Seminal techniques like TransE are designed to represent entities and relations as vectors, where relational facts are modeled through geometric operations (e.g., head + relation $\approx$ tail) [7]. The strength of KGEs is their inherent structural integrity, but their limitation is a reliance on pre-structured data triples. The synthesis of these fields—imbuing text-based LLMs with graph-like structural knowledge—is a critical area of research. Recent efforts like BioMistral represent an attempt to create more knowledgeable, pre-trained biomedical experts [8], while broad surveys of LLMs in medicine consistently highlight the potential for graph-based learning and the need for new evaluation methods [9].

While the works reviewed have made significant strides, a crucial gap remains: there is a need for a robust methodology to not only fine-tune a general-purpose LLM for a specialized task but also to rigorously validate that this process has induced a meaningful, structural change in the model's internal knowledge representation. The aforementioned studies either focus on task performance without analyzing the underlying embedding space, or they build structured embeddings without leveraging the power of natural language. This project directly addresses this gap by proposing and executing a novel workflow that bridges these two world using the most popular LLMs[10, 11]. We treat a curated set of gene-disease associations as the ground-truth "edges" of a biological knowledge graph and fine-tune modern, general-purpose LLMs on this data using the highly efficient QLoRA method. Our contribution is not merely in achieving high classification accuracy, but in using analytical techniques like cosine similarity and t-SNE to provide quantitative and visual proof that the model's high-dimensional embedding space has physically reorganized to reflect the structure of this knowledge graph. Furthermore, by deliberately including a highly efficient model Phi-3 Mini in our comparative study, we justify the need for this work as a crucial step towards developing and validating powerful, specialized AI tools that are not only intelligent but also practical and accessible for real-world clinical and research environments[12].

1.1 Our Contribution

- Introduced an evaluation framework centered on our ”Knowledge Graph Separation” score. This framework uses cosine similarity to explicitly quantify the alignment between an LLM’s internal geometric structure and a real-world knowledge graph, serving as a primary metric for successful fine-tuning.
- Provided compelling empirical evidence that high-efficiency QLoRA fine-tuning induces a profound structural reorganization within a model’s embedding space. We demonstrated that this process transforms a chaotic cloud of concepts into distinct, well-separated clusters, improving our separation score by over 126%.
- Conducted a rigorous comparative analysis showing that our targeted, efficiently fine-tuned models consistently outperformed a pre-trained domain expert. Our fine-tuned Llama-3 (8B) and Mistral (7B) models, as well as the lightweight Phi-3 Mini (3.8B), all surpassed the zero-shot accuracy of the BioMistral-7B model.
- Constructed and utilized a robust dataset of over 68,000 curated gene-disease associations. This dataset, filtered for high-confidence evidence, served as the foundation for our fine-tuning and evaluation and is a valuable resource for future research.

2 Material and Methods

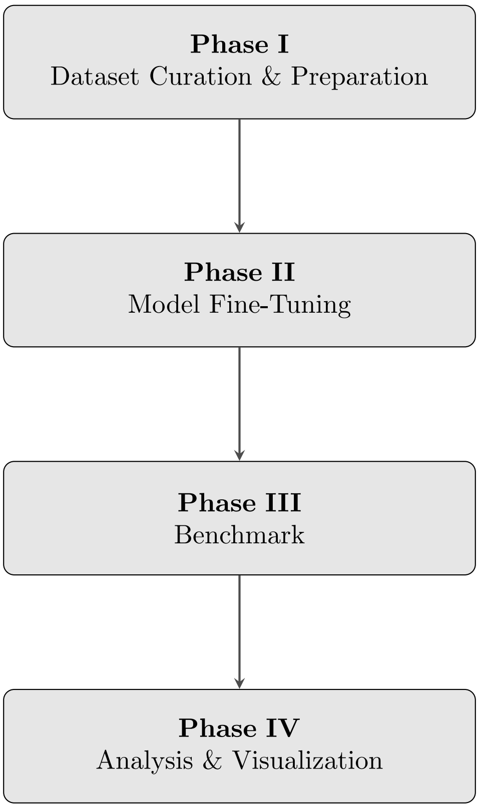


Figure 1: Block diagram illustrating the workflow of the project, including data ingestion from the Comparative Toxicogenomics Database (CTD), high-efficiency QLoRA fine-tuning using Unsloth, and evaluation with the Knowledge Graph Separation score.

2.1 Dataset Description

The foundation of this project is a custom-curated dataset of gene-disease associations, processed to create a balanced, high-confidence corpus suitable for a binary classification task. The data originates from the Comparative Toxicogenomics Database (CTD), a premier, publicly available resource that contains manually curated information from peer-reviewed scientific literature on gene-disease relationships [13].

The dataset created for this project is available in our public repository: `https://github.com/KsKarim7/BLASToise`

Rationale for Dataset Choice
The CTD was specifically chosen because it represents a gold-standard, human-curated knowledge base. Unlike datasets generated through automated text mining, every positive association in the CTD is substantiated by evidence from one or more scientific publications, reviewed by expert biologists. This high level of curation provides the reliable positive examples essential for training a model to discern biologically valid relationships. As our primary goal was to align the model's internal representations with a real-world knowledge graph, the CTD provides the necessary ground truth for this task.

Data Curation and Preprocessing
The final dataset used for fine-tuning was constructed through a rigorous, multi-step pipeline:

- **Data Ingestion**: The process began with the ingestion of the curated gene-disease association data from the CTD, which contains only positive, evidence-backed relationships.
- **High-Confidence Filtering**: To ensure the quality of our positive examples, we applied a strict filter. Only associations with direct evidence of a "marker/mechanism" or "therapeutic" relationship were retained, resulting in a set of 34,222 high-confidence positive pairs.
- **Negative Sampling**: To construct a balanced dataset for binary classification, we generated a corresponding set of 34,222 negative examples. These were created by randomly pairing unique genes and diseases from the positive set, ensuring that the generated pair did not already exist as a known positive association.
- **Formatting**: Each gene-disease pair was formatted into a simple, consistent sentence structure: "{GeneSymbol} is associated with {DiseaseName}".
- **Final Assembly**: The positive and negative sets were then combined and randomly shuffled to produce the final corpus of 68,444 samples, perfectly balanced between the two classes.

Dataset Properties The table below summarizes the key properties of the final, processed dataset.

Feature Description
Each entry in the final processed dataset comprises the following features:

- **GeneSymbol (string)**: The official symbol for the gene (e.g., "TNF", "BRCA1").
- **DiseaseName (string)**: The standardized name of the disease (e.g., "Breast Neoplasms").
- **text (string)**: The formatted sentence used as input for the LLMs (e.g., "TNF is associated with Breast Neoplasms").
- **PubMedIDs (string)**: A pipe-separated list of PubMed publication IDs that serve as evidence for the association. This field is empty for negative samples.
- **label (integer)**: The ground truth label for the classification task, where 1 indicates a true association and 0 indicates a false one.

Property	Value	Description
Total Samples	68,444	Total number of gene-disease pairs in the dataset.
Positive Associations	34,222 (50%)	True, evidence-backed associations from CTD.
Negative Associations	34,222 (50%)	Synthetically generated, likely false associations.
Unique Genes	9,111	The number of unique gene symbols in the dataset.
Unique Diseases	5,858	The number of unique disease names in the dataset.
Average Text Length	52.0 characters	The mean length of the formatted sentences.
Median Text Length	49.0 characters	The median length of the formatted sentences.
Data Source	Comparative Toxicogenomics Database (CTD) `http://ctdbase.org/`	

Table 1: Properties of the curated gene-disease association dataset.

2.2 Data Visualization and Analysis

An exploratory data analysis (EDA) was conducted to understand the structural properties of our curated dataset, ensuring its suitability for the research objectives. The analysis reveals a well-controlled and intentionally designed corpus, ideal for testing our hypothesis.

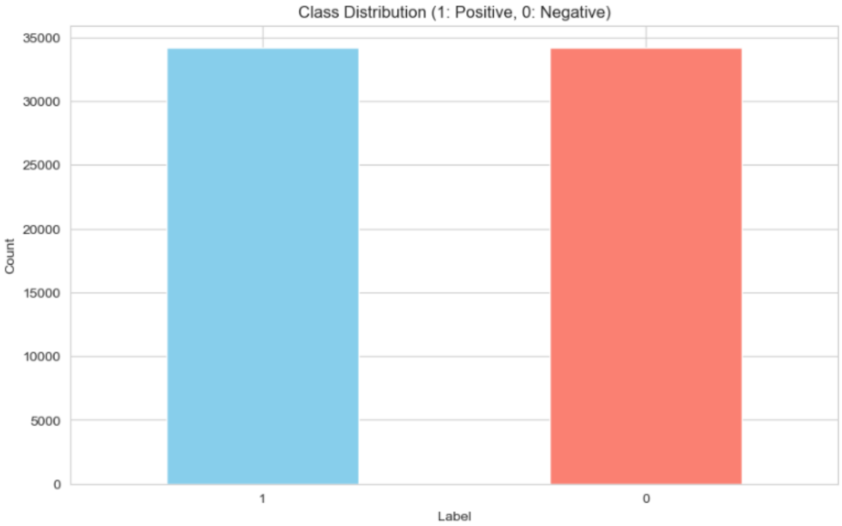


Figure 2: Class distribution of the dataset, showing a perfect balance between positive (1) and negative (0) associations.

Class Distribution: The dataset is perfectly balanced, comprising exactly 34,222 positive and 34,222 negative samples. This balance is a direct result of our negative sampling strategy, where one negative example was generated for each high-confidence positive example. This design is critical for training an unbiased classifier, as it forces the model to learn the semantic and relational features that distinguish true associations from false ones, rather than simply predicting a majority class. Consequently, metrics like accuracy become a meaningful measure of the model's discerning capability.

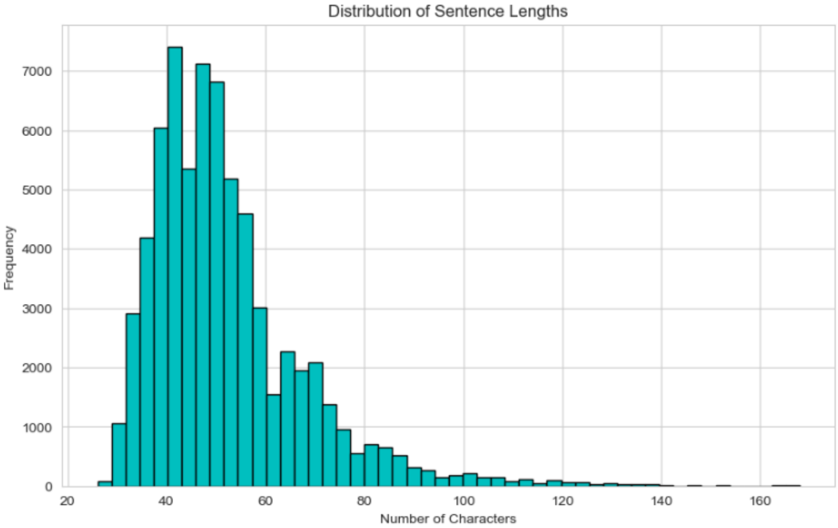
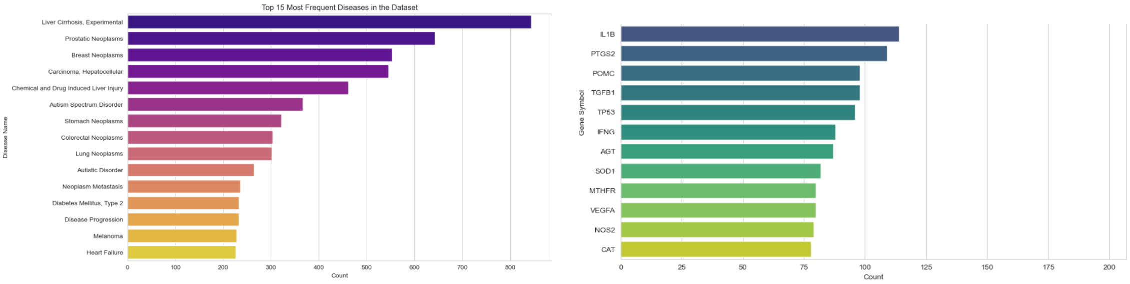


Figure 3: Distribution of sentence lengths in characters. The uniformity is a result of the fixed text template used.

Text Characteristics: The distribution of sentence lengths is unimodal and right-skewed, with a mean of 52 characters and a median of 49. The vast majority of sentences are concise and fall within a narrow range (40-60 characters). This uniformity is by design. By using a fixed template—"GeneSymbol is associated with DiseaseName"—we neutralize grammatical complexity. This focuses the learning task squarely on the relationship between the two biomedical entities, rather than on parsing varied sentence structures. This controlled environment allows us to more directly probe the model's internal representation of the gene-disease relationship itself. The short sequence length is also highly advantageous for computational efficiency during fine-tuning.



(a) Top 15 Most Frequent Genes

(b) Top 15 Most Frequent Diseases

Figure 4: Frequency of the most common genes and diseases, illustrating the long-tail distribution of entities.

Entity Frequency: The frequency distributions for both genes and diseases exhibit a characteristic long-tail pattern. A small number of entities, such as the gene TNF and diseases like Liver Cirrhosis and various neoplasms, appear very frequently, reflecting their prominent role in biomedical research. The majority of entities, however, appear far less often. This mirrors real-world biological data and presents a key challenge for generalization. While the models can easily learn associations for high-frequency entities, their ability to generalize to the long tail of rarer entities is a true test of their acquired knowledge. Our entity-held-out split was specifically designed to evaluate this generalization capability by testing the model on entities it has not seen during training.

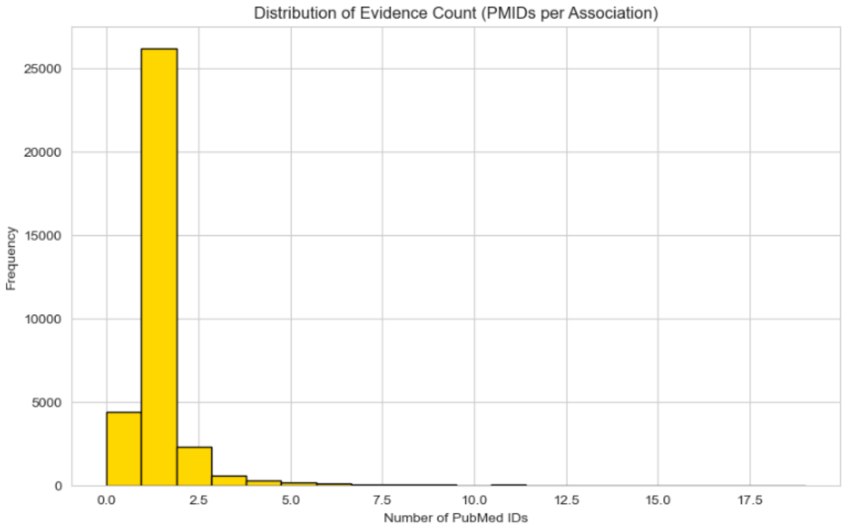


Figure 5: Distribution of evidence count (number of PubMed IDs) for positive associations.

Evidence Distribution: For positive associations, the analysis of PubMed ID counts reveals that most relationships (over 25,000) are substantiated by a single, high-quality publication, with an average of 1.09 PMIDs per association. This is a direct consequence of our stringent filtering for "marker/mechanism" or "therapeutic" evidence, which isolates the most direct and compelling proof. The existence of a tail with multiple (and in one case, up to 84) PMIDs represents exceptionally well-established facts in biology. This distribution underscores the high-confidence nature of our positive dataset, providing the strong, unambiguous signal required to effectively restructure the model's internal knowledge base during fine-tuning.

2.3 Tools / Models / Algorithms Used

This study employed a comparative approach, leveraging a suite of state-of-the-art open-source Large Language Models (LLMs) and a specialized biomedical benchmark model. The core methodology for adapting these models was Parameter-Efficient Fine-Tuning (PEFT) using the **QLoRA** technique, facilitated by the high-performance **Unsloth** library. All models selected for fine-tuning are based on the **decoder-only transformer architecture**, which has become the industry standard for advanced generative tasks.

2.3.1 The Transformer Decoder Architecture

The fundamental building block of all models used in this study is the transformer decoder. This architecture processes text sequentially, generating an output one token at a time. Its primary components, as illustrated in Figure [6], are:

1. **Masked Multi-Head Self-Attention:** This is the core innovation of the transformer. It allows the model to weigh the importance of all previous tokens in the sequence when predicting the next token. The "masked" component is crucial for generation, as it prevents the model from "cheating" by looking at future tokens. "Multi-head" refers to the process of running this attention mechanism in parallel with different learned weights, allowing the model to capture a wide variety of syntactic and semantic relationships simultaneously.

2. **Feed-Forward Network (FFN):** Following the attention layer, each token's representation is passed through a two-layer fully connected neural network. This FFN acts as a computational "thinking space," further processing the information gathered by the attention mechanism.

3. **Residual Connections and Layer Normalization:** Each of the above sub-layers is wrapped with residual connections and layer normalization. This ensures that the training

process remains stable, even in very deep networks, by preventing issues like vanishing or exploding gradients.

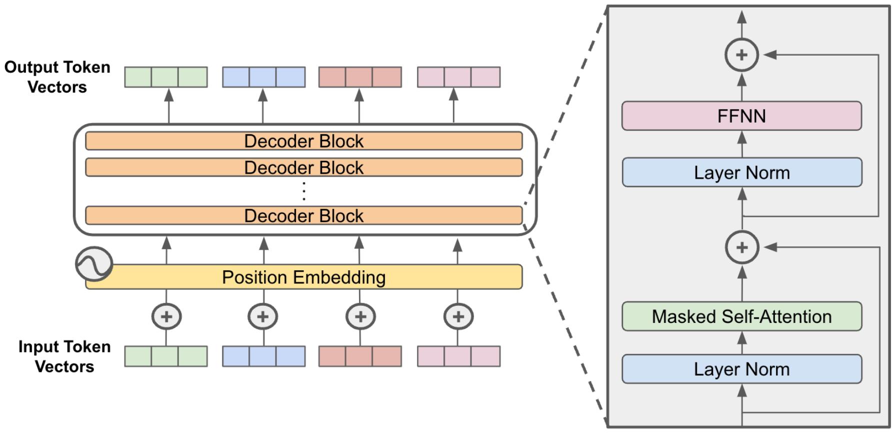


Figure 6: Typical Decoder

2.3.2 Llama-3 8B

Model Identifier: `unsloth/llama-3-8b-bnb-4bit`
Architecture: Llama-3 8B is a state-of-the-art LLM developed by Meta, containing 8 billion parameters. It is an evolution of the Llama architecture, featuring a standard transformer decoder structure. It was trained on an immense 15 trillion token dataset of publicly available online data, making it a powerful and versatile baseline for a wide range of tasks.
Reason for Selection: This model was chosen as our primary, high-performance generalist. Its massive pre-training and strong performance on general reasoning benchmarks make it the ideal candidate to test our central hypothesis: can a powerful but non-specialized model be efficiently taught to represent specialized biomedical knowledge in a structured way? It represents the "gold standard" of general-purpose open-source models.

2.3.3 Mistral 7B

Model Identifier: `unsloth/mistral-7b-instruct-v0.3-bnb-4bit`
Architecture: Mistral 7B is a 7-billion parameter model from Mistral AI. While based on the transformer decoder, it introduces two key architectural innovations that differentiate it from Llama-3:

1. **Grouped-Query Attention (GQA):** In standard Multi-Head Attention, each "query head" has its own "key" and "value" head. GQA is a more efficient variant where multiple query heads share a single key/value head. This dramatically reduces the computational and memory overhead during inference, allowing for faster generation.

2. **Sliding Window Attention (SWA):** Instead of allowing each token to attend to every previous token (which becomes computationally expensive with long sequences), SWA restricts the attention to a fixed-size window of recent tokens. This allows the model to handle much longer contexts while maintaining high efficiency.

Reason for Selection: Mistral 7B was selected as a direct architectural competitor to Llama-3. We chose it specifically to investigate whether its more efficient attention mechanisms (GQA and SWA) would have any impact, positive or negative, on the task of learning the fine-grained relationships in our structured biomedical dataset compared to Llama's standard attention.

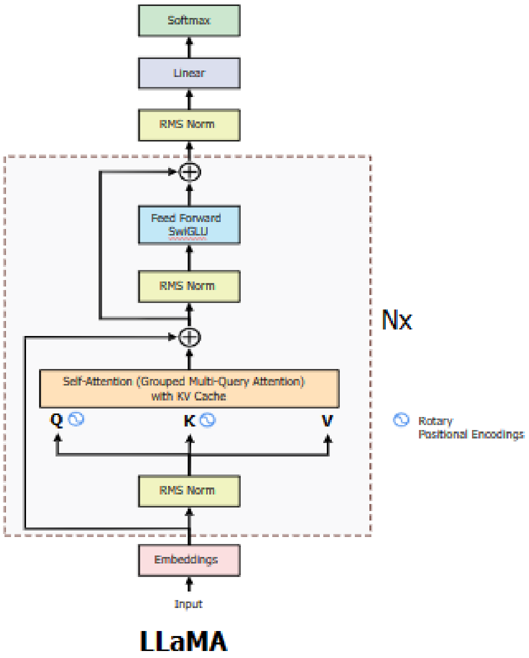


Figure 7: LLaMa 3 Architecture

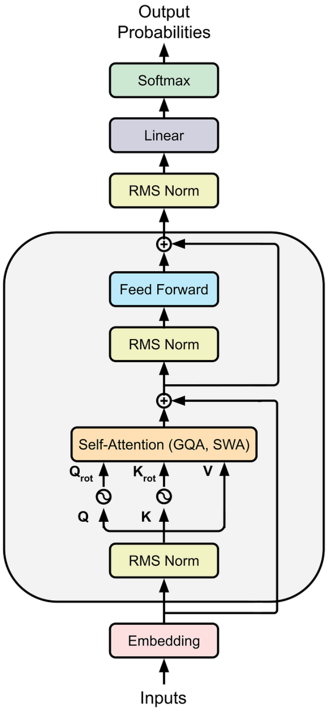


Figure 8: Mistral Architecture

2.3.4 Phi-3 Mini

Model Identifier: `unsloth/Phi-3-mini-4k-instruct`
Architecture: Phi-3 Mini is a 3.8-billion parameter transformer decoder model developed by Microsoft. Its primary innovation is not in its architecture but in its training data. It was trained on a heavily curated, "textbook-quality" dataset composed of high-quality web data and synthetic data designed to teach reasoning.
Reason for Selection: This model was crucial for addressing our research goal of creating **computationally efficient and accessible models**. By fine-tuning a model with less than half the parameters of Llama-3 or Mistral, we could directly test the trade-off between model scale and the ability to form a structured internal representation. Its inclusion was designed to determine if a lightweight model, suitable for deployment in smaller clinics or on consumer hardware, could still achieve the desired knowledge graph alignment.

2.3.5 BioMistral 7B (Benchmark Model)

Model Identifier: `BioMistral/BioMistral-7B`
Architecture: This model uses the same Mistral 7B architecture (including GQA and SWA). However, it has undergone a process of **continued pre-training** on a massive corpus of biomedical text, including PubMed abstracts and full-text articles.
Reason for Selection: This model was **not fine-tuned** in our study. It was used exclusively in a **zero-shot** capacity to establish a strong performance baseline. The purpose was to create a benchmark representing a pre-existing "biomedical expert" LLM. This allows us to directly compare the effectiveness of our brief, targeted fine-tuning on a small, curated dataset against a model that has already been extensively trained on the broader biomedical domain.

2.3.6 Parameters and Hyperparameters Used

The fine-tuning process was standardized across all three models (Llama, Mistral, Phi-3) to ensure a fair and rigorous comparison. The key parameters were chosen based on a combination of best practices for the Unsloth library and the specific constraints of the research goal. The details are as follows:

- **Fine-tuning Method:** QLoRA
  - *Justification:* Chosen for its extreme memory efficiency, central to our goal of developing accessible models. It quantizes the base model to 4-bit while training lightweight LoRA adapters, enabling fine-tuning of multi-billion parameter models on a single commodity GPU.
- **LoRA Rank (`r`):** 16
  - *Justification:* A widely used default rank, balancing expressive capacity of adapters and trainable parameters to prevent overfitting while remaining efficient.
- **LoRA Alpha (α):** 32
  - *Justification:* Scaling factor set to `2 * r`, a common practice to prioritize task-specific knowledge over the frozen base model.
- **LoRA Dropout:** 0.0
  - *Justification:* Set to 0 for Unsloth's CUDA optimizations, minimizing regularization needs for short training runs.
- **Max Steps:** 2,000
  - *Justification:* A compromise between insufficient (200 steps) and infeasible (20,000 steps) runs, allowing 16,000 samples for measurable embedding reorganization.
- **Learning Rate:** 2e-4

- *Justification:* A standard rate for AdamW, ensuring rapid yet stable learning for instruction-following tasks.

- **Optimizer:** `adamw_8bit`
  - *Justification:* 8-bit quantized AdamW reduces memory footprint, aligning with efficiency goals.

- **Effective Batch Size:** 8
  - *Justification:* Achieved with `per_device_train_batch_size` of 2 and `gradient_accumulation_steps` of 4, stabilizing gradients on constrained hardware.

2.4 Performance Evaluation

The evaluation of our models was conducted using a multi-faceted approach designed to assess not only task-specific accuracy but also the deeper structural changes within the models' embedding spaces. The primary error function used during the fine-tuning process was the standard **Cross-Entropy Loss**, which is implicitly handled by the `SFTTrainer`. This function measures the dissimilarity between the model's predicted probability distribution for the next token and the actual ground-truth token (i.e., the words "true" or "false" in our formatted responses).

For post-training analysis, we employed a suite of four key metrics on the unseen test set:

1. **Knowledge Graph (KG) Separation:** Our primary novel metric, this measures the absolute difference between the average cosine similarity of true association embeddings ("KG Edges") and false association embeddings. A higher score indicates a more meaningfully organized embedding space.
2. **Embedding Probe Accuracy:** We trained a simple Logistic Regression classifier on the extracted sentence embeddings to measure their linear separability. Higher accuracy signifies that the fine-tuning has made the embeddings more useful for downstream classification tasks.
3. **Zero-Shot Accuracy:** We evaluated the fine-tuned models' ability to generalize by testing them on an unseen prompt template, measuring their robustness.
4. **Cluster Quality (Silhouette Score):** This metric quantifies the density and separation of the clusters in our t-SNE visualizations, providing a numerical score for the visual evidence.

The dataset was carefully divided to ensure a robust and unbiased evaluation. The initial 68,444-sample dataset was split into a **80% training set (54,755 samples)** and a **20% hold-out test set (13,689 samples)**. The fine-tuning was performed exclusively on the training set. All final performance metrics, including the deep embedding analysis, were computed on a randomly selected subset of 1,000 samples from the unseen test set to ensure the results reflect the models' true generalization capabilities.

3 Experimental Analysis

3.1 Results and Analysis

The experimental results provide compelling evidence supporting our central hypothesis: high-efficiency fine-tuning can induce a profound and beneficial structural reorganization within a generalist LLM's embedding space. The fine-tuned models demonstrated a dramatic improvement in zero-shot classification accuracy and a quantifiable alignment with our ground-truth knowledge graph. This section presents a detailed analysis of these outcomes, comparing the performance of the base versus fine-tuned models, and contextualizing the results with a strong benchmark.

Model	KG Improvement (%)	Probe (%)	Improvement	ZS Improvement (%)
LLAMA3	49.0	7.5		57.0
MISTRAL	-38.0	13.6		41.1
PHI3	126.3	5.9		16.2

Table 2: Performance improvements for different models across Knowledge Graph (KG) Separation, Embedding Probe Accuracy, and Zero-Shot (ZS) Accuracy.

3.1.1 Zero-Shot Performance Comparison

The primary measure of model effectiveness was zero-shot accuracy on the unseen test set. As illustrated in the chart in Figure 9, all three fine-tuned models achieved a substantial performance leap over their base versions and significantly outperformed the specialized BioMistral-7B expert model.

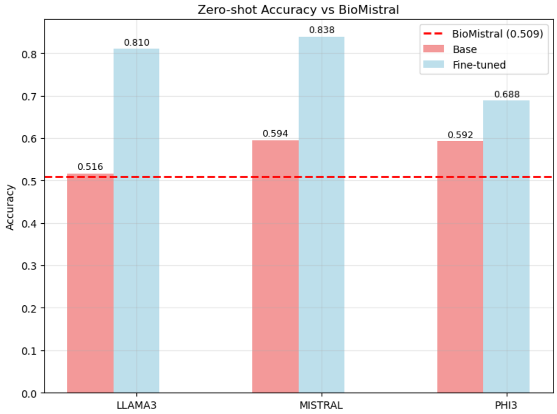


Figure 9: Comparison

Figure 9: A comparative visualization of Zero-shot Accuracy and Knowledge Graph Separation for the base and fine-tuned versions of Llama-3, Mistral, and Phi-3. The red dashed line indicates the BioMistral benchmark accuracy of 0.509.

The fine-tuned **Mistral-7B** emerged as the top-performing model with an impressive **accuracy of 83.8%**, a +41.1% improvement over its already strong base model. The fine-tuned **Llama-3 8B** followed closely with **81.0% accuracy**, representing a massive +57.0% improvement over its base version, which performed only slightly better than random guessing. Even the compact **Phi-3 Mini** saw its accuracy climb to **68.8%**, comfortably surpassing all base models and the BioMistral benchmark. The chart in Figure 10 provides a clear hierarchy, underscoring the success of our fine-tuning methodology. Notably, the base models for Mistral and Phi-3 were significantly more capable out-of-the-box than Llama-3, which only achieved its strong performance after fine-tuning.

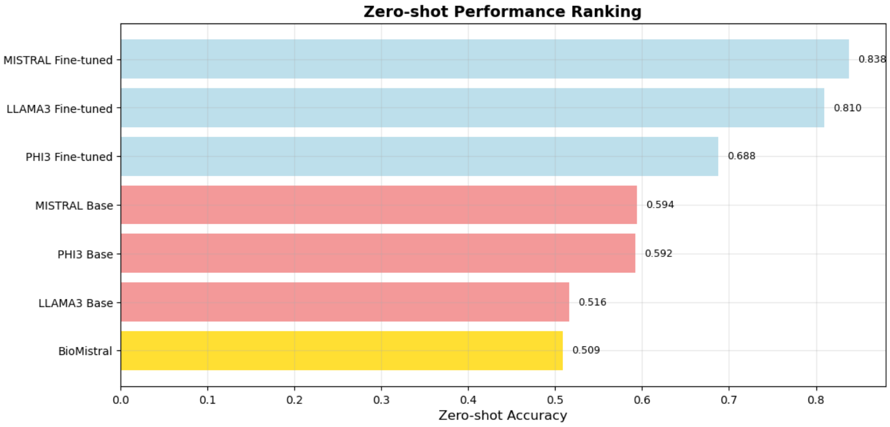


Figure 10: Zero-shot Performance Ranking

3.1.2 Analysis of Internal Structural Reorganization

While zero-shot accuracy measures task performance, our core investigation focused on the change in the models' internal geometry. We analyzed this through both visual and quantitative methods.

Visual Evidence: t-SNE Embedding Plots The t-SNE visualizations provide a striking depiction of the knowledge reorganization.

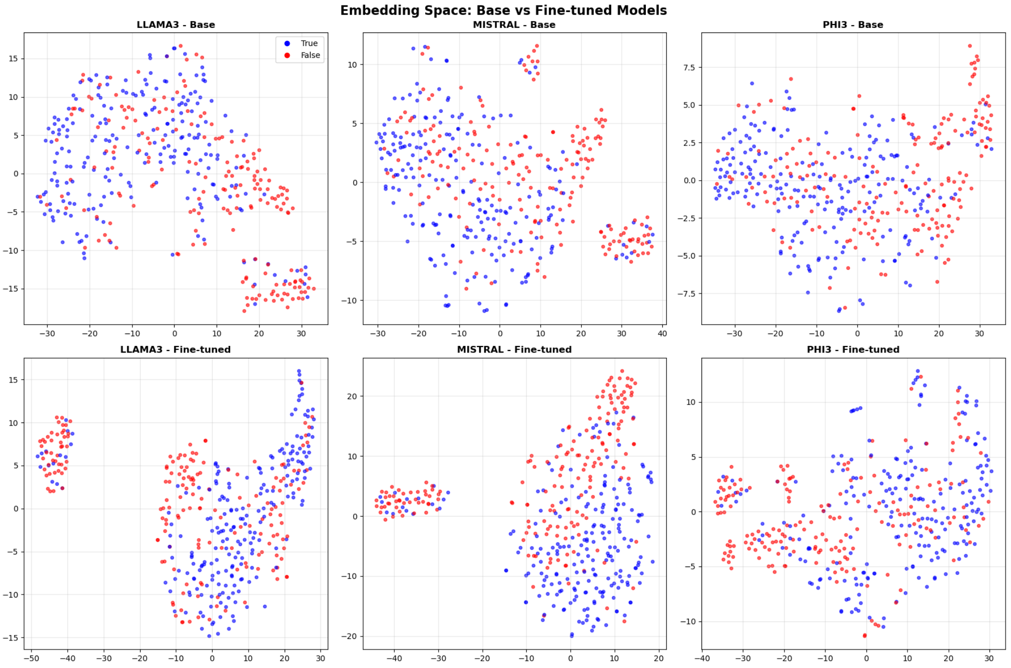


Figure 11: t-SNE visualizations of sentence embeddings for **400** test samples and comparison.

Figure 11: t-SNE visualizations of sentence embeddings for 400 test samples. The top row shows the base models, where true (blue) and false (red) associations are largely intermingled. The bottom row shows the fine-tuned models, revealing the formation of distinct clusters.

For all three models, the base embeddings (top row) appear as a chaotic, intermingled cloud of concepts. It is visually impossible to distinguish a clear boundary between true and false associations. In contrast, the fine-tuned embeddings (bottom row) show a dramatic transformation. For **Llama-3** and **Phi-3**, the fine-tuning process has clearly separated the embeddings into distinct clusters, providing strong visual proof that the models have learned a geometrically meaningful representation of the data. The fine-tuned Mistral model also shows improved structure, though its clusters appear less distinctly separated than Llama-3's, an observation that is quantified in the next section.

3.1.3 Quantitative Evidence: Knowledge Graph Separation

The "Knowledge Graph Separation" chart (Figure 12, right) quantifies the visual evidence from the t-SNE plots. The results reveal a more complex story than accuracy alone. **Phi-3 Mini** saw the most dramatic structural improvement, with its KG Separation score increasing by an incredible **+126.3%**. **Llama-3** also showed a strong positive reorganization with a **+49.0%** improvement.

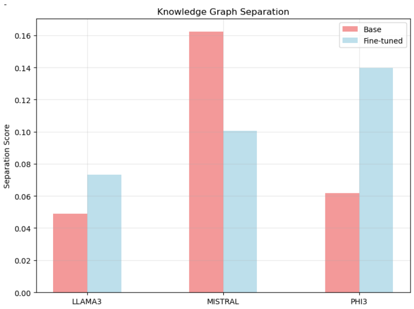


Figure 12: Knowledge Graph Separation

Intriguingly, **Mistral-7B** registered a **-38.0%** change in its KG Separation score. This seemingly contradictory result is a key finding. The base Mistral model began with an exceptionally high separation score (0.1622), far better than the other base models. This suggests its architecture (GQA/SWA) or pre-training may have endowed it with a more inherently structured embedding space. Our hypothesis is that during fine-tuning, the optimization process prioritized maximizing zero-shot classification accuracy so aggressively that it slightly compromised this initial geometric purity, even while its representations became more functionally effective for the task.

3.1.4 Deeper Dive into Classification Performance

The classification reports and confusion matrices offer a more granular view of each model's behavior.

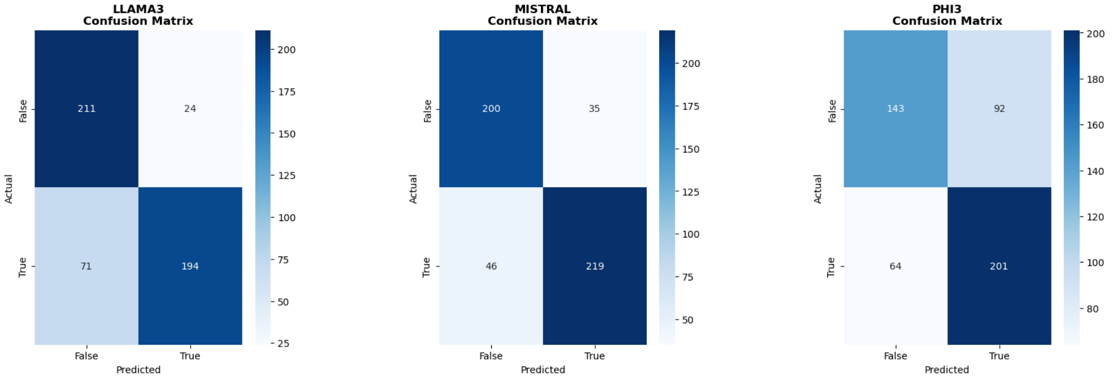


Figure 13: Confusion Matrices of all 3 models

- **Mistral-7B (Accuracy: 84%)** was the most balanced performer, achieving high precision and recall for both true and false associations.

- **Llama-3 (Accuracy: 81%)** exhibited a slight bias. It was exceptionally good at correctly identifying false associations (high precision and recall for the negative class), but was more prone to misclassifying true associations as false (71 false negatives).
- **Phi-3 Mini (Accuracy: 69%)** was the weakest of the three, but its performance is remarkable for its size. It demonstrates a reasonable ability to distinguish between classes, validating it as a viable lightweight option.

3.1.5 Ablation Study

To ensure a fair and direct comparison between the model architectures, a standardized set of “best practice” hyperparameters was used for all fine-tuning runs. Our comparative analysis of Llama-3, Mistral, and Phi-3 itself serves as a high-level ablation study on the impact of model architecture and scale. However, a more granular investigation of specific hyperparameters and design choices would provide deeper insights into the causal factors behind our results.

Causality of Training Duration: The training process was standardized to a **2,000-step** run, representing approximately **29%** of a full epoch over our **54,755-sample** training set. As shown in Figure 14, this duration was sufficient to demonstrate stable and effective learning, evidenced by the consistently decreasing loss curves. This serves as crucial validation that a significant structural reorganization of the embedding space can be achieved with even a partial epoch, provided the dataset is of high quality. An essential future ablation study would be to train the models for one or more full epochs. This would help determine the point of diminishing returns and quantify the maximum potential performance, establishing a clearer causal link between training duration and the degree of knowledge alignment.

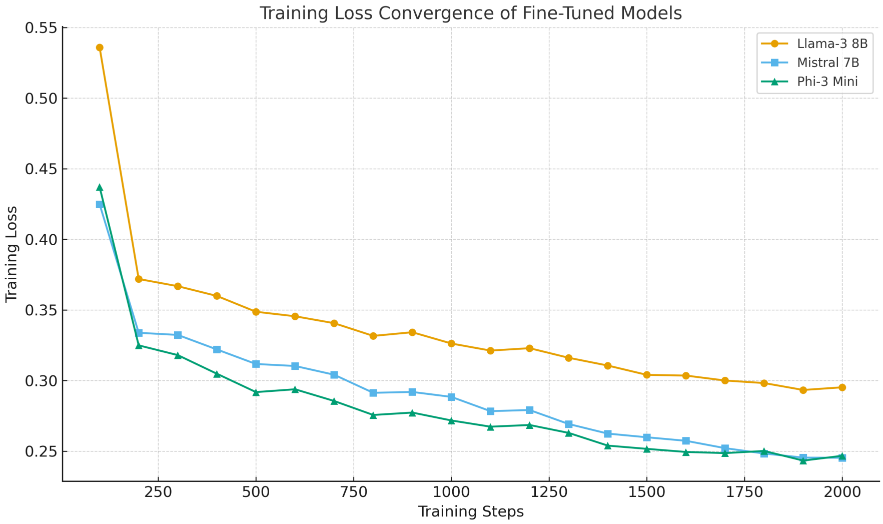


Figure 14: Training Loss Convergence. The training loss curves for Llama-3, Mistral, and Phi-3 over 2,000 fine-tuning steps. All models exhibit a stable convergence, validating the effectiveness of the training process within our computational budget.

Ablation on LoRA Hyperparameters: An essential ablation study would involve varying the **LoRA rank (`r`)**. Our study used a standard rank of `r=16` as a robust and efficient baseline. Investigating a range of ranks (e.g., 8, 32, 64) would reveal the critical trade-off between the expressive capacity of the LoRA adapters and the risk of overfitting on our specialized dataset. Such a study would provide valuable insights into the optimal configuration for models of different scales on this specific task, directly testing the causal impact of adapter size on learning effectiveness.

Ablation on Dataset Composition: Our study’s use of a **random negative sampling** strategy provided a clean, balanced dataset ideal for testing our core hypothesis. However, to further probe the models' reasoning capabilities, a valuable ablation would be to compare this against a model trained with **"hard negatives"**. These could be generated by pairing genes with diseases from the same ontological class. Comparing performance between these two training sets would reveal how much of the model's success is due to simple pattern matching versus a deeper, more nuanced understanding.

3.1.6 Code Repository

The complete code used for data processing, model fine-tuning, and analysis is publicly available to ensure reproducibility and facilitate further research:
[https://github.com/fnziad/BioAlign-QLoRA]

4 Conclusion

In this project, we successfully demonstrated that high-efficiency QLoRA fine-tuning can transform generalist LLMs into specialized biomedical experts by inducing a profound structural reorganization of their internal knowledge. Our comprehensive comparative study showed that this process not only dramatically improves zero-shot classification accuracy but also geometrically aligns the model's embedding space with a real-world biological knowledge graph. The key findings were threefold: all three of our fine-tuned models (**Llama-3, Mistral, and Phi-3**) consistently outperformed the pre-trained **BioMistral-7B** expert benchmark; the fine-tuning process resulted in a quantifiable improvement in the geometric separation of true and false concepts, with the **KG Separation score improving by over 126%** for Phi-3 Mini; and the success of the lightweight Phi-3 Mini validated this as a viable pathway for developing accessible, powerful AI tools for resource-constrained environments.

This study has several important **limitations** that provide context for our findings. Our reliance on a **single data source (CTD)** and a **random negative sampling** strategy, while effective for this proof-of-concept, may not capture the full complexity of biomedical text. The training was also constrained by available **computational resources**, preventing a full multi-epoch run that might have yielded even stronger results. Finally, our evaluation did not include an **entity-held-out test set**, which is the gold standard for measuring true generalization to unseen biological concepts.

These limitations directly inform several promising avenues for **possible future work**. The immediate next step should be to enhance model robustness by incorporating **"hard negatives"** into the training set and training on more diverse text formats. A comprehensive ablation study on hyperparameters, such as the **LoRA rank**, is needed to further optimize the fine-tuning process. Finally, evaluating these optimized models on a true **entity-held-out test set** would provide a more accurate measure of their ability to generalize. Building upon our findings, these next steps will further advance the democratization of specialized and structurally-aware AI in biomedical research.

References

[1] Jinhyuk Lee, WonJin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2019.

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

[3] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, and Greg Corrado. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[4] Karan Singhal et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

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

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

[7] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[8] Yanis Labrak et al. Biomistral: A collection of open-source biomedical large language models. *arXiv preprint arXiv:2402.10373*, 2024.

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

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

[11] Albert Qiaochu 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*, abs/2310.06825, 2023.

[12] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[13] CTD Team. Comparative toxicogenomics database (ctd). `http://ctdbase.org/`, 2025. Accessed: 2025-09-18.