Title

Evaluating the Diagnostic Potential of Large Language Models in Fetal Alcohol Spectrum Disorder

Author

Yuan, Faith Tsyr-Huey

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluating the Diagnostic Potential of Large Language Models in Fetal Alcohol Spectrum Disorder

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioinformatics

by

Faith Tsyr-Huey Yuan

2024

ABSTRACT OF THE THESIS

Evaluating the Diagnostic Potential of Large Language Models in
Fetal Alcohol Spectrum Disorder

by

Faith Tsyr-Huey Yuan
Master of Science in Bioinformatics
University of California, Los Angeles, 2024
Professor Shantanu H. Joshi, Chair

Large language models enjoy wide spread applications in both general and more personalized use cases. These models can be dynamically trained on well defined clinical data. However, several pre-existing models that have not been trained to provide diagnostic information for disorders with clinical heterogeneity. Specifically, our preliminary analysis showed that existing models such as BioMistral are generalized on publicly available PubMed data but are unable to accurately take in clinical symptoms for accurate characterization of fetal alcohol spectrum disorder (FASD). To overcome this challenge, we propose to retrain the pre-existing BioMistral model on a synthetic FASD-specific training set to correctly categorize symptoms into diagnostic codes. By changing the learning rates and epochs, we are able to evaluate the performance of both overfitted or poorly trained models and a highly trained model on a test set containing synthetic clinical notes. We demonstrate evaluation performance using confusion matrices and the Kullback-leibler divergence (on the log-odds probabilities) and show that retraining BioMistral model has the capability to correctly diagnose individuals with fetal alcohol spectrum disorder over a poorly trained model.

The thesis of Faith Tsyr-Huey Yuan is approved.

Yingnian Wu

Ricky K. Taira

Shantanu H. Joshi, Committee Chair

University of California, Los Angeles

2024

To everyone involved in this journey . . .

especially— my parents and family —

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I wish to express my deepest gratitude to my committee chair Dr. Shantanu Joshi. His invaluable support and guidance and contributions have been fundamental to the completion of this thesis. I am profoundly grateful for his dedication and mentorship throughout this journey.

I would also like to convey my gratitude to Professor Ying Nian Wu and Professor Ricky Taira for their involvement in the review process of my thesis.

I am especially grateful to Andrew Lizarraga for his exceptional mentorship and support along the way.

Finally, I would like to thank my family, friends, and my cat for their unconditional love and support who make this all possible.

CHAPTER 1

Introduction

The wide-spread public introduction of large language models with Chat-GPT in 2022 [4] brought about controversy and excitement. The general public was skeptical of AI and machine-learning methods, but the scientists were excited to see the possibilities of advancement using these models. A large-area of need for scientific research and improvement has long been clinical diagnostics. Large language models can be tailored and trained to improve the diagnostic space, aiming to aid physicians with more structure and efficiency. Diagnosing neurodevelopmental disorders has long posed a variety of challenges due to the complex etiology and behavior heterogeneity witnessed in patients. There is a crucial need for a standardized diagnostic criteria to be prioritized.

In this study, we focused on fetal alcohol spectrum disorder, which is a leading cause of developmental disabilities worldwide. The prevalence of fetal alcohol spectrum disorder is difficult to estimate because of the variability and lack-there-of routine screening procedures [5]. A recent study that took place in the US suggested that there is a missed diagnosis rate of 80.1% as well as a misdiagnosis rate of 6.4%. This study suggests that the extremely high rates of missed diagnosis can have significant implications for intervention and therapeutic services [6].

Our study aims to tackle the high missed diagnosis rates seen across the United States. By training pre-existing large language models, we hope to see an increase in efficiency for

diagnosing individuals with fetal alcohol spectrum disorder. For diagnostic terminology, we utilize the International Classification of Diseases (ICD) codes, which is part of a globally used system for medical classification. Included within ICD codes are diseases, symptoms, and procedures. The purpose of ICD codes are to standardize and integrate classifications of medical-related terminologies globally and systematically [7]. For the purposes of our study, we leverage the utility of ICD-10 [8] codes to train our model to be equipped with a standardized way of approaching diagnostics.

We utilize, BioMistral an open-source large language model that is tailored specifically for the medical domain and is pre-trained on PubMed Central [9]. By leveraging BioMistral in our study, we further fine-tuned and trained the model to become more specialized in fetal alcohol spectrum disorder. Previous literature has shown clinical diagnostic testing on GPT-4, a generalized model that has not been pre-trained on clinical data, being extremely unstable and sensitive to prompts [10]. These limitations demonstrate that a generalized model does not have the current capabilities to deliver meaningful diagnoses.

Overall, we provide a pipeline that trains BioMistral on standardized fetal alcohol spectrum disorder and related developmental disorder International Classification of Disease codes to test among mock clinician notes. By varying hyperparameters in our test stages, we demonstrate that there are optimal improvements to the model when specific hyperparameters are selected. These results are encouraging and reveal the capabilities of fine-tuning a model to aid in standardized diagnostics in fetal alcohol spectrum disorder.

Our long term goal is to achieve accurate and efficient categorization of FASD symptoms into ICD codes. This will not only aid physicians in speedier diagnoses, but will also help to ensure that diagnosis and potentially interventions will be available to children with FASD, especially in cases where physician care is not readily accessible.

CHAPTER 2

Background & Motivation

Fetal Alcohol Spectrum Disorder (FASD) is a lifelong disorder that affects both physiological functions of the body and the psycho-physiological processes of brain [11]. The diagnosis of fetal alcohol syndrome (FAS) was formally suggested in a 1973 publication in *The Lancet* [12] [13]. A second article was published in *The Lancet* [12] in which David W. Smith and Kenneth L. Jones presented an anecdotal association between prenatal alcohol abuse and the effects of FAS [14]. Over the years, a group of pediatricians and psychiatrists helped to define the morphological defects and developmental delays that can affect children whose mother had consumed alcohol during the pregnancy [15].

FASD is an "umbrella term" for the spectrum of disorders due to prenatal alcohol exposure. Impacts from this disorder on social, behavioral, physical, and cognitive aspects of development vary greatly as there is no *singular* presentation [16]. The constellation of symptoms include pre- and/or postnatal growth retardation, central nervous system disorders that include developmental delay, intellectual impairment, and characteristic craniofacial abnormalities [15].

Beyond the mental and physical challenges in a person on the FASD spectrum, there are social complexities to individuals living with FASD. Individuals living with FASD often benefit from early diagnosis, the support from stable caregivers, and the provision of educational, physical, and mental systems put in place in order to successfully conduct personal and

professional activities [16]. FASD is a socially-rooted disability, wherein it is theoretically preventable. However this is a simplistic understanding of the disorder as there are many factors that may lead to the consumption of alcohol by pregnant women [17]. For example traumatic life events, stressful life events, intimate partner abuse, mental health challenges, and overall limited awareness of the harms associated with drinking while pregnant [17]. Additionally, FASD may be viewed as an inter-generational disability as a person having FASD may themselves experience an increased likelihood of substance abuse and other risky behavior [17]. These social determinants are an important consideration when evaluating the "preventability" of FASD. The occurrence of FASD can be seen in all socioeconomic and ethnic groups. The manifestations of FASD are complex and affect the family dynamic because individuals with FASD typically require long-term healthcare and social and vocational support [18]. The hope is that by increasing the knowledge of the experiences of people living with FASD it can reduce the stigmatization that comes with this neurodevelopmental disorder and increase accessibility and access to them receiving the proper care. Overall, the emphasis should be placed on creating accessibility for individuals with FASD or a family history of FASD to address both a missed diagnosis and misdiagnosis of patients.

CHAPTER 3

Related Work in Language Processing

Artificial intelligence (AI) has the capability to improve access and standard of care to patients all around the world. This application both in and outside of hospitals has provided solutions to predict and subsequently prevent harm. AI has been shown to provide decision support in identifying patients of high risk and offer preventative solutions [19].

Before we discuss, latest advances in AI for textual and language processing, we provide a brief history of language processing starting with linguistic modeling. Natural Language Processing (NLP) is a field of artificial intelligence that focuses on the interaction between computers and human language. In 1950, Alan Turing introduced the Turing Test, which asked the question "Can machines think?" [20]. He then formulated the principles of a three-person game, the Imitation Game, where an interrogator asks questions to a man and woman to determine the sex of the player in another room. Turing then reformulated the question to ask if imaginary computers that would do well at the Imitation Game, in which he believed would be true [20]. This paper by Turing set up the stage for exploring computer intelligence as it relates to human intelligence.

These developments were also soon applied to language processing. Early simplistic approaches for NLP included word-for-word Russian-to-English translations as demonstrated by the Georgetown-IBM system in 1954 [21]. Knowledge-based approaches to natural language processing began dominating the field in the 1960s until the 1970s. Noam Chomsky's

1956 theoretical analysis of formal English grammar provided an approximation of "n-order statistical approximations" [22]. This directly influenced the creation of Backus-Naur Form (BNF), which is used to specify "context-free-grammar". The usage of BNF, allows for one to define the sequences of symbols that make up a syntactically valid program in a particular programming language [23]. A language's BNF specification make up the set of rules that are used to correctly describe the structure of the code in the language, such as the rules and how each of the elements are combined to form syntactically correct programs [23]. In the 1970s, lexical analyzers (also known as lexers) became prominent. Lexical analysis is deployed when two tasks occur repeatedly to divide the structured input from programs into meaningful units and and discovers the relationship among units [24]. This division resulting into units also known as tokens, is a process of lexical analysis or lexing. These developments led to the creation of the Lex computer program, which can take in a set of descriptions of possible tokens and produce a C routine which are called a lexical analyzer, lexer, or scanner, that can identify tokens [24]. These lexer generators simplify the programming-language implementation by taking in Backus-Naur Form as input and generate code and lookup tables [25]. However, the rise of the symbolic, knowledge-based rules failed to extract the meaning, the exact semantics, from text. The specific relationship between textual units was not successfully captured by this method. Thus, there was a need to shift towards statistical processing in NLP.

In the 1980s, the reorientation of Natural Language Processing resulted in the beginning of statistical NLP methods. The fundamental acknowledgement of the complexity of language brought about novel implementations of NLPs. The simplest kind of language model is the "n-gram" language model. An n-gram takes in sequence of n words and and estimates the probability of a word based on its preceding context [26]. Building off of "n-gram" model is the Hidden Markov Models (HMM), which is a statistical technique that involves temporal sequence prediction, recognition, or processing [27]. HMM's ingest an entire sequence of past events into account in prediction [27]. In these models, the system can exist in multiple

hidden states, which cannot be observed directly. Instead, the hidden states are inferred from the observable outputs generated by the system. These statistical approaches give good results by learning from the pre-existing data [25].

As NLP has advanced and been brought into the clinical diagnostic space, there have been several challenges. Generally the problem of clinical text inference is split into low-level and high-level NLP tasks. Low-level tasks include sentence boundary detection with abbreviation and titles, identifying individual tokens within a sentence since often times in biomedical texts there are hyphens and forward slashes, morphological decomposition of compound words, shallow parsing, and problem-specific segmentation [25]. NLP's fall short of high-level tasks in spelling/grammatical errors such as correct words being flagged as errors and named entity recognition (NER) in failure to identify specific words and entities and categorizing them [25].

Recently, with the abundance of text data, the problem of textual inference has transitioned from rule and model-based techniques to large-data intensive large language models (LLMs) [28] [29] [30] [31]. However, there are significant limitations to the personalized diagnosis from the direct use of large language models such as GPT-4. ChatGPT, which is built on GPT-4 can offer solutions in regards to patient enquires, note-taking, decision making, and research support. However, it falls short of originality, privacy, correctness, bias, and legality [32]. The ChatGPT tool is shown to provide enhanced productivity and expedite as a clinical assistant but the information provided must continually be vetted by humans before it can provide accurate and reliable information [32]. Beyond generalized tools like ChatGPT, there are other models that are trained in the generalized biomedical domain. Pre-existing models such as BioMistral, PubMedBERT, SciBERT, BioMegatron, and ClinicalBERT all appear to be fine-tuned to a certain degree on biomedical information and in some cases rare disease models [9] [33] [34] [35] [36].

Web platforms such as DxGPT exist to assist healthcare professionals in the diagnostic pro-

cess for rare diseases [37]. DxGPT is tested with both real-world data from RAMEDIS and Peking Union Medical College Hospital (PUMCH) and synthetic datasets from Chat-GPT prompts. It is found that current LLMs can often effectively leverage symptom descriptions found in synthetic prompts and generate accurate diagnostic suggestions for diseases [37]. When assessing the performance of LLMs on real-world datasets, the accuracy notably dropped [37]. Both closed and open models were used to test between synthetic and real-world datasets, with closed models performing more accurately on a scale of the rate of which the top diagnostic suggestion matched the actual diagnosis [37]. The fundamental understanding and takeaway is that the lack of direct clinical validation means that the utility of AI models in the real-world diagnostic space remains untested.

CHAPTER 4

Methods

4.1 BioMistral

BioMistral [9] is an open-source Mistral-based model [38] that is tailored for the biomedical domain. It is a large language model that has been pre-trained on PubMed Central [9]. The choice to select PMC Open Access Subset as the training set is due to its comprehensive and freely accessible medical research papers [9]. The model architecture is inherited from the standard transformer architecture [39] from Mistral. These include features such as Grouped-Query Attention, Sliding Window Attention, and Rolling Buffer Cache [9] [40] [41]. Specific optimization parameters were selected as well as efforts to improve pre-training efficiency. Quantization techniques were also implemented that would enable the execution of these LLMs on smaller devices due to the minimization of the memory requirements [9]. By pre-training Mistral (7.3 Billion parameters) on high quality PuBMed data and applying efficient quantization and merged model variants, BioMistral demonstrates its efficiency and strength in evaluating medical benchmarks.

4.2 Model Selection

For the purposes of this study, we selected BioMistral, a pre-trained, open-source model in the biomedical domain. This selection is based off of the superior performance BioMistral has

in comparison to other large language model's that are trained in the medical domain. When comparing BioMistral's performance on biomedical baseline task performances, it shows a significant improvement over MedAlpaca 7B, MediTron-7B, and PMC-LLaMA 7B [9] [42] [43] [44]. BioMistral takes an unique approach by incorporating quantized and merged model variants. Quantization is a technique that is used to reduce the size of models and make them more efficient by lowering the precision of weights [9]. Merging techniques are also used to combine different models to enhance the overall performance by isolating strengths of individual models [9]. We opted out of not training an entire model from scratch because BioMistral had been already pre-trained on a large corpus of data with promising results [45] [46].

4.3 Fine-Tuning the Model

In order to train BioMistral on our ICD-10 [8] codes dataset, we first needed to fine-tune the model. Several methods exist to fine-tune models and tailor to fit specific needs. Full-parameter fine-tuning is a commonly chosen method that involves updating all of the parameters on a pre-trained model on a new dataset [47]. It can achieve high performance in a given task however it is extremely computationally expensive. Feature-based fine-tuning is an approach that transforms the original features to create a new feature representation [48]. It is less computationally intensive, however, it also fails to capture full complexity of the target task. Many applications of natural language processing will rely on the adaptation of one large-scale, pre-trained language model to multiple downstream applications [1]. The downside of this common approach is that the new models contains the same amount of parameters as the original model and often become critical deployment challenges [49]. For example, GPT-3 required 175 billion trainable parameters [50].

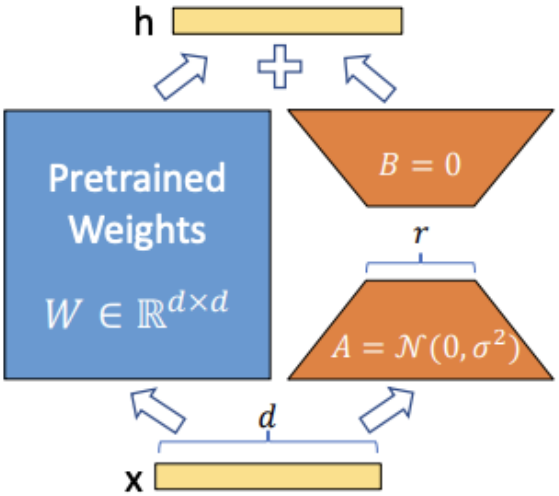

Figure 1: LoRA Architecture: Reparametrization diagram displays only A and B being Trained instead of all Pretrained Weights. Image taken from [1].

We forego the above methods of full-parameter fine-tuning and feature-based fine-tuning and instead select the Low-Rank Adaptation (LoRA) technique to fine-tune our generalized BioMistral LLM. LoRA is a method that keeps pre-trained model weights fixed and adds trainable rank decomposition matrices to each layer of the Transformer architecture. This approach significantly decreases the number of trainable parameters needed for downstream tasks [1]. By adapting only some parameters for new tasks, only a small number of task-specific parameters is needed in addition to the pre-trained model for each task [1]. LoRA is also selected over other pre-existing fine-tuning methods due to it's efficiency, scalability, ability to avoid over-fitting, and flexibility. This idea is illustrated in Figure 1, in which we can see that efficiency is achieved by only training A and B layers. To execute the fine-tuning and training, we imported specific modules as shown in Figure 2.

```
from transformers import TrainingArguments, AutoTokenizer, TrainingArguments, MistralForCausalLM
from trl import SFTTrainer
from peft import LoraConfig

from datasets import load_dataset
```

Figure 2: Modules Imported to Fine-Tune and Train Model.

Following the importing of these modules, we also defined the training arguments as shown in Figure 3.

```
training_args = TrainingArguments(
    output_dir="output",
    per_device_train_batch_size=8,
    gradient_accumulation_steps=8,
    gradient_checkpointing=True,
    learning_rate= 2.0e-6,
    logging_steps=5,
    num_train_epochs=3,
    max_steps=-1,
    save_steps=1000,
    save_total_limit=10,
    bf16=True,
    lr_scheduler_type="cosine",
    warmup_ratio=0.1,
    evaluation_strategy="epoch",
    logging_first_step=True,
    neftune_noise_alpha=5,
)
```

Figure 3: Training Arguments used for the Model.

In Figure 3, the **per_device_train_batch_size** = 8 defines the number of training samples processed in parallel on each device per training step. This batch size of 8 indicates that each device will process 8 samples before it will update the model weights. The next argument **gradient_accumulation_steps** = 8 controls how many steps are accumulated before performing a weight update. This will essentially increase the batch size without needing

more GPU memory. **gradient_checkpointing = True** is also set to enable gradient checkpointing which allows for saving memory by not storing intermediate activation during the forward pass but recomputes them instead during the backward pass. **learning_rate =** 2.0e−6 is a hyperparameter that is adjusted throughout this process to obtain the most optimal model to train the model. We will explore learning rates ranging from 2.0e−2 to 2.0e−7 throughout this study. **logging_steps =** 5 will log metrics like loss and learning rate every 5 steps, allowing us to monitor the training process. The parameter **num_train_epochs =** 3 indicates the number of times the entire training dataset will be passed through the model. This is another hyperparameter that will be adjusted throughout the process ranging from 1 to 3. The parameter **max_steps =** −1 will set the maximum number of training steps and a value of −1 means there is no limit and training will run for the 'num_train_epochs' steps. **save_steps =** 1000 defines that the model will save every 1000 steps to have intermediate checkpoints. The parameter **save_total_limit =** 10 will limit the total number of saved checkpoints to 10, ensuring that there is no excessive storage use and it only keeps the most recent checkpoints. **bf16 = True** will enable the use of bfloat16 precision, which can speed up the training and reduce memory usage while maintaining numerical stability. The parameter **lr_scheduler_type = "cosine"** specifies the type of learning rate schedule to use. In this case, a cosine scheduler gradually decreases the learning rate by following a cosine function and it can help achieve better convergence by reducing the learning rate as training will progress. The parameter **warmup_ratio = 0.1** sets the ratio of warm up steps relative to the total number of steps. This warm up value will help increase the learning rate at the beginning of training to avoid large updates that could derail and destabilize the rest of the training process. The parameter **evaluation_strategy = "epoch"** ensures the model evaluates at the end of each epoch. The parameter **logging_first_step = True** denotes that logging metrics will be set up for the first training step, and so that it is set up properly from the start. **neftune_noise_alpha =** 5 refers to the hyperparamter related to the NEFTune (Neural Evolution of Sparse Networks) [51] methods where noise_alpha con-

trols the amount of noise during optimization. Collectively these arguments control various aspects of the training process, including resource management, efficiency, monitoring, and checkpointing.

In addition to setting these argument we also implemented fine-tuning in LoRA as shown in Figure 4.

```
peft_config = LoraConfig(
    r=16,
    lora_alpha=32,
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM",
)
```

Figure 4: LoRA Configuration Settings.

Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA is designed to fine-tune large language models with fewer parameters, making this process more streamlined and efficient. $\mathbf{r} = 16$ is a parameters that specifies the rank of the low-rank adaption matrices used in LoRA. In LoRA the weight matrices in the neural networks are factorized into two smaller matrices, where **r** determines the dimensionality of these smaller matrices [1]. A high **r** would mean more parameters are added, allowing the model to capture more complexity during fine-tuning, a setting of 16 is a choice that captures both expressiveness and efficiency. $\mathbf{lora\_alpha} = 32$ is a parameter that is a scaling factor that is applied to the output of the low-rank matrices before they are added back to the original weights. This scaling will help control the influence of the low-rank adaptation on the original model [1]. A value of 32 means that this adaptation will have moderate influence and will allow the fine-tuned model to adjust its behavior without drastically altering the original pre-trained weights.

lora_dropout = 0.05 is a regularization technique that will be used to prevent overfitting by randomly setting a fraction of the units to zero during training. A rate of 0.05 means that 5% of the units will be dropped out in the LoRA adaptation layer and will ensure that this fine-tuning process will not rely too heavily on any single feature [1]. **bias = none** means that no biases will be modified during the LoRA adaptation and this is done to simplfiy the fine-tuning process because biases generally have less impact on the model performance than weights [1]. The parameter **task_type = CAUSAL_LM** parameter specifies the type of task the model is being fine-tuned for "causal language modeling", where the task is the model predicts the next token in a sequence given the previous tokens. This will generate coherent text sequences in a left-to-right fashion [1]. This configuration that is used to fine-tune BioMistral is set up in a way that does not require a large number of parameters to be updated, thus leading to a more efficient but effective model.

4.4 Training Data

The training set included a total of 103 fetal alcohol spectrum disorder or related developmental disorder International Classification of Disease, Tenth Revision (ICD-10) codes [8]. ICD is a standardized system that is used to code diseases and medical conditions. It is common practice for medical providers to provide ICD codes when diagnosing patients [7]. The exposure to a larger array of ICD codes related to FASD would allow BioMistral to get a robust sense of relevant ICD diagnostic codes.

Diagnostic to ICD Code
Newborn (suspected to be) affected by maternal use of alcohol (Excludes Fetal Alcohol Syndrome), P04.3
Fetal alcohol syndrome (dysmorphic), Q86.0
Mood disorder due to known physiological condition, unspecified, F06.30
Newborn (suspected to be) affected by maternal nutritional disorders, P00.4
Newborn (suspected to be) affected by maternal complication of pregnancy, unspecified, P01.9
Encephalopathy, other and unspecified (static), G96.8
Other specified disorders of central nervous system, G93.4
Disorder of central nervous system, unspecified, G96.9
Microphthalmos, Q11.2
Other general symptoms and signs (eg, dysmorphic features), R68.89
Underweight, R63.6
Feeding difficulties, R63.3
Failure to thrive (child), R62.51
Short stature (child), R62.52
Lack of expected normal physiological development in childhood, unspecified, R62.50
Delayed milestone in childhood, R62.0
Mild cognitive impairment, so stated, G31.84
Mild intellectual disabilities, F70
Moderate intellectual disabilities, F71
Severe intellectual disabilities, F72

Figure 5: Mapping between 20 diagnostic terms to ICD codes.

Figure 5 displays a sample of 20 ICD-10 codes that were used to as part of the overall training set.

4.5 Training the Model

Using the ICD training set (Appendix Figure 30, 31, we allow BioMistral to get a robust sense of ICD codes that are relevant to FASD. Then, With LoRA, we focused on varying two specific hyperparameters, learning rate and epoch. This is a global optimization problem where during each iteration a loss function is employed that measures the model's deviation from the ground truth and updates the parameter θ, to minimize the loss function Lθ [52]. A learning rate is dynamically changed in response to the estimated error. A learning rate that

is too large can cause the model to be overfitted and the loss function to get stuck at a local minimum or diverge. A learning rate that is too small may lead to a time-heavy convergence. It is imperative that an optimal learning rate is chosen to produce stable results with each model update and provides smooth convergence. An epoch is defined as the complete pass through all of the datasets in one cycle. By varying between epoch one, two, and three, we were able to successfully demonstrate how with a consistent epoch the model generalizes well to novel data.

4.6 Queries

The queries are inputs to the testing script. These queries are examples of mock clinical symptoms and or notes. We generated synthetic datasets based off of example MedQA Questions (Figure 6).

Table 1. Example MedQA questions.

Example Question 1

Shortly after undergoing a bipolar prosthesis for a displaced femoral neck fracture of the left hip acquired after a fall the day before, an 80-year-old woman suddenly develops dyspnea. The surgery under general anesthesia with sevoflurane was uneventful, lasting 98 min, during which the patient maintained oxygen saturation readings of 100% on 8 l of oxygen. She has a history of hypertension, osteoporosis, and osteoarthritis of her right knee. Her medications include ramipril, naproxen, ranitidine, and a multivitamin. She appears cyanotic, drowsy, and is oriented only to person. Her temperature is 38.6 °C (101.5 °F), pulse is 135/min, respirations are 36/min, and blood pressure is 155/95 mm Hg. Pulse oximetry on room air shows an oxygen saturation of 81%. There are several scattered petechiae on the anterior chest wall. Laboratory studies show a hemoglobin concentration of 10.5 g/dl, a leukocyte count of 9000/mm^3, a platelet count of 145,000/mm^3, and a creatine kinase of 190 U/l. An ECG shows sinus tachycardia. What is the most likely diagnosis?

Example Question 2

A 55-year-old man comes to the emergency department because of a dry cough and severe chest pain beginning that morning. Two months ago, he was diagnosed with inferior wall myocardial infarction and was treated with stent implantation of the right coronary artery. He has a history of hypertension and hypercholesterolemia. His medications include aspirin, clopidogrel, atorvastatin, and enalapril. His temperature is 38.5Â°C (101.3 °F), pulse is 92/min, respirations are 22/min, and blood pressure is 130/80 mm Hg. Cardiac examination shows a high-pitched scratching sound best heard while sitting upright and during expiration. The remainder of the examination shows no abnormalities. An ECG shows diffuse ST elevations. Serum studies show a troponin I of 0.005 ng/ml ($N < 0.01$). What is the most likely cause of this patient's symptoms?

Example questions used in all MEDQA prompts provided in Table 2.

Figure 6: Example of MedQA Questions. Reproduced from [2].

Using these example questions as inputs into ChatGPT with the prompt "Diagnostic reasoning prompt for fetal alcohol spectrum disorder, frame it like this", we compiled a list of synthetic questions regarding fetal alcohol spectrum disorder. Figure 7 shows examples of

several queries that were used to test the model that have been generated from ChatGPT with the above input (See Appendix Figures 24, 25, 26 for all queries). The last prompt shown is a "human-made" prompt based off of common symptoms associated with FASD. It was important for these queries to include both synthetic, wordier prompts as well as more succinct prompts to compare how the model reacts to both styles of prompts.

Example
A 7-year-old boy is brought in by his foster parents for evaluation due to persistent learning and behavioral difficulties at school. He has a history of hyperactivity, impulsivity, and trouble concentrating. The foster parents note that he struggles with memory, has difficulty following instructions, and often acts out in frustration. Physically, he is smaller in stature compared to his peers, and his facial features include a smooth philtrum, thin upper lip, and small palpebral fissures. The boy was adopted from a background with limited prenatal care, and there is a suspected history of maternal alcohol use during pregnancy. On examination, his head circumference is below the 10th percentile, and he exhibits poor coordination and fine motor skills.
A 9-year-old girl is brought to the clinic by her parents due to ongoing issues with attention and behavior at school. They report that she has trouble focusing on tasks, is easily distracted, and frequently gets into trouble for being disruptive. Her teachers have noted that she struggles with academic tasks and often seems confused by instructions. The parents mention that she has difficulty making friends and tends to be impulsive and hyperactive. The girl has a history of growth delays and exhibits facial features such as a smooth philtrum, a thin upper lip, and small palpebral fissures. They recently discovered that the biological mother consumed alcohol throughout her pregnancy.
The symptoms are distinctive facial features, deformities of joints, limbs, and fingers. The patient also has social and behavioral issues.

Figure 7: Example of Queries given to the model.

4.7 Testing the Model

In order to properly test the model, there was an instantiation for the model as shown in Figure 8.

```
{"role": "assistant", "content": "You are a medical professional diagnosing a patient.
You will be given patient symptoms and must categorize the symptoms into ICD codes.
Only list the ICD codes first. If you don't have enough information or don't know,
then say so and don't attempt to answer.\n"}
```

Figure 8: Example of the user prompt given to the model.

Allowing the model to receive context of it's role was the appropriate way to approach testing the model. The model takes in the appropriate trained model with a specific learning rate and epoch and tests it using PyTorch and Hugging Face's 'transformers' for loading models and tokenizers. The queries are processed and responses are then generated into a separate text file. In addition to qualitative results we also generate logits, the raw and unnormalized scores. Logits represent the model's confidence in predicting each possible token in the vocabulary in the given position in the sequence. Building off of this, we used the softmax function to softmax our logits to compute probabilities referring to Figure 4.1.

$$f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{4.1}$$

The softmax function is an activation function that is used in neural computing [53]. It is used to compute the probability distribution vector of real numbers and it produces an output of a range of values between 0 and 1 with the sum of these probabilities totaling to 1 [53]. Previous literature has suggested that by softmaxxing the logits, we receive probabilities that predict large language model correctness on multiple-choice Q&As [54]. While our study does not use multiple-choice Q&As this paper sufficiently justifies the usage of softmaxxing logits

to receive probabilities. Here, we explored the extrapolation of this method in softmaxxing logits from the model generated ICD code response to a fetal alcohol spectrum disorder diagnosis.

CHAPTER 5

Results

This chapter describes the experimental results from the retrained model on a synthetic dataset that was a mix of manually generated queries and those that were generated using a LLM such as ChatGPT. The initial results from fine-tuning the model using Low-Rank Adaptation of Large Language Models (LoRA) confirm the properties expected from LoRA tuning. We expect that the newly trained models will not deviate too much from the initial set of training [1].

5.1 Evaluation Loss Curves and Visualization

We started with a training set that included a total of 103 fetal alcohol spectrum disorder or related neurodevelopmental disorder International Classification of Disease (ICD) codes (Figure 5). It was important to include disorders whose symptoms were similar to that of FASD to properly train the model to avoid misclassification or misdiagnosis. We first examined the differences in loss curves from varying learning rates and epochs to determine which model, if any, would be over-fitted from this modification. Initially, the evaluation loss curves that were generated from LoRA fine-tuning provided an intuitive idea as to whether the learning rates promoted over-fitting or were optimally chosen.

We observed that there was an empirical threshold for the model, after which it went from

performing extremely well to extremely poorly, and particularly over-fitted. While this was an important discovery, it was an expected result as higher learning rates and greater number of epochs appear to be associated with more poorly performing and over-fitted models. With a learning rate of 2.0e−2 and an epoch of two and three, we can see a significant jump in the poor performance in comparison to an epoch of one (Figure 9). This large increase in the evaluation loss is attributed to a high learning rate and an epoch that is greater than one. In contrast, if we choose a much smaller learning rate of 2e-6, we can see that the evaluation loss remains constant throughout the increase of epochs (Figure 10).

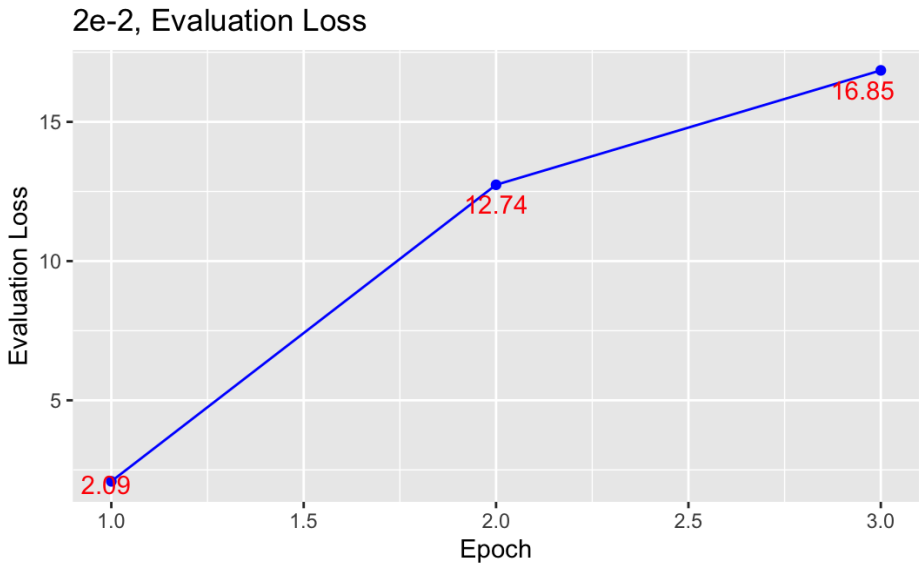

Figure 9: Demonstrating the evaluation loss for a learning rate of 2.0e−2 across epochs at 1, 2, and 3.

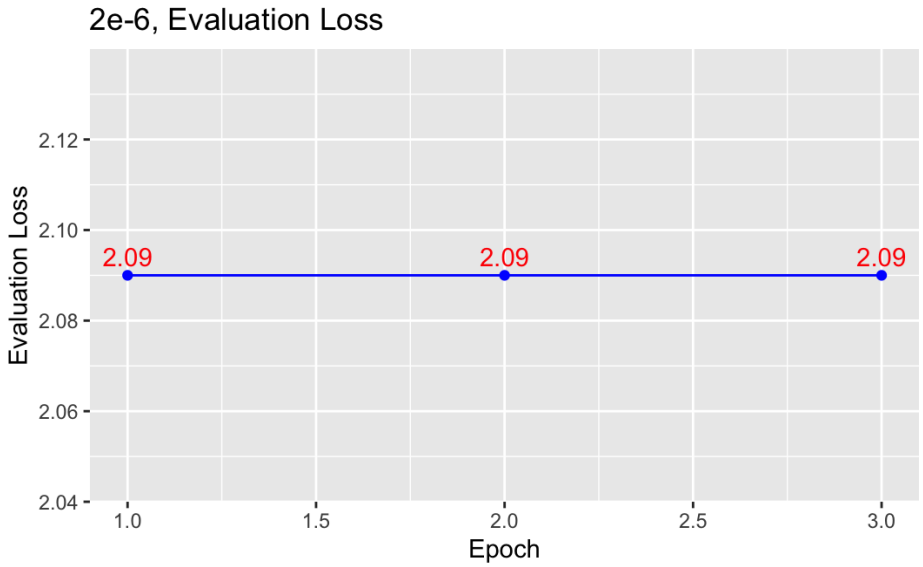

Figure 10: Demonstrating the evaluation loss at a learning rate of 2.0e−6 across epochs at 1, 2, and 3.

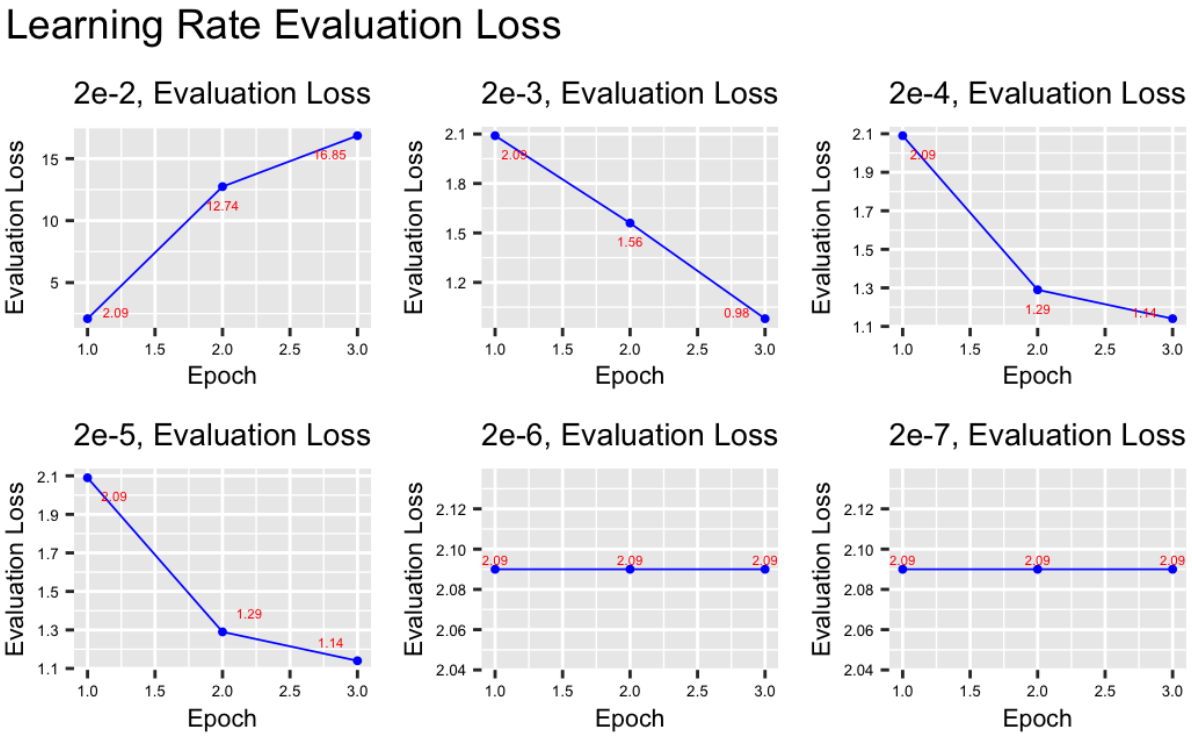

Figure 11: Demonstrating the evaluation loss associated with different learning rates across epochs at 1, 2, and 3.

We also plotted the evaluation losses for different learning rates across epochs used to fine-tune the model as shown in Figure 11. We observed that as the learning rate begins to decrease, we see a stabilization in the evaluation loss obtained from increasing the epochs. These results demonstrate that the performance becomes more stable and the model is more appropriately fitted for lower learning rates. This implies that a higher learning rate is associated with a greater evaluation loss and that as learning rates decrease we see stabilization in the evaluation losses.

By varying across epochs one, two, and three, we were able to empirically demonstrate abrupt changes in the evaluation losses. Particularly, we see that the model performs well with a consistent evaluation loss as learning rate decreases.

5.2 Probability Distribution Models

We also visualized the probability distribution of the models trained on each various learning rates and epochs. Visualization of probability distributions helped to identify if the model underwent appropriate learning from the training data based on the distribution. To obtain the probabilities, we first calculated the logits (log-odds) from the raw and un-normalized scores. We applied a softmax function to the logits to obtain the probabilities. The visualization of these probabilities clarified the likelihood of the model predicting the next token and the level of confidence of the model. We can see that with the trained model of 2.0e−2 and epoch two (Figure 12) and epoch three have a nearly uniform logged distribution. This uniformity implies that model has less certainty about which token should come next and that it may see tokens as having equal probabilities of being the next correct token. The model is repeating itself over and over again. The patterned distribution ultimately suggested that the model was overfitted. Beyond certainty, a uniform probability distribution model also revealed higher entropy in the predictions for what is an indication of a lack of preference for a particular token. Further, in the logged, softmax, logits (probabilities) of 2.0e−2, 2.0e−3, 2.0e−4, 2.0e−5, 2.0e−6, and 2.0e−7, we saw a less uniform distribution which implies more certainty in which token should come next. The model assigned higher probabilities to a smaller number of tokens. This non-pattern suggested that the model was preferential towards certain tokens (Figure 13).

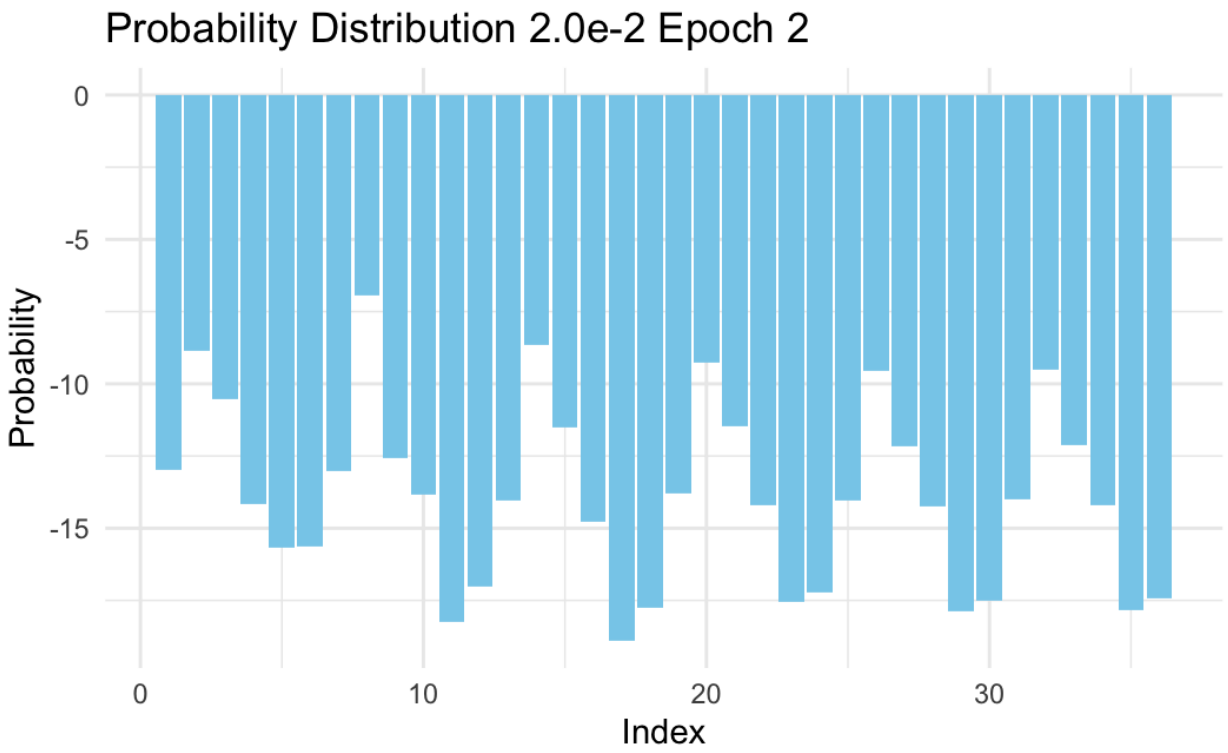

Figure 12: Learning Rate 2.0e−2 Epoch 2 Probability Distribution.

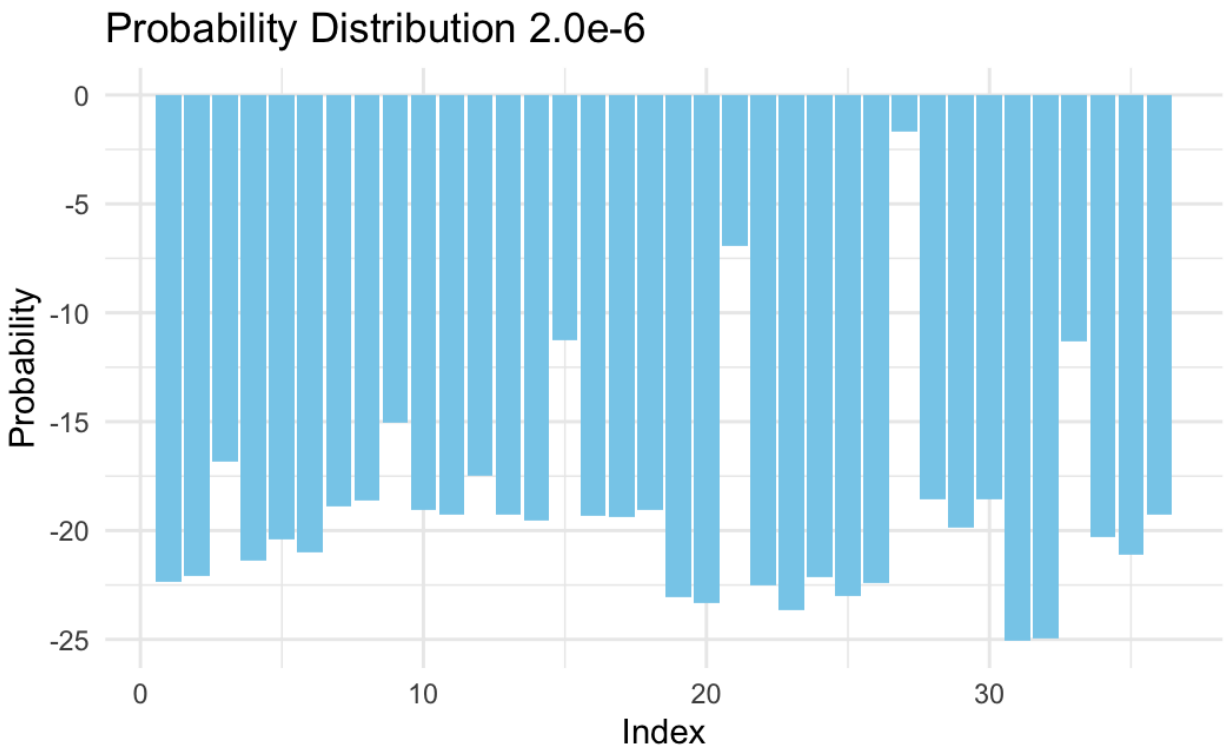

Figure 13: Learning Rate 2.0e−6 Epoch 1 Probability Distribution.

By visualizing all the probability distributions on a single plot, we observed that there were very minimal differences in the probability distribution among all the learning rates except for 2.0e−2 with an epoch of two in Figure 14. These empirical results demonstrated that models with lower learning rates have learned and predicted the next token more appropriately than a model of a higher learning rate across larger epochs.

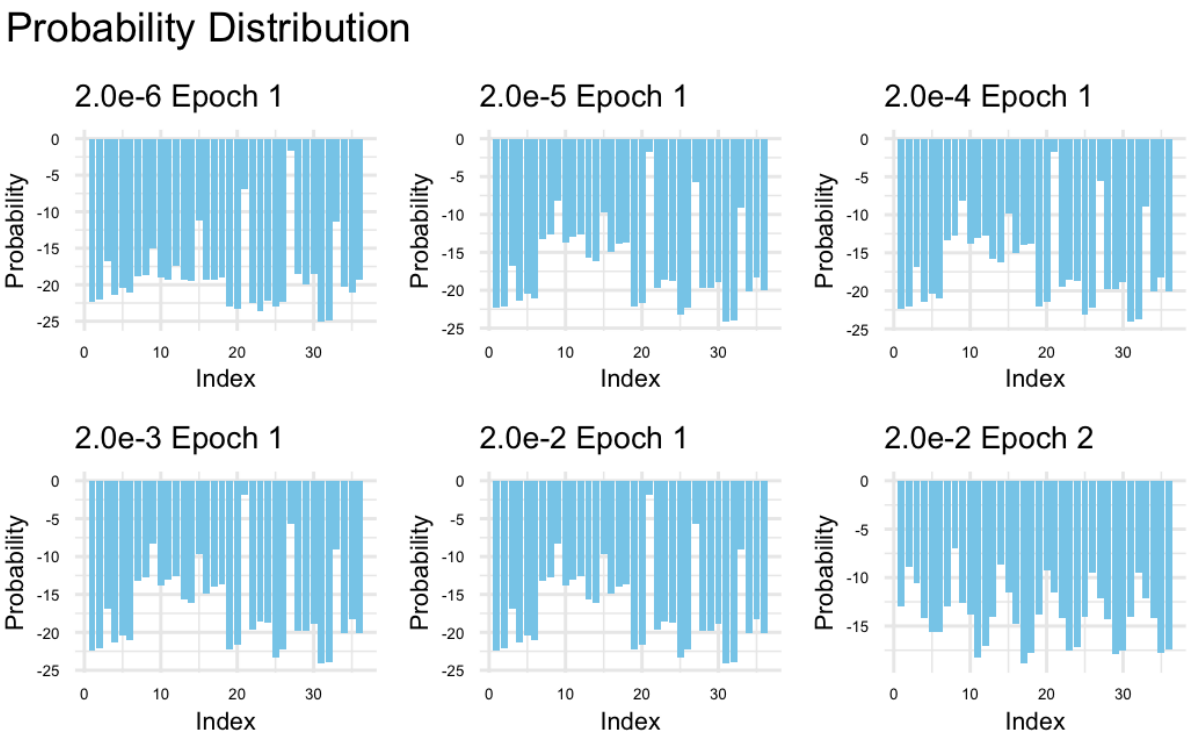

Figure 14: Probability Distribution Among All Learning Rates.

The softmax function was applied to the raw logit values to yield the probability distributions in Figure 14. We observed a cyclical nature of a large learning rate with an epoch greater than one. This emphasizes the importance of an appropriate learning rate and epoch selection when fine-tuning a large language model.

5.3 Kullback-Leibler Divergence Matrix

Kullback-Leibler (KL) divergence is a statistical measure that quantifies in bits how close a probability distribution is to a model distribution [55]. Here, we also utilized the KL divergence to demonstrate the perturbations between models.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)} \tag{5.1}$$

KL divergence is also referred to as relative entropy and plays a key role in machine learning, information theory, and statistics [56]. KL divergence is essentially the expectation of the log difference between the probability of the data from the original distribution and the approximated distribution. In our case, we calculated the KL divergence between the different epoch selections in all of the learning rates we chose. If the mutual information is zero, it means that the variables are statistically independent. [55]. From comparing the KL divergence graphs between a learning rate of 2.0e−2 to 2.0e−6, we saw large differences in magnitude of the KL values (Figures 15, 16).

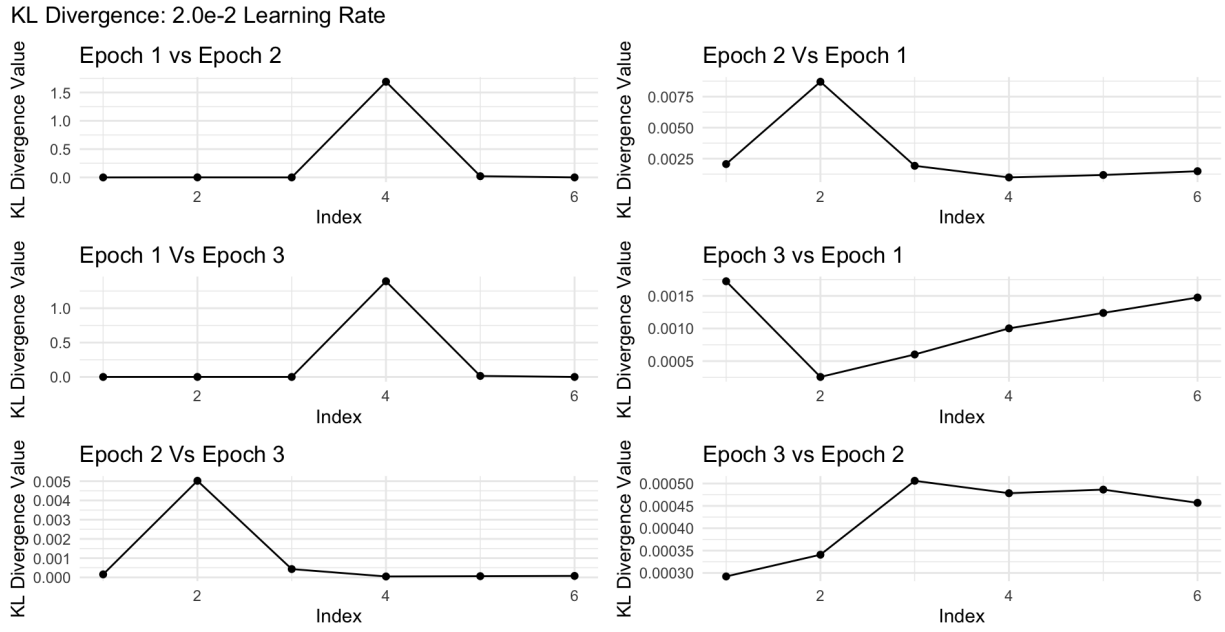

Figure 15: KL Divergence 2.0e−2. Comparing between all combinations of epoch's to determine statistical independence.

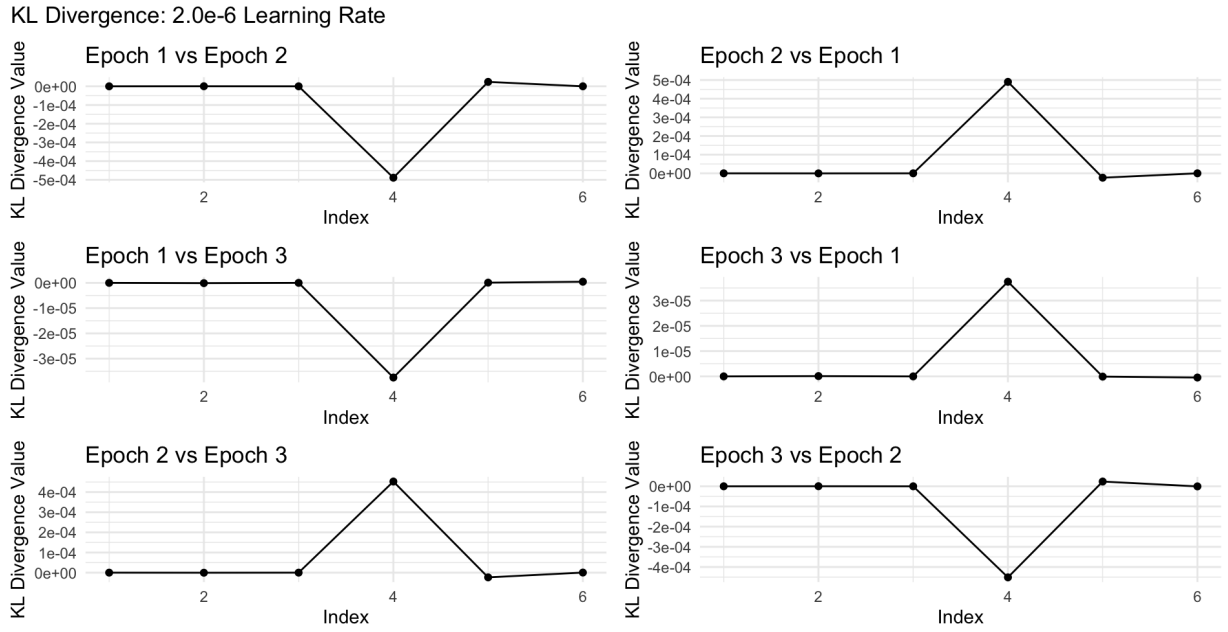

Figure 16: KL Divergence 2.0e−6. Comparing between all combinations of epoch's to determine statistical significance.

We observed slight perturbations in the distributions. We also observed that for large perturbations at a higher learning rate of 2.0e−2, the results were less meaningful, and higher learning rates were associated with larger perturbations. Further experimental results that showed larger perturbations with increased learning rates for different epochs are included in the Appendix.

5.4 Qualitative Analysis

In addition to evaluating the model performance using KL divergence, we also performed qualitative evaluations. The textual results from the various large language models that have been trained on varying learning rates revealed interesting characteristics on how the models performed with different prompts. The same models that showed less meaningful empirical

results also showed less promising qualitative results. The model that was trained with 2.0e−2 learning rate with epoch two and epoch three produced indistinguishable, meaningless repetitive text (Figure 17). The higher learning rate in combination with more than one epochs cause the optimization process to become unstable, thus making the model less likely to converge to an optimal solution. This process is what produces the repeated string "INSTINST", indicating that the model is stuck in a loop (Figure 17). These qualitative results also reveal the importance of an optimal learning rate. A learning rate that is too large will result in a model that cannot converge and thus will produce meaningless results.

Learning Rate	Epoch	Result
2.00E-02	2	INSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN STINSTINSTINSTINSTINSTINSTINSTINSTINSTINS TINSTINSTINSTINSTINSTINSTINSTINSTINSTINSTI NSTINSTINSTINSTINSTINSTINSTINSTINSTINSTIN ST

Figure 17: Learning Rate 2.0e−2 with Epoch 2. Qualitative results from query number one Appendix Figure 24.

After lowering the learning rate to 2.0e−6 and selecting an appropriate number of epochs, we qualitatively observed more appropriate ICD codes generated by the model (Figure 18).

Learning Rate	Epoch	Result
2.00E-06	1	1. F70.0: Alcohol-related birth defects of nervous system 2. F70.1: Alcohol-related birth defects of circulatory system 3. F70.2: Alcohol-related birth defects of digestive system 4. F70.3: Alcohol-related birth defects of genitourinary system 5. F70.4: Alcohol-related birth defects of musculoskeletal system 6. F70.5: Alcohol-related birth defects of respiratory system 7. F70.6: Alcohol-related birth defects of skin, hair, and nails 8. F70.8: Alcohol-related birth defects, unspecified 9. F70.9: Alcohol-related birth defects, not elsewhere classified 10. F71.0: Alcohol-related birth defects of nervous system, unspecified 11. F71.1: Alcohol-related birth defects of circulatory system, unspecified 12. F71.2: Alcohol-related birth defects of digestive system, unspecified 13. F71.3: Alcohol-related birth defects of genitourinary system, unspecified 14. F71.4: Alcohol-related birth defects of musculoskeletal system, unspecified 15. F71.5: Alcohol-related birth defects of respiratory system, unspecified 16. F71.6: Alcohol-related birth defects of skin, hair, and nails, unspecified

Figure 18: Learning Rate 2.0e−6 with Epoch 1. Qualitative result from query number one Appendix Figure 24.

5.5 Confusion Matrix Analysis

We also demonstrate qualitative results through the generation of confusion matrices that evaluate model performance. A confusion matrix (Figure 19) is used to evaluate the quality of a classifier via true positives and false negatives and the cross-classification performance between the predicted and actual decisions in a *tabular form* [3].

		True value	
		P	N
Predicted value	$\hat{P}$	True Positive	False Positive
	$\hat{N}$	False Negative	True Negative

Figure 19: 2-Class Confusion Matrix [3].

A confusion matrix does not assume distributional parameters and will instead directly use observed counts of predictions and actual outcomes to summarize how well a model performs. It will only include knowledge up to a certain level of granularity, the level of detail or precision in the data used to make predictions, affirming that the counts of True Positives, False Positives, True Negatives, and False Negatives is based on the available set of data [3].

When we tested each individual query, we noticed that the model's response of the number of ICD code outputs varied depending on the query. This is shown in Figure 23, where the summation of the columns for each row is the number of ICD-10 codes returned by the model.

As an example, we show results from a single query by displaying a confusion matrix result from a learning rate of 2.0e−6 in Figure 20.

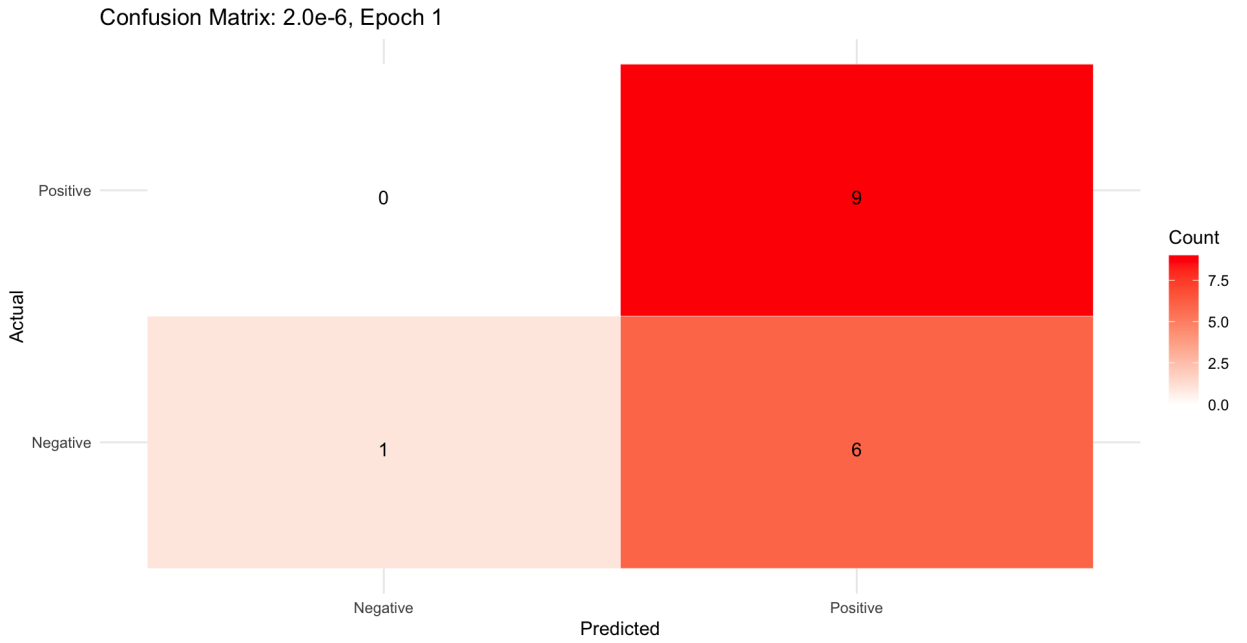

Figure 20: Confusion Matrix with True and Predicted Labels for 2.0e−6, Epoch 1. The results that match positive and positive, and negative and negative are considered to be a match between the actual and predicted model.

The confusion matrix in Figure 20 shows that out of the 16 ICD-10 code responses, 10 match the true predicted labels. There are 6 false positives. This is an interesting result because it demonstrates that the trained model will tend towards over-fitting. It attempts to predict an ICD-10 diagnostic code even if it isn't entirely fighting. However, we do not see any results in the false negative realm.

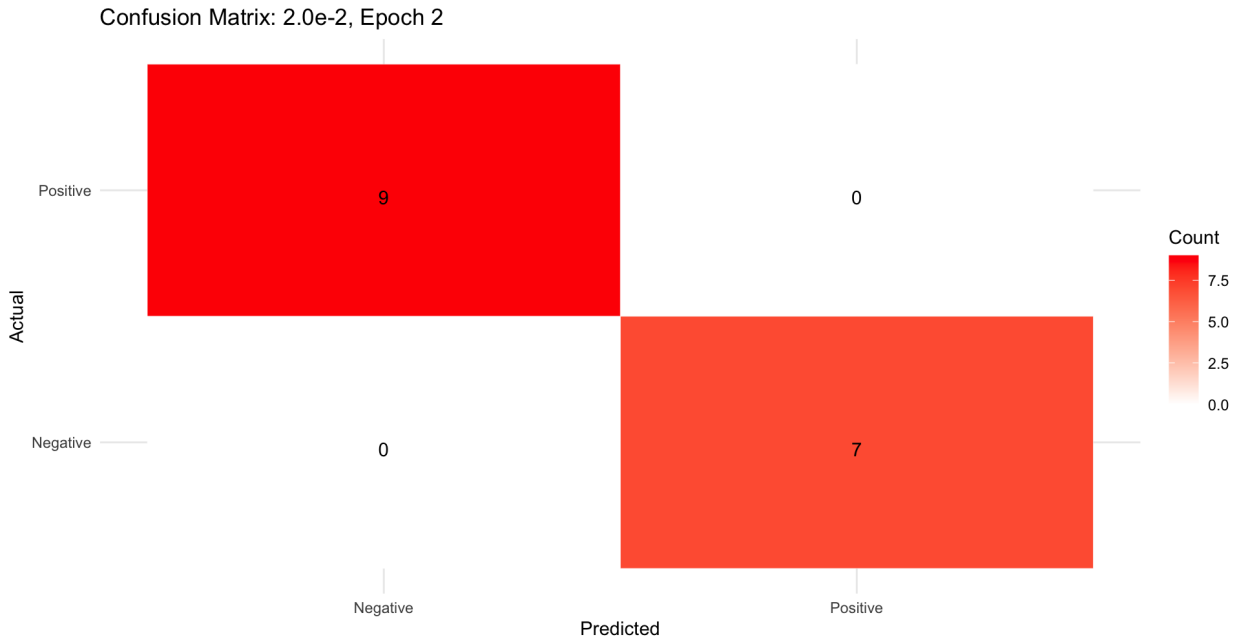

Figure 21: Confusion Matrix with True and Predicted Labels for 2.0e−2, Epoch 2. The results that match positive and positive, and negative and negative are considered to be a match between the actual and predicted model.

For a higher learning rate of 2.0e−2, the confusion matrix in Figure 21 does not show any correct or predicted labels. This is expected with such a higher learning rate that causes model overfitting and results in the failure to produce coherent diagnostic ICD-10 codes.

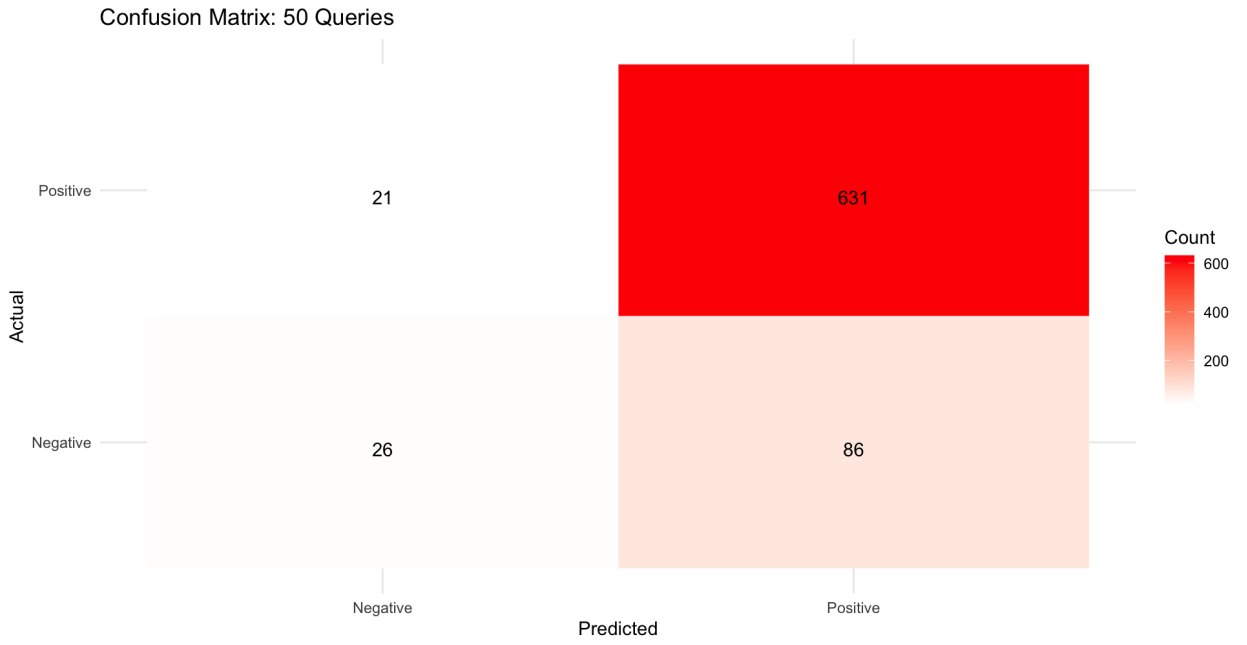

Figure 22: Confusion Matrix for All 50 Queries. The results that match positive and positive, and negative and negative are considered to be a match between the actual and predicted model.

Finally, we evaluated the model for 50 generated queries. From the confusion matrix shown in Figure 22, we observed that overall the model performed quite well. We suggest that the false-positive results may occur due slight over-fitting of the model. However, a key takeaway of this large-set confusion matrix is that this model often predicts the correct ICD code with room for over-fitting (Figure 23).

Queries	True Positive	False Positive	False Negative	True Negative
Query 1	7	4		1
Query 2		6		
Query 3	10		2	
Query 4	17			
Query 5	18		1	
Query 6	14	1		1
Query 7		7		1
Query 8	7	4		1
Query 9		6		
Query 10	12			
Query 11	10		8	
Query 12	18			1
Query 13	17			
Query 14	17			
Query 15	19			
Query 16	12			
Query 17	13			1
Query 18	16			
Query 19	11			1
Query 20	12			
Query 21	11			
Query 22	16			1
Query 23	19			1
Query 24	17			1
Query 25	10		10	1
Query 26	7	5		
Query 27	16			
Query 28	11			1
Query 29	15			1
Query 30	5	10		
Query 31	17			1
Query 32	16			1
Query 33	12	10		1
Query 34	16			1
Query 35	10			
Query 36	12			
Query 37	16			1
Query 38	19			
Query 39		19		1
Query 40	18			1
Query 41	13			1
Query 42	16			
Query 43	11			1
Query 44	17			1
Query 45	15	4		1
Query 46	11			1
Query 47	16	2		
Query 48	8	8		1
Query 49	11			
Query 50	20			
Sum	631	86	21	26

Figure 23: Confusion Matrix Table.

CHAPTER 6

Discussion

In this study we investigated the use of large language models for prediction of ICD-10 codes for fetal alcohol spectrum disorder. We leveraged online resources of ICD-10 codes to create a robust dataset to train and fine-tune BioMistral [9]. The parent model of BioMistral is Mistral 7B [38], a pre-existing large language model that has been tailored for the medical domain. By training BioMistral with our curated set of ICD-10 codes and testing on both synthetic clinician notes and generalized symptoms, we are able to demonstrate the importance of choosing an optimal learning rate and epoch selection when using LoRA to fine-tune a model. Our study goal was to use both statistical frameworks and qualitative analysis to emphasize the important of the hyperparameter selection behind using LoRA to fine-tune a large language model in the neuroscience diagnostic space. A model that has been optimally trained will have a better success at identifying ICD codes and that has been demonstrated both quantitatively and qualitatively.

The current study is limited by our training set, relying on already fixed ICD-10 codes. They are binary, i.e. either a patient displays symptoms or they do not. Fetal alcohol spectrum disorder's symptoms and behaviors are incredibly heterogeneous and exhibit greater range, so ICD codes may fail to capture the significance of each symptom. ICD codes are also not weighted, in the case of individuals with FASD, there are certain markers and symptoms that are incredibly indicative of FASD. Our test set is also composed of synthetic diagnostics and human-made generalizable symptoms. Previous literature has shown that open-sourced

large language models tend to do a better job at deciphering diagnostics from a synthetic dataset in comparison to a real-world dataset [16]. Because our training set was limited by the characteristics of ICD codes, we note that our synthetic datasets may not have fully exploited the large language model's capabilities.

In the future, access to real clinician notes for patients with fetal alcohol spectrum disorder will greatly improve the testing of the model. The limitations of synthetic and human-made datasets need to be considered to accurately and efficiently determine whether or not the fine-tuned BiomMistral large language model is capable of robust diagnostics. This work presented preliminary results showing the importance of the hyperparameters related to LoRA fine-tuning of these models as it relates to synthetic data.

Overall, we suggest for more large language models to be trained on ICD codes and physician notes related to the disease of interest to further push for a standardized and efficient diagnosing process for physicians. The high miss-diagnosed rate of fetal alcohol spectrum disorder can be tackled through actively leveraging the use of large language models to aid in diagnostics as well as being used as an AI-powered healthcare platform for parents with children suspected of having fetal alcohol spectrum disorder.

Appendices

Queries
A 7-year-old boy is brought in by his foster parents for evaluation due to persistent learning and behavioral difficulties at school. He has a history of hyperactivity, impulsivity, and trouble concentrating. The foster parents note that he struggles with memory, has difficulty following instructions, and often acts out in frustration. Physically, he is smaller in stature compared to his peers, and his facial features include a smooth philtrum, thin upper lip, and small palpebral fissures. The boy was adopted from a background with limited prenatal care, and there is a suspected history of maternal alcohol use during pregnancy. On examination, his head circumference is below the 10th percentile, and he exhibits poor coordination and fine motor skills.
A 9-year-old girl is brought to the clinic by her parents due to ongoing issues with attention and behavior at school. They report that she has trouble focusing on tasks, is easily distracted, and frequently gets into trouble for being disruptive. Her teachers have noted that she struggles with academic tasks and often seems confused by instructions. The parents mention that she has difficulty making friends and tends to be impulsive and hyperactive. The girl has a history of growth delays and exhibits facial features such as a smooth philtrum, a thin upper lip, and small palpebral fissures. They recently discovered that the biological mother consumed alcohol throughout her pregnancy.
A 10-year-old boy presents with behavioral issues and poor academic performance. His teacher reports difficulty with attention and impulsivity. His mother had a history of alcohol use during pregnancy. On examination, he has a smooth philtrum, and his height and weight are in the 3rd percentile for his age.
A 7-year-old girl is referred for evaluation of speech and language delay. Her birth history reveals exposure to alcohol during pregnancy. She has a flat midface, epicanthal folds, and microcephaly. Her IQ is significantly below average.
A 3-year-old boy with a history of prenatal alcohol exposure is brought in for evaluation of poor growth and developmental delays. His facial features include a smooth philtrum, thin upper lip, and small eye openings. He also has behavioral problems, including irritability and difficulty with transitions.
A 9-year-old boy is evaluated for aggressive behavior and poor school performance. He has a history of prenatal alcohol exposure. Physical examination reveals a smooth philtrum, micrognathia, and growth retardation. Neurocognitive testing shows deficits in memory and executive function.
A 5-year-old girl presents with difficulties in attention and hyperactivity. She was born prematurely to a mother who consumed alcohol during pregnancy. She has facial dysmorphisms, including a thin upper lip and a flat midface. Her growth is below the 5th percentile, and she has a mild intellectual disability.
A 2-year-old boy is brought in for evaluation of developmental delays. His mother reports drinking alcohol during pregnancy. He has a history of poor weight gain and has facial features such as a smooth philtrum and small palpebral fissures. He is also irritable and has difficulty sleeping.
A 12-year-old girl is seen for persistent learning difficulties and social challenges. Her birth mother had a history of alcohol use during pregnancy. Physical examination reveals a thin upper lip, short palpebral fissures, and a smooth philtrum. She has difficulty with tasks requiring planning and organization.
A 6-year-old boy is brought in for evaluation of hyperactivity and speech delay. He was exposed to alcohol in utero, and his facial features include microcephaly, smooth philtrum, and small eye openings. His growth is below the 10th percentile, and he has significant delays in motor skills.
A 4-year-old girl is evaluated for behavioral problems and delayed milestones. She has a history of prenatal alcohol exposure. Her facial features include a smooth philtrum and a thin upper lip. She also has a low birth weight and short stature. Neurodevelopmental testing reveals cognitive deficits.
A 7-year-old boy with a history of prenatal alcohol exposure is evaluated for poor academic performance and behavioral issues. He has a flat midface, a smooth philtrum, and short palpebral fissures. He also exhibits hyperactivity and difficulty with attention.
A 10-year-old girl presents with difficulties in school and social interactions. Her birth mother consumed alcohol during pregnancy. She has facial features consistent with FASD, including a thin upper lip and smooth philtrum. Her growth parameters are below the 5th percentile, and she has significant learning disabilities.
A 5-year-old boy is brought in for evaluation of speech and language delay. He has a history of prenatal alcohol exposure. His physical examination reveals microcephaly, a smooth philtrum, and small palpebral fissures. His cognitive development is below age expectations, and he has difficulty with attention and impulse control.
A 2-year-old girl is evaluated for failure to thrive and developmental delays. Her mother drank alcohol during pregnancy. She has dysmorphic features, including a thin upper lip, smooth philtrum, and microcephaly. She also exhibits irritability and poor sleep patterns.
A 9-year-old boy is referred for evaluation of academic difficulties and behavioral problems. His birth mother had a history of heavy alcohol use during pregnancy. On examination, he has short palpebral fissures, a flat midface, and a smooth philtrum. His cognitive assessment shows deficits in attention, memory, and executive function.
A 6-year-old girl presents with hyperactivity and poor social skills. She was exposed to alcohol in utero, and her physical examination reveals a smooth philtrum, small eye openings, and a thin upper lip. Her growth is below the 5th percentile, and she has a mild intellectual disability.
A 3-year-old boy is evaluated for delayed milestones and poor growth. His mother reports consuming alcohol during pregnancy. His facial features include a smooth philtrum, small palpebral fissures, and micrognathia. He also has behavioral issues, including irritability and difficulty with transitions.
A 12-year-old girl is seen for persistent learning difficulties and social challenges. Her birth mother consumed alcohol during pregnancy. Physical examination reveals a thin upper lip, short palpebral fissures, and a smooth philtrum. She has difficulty with tasks requiring planning and organization.

Figure 24: Queries Part 1.

A 6-year-old boy is brought in for evaluation of hyperactivity and speech delay. He was exposed to alcohol in utero, and his facial features include microcephaly, smooth philtrum, and small eye openings. His growth is below the 10th percentile, and he has significant delays in motor skills.
A 4-year-old girl is evaluated for behavioral problems and delayed milestones. She has a history of prenatal alcohol exposure. Her facial features include a smooth philtrum and a thin upper lip. She also has a low birth weight and short stature. Neurodevelopmental testing reveals cognitive deficits.
A 7-year-old boy with a history of prenatal alcohol exposure is evaluated for poor academic performance and behavioral issues. He has a flat midface, a smooth philtrum, and short palpebral fissures. He also exhibits hyperactivity and difficulty with attention.
A 10-year-old girl presents with difficulties in school and social interactions. Her birth mother consumed alcohol during pregnancy. She has facial features consistent with FASD, including a thin upper lip and smooth philtrum. Her growth parameters are below the 5th percentile, and she has significant learning disabilities.
A 5-year-old boy is brought in for evaluation of speech and language delay. He has a history of prenatal alcohol exposure. His physical examination reveals microcephaly, a smooth philtrum, and small palpebral fissures. His cognitive development is below age expectations, and he has difficulty with attention and impulse control.
A 2-year-old girl is evaluated for failure to thrive and developmental delays. Her mother drank alcohol during pregnancy. She has dysmorphic features, including a thin upper lip, smooth philtrum, and microcephaly. She also exhibits irritability and poor sleep patterns.
A 9-year-old boy is referred for evaluation of academic difficulties and behavioral problems. His birth mother had a history of heavy alcohol use during pregnancy. On examination, he has short palpebral fissures, a flat midface, and a smooth philtrum. His cognitive assessment shows deficits in attention, memory, and executive function.
A 6-year-old girl presents with hyperactivity and poor social skills. She was exposed to alcohol in utero, and her physical examination reveals a smooth philtrum, small eye openings, and a thin upper lip. Her growth is below the 5th percentile, and she has a mild intellectual disability.
A 3-year-old boy is evaluated for delayed milestones and poor growth. His mother reports consuming alcohol during pregnancy. His facial features include a smooth philtrum, small palpebral fissures, and micrognathia. He also has behavioral issues, including irritability and difficulty with transitions.
A 12-year-old girl is seen for persistent learning difficulties and social challenges. Her birth mother consumed alcohol during pregnancy. Physical examination reveals a thin upper lip, short palpebral fissures, and a smooth philtrum. She has difficulty with tasks requiring planning and organization.
A 5-year-old boy is brought in for evaluation of hyperactivity and speech delay. He was exposed to alcohol in utero, and his facial features include microcephaly, smooth philtrum, and small eye openings. His growth is below the 10th percentile, and he has significant delays in motor skills.
A 7-year-old boy with a history of prenatal alcohol exposure is evaluated for poor academic performance and behavioral issues. He has a flat midface, a smooth philtrum, and short palpebral fissures. He also exhibits hyperactivity and difficulty with attention.
A 10-year-old girl presents with difficulties in school and social interactions. Her birth mother consumed alcohol during pregnancy. She has facial features consistent with FASD, including a thin upper lip and smooth philtrum. Her growth parameters are below the 5th percentile, and she has significant learning disabilities.
A 2-year-old girl is evaluated for failure to thrive and developmental delays. Her mother drank alcohol during pregnancy. She has dysmorphic features, including a thin upper lip, smooth philtrum, and microcephaly.
A 9-year-old boy is referred for evaluation of academic difficulties and behavioral problems. His birth mother had a history of heavy alcohol use during pregnancy. On examination, he has short palpebral fissures, a flat midface, and a smooth philtrum. His cognitive assessment shows deficits in attention, memory, and executive function.
A 3-year-old boy with a history of prenatal alcohol exposure is brought in for evaluation of poor growth and developmental delays. His facial features include a smooth philtrum, thin upper lip, and small eye openings. He also has behavioral problems, including irritability and difficulty with transitions.
A 4-year-old girl is evaluated for developmental delay. She was born to a mother with a history of alcohol abuse. The child has a history of poor weight gain, and her facial features include a short palpebral fissure and a thin upper lip. Cognitive assessment shows significant deficits in executive function.
A 5-year-old girl presents with difficulties in attention and hyperactivity. She was born prematurely to a mother who consumed alcohol during pregnancy. She has facial dysmorphisms, including a thin upper lip and a flat midface. Her growth is below the 5th percentile, and she has a mild intellectual disability.
A 10-year-old boy is evaluated for aggressive behavior and poor school performance. He has a history of prenatal alcohol exposure. Physical examination reveals a smooth philtrum, micrognathia, and growth retardation. Neurocognitive testing shows deficits in memory and executive function.
A 2-year-old boy is brought in for evaluation of developmental delays. His mother reports drinking alcohol during pregnancy. He has a history of poor weight gain and has facial features such as a smooth philtrum and small palpebral fissures. He is also irritable and has difficulty sleeping.
A 6-year-old boy is brought in by his foster parents due to concerns about hyperactivity and learning difficulties. He was adopted at 2 years of age. His birth mother reportedly drank alcohol heavily during pregnancy. He has dysmorphic facial features, including a smooth philtrum and thin upper lip. His growth parameters are below the 5th percentile for age.
A 7-year-old girl is referred for evaluation of speech and language delay. Her birth history reveals exposure to alcohol during pregnancy. She has a flat midface, epicanthal folds, and microcephaly. Her IQ is significantly below average.
A 6-year-old boy is brought in for evaluation of speech delay and learning difficulties. He was exposed to alcohol in utero, and his facial features include microcephaly, smooth philtrum, and small palpebral fissures. His growth is below the 10th percentile, and he has significant delays in motor skills.

Figure 25: Queries Part 2.

A 9-year-old girl is seen for learning difficulties and social challenges. Her birth mother had a history of alcohol use during pregnancy. Physical examination reveals a thin upper lip, short palpebral fissures, and a smooth philtrum. She has difficulty with tasks requiring planning and organization.
A 4-year-old girl presents with developmental delays and behavioral issues. She has a history of prenatal alcohol exposure. Her facial features include a smooth philtrum and a thin upper lip. She also has a low birth weight and short stature. Neurodevelopmental testing reveals cognitive deficits.
A 10-year-old boy is evaluated for hyperactivity and poor school performance. His teacher reports difficulty with attention and impulsivity. His mother had a history of alcohol use during pregnancy. On examination, he has a smooth philtrum, and his height and weight are in the 3rd percentile for his age.
A 5-year-old girl is brought in for evaluation of speech delay and behavioral problems. She was exposed to alcohol in utero, and her facial features include microcephaly, smooth philtrum, and small palpebral fissures. Her growth is below the 10th percentile, and she has significant delays in motor skills.
A 7-year-old boy with a history of prenatal alcohol exposure is evaluated for poor academic performance and behavioral issues. He has a flat midface, a smooth philtrum, and short palpebral fissures. He also exhibits hyperactivity and difficulty with attention.
A 12-year-old girl presents with difficulties in school and social interactions. Her birth mother consumed alcohol during pregnancy. She has facial features consistent with FASD, including a thin upper lip and smooth philtrum. Her growth parameters are below the 5th percentile, and she has significant learning disabilities.
A 3-year-old boy is evaluated for delayed milestones and poor growth. His mother reports consuming alcohol during pregnancy. His facial features include a smooth philtrum, small palpebral fissures, and micrognathia. He also has behavioral issues, including irritability and difficulty with transitions.
A 6-year-old girl presents with hyperactivity and poor social skills. She was exposed to alcohol in utero, and her physical examination reveals a smooth philtrum, small eye openings, and a thin upper lip. Her growth is below the 5th percentile, and she has a mild intellectual disability.
A 4-year-old girl is evaluated for speech delay and developmental issues. She was born to a mother with a history of alcohol abuse. The child has a history of poor weight gain, and her facial features include a short palpebral fissure and a thin upper lip. Cognitive assessment shows significant deficits in executive function.
A 5-year-old boy presents with difficulties in attention and hyperactivity. He was born prematurely to a mother who consumed alcohol during pregnancy. He has facial dysmorphisms, including a thin upper lip and a flat midface. His growth is below the 5th percentile, and he has a mild intellectual disability.
A 9-year-old boy is evaluated for poor academic performance and behavioral issues. His birth mother had a history of alcohol use during pregnancy. On examination, he has short palpebral fissures, a flat midface, and a smooth philtrum. His cognitive assessment shows deficits in attention, memory, and executive function.
A 2-year-old girl is evaluated for failure to thrive and developmental delays. Her mother drank alcohol during pregnancy. She has dysmorphic features, including a thin upper lip, smooth philtrum, and microcephaly. She also exhibits irritability and poor sleep patterns.
A 3-year-old boy is brought in for evaluation of developmental delays. His mother reports drinking alcohol during pregnancy. He has a history of poor weight gain and has facial features such as a smooth philtrum and small palpebral fissures. He is also irritable and has difficulty sleeping.
A 6-year-old boy is brought in by his foster parents due to concerns about hyperactivity and learning difficulties. He was adopted at 2 years of age. His birth mother reportedly drank alcohol heavily during pregnancy. He has dysmorphic facial features, including a smooth philtrum and thin upper lip. His growth parameters are below the 5th percentile for age.
A 7-year-old girl is referred for evaluation of speech and language delay. Her birth history reveals exposure to alcohol during pregnancy. She has a flat midface, epicanthal folds, and microcephaly. Her IQ is significantly below average.
A 10-year-old boy presents with aggressive behavior and poor school performance. He has a history of prenatal alcohol exposure. Physical examination reveals a smooth philtrum, micrognathia, and growth retardation. Neurocognitive testing shows deficits in memory and executive function.
A 2-year-old girl is evaluated for failure to thrive and developmental delays. Her mother drank alcohol during pregnancy. She has dysmorphic features, including a thin upper lip, smooth philtrum, and microcephaly. She also exhibits irritability and poor sleep patterns.
A 12-year-old girl is seen for persistent learning difficulties and social challenges. Her birth mother consumed alcohol during pregnancy. Physical examination reveals a thin upper lip, short palpebral fissures, and a smooth philtrum. She has difficulty with tasks requiring planning and organization.

Figure 26: Queries Part 3.

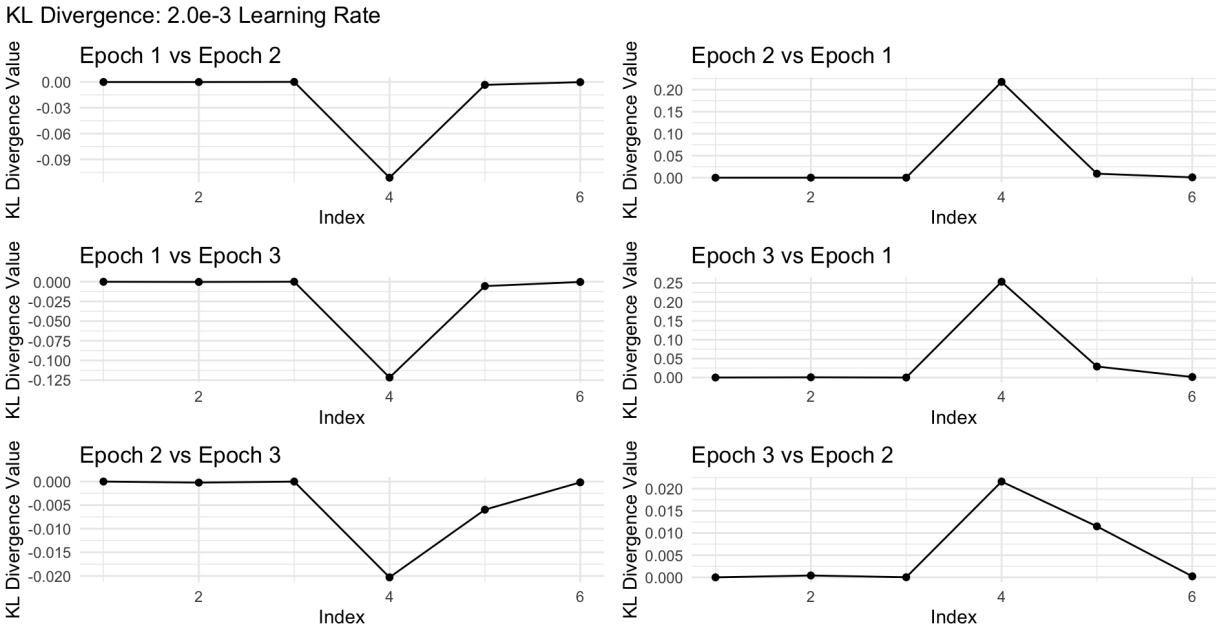

Figure 27: KL Divergence 2e-3.

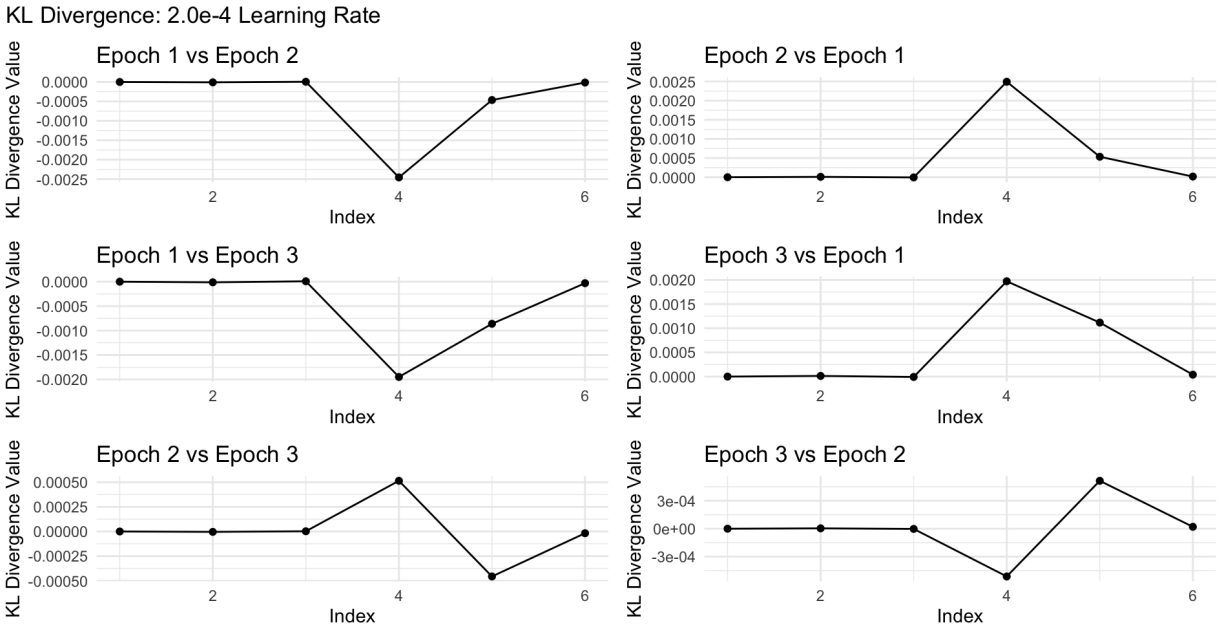

Figure 28: KL Divergence 2e-4.

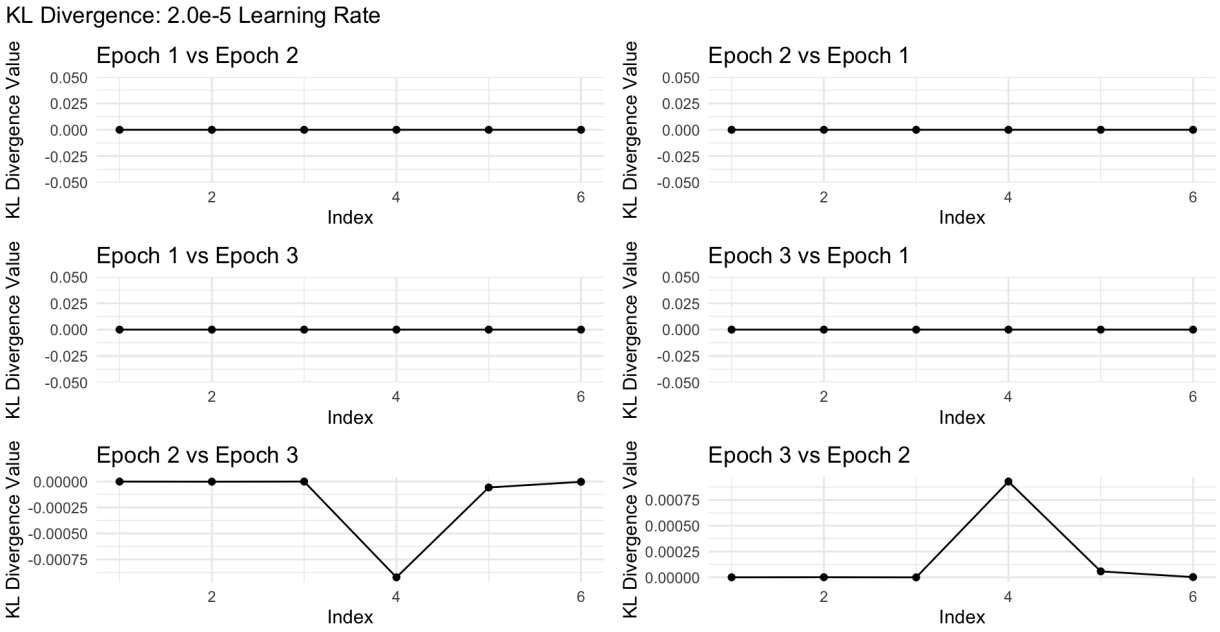

Figure 29: KL Divergence 2e-5.

Diagnostic to ICD Code
Newborn (suspected to be) affected by maternal use of alcohol (Excludes Fetal Alcohol Syndrome), P04.3
Fetal alcohol syndrome (dysmorphic), Q86.0
Mood disorder due to known physiological condition, unspecified, F06.30
Newborn (suspected to be) affected by maternal nutritional disorders, P00.4
Newborn (suspected to be) affected by maternal complication of pregnancy, unspecified, P01.9
Encephalopathy, other and unspecified (static), G96.8
Other specified disorders of central nervous system, G93.4
Disorder of central nervous system, unspecified, G96.9
Microphthalmos, Q11.2
Other general symptoms and signs (eg, dysmorphic features), R68.89
Underweight, R63.6
Feeding difficulties, R63.3
Failure to thrive (child), R62.51
Short stature (child), R62.52
Lack of expected normal physiological development in childhood, unspecified, R62.50
Delayed milestone in childhood, R62.0
Mild cognitive impairment, so stated, G31.84
Mild intellectual disabilities, F70
Moderate intellectual disabilities, F71
Severe intellectual disabilities, F72
Profound intellectual disabilities, F73
Intellectual disabilities, Other specified, F78
Intellectual disabilities, Unspecified, F79
Toxic encephalopathy (code first (T51-T65)) to identify toxic agent, G92
Post-traumatic stress disorder, unspecified, F43.10
Post-traumatic stress disorder, acute, F43.11
Post-traumatic stress disorder, chronic, F43.12
Reactive attachment disorder of childhood, F94.1
Intermittent explosive disorder, F63.81
Expressive language disorder, F80.1
Mixed receptive-expressive language disorder, F80.2
Developmental disorder of scholastic skills, unspecified, F81.9
Disorder of psychological development, unspecified, F89
Attention-deficit hyperactivity disorder, predominantly inattentive type, F90.0
Attention-deficit hyperactivity disorder, predominantly hyperactive type, F90.1
Attention-deficit hyperactivity disorder, other type, F90.8
Specific reading disorder, F81.0
Alexia/dyslexia, NOS, R48.0
Developmental Dyslexia, F81.0
Mathematics disorder, F81.2
Disorder of written expression, F81.81
Ataxia, unspecified, R27.0
Other lack of coordination, R27.8
Unspecified lack of coordination, R27.9
Symbolic dysfunction, unspecified, R48.9
Alexia/dyslexia, NOS, R48.0
Agnosia, R48.1
Apraxia, R48.2
Visual agnosia, R48.3
Other symbolic dysfunctions, R48.8
Attention and concentration deficit (Excludes attention deficit disorder), R41.840

Figure 30: 51 ICD Codes.

Cognitive communication deficit, R41.841
Visuospatial deficit, R41.842
Psychomotor deficit, R41.843
Frontal lobe and executive function deficit, R41.844
Other symptoms and signs involving cognitive functions and awareness, R41.89
Developmental disorder of scholastic skills, unspecified, F81.9
Other symptoms and signs involving appearance and behavior, R46.89
Nonpsychotic mental disorder, unspecified, F48.9
Encounter for observation for other suspected diseases and conditions ruled out (eg, mental health), Z03.89
Encounter for blood-alcohol and blood-drug test, Z02.83
Finding of alcohol in blood, R78.0
Alcohol use, unspecified, F10.9
Alcohol abuse, F10.1
Alcohol dependence, F10.2
Personal history of other mental and behavioral disorders, Z86.59
Family History of Alcohol Abuse and Dependence, Z81.1
Alcohol abuse counseling and surveillance of alcoholic, Z71.41
Encounter for antenatal screening for chromosomal anomalies, Z36.0
Encounter for antenatal screening for malformations, Z36.3
Encounter for antenatal screening for fetal growth retardation, Z36.4
Encounter for antenatal screening for congenital cardiac abnormalities, Z36.83
Encounter for antenatal screening for other specified antenatal screening, Z36.89
Encounter for antenatal screening for other genetic defects, Z36.8A
Maternal care for known or suspected fetal abnormality and damage, 035.4
Alcohol use complicating pregnancy, childbirth, and the puerperium, O99.31
Other congenital malformation syndromes due to unknown exogenous causes, Q86.8
Other specified congenital malformation syndromes affecting multiple systems, Q87
Congenital malformation syndromes predominantly affecting facial appearance, Q87.0
Congenital malformation syndromes predominantly associated with short stature, Q87.1
Prader-Willi syndrome, Q87.11
Other congenital malformation syndromes predominantly associated with short stature, Q87.19
Congenital malformation syndromes predominantly involving limbs, Q87.2
Congenital malformation syndromes involving early overgrowth, Q87.3
Marfan syndrome, Q87.4
Marfan syndrome with cardiovascular manifestations, Q87.41
Marfan syndrome with aortic dilation, Q87.410
Marfan syndrome with other cardiovascular manifestations, Q87.418
Marfan syndrome with ocular manifestations, Q87.42
Marfan syndrome with skeletal manifestation, Q87.43
Other congenital malformation syndromes with other skeletal changes, Q87.5
Other specified congenital malformation syndromes, not elsewhere classified, Q87.8
Alport syndrome, Q87.81
Arterial tortuosity syndrome, Q87.82
Bardet-Biedl syndrome, Q87.83
Laurence-Moon syndrome, Q87.84
MED13L syndrome, Q87.85
Other specified congenital malformation syndromes, not elsewhere classified, Q87.89
Williams Syndrome, Q93.82
Cornelia de Lange Syndrome, Q87.19
Fetal hydantoin syndrome, Q86.1
Dwarfism due to warfarin, Q86.2
Newborn affected by opiates, P04.14

Figure 31: 52 Additional ICD Codes.

REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[2] T. Savage, A. Nayak, R. Gallo, *et al.*, "Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine," *NPJ Digital Medicine*, vol. 7, p. 20, 2024.

[3] I. Düntsch and G. Gediga, "Confusion matrices and rough set data analysis," *Journal of Physics: Conference Series*, vol. 1229, p. 012055, May 2019.

[4] OpenAI, "Chatgpt," 2024.

[5] P. A. May, C. D. Chambers, W. O. Kalberg, J. Zellner, H. Feldman, D. Buckley, D. Kopald, J. M. Hasken, R. Xu, G. Honerkamp-Smith, H. Taras, M. A. Manning, L. K. Robinson, M. P. Adam, O. Abdul-Rahman, K. Vaux, T. Jewett, A. J. Elliott, J. A. Kable, N. Akshoomoff, D. Falk, J. A. Arroyo, D. Hereld, E. P. Riley, M. E. Charness, C. D. Coles, K. R. Warren, K. L. Jones, and H. E. Hoyme, "Prevalence of Fetal Alcohol Spectrum Disorders in 4 US Communities," *JAMA*, vol. 319, pp. 474–482, 02 2018.

[6] I. J. Chasnoff, A. M. Wells, and L. King, "Misdiagnosis and missed diagnoses in foster and adopted children with prenatal alcohol exposure," *Pediatrics*, vol. 135, no. 2, pp. 264–270, 2015.

[7] World Health Organization, "International classification of diseases (icd)." `https://www.who.int/standards/classifications/classification-of-diseases`, 2024. Accessed: 2024-07-22.

[8] K. W. Fung, J. Xu, and O. Bodenreider, "The new international classification of diseases 11th edition: a comparative analysis with icd-10 and icd-10-cm," *Journal of the American Medical Informatics Association*, vol. 27, no. 5, pp. 738–746, 2020.

[9] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pretrained large language models for medical domains," 2024.

[10] J. T. Reese, D. Danis, J. H. Caufield, T. Groza, E. Casiraghi, G. Valentini, C. J. Mungall, and P. N. Robinson, "On the limitations of large language models in clinical diagnosis," *medRxiv : the preprint server for health sciences*, vol. 2023.07.13.23292613, 2024.

[11] J. D. Thomas, K. R. Warren, and B. G. Hewitt, "Fetal alcohol spectrum disorders: From research to policy," *Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism*, vol. 33, no. 1-2, pp. 118–126, 2010.

[12] K. L. Jones, D. W. Smith, C. N. Ulleland, and A. P. Streissguth, "Recognition of the fetal alcohol syndrome in early infancy," *The Lancet*, vol. 302, no. 7836, pp. 999–1001, 1973.

[13] E. M. Armstrong and E. L. Abel, "Fetal alcohol syndrome: The origins of a moral panic," *Alcohol and alcoholism*, vol. 35, no. 3, pp. 276–282, 2000.

[14] K. Jones and D. Smith, "Recognition of the fetal alcohol syndrome in early infancy," *The Lancet*, vol. 302, no. 7836, pp. 999–1001, 1973. Originally published as Volume 2, Issue 7836.

[15] E. M. Armstrong, "Diagnosing moral disorder: the discovery and evolution of fetal alcohol syndrome," *Social Science & Medicine*, vol. 47, no. 12, pp. 2025–2042, 1998.

[16] E. Hargrove, C. J. Lutke, K. Griffin, M. Himmelreich, J. Mitchell, A. Lutke, and P. Choate, "FASD: The Living Experience of People with Fetal Alcohol Spectrum Disorder—Results of an Anonymous Survey," *Disabilities*, vol. 4, no. 2, pp. 332–347, 2024.

[17] K. Flannigan, K. Harding, J. Pei, K. McLachlan, M. Mela, J. Cook, and A. McFarlane, "The unique complexities of fetal alcohol spectrum disorder," *Canada FASD Research Network*, December 2020.

[18] S. Popova, M. E. Charness, L. Burd, *et al.*, "Fetal alcohol spectrum disorders," *Nature Reviews Disease Primers*, vol. 9, p. 11, 2023. Accepted 16 January 2023, Published 23 February 2023.

[19] D. W. Bates, D. Levine, A. Syrowatka, M. Kuznetsova, K. J. Craig, A. Rui, G. P. Jackson, and K. Rhee, "The potential of artificial intelligence to improve patient safety: a scoping review," *NPJ Digital Medicine*, vol. 4, no. 1, p. 54, 2021.

[20] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[21] W. J. Hutchins, "The Georgetown-IBM experiment demonstrated in January 1954," in *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas: Technical Papers* (R. E. Frederking and K. B. Taylor, eds.), (Washington, USA), pp. 102–114, Springer, Sept. 28 - Oct. 2 2004.

[22] N. Chomsky, "Three models for the description of language," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113–124, 1956.

[23] D. McCracken and E. Reilly, "Backus-naur form (bnf)," *Encyclopedia of Computer Science*, pp. 129–131, 01 2003.

[24] J. R. Levine, T. Mason, and D. Brown, *Lex & Yacc*. Sebastopol, CA: O'Reilly Media, 2nd ed., 1992.

[25] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–551, 09 2011.

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. New York, NY: Pearson, 3rd ed., 2024.

[27] M. A. Rohrmeier and S. Koelsch, "Predictive information processing in music cognition. a critical review," *International Journal of Psychophysiology*, vol. 83, no. 2, pp. 164–175, 2012. Predictive information processing in the brain: Principles, neural mechanisms and models.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[29] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, 2018.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI*, 2019.

[31] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," *OpenAI*, 2022.

[32] R. K. Garg, V. L. Urs, A. A. Agarwal, S. K. Chaudhary, V. Paliwal, and S. K. Kar, "Exploring the role of chatgpt in patient care (diagnosis and treatment) and medical research: A systematic review," *Health Promotion Perspectives*, vol. 13, no. 3, pp. 183–191, 2023.

[33] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[34] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Association for Computational Linguistics, 2019.

[35] H. Shin, P. Ramanan, A. Beutel, H. Yun, X. Chen, M. Huang, C. Sheu, S. Merity, J. McAuley, and S. Yun, "Biomegatron: Larger biomedical domain language model," *arXiv preprint arXiv:2010.06060*, 2020.

[36] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," in *Proceedings of the 3rd Machine Learning for Healthcare Conference (MLHC)*, pp. 117–136, PMLR, 2019.

[37] J. do Olmo, J. Logroño, C. Mascías, M. Martínez, and J. Isla, "Assessing dxgpt: Diagnosing rare diseases with various large language models," *medRxiv*, 2024.

[38] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, Curran Associates, Inc., 2017.

[40] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[41] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[42] J. Han and et al., "Medalpaca: An open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[43] K. Zhao and et al., "Alpacare: Instruction-tuned large language models for medical application," *arXiv preprint arXiv:2310.14558*, 2023.

[44] Y. Wu and et al., "Pmc-llama: Biomedical domain instruction tuning of llama models," *arXiv preprint arXiv:2310.14558*, 2023.

[45] J. Wu, H. Dong, Z. Li, A. Patra, and H. Wu, "Retrieving and refining: A hybrid framework with large language models for rare disease identification," 2024.

[46] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," 2024.

[47] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, and X. Qiu, "Full parameter fine-tuning for large language models with limited resources," 2024.

[48] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[49] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning (ICML)*, PMLR, 2019.

[50] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[51] D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta, "Neural evolution of sparse networks: A randomized training algorithm for deep learning," *Neural Computing and Applications*, vol. 31, no. 7, pp. 2075–2090, 2019.

[52] Y. Wu, L. Liu, J. Bae, K.-H. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, and Q. Zhang, "Demystifying learning rate polices for high accuracy training of deep neural networks," *ArXiv*, vol. abs/1908.06477, 2019.

[53] C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.

[54] B. Plaut, K. Nguyen, and T. Trinh, "Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a," 2024.

[55] J. Shlens, "Notes on kullback-leibler divergence and likelihood," 2014.

[56] Y. Zhang, W. Liu, Z. Chen, J. Wang, and K. Li, "On the properties of kullback-leibler divergence between multivariate gaussian distributions," *Entropy*, vol. 25, no. 6, p. 871, 2023.