IDS 705 Final Report

Text Classification For Medical Specialty

Natural Language Processing Applied to Patient Pre-Screening

Team Members:
Chen, Sizhe
Haeussler, Kayla
Mammadov, Ramil
Paredes La Torre, Alejandro
Xiao, Zihan

Team Identifier: Raccoon

Abstract

Efficiently routing patients to the appropriate medical specialty remains a critical bottleneck in healthcare intake, often leading to misdiagnosis, delayed care, and increased costs. In this project, we develop and evaluate a comprehensive transformer-based NLP pipeline for medical pre-screening that combines symptom classification, embedding interpretability, cross-lingual generalization, and end-to-end clinical integration. We fine-tuned three pretrained transformer models — BERT-base, RoBERTa-base, and DistilBERT-base — on a unified dataset of 2,300 patient-reported symptom descriptions. Across five evaluation runs, RoBERTa achieved the highest mean F1 score (0.918), followed by BERT-base (0.909) and DistilBERT (0.799). Robustness testing under insertion- and spelling-noise perturbations demonstrated over 81 % prediction consistency, highlighting resilience to real-world input variability. In a zero-shot evaluation on Chinese translations, multilingual BERT achieved 88.7 % accuracy, underscoring the value of multilingual pretraining for low-resource contexts. Embedding analyses via t-SNE revealed clear specialty-specific clusters, while attention and attribution methods (BertViz, SHAP, Integrated Gradients) provided transparent, token-level explanations of model decisions. Finally, we showcase an end-to-end framework by integrating T5-based medical report summarization (ROUGE-L ≈ 0.77) and GPT-2 question-answering, illustrating practical deployment for streamlined pre-diagnosis and enhanced clinician support.

Introduction

Accurately guiding patients to the appropriate medical specialist remains a significant challenge in healthcare. Delays or inaccuracies during the intake process can lead to misdiagnosis, longer wait times, and increased strain on medical resources. Automating the classification of medical specialties based on patients' reported symptoms could help streamline care and reduce the burden on providers.

Deep learning techniques have gained rapid traction in the medical space in recent years [1,2,3], driven primarily by their improvement from conventional machine learning models [4,5,6] and effectiveness in dealing with alternative data formats ranging from text [7] to image [8] and audio [9]. Robust deep-learning-based systems can even achieve similar performance to domain experts in the diagnosis of specific diseases [16, 17, 18]. In the field of natural language processing, effective models need massive amounts of high quality and diverse data in order to capture the complex semantics of language and achieve generalization comparable with a human understanding of language [19].

Transfer learning addresses the issues of data scarcity by pre-training deep learning models on large unannotated corpora and then fine tuning on smaller, task-specific datasets. In the natural language processing field, this approach has led to the development of large language models like BERT (bidirectional encoder representations from transformers) which leverages transformer architecture to understand linguistic semantics [21,22]. BERT has shown strong performance on clinical tasks such as medical report analysis and text classification [12,13,15,23,24], demonstrating the effectiveness of transfer learning in the medical domain.

This study proposes a set of experiments related to pre-screening and this model integration in a workflow of efficient medical attention. The main contribution consists of a deep learning-based model using transformer architectures to pre-screen and predict medical specialties from patient-reported symptoms, aiming to improve efficiency in medical intake and specialist referrals. Through a series of experiments, the project will assess the model's robustness, model ability to handle a different language, analyze embedding distributions, and explore the model's decision patterns. This robust

model gets integrated in our proposed workflow to demonstrate how the model can fit into a full medical framework, with the broader goal of streamlining pre-screening and patient support, reducing decision-making time, and supporting healthcare providers by automating parts of the intake and pre-diagnosis process.

Related Work

Founding work such as [25] shows that expanding clinical abbreviations using task-oriented word embeddings is effective, and so using deep learning techniques could be leveraged for various tasks. Related work using deep learning methods for analyzing medical text discusses that neural networks can capture semantic relationships in biomedical language, aiming to enhance tasks like information retrieval and knowledge discovery in healthcare [26].

Text classification is a fundamental task in natural language processing (NLP) that involves identifying the underlying class behind a given text input [1]. This task is particularly relevant in applications such as intent classification [3], sentiment analysis [4], and Question Answer selection [5]. The advancements made in transformer-based architectures have greatly improved the accuracy of text classification models [2], with RoBERTa (Robustly Optimized BERT Pre Training Approach) emerging as a highly effective model for such tasks.

Various modeling approaches have been proposed for medical specialty classification using patient-reported text data.. Habib et al. [10] presented a novel predictive text system tailored for medical recommendations within telemedicine, leveraging deep learning methodologies specifically adapted to the Arabic language context. Weng et al. [11] developed a medical subdomain classification model for clinical notes using convolutional neural networks. BioBERTL [27] helps researchers with tasks such as named entity recognition, lation extraction, and question answering.

Recently, multilingual transformer models such as multilingual BERT (mBERT) [35] and XLM-R [36] have opened new directions for medical NLP, especially in scenarios involving cross-lingual transfer or zero-shot generalization. These models are pre-trained across 100+ languages and have demonstrated strong performance in tasks requiring language transfer without target-language training data. Their ability to generalize across languages has made them increasingly relevant in multilingual healthcare applications, where clinical data may not always be available in English.

Datasets

A collection of datasets have been used to enhance the model capabilities, test robustness, evaluate the model’s capability for zero-shot learning in Chinese and build associated models for this model to be integrated in an end-to-end framework.

For the main classification task, three datasets from Hugging Face and Kaggle were combined, totaling approximately 2,300 records of medical domain classification and common patient-reported symptom descriptions [29,30,31]. These three datasets were combined as they each contained the style of data needed for the task (patient described symptoms and condition labels), but individually lacked a sufficiently large number of rows needed for fine-tuning a natural language model. Additionally, although all three sources had comparable patient-reported symptom formats, their condition labels showed little overlap. To enhance model performance, we manually consolidated the original labels into

eight broader categories for our unified dataset. For example, labels like “Skin issue” in one data set, “Psoriasis” in the second dataset and “Dermatitis due to Sun Exposure” in the third dataset were all mapped to the label of "Dermatological & Skin Conditions" in our final go forward dataset. Additional information on dataset consolidation and the final combined dataset can be found in appendix A.

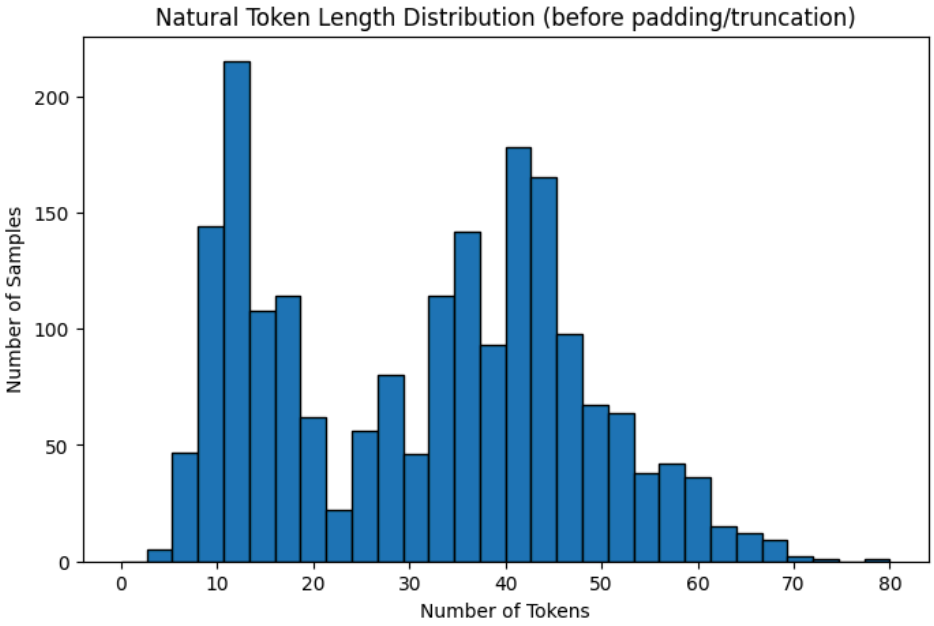


Figure 1. *Distribution of Token Length for Patient Phrases in Our Final Combined Dataset*

For the robustness testing experiment, adversarial variants of the classification dataset were constructed, including insert-attack and spelling-noise examples, to assess the model’s resilience to perturbed inputs.

In order to evaluate the model’s zero-shot classification capabilities in a foreign language, we translated our combined English classification dataset into Chinese. Given the difficulty of sourcing high-quality Chinese medical text datasets—such as those available through platforms like Alibaba Cloud, which require an application process and often involve a usage fee—we opted to translate our own dataset. The translation was performed automatically using Google Translate via the `deep_translator` library[45], with a fallback mechanism to retain the original text if any translation failed. To verify whether the automatic translation process introduced significant errors, we conducted a back-translation quality analysis. A random subset of the translated Chinese symptom descriptions was translated back into English using the Google Translate API. The back-translated sentences were then compared against the original English phrases. Based on this quality check, we conclude that the translation process introduced minimal semantic errors and did not materially impact the integrity of the dataset for evaluating cross-lingual model performance.

For the medical framework integration experiment, a dataset composed of question answering [34] and a dataset of medical reports and their corresponding summarizations [33] have been used to train a question-answering model and a medical summarization model respectively.

Experiments

Experiment 1: Pre-trained Model Comparisons

The goal of this experiment was to compare performance of three pretrained transformer models —BERT-base, RoBERTa-base, and DistilBERT-base—on a multi-class medical symptom classification task. The dataset consisted of free-text patients-reported symptom descriptions labeled by condition category (e.g., “Infections”, “Chronic Conditions”). To ensure consistency, we used an 80/10/10 stratified split of the dataset into training, validation, and test sets. The training set was used to fit model parameters, the validation set used to manually tune hyperparameters (specifically epochs, see Figure X) without contaminating the test set, and the test set provided an unbiased estimate of the model’s final generalization performance. All models were trained and evaluated on these same splits, with raw text re-tokenized using each model’s specific tokenizer and dynamic padding/truncation applied via Hugging Face’s DataCollatorWithPadding.

Each model was fine-tuned using the Hugging Face Trainer API with consistent weight decay of 0.01, and model-specific numbers of epochs selected based on observed validation performance. BERT-base and RoBERTa both achieved early convergence, stabilizing after approximately 5 epochs without signs of overfitting. DistilBERT showed slower improvement and was allowed to train for 8 epochs to maximize performance before plateauing. For additional information on hyperparameter selection, see Appendix B.

Experiment 2: Robustness Testing

This experiment evaluates the robustness of our fine-tuned BERT-base model to input-level noise typical in clinical settings. Unlike prior work focused on clean inputs, we assess the model’s stability under minor adversarial changes—reflecting real-world scenarios like patient chatbots, where inputs may be noisy or contain typos. A robust model should yield consistent predictions despite such surface-level variations.

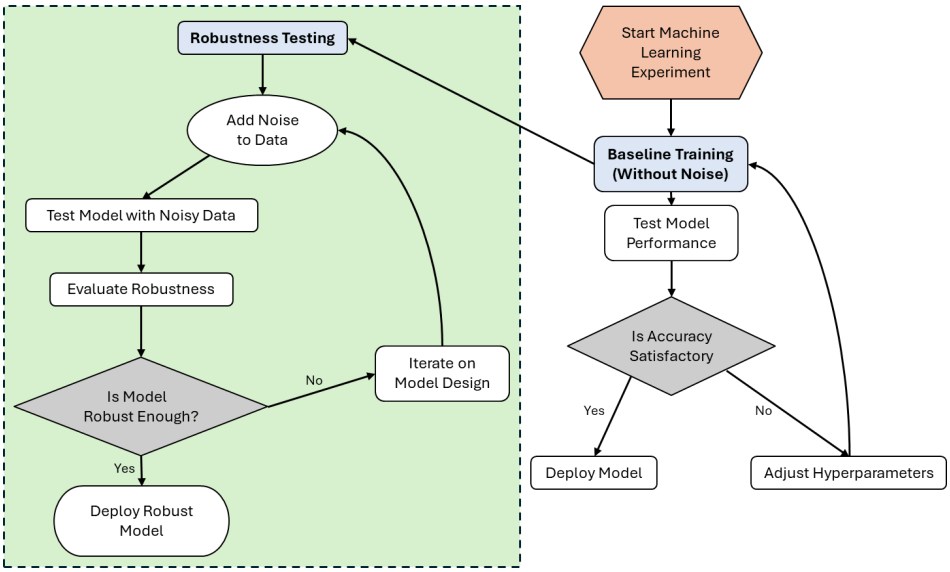


Figure 2. *Framework for Adversarial Robustness Evaluation*

To simulate realistic adversarial conditions, we designed two categories of input perturbations based on common sources of noise observed in NLP and clinical text processing. First, we introduced irrelevant insertions into the input sequence by adding the phrase “I have insurance” at random positions. This type of filler or off-topic phrase is frequently found in patient-provider dialogues and has been explored in prior robustness studies such as Belinkov & Bisk, 2018 and Ribeiro et al., 2020, which emphasize the impact of distractor tokens on model behavior. Second, we introduced spelling noise by applying character-level modifications—such as random deletions, swaps, or substitutions—to selected words within each sentence. This simulates common typographical errors, similar to settings explored in Michel et al., 2019 and other robustness literature addressing real-world textual noise. These two perturbation types reflect the kinds of noise likely to arise in clinical chatbot interactions, where patients may submit informal, unstructured, or typo-ridden inputs.

The same BERT-base model, previously fine-tuned on our final_data.csv dataset, was reused for this experiment without additional training. Instead of using the Hugging Face Trainer API, we implemented a custom PyTorch-based evaluation pipeline to allow fine-grained control over the input transformations and consistency measurement. A subset of 48 examples was selected from the test set to ensure the interpretability of prediction shifts. Each original input was paired with its two perturbed counterparts—one with an insertion and one with spelling distortion.

We evaluated the model using two key metrics. The first was **classification accuracy** on the clean test subset, used as a baseline for model performance. The second was **prediction consistency**, defined as the proportion of perturbed examples for which the model’s predicted label remained the same as its original prediction. This metric quantifies stability across input variants and serves as a proxy for robustness under noise.

Experiment 3: Cross-Lingual Generalization and Adaptation Testing

While prior experiments evaluated model performance on clean English symptom descriptions, this experiment investigates the ability of transformer-based models to generalize across languages by classifying Chinese symptom inputs. Building upon the final dataset curated and labeled in earlier stages, this section assesses both zero-shot generalization and supervised fine-tuning adaptation across language boundaries.

The motivation arises from real-world multilingual applications, where models must either generalize across languages without supervision or adapt efficiently with limited target-language data. Evaluating these conditions clarifies the limits of multilingual architectures and the advantages of language-specific pretraining.

To simulate cross-lingual conditions, the English symptom classification dataset was automatically translated into Chinese using Google Translate via the `deep_translator` library. We evaluated four distinct model configurations to systematically explore generalization and adaptation effects. First, a monolingual English BERT model (`bert-base-uncased`) fine-tuned on English data was evaluated directly on the translated Chinese test set under strict zero-shot conditions. Second, a multilingual BERT model (`bert-base-multilingual-cased`) pretrained on over 100 languages, including Chinese, was also fine-tuned on English and tested zero-shot on the Chinese inputs. Third, the multilingual BERT model was fine-tuned further on the translated Chinese dataset to measure gains from explicit supervised adaptation. Lastly, a Chinese-specific BERT model (`bert-base-chinese`)

pretrained solely on Chinese corpora was fine-tuned on the translated dataset to isolate the advantages of language-specific pretraining.

All models were evaluated on the same translated Chinese test set. Performance comparison was conducted using classification accuracy, macro-averaged F1 score, and confusion matrix analysis across the eight consolidated symptom categories. This design allows a direct comparison between strict zero-shot transfer and supervised fine-tuning, as well as between multilingual and monolingual pretraining strategies, providing insight into cross-lingual robustness in medical text classification.

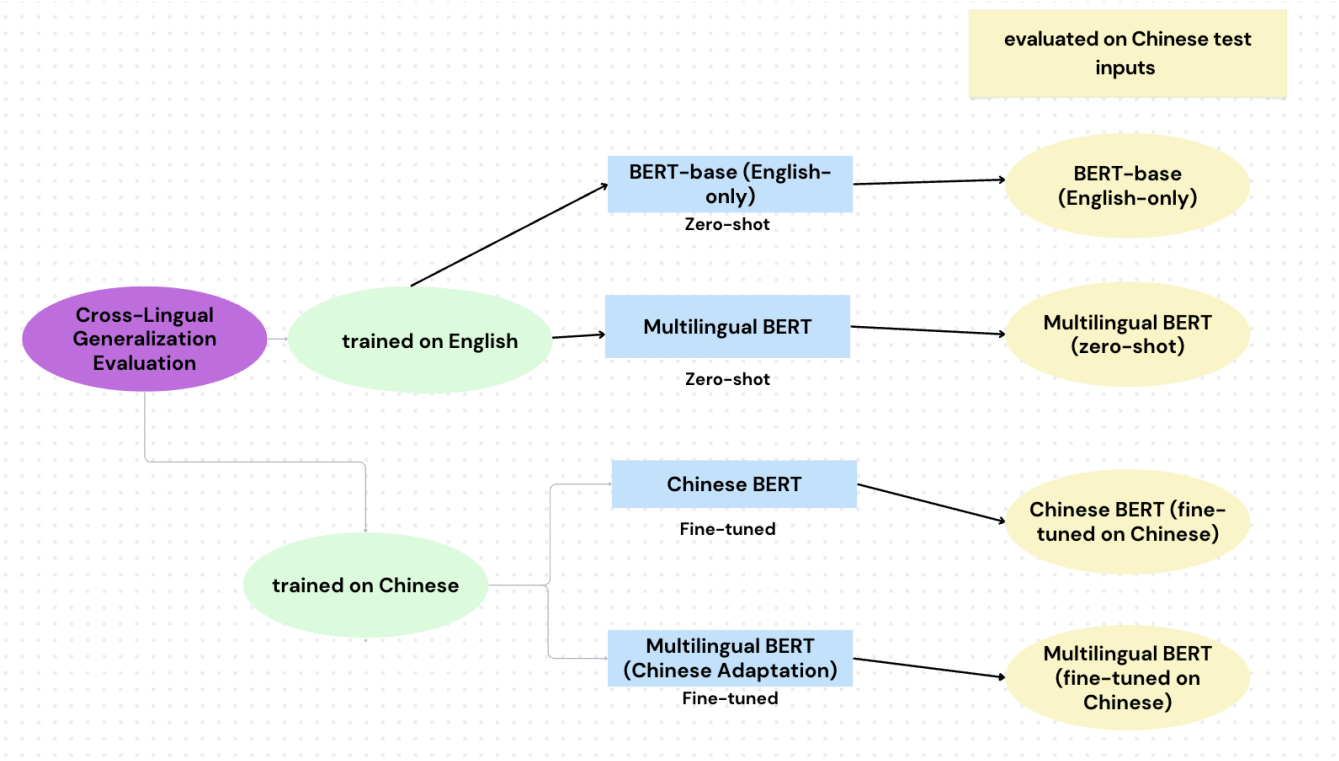


Figure 3. *Experimental Design for Cross-Lingual Generalization Testing*

Experiment 4: Embedding Analysis & Interpretation

To analyze our fine-tuned BERT base model's internal representations, we first extracted the 768-dimensional [CLS] embeddings for all 394 test sentences and reduced them to two dimensions with t-SNE. The resulting scatter plot shows natural clusters by medical specialty - such as distinct dermatology and neurology groups - confirming that the model encodes meaningful clinical distinctions. Any overlapping clusters or outliers immediately flag ambiguous or mislabeled examples, guiding us to refine labels or gather more data [37].

Next, we evaluated per-class precision, recall, and F1-scores on a held-out test set and examined the confusion matrix. This detailed view highlights strengths - like high recall for infections - and weaknesses - such as lower precision for chronic conditions. It also exposes systematic misclassification patterns (for example, chronic cases confused with neurological symptoms), which informs targeted error analysis: augmenting underperforming classes, adjusting class boundaries, or revisiting label definitions. By focusing on each specialty rather than overall averages, we gain a clearer understanding of category-level performance [38].

Finally, we used a dataset-level SHAP explainer to create a privacy-safe summary of token importance across the validation set. Aggregating SHAP values for each word, we generated a summary plot that reveals which terms most consistently drive model predictions without exposing individual sentences. This global feature-importance view, together with our clustering and per-class metrics, guides improvements to our tokenizer, emphasizes key clinical vocabulary in preprocessing, and uncovers potential biases in how the model weights different terms [39].

Experiment 5: Pre-Screening Integration on an end-to-end medical framework

For this experiment we propose a workflow consisting of the initial robust pre-screening model and two models to help patients better understand the results and considerations they have to take to improve their health. The overall workflow is presented below where in red is the main pre-screening model, in blue is a physician's attention (the actual visit to the doctor) and afterwards two models are proposed for users to better understand the implications of their diagnosis. One model is trained to be an expert on medical report summarization by taking the medical report provided by the physician and generating a semantically equivalent but shorter version of the text. The second model is trained to be an expert on Q&A on medical topics that patients may have.

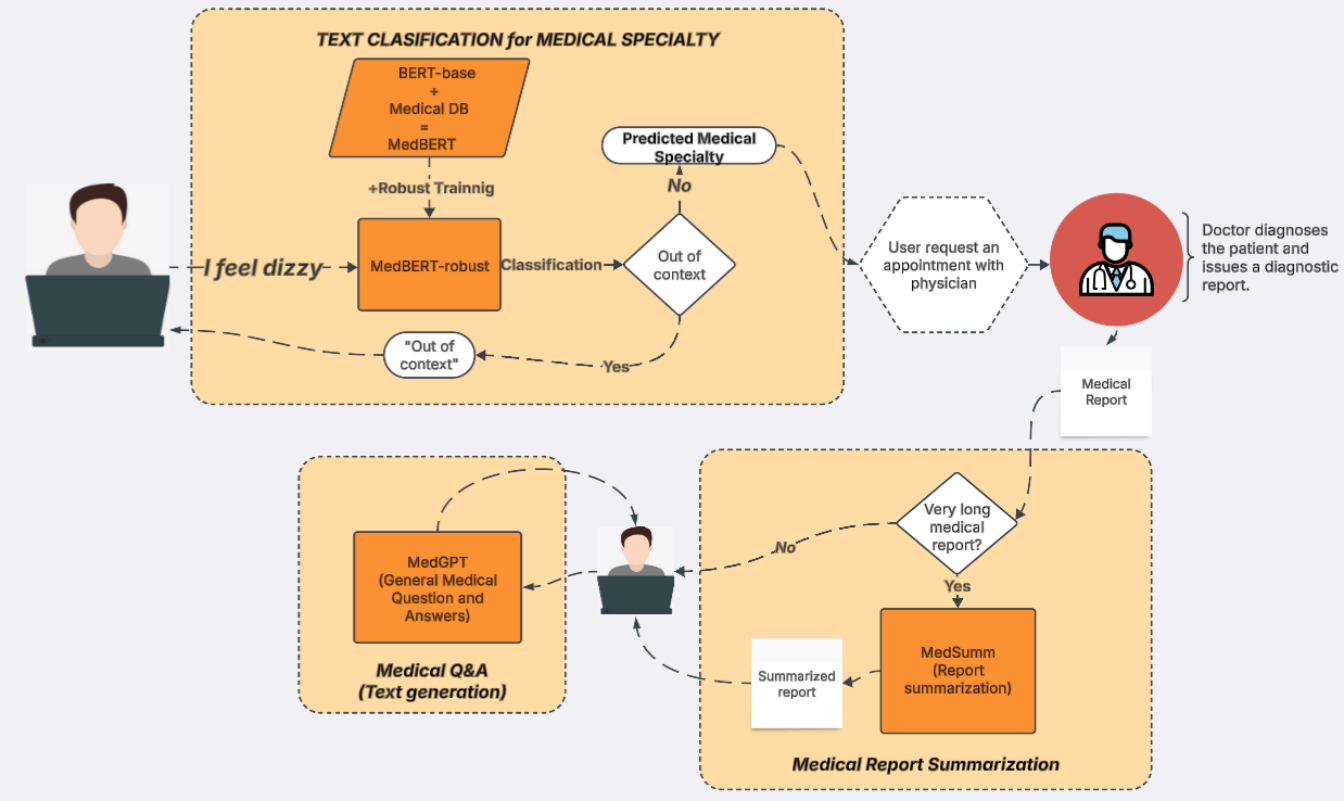


Figure 4. *Sequence of Patient interaction with the proposed workflow. Final models are orange rectangles with their corresponding names. 1. The patient prompts the MedBERT robust model to get the model's suggestion on what kind of specialty they should book an appointment with. 2. The patient then Makes an appointment with the predicted specialty. 3. The doctor diagnoses the patient and issues a report. 4.If the report is too long the user can submit the document to the MedSummarization model to get a summarized version of the report. 5. The user can ask questions about the diagnosis to MedGPT our proposed model for medical Q&A.*

5.1 Medical Report Summarization (MedSumm)

Using google's pre-trained encoder-decoder T5-small, a transformers based model, a summarization training task was performed. This model has as input a very long sequence of text and it is trained to extract the most relevant and semantically equivalent text effectively generating a summarized version of the original input. The model is proposed in this scenario to be used for patients to get the most important aspects of their diagnosis. Further mathematical specifications and architecture details can be found in ***appendix D.2***.

5.2 Medical Question and answer (MedGPT)

Following the previous task, a gpt2 model was fine tuned using Low Rank Adaptation for Large language models for sequence generation using a question and answer template. The dataset [34] contains pairs of general questions and answers scraped from reputed web sources on medical conditions and terminology. Further mathematical specifications and architecture details can be found in ***appendix D.3.***

Results

Experiment 1: Pre-trained Model Comparisons

Performance was assessed using accuracy, precision, recall, F1 score, and confusion matrices. After averaging results across five random seeds, RoBERTa and BERT-base achieved similar performance, with RoBERTa slightly outperforming BERT-base across all metrics. RoBERTa achieved the highest evaluation accuracy (0.9182 ± 0.0029), precision (0.9203 ± 0.0029), recall (0.9182 ± 0.0029) and F1 score (0.9183 ± 0.0029). BERT-base trailed slightly with lower mean values and higher standard deviations across all metrics compared to RoBERTa. In contrast, DistilBERT underperformed relative to both, with all its mean evaluation metrics falling substantially below the others and comparatively larger variability.

Table 1. *Mean and standard deviation of evaluation metrics (Accuracy, Precision, Recall, F1 Score) for pretrained models. Each model was trained and evaluated five times using different random seeds, and the mean and standard deviation across these runs were calculated for each metric.*

	Accuracy	Precision	Recall	F1 Score
BERT-base	(0.9091 ± 0.0150)	(0.9121 ± 0.0133)	(0.9091 ± 0.0150)	(0.9088 ± 0.0150)
RoBERTa	**(0.9182 ± 0.0029)**	**(0.9203 ± 0.0029)**	**(0.9182 ± 0.0029)**	**(0.9183 ± 0.0029)**
DistilBERT	(0.7990 ± 0.0160)	(0.8062 ± 0.0142)	(0.7990 ± 0.0160)	(0.7994 ± 0.0160)

Confusion matrices show that BERT-base and RoBERTa made few misclassifications and generally performed similarly, whereas DistilBERT struggled, with especially low performance accurately predicting the "Chronic Conditions" and "Allergy & Immune Reactions" categories. See Appendix B for BERT base and distilBERT confusion matrices.

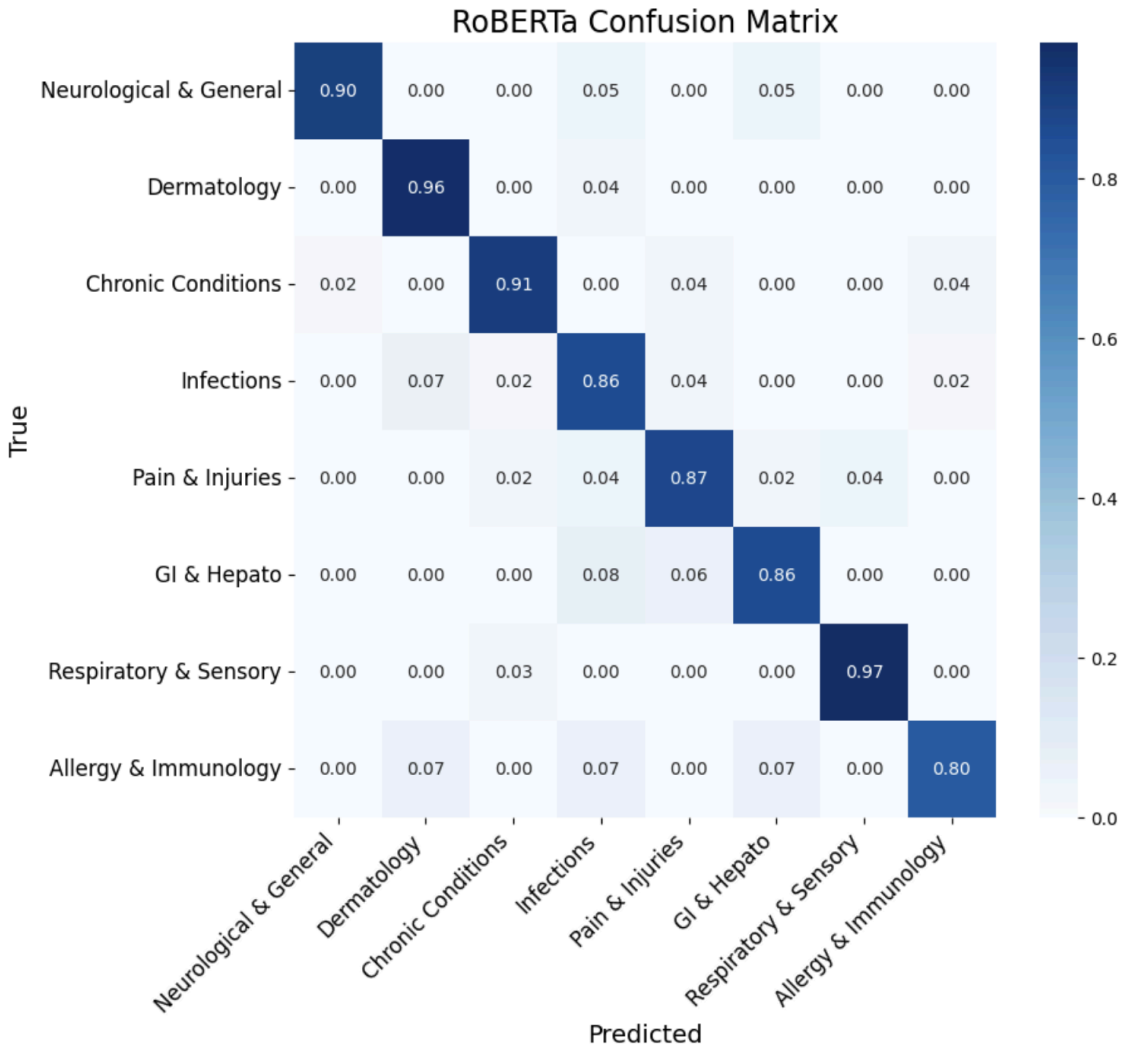


Figure 5. *RoBERTa Confusion Matrix*

Comparing the three models shows roBERTa as the strongest performer, closely followed by BERT base, while DistilBERT falls behind. For experiments 2-4, BERT-base is used as it is much less computationally expensive and is able to be trained and evaluated approximately 15x faster than roBERTa. BERT-base was originally selected as our baseline for use in the following experiments and performs similarly well to more complicated models as shown here.

Experiment 2: Robustness Testing

We evaluated model robustness by comparing two settings: baseline training without perturbations and robustness testing with controlled input noise. During robustness testing, adversarial noise was introduced to examine prediction stability and performance degradation.

The evaluation revealed a notable contrast between the model's overall predictive performance and its robustness to input perturbations. On the full clean test set, the fine-tuned model achieved high performance, with weighted F1 scores exceeding 0.9. However, for a robustness analysis, we focused on a smaller adversarial subset of the data specifically designed to introduce controlled perturbations. On this perturbed subset, the model’s baseline accuracy on the original (unperturbed) examples was

only 22.9%, reflecting the increased difficulty of classification under constrained sample diversity and imbalanced class distributions.

Despite the low absolute accuracy on this subset, the model exhibited substantial stability to input noise. We define prediction consistency as the proportion of examples where the predicted class remained unchanged after the input was perturbed. Under insertion attacks, prediction consistency was 81.3%, and under spelling noise perturbations, consistency reached 95.8%. These metrics suggest that although the model frequently misclassifies, it tends to misclassify consistently across similar inputs, indicating a degree of robustness to linguistic perturbations. However, it is important to note that prediction consistency, particularly when baseline accuracy is low, should be interpreted cautiously. High consistency does not necessarily imply high correctness; rather, it reflects that model outputs are stable under small perturbations, regardless of their correctness. In future work, a more nuanced robustness evaluation could involve progressively increasing the noise level and observing the corresponding impact on prediction accuracy, providing a clearer characterization of the model's tolerance to perturbations.

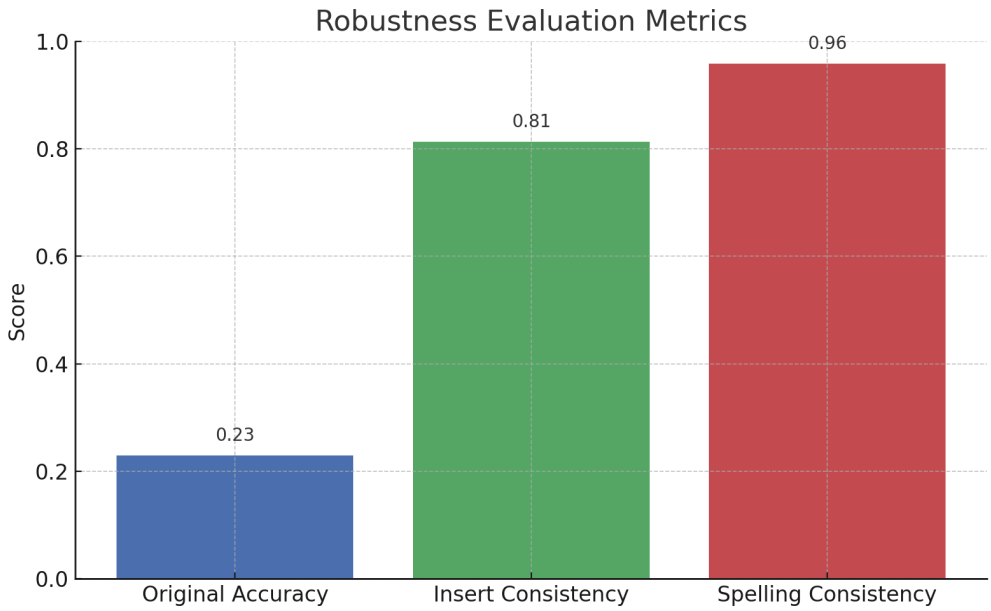


Figure 6. *Robustness evaluation metrics for the BERT-based classification model*

The chart compares model performance across three conditions: original test set accuracy (0.23), prediction consistency under insertion-based perturbation (0.81), and prediction consistency under spelling noise (0.96). Results indicate that while the model shows stability under minor perturbations, this stability must be contextualized by the low baseline accuracy. In particular, minor spelling errors had little effect on predicted labels, suggesting that the tokenizer and transformer layers are tolerant to orthographic variation. Insertion of unrelated phrases had a more pronounced impact, lowering consistency by approximately 15%, although prediction alignment was still preserved in the majority of cases. Overall, these results provide an initial view of model robustness, with the caveat that a broader range of noise intensities should be explored in future evaluations.

To gain a finer-grained understanding of how predictions shifted under perturbation, we examined confusion matrices for each of the three input conditions—clean, insertion-attacked, and spelling-distorted. These are visualized in the figures below.

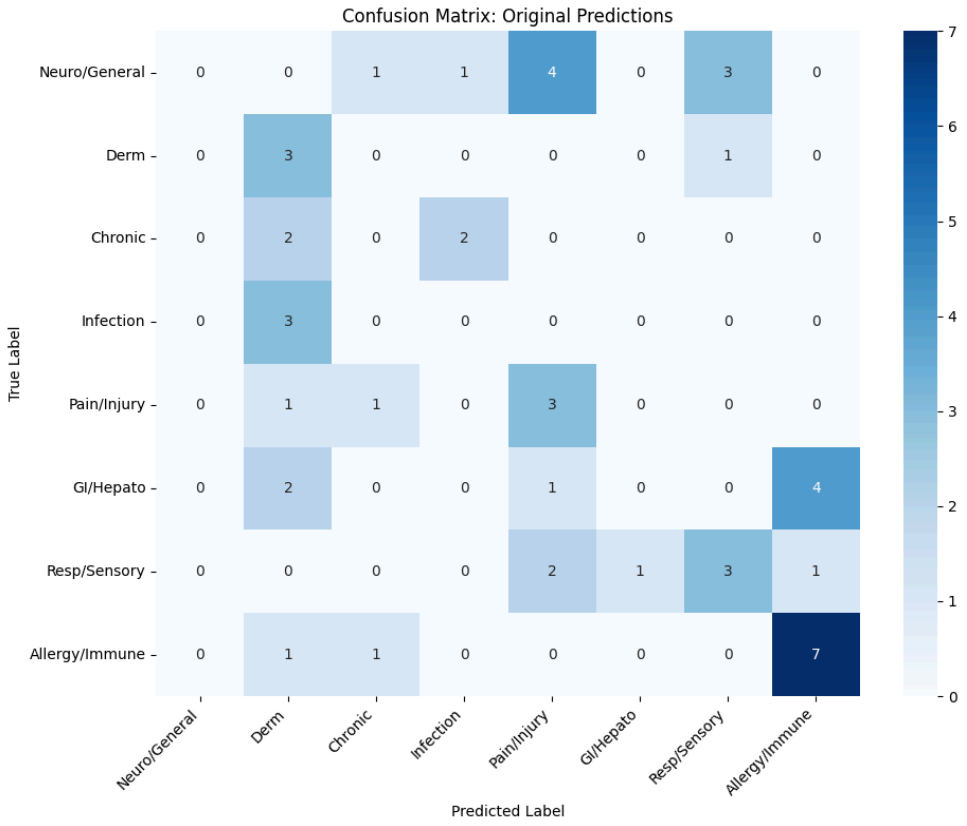


Figure 7. *Confusion matrix for original inputs*

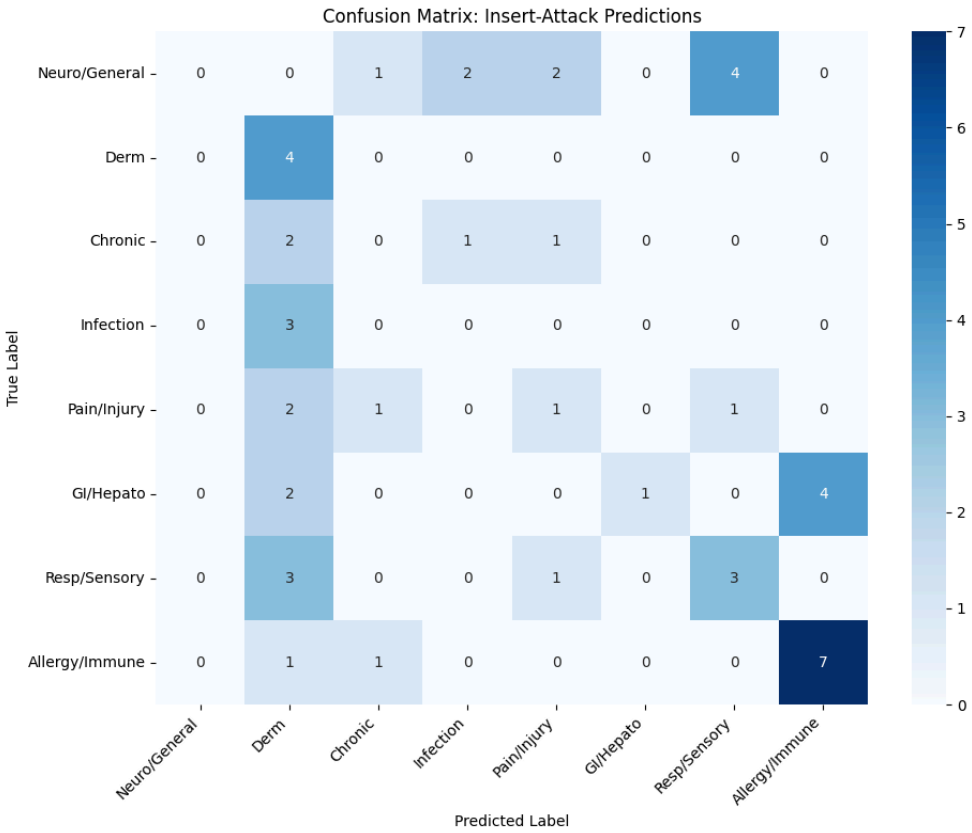


Figure 8. *Confusion matrix for insert-attack perturbed inputs*

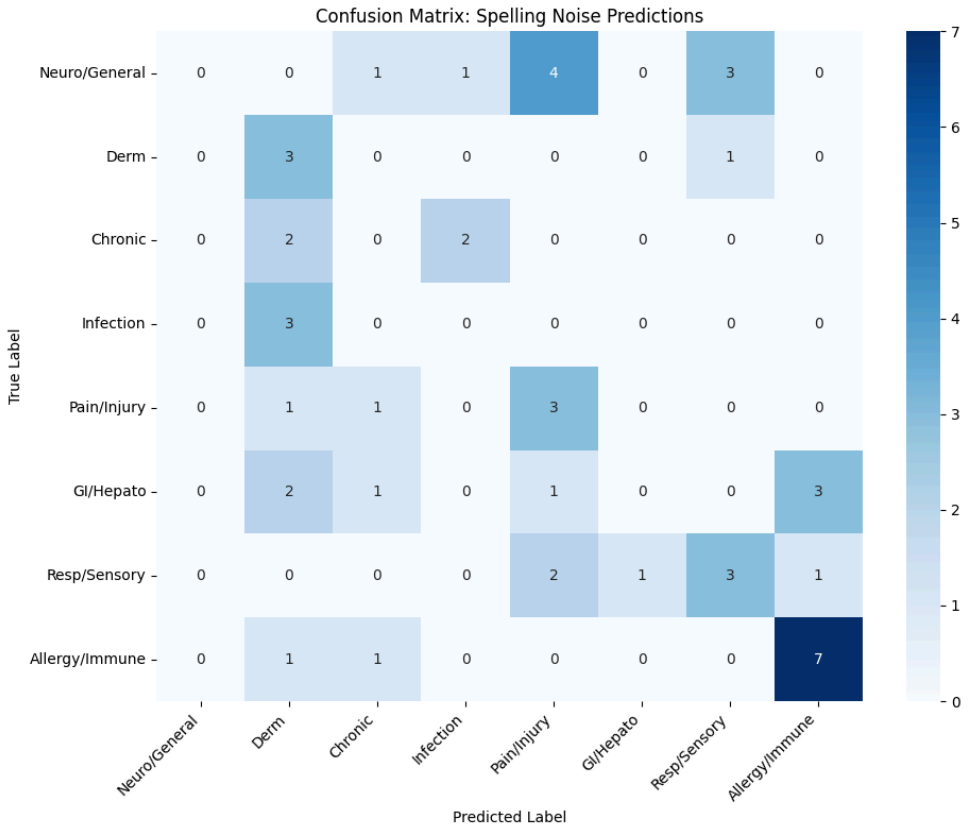


Figure 9. Confusion matrix for spelling-noise perturbed inputs.

Across all conditions, a subset of classes—particularly “Chronic Conditions” and “Dermatological & Skin Conditions”—were frequently misclassified or confused with one another. Under insert-attacked inputs, the model exhibited a mild tendency to overpredict “Neurological & General Symptoms,” likely due to its semantically ambiguous or generic nature. In contrast, predictions under spelling noise closely mirrored those of the original inputs, with minimal disruption to class distributions.

These results illustrate an important distinction: robustness does not necessarily imply correctness. The model is stable, but often confidently incorrect. This indicates that further improvements to classification accuracy are necessary before deployment, but that the current model architecture provides a promising foundation for real-world scenarios where input inconsistency is unavoidable. Future work may include adversarial training, augmentation with noisy data, or calibration methods to balance robustness and accuracy.

Experiment 3: Zero-shot Generalization Testing

Due to the worse performance of the Bert-base model, I added the Chinese-bert model which was specifically pre trained by chinese. The performance of the four evaluated models on the translated Chinese test set is summarized in Table 1. The Chinese-BERT model achieved the best results, with 94.05% accuracy and 93.96% F1. Fine-tuning Multilingual-BERT on translated Chinese data led to lower scores—likely due to the small size and limited quality of the translated dataset. Noise from machine translation and the challenge of both training and testing on synthetic data may have constrained the model's adaptation. BERT-Base, trained only on English, failed to generalize to Chinese inputs. These results emphasize the importance of selecting models pre-trained with multilingual corpora and, when feasible, further fine-tuning on in-domain language-specific data.

Table 1. Accuracy and F1 Score Comparison Across Models

Model	**Accuracy**	**F1 Score**
BERT-Base (Zero-shot)	0.0000	0.0000
Multilingual-BERT (trained on English)	0.9297	0.9293
Multilingual-BERT (fine-tuned on Chinese)	0.7757	0.7447
Chinese-BERT (fine-tuned on Chinese)	0.9405	0.9396

The training dynamics of the models were analyzed by tracking the validation loss across epochs. The learning curves indicate that the fine-tuned models, specifically Chinese-BERT and Multilingual-BERT fine-tuned on Chinese, exhibited steadily decreasing validation loss over the course of training. In contrast, BERT-Base and zero-shot Multilingual-BERT, evaluated without further training on Chinese data, maintained low but relatively static validation loss, as no fine-tuning updates were performed. Notably, Chinese-BERT demonstrated the most consistent improvement across epochs, reaching the lowest final validation loss among all models, aligning with its superior classification metrics. The trend observed in the learning curves supports the conclusion that fine-tuning significantly enhances model adaptation to new language domains, even when using a modest amount of translated training data.

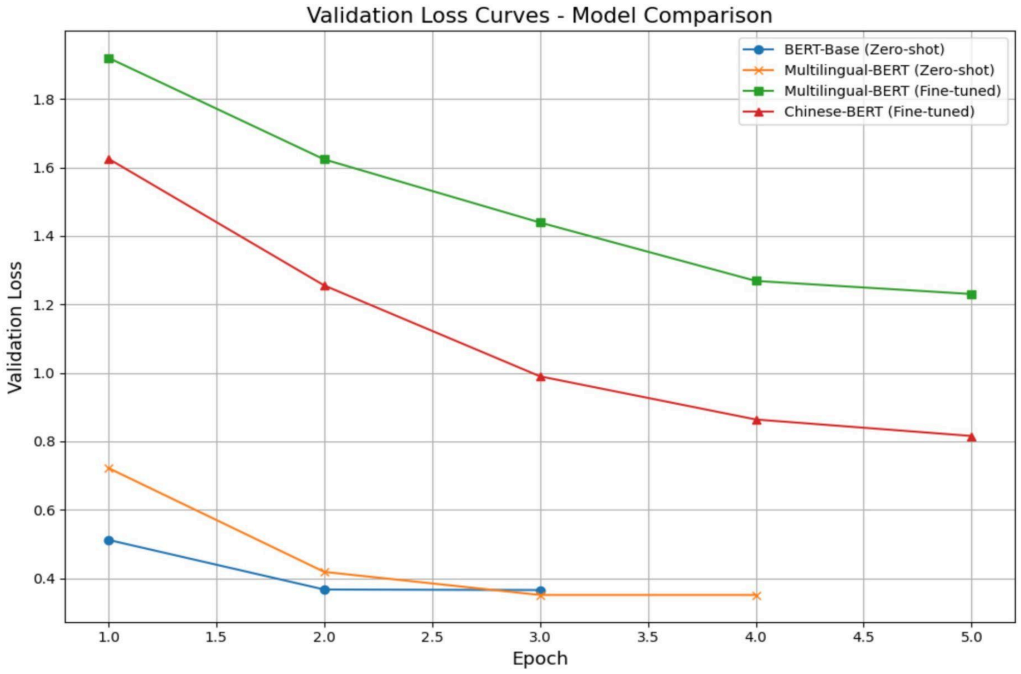


Figure 10. *Validation loss curves across epochs for four models evaluated on Chinese symptom classification.*

In confusion matrices (Appendix C), the Chinese-BERT fine-tuned model demonstrated the strongest overall performance, with nearly perfect diagonal alignment in its confusion matrix. Categories like *Chronic Conditions*, *Dermatological & Skin Conditions*, and *Respiratory & Sensory Issues* were classified with near-perfect precision and recall. Multilingual-BERT in the zero-shot setting also performed well, showing strong diagonal dominance with minor misclassifications. The confusion matrix indicates the model generalized across most categories, including *Infections* and *Chronic Conditions*, though slight overlap appeared in semantically close categories. Fine-tuning Multilingual-BERT on the translated Chinese data led to reduced performance, with more off-diagonal entries. Particularly, there was notable confusion in predicting *Neurological & General Symptoms*, *Allergic Reactions*, and *Gastrointestinal Conditions*, suggesting limited benefit from fine-tuning on noisy translated data. In contrast, the BERT-Base model failed to predict meaningfully in Chinese. Its confusion matrix was entirely empty, confirming 0% classification accuracy and F1-score, as it was neither multilingual nor trained on any Chinese input.

The evaluation results demonstrate that fine-tuning Chinese-specific pre-trained models such as Chinese-BERT yields the best performance for medical symptom classification in Chinese. Zero-shot transfer using multilingual models like Multilingual-BERT provides a viable alternative with high baseline accuracy, but fine-tuning remains critical for achieving optimal results. The overall findings underscore the value of language-specific pre-training and domain adaptation, particularly when translating tasks across languages with distinct linguistic characteristics.

Experiment 4: Embedding Analysis & Interpretation

The 2D t-SNE projection of [CLS] embeddings reveals distinct clusters for eight clinical categories (Figure 15).

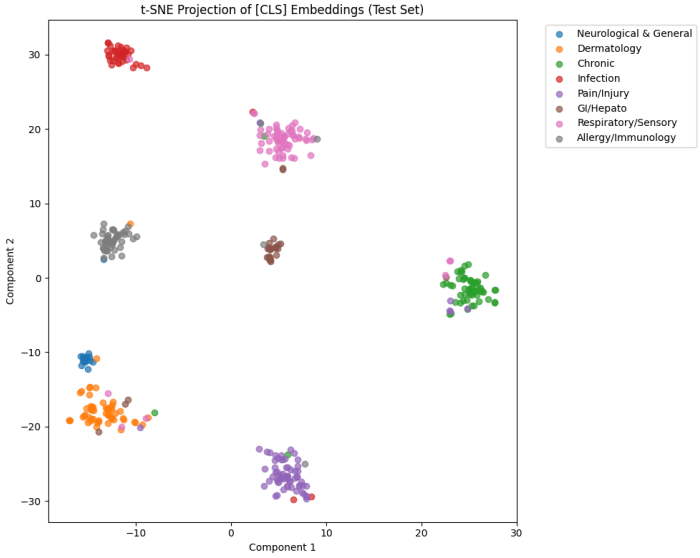


Figure 11. *Two-Dimensional t-SNE Projection of [CLS] Embeddings, Colour-Coded by Symptom Category*

Infections (red) and Allergy & Immunology (grey) occupy the upper-left, while Neurological & General (blue) and Dermatology (orange) group below. Gastrointestinal/Hepatobiliary (brown) centers around

the origin, Respiratory & Sensory (pink) forms a band above, and Pain/Injury (purple) clusters just below. Chronic Conditions (green) stand apart on the right. These separations confirm semantically distinct representations, though the close proximity of Neurological & General, Dermatology, and Allergy & Immunology suggests areas for further error analysis.

Table 3 reports class-level precision, recall, F1-score, and support, providing the quantitative complement to the qualitative structure observed in the t-SNE projection. Such a breakdown is essential for verifying that apparent cluster separations translate into concrete predictive gains and for identifying category-specific weaknesses that aggregate metrics can obscure.

Table 3. *Classification Report Showing Per-Class Precision, Recall, F1-Score, and Support for the Symptom-Category Model*

	precision	recall	f1-score	support
Neurological & General	0.7333	1.0000	0.8462	11
Dermatology	0.8644	0.9444	0.9027	54
Chronic	0.9167	0.9016	0.9091	61
Infection	0.8889	0.9231	0.9057	26
Pain/Injury	0.9298	0.8833	0.9060	60
GI/Hepato	0.9000	0.7826	0.8372	23
Respiratory/Sensory	0.8889	0.8511	0.8696	47
Allergy/Immunology	0.9375	0.9091	0.9231	33
accuracy	0.8952	0.8952	0.8952	0.8952
macro avg	0.8824	0.8994	0.8874	315
weighted avg	0.8983	0.8952	0.8952	315

Chronic Conditions achieve precision 0.92, recall 0.90 (F1 0.91), and Allergy & Immunology score P 0.94, R 0.91 (F1 0.92). Dermatology recalls 0.94 (precision 0.86, F1 0.90), while Infection balances P 0.89, R 0.92 (F1 0.91) and Pain/Injury P 0.93, R 0.88 (F1 0.91). GI/Hepato lags with R 0.78, P 0.90 (F1 0.84), and Respiratory & Sensory sits at P 0.89, R 0.85 (F1 0.87). Neurological & General, though perfectly recalled (1.00), has lower precision 0.73 (F1 0.85). Overall accuracy is 0.90 (macro F1 0.89, weighted F1 0.90), flagging GI/Hepato and Neurological & General for further tuning. Together, these numbers confirm that the model's well-separated embeddings translate into high, balanced performance - while flagging GI/Hepato and Neurological & General as prime candidates for further tuning.

The validation confusion matrix (Figure 16) given below is strongly diagonal, yet several systematic misclassifications persist.

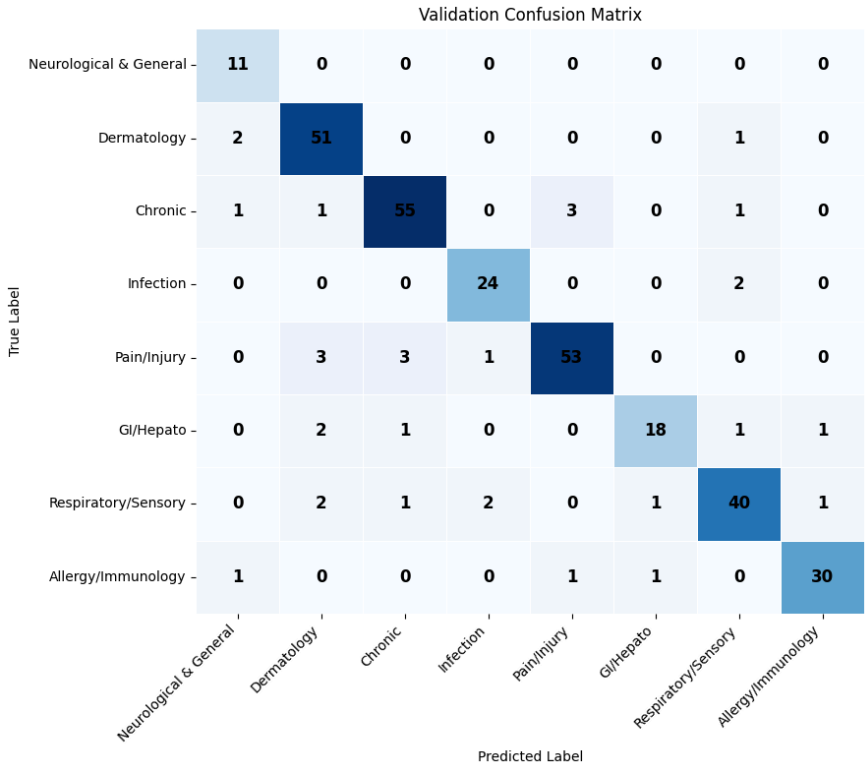


Figure 12. *Confusion Matrix of Predicted versus True Symptom Categories on the Validation Set*

The confusion matrix shows most specialties are well separated: Neurological & General is perfect (11/11) and Dermatology nearly so (51/54). Chronic Conditions correctly classifies 55/61 but leaks into Neurological & General, Dermatology, Pain/Injury, and Respiratory/Sensory. Infection labels 24/26 accurately, with two misrouted to Respiratory/Sensory. Pain/Injury gets 53/60 right but spills into Dermatology, Chronic, and Infection, while GI/Hepato captures 18/23 yet leaks into Dermatology, Chronic, Respiratory/Sensory, and Allergy/Immunology. These off-diagonal errors mirror t-SNE overlaps and pinpoint Pain/Injury and GI/Hepato for targeted data or boundary refinement.

The bar chart (Figure 17) provides a global view of lexical importance by plotting the mean absolute SHAP value for each token across the validation set. Longer bars mark words that most strongly shift the model's log-odds, regardless of direction. "vomit" leads, followed by "infected," "calves," and the WordPiece fragment "el," all above 1.8 mean |SHAP|. Close behind are "brownish," "stomach," and "heart," with symptom-related terms like "legs," "chill," and "nails" also prominent. Sub-token pieces such as "us," "ch," and "bel" rank highly, highlighting that even fragments can carry significant predictive weight. This view reveals which terms move the model most, though not their directional effect.

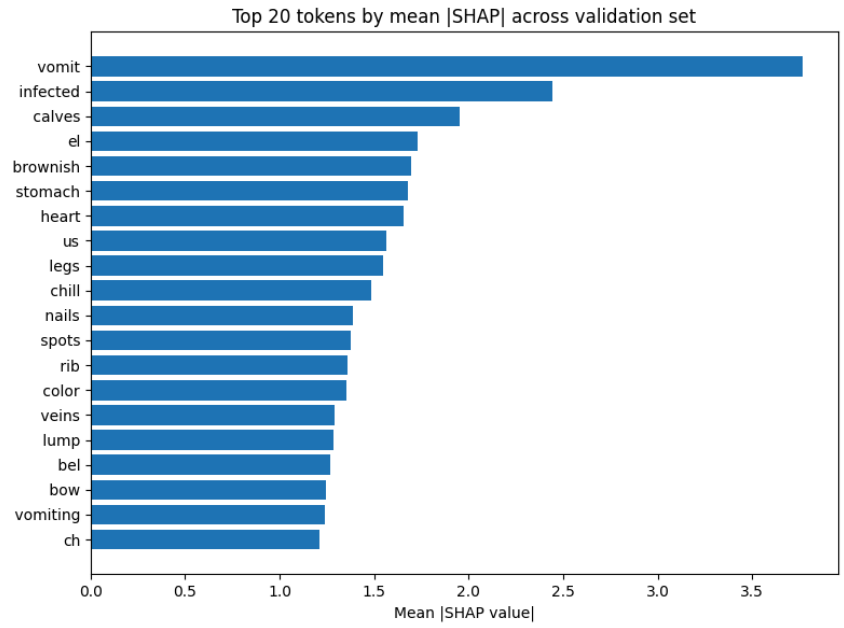


Figure 13. *Mean Absolute SHAP Values for the 20 Most Influential Tokens in the Validation Set*

The dot-plot in Figure 18 lays out every SHAP value for our top twenty tokens, with values to the left of zero pulling the model's confidence down and those to the right boosting it. Tokens like "vomit," "el," "brownish," "stomach," "chill," "nails," "spots," "bel," "bow," "vomiting," and "ch" cluster almost entirely left, acting as consistent suppressors. In contrast, "heart," "lump," "us," and "veins" lie mostly to the right, serving as reliable positive indicators.

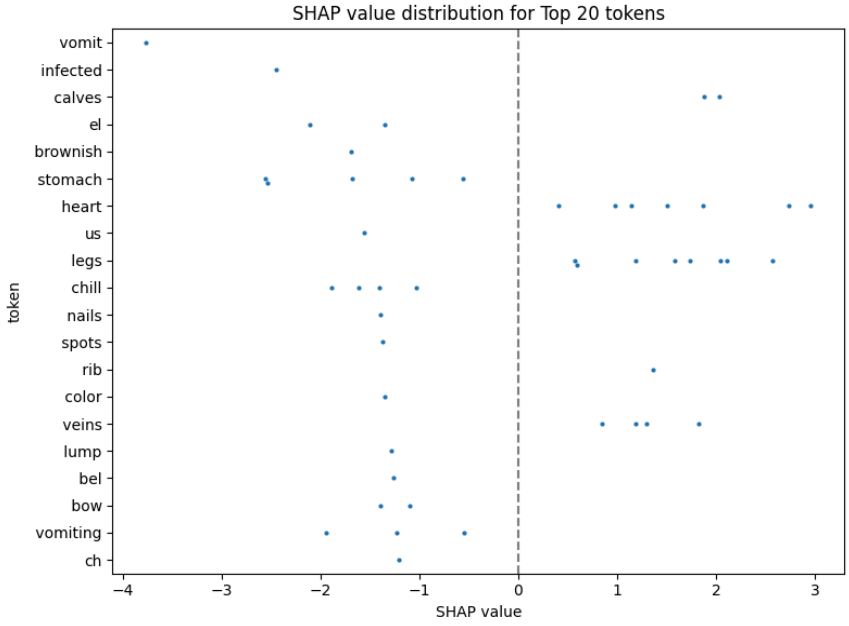


Figure 14. *Signed SHAP Value Distribution for the 20 Most Influential Tokens*

Tokens such as "infected," "calves," "legs," "rib," and "color" scatter on both sides, indicating context-dependent effects, with outliers revealing rare role reversals. Together with the bar chart, this

fine-grained SHAP distribution confirms the model's reliance on meaningful medical cues and pinpoints which tokens and contexts warrant further scrutiny or targeted augmentation.

Experiment 5: End-to-end integration

Medical Report Summarization

Figure 18 reports the results of the training sequence of medical summarization using the standard metrics for text summarization, ROUGE (Recall-Oriented Understudy for Gisting Evaluation scores) and BLUE score. A detailed mathematical formulation of these metrics can be found on **appendix D2**. The model achieved an evaluation loss of 1.0714, indicating a moderate degree of error in prediction. In terms of content overlap with reference summaries, the model obtained ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores of 0.7961, 0.6556, 0.7681, and 0.7673, respectively. These metrics suggest a relatively high level of lexical and structural similarity between the generated and reference summaries, particularly for unigram and longest common subsequence matches. Additionally, the model's BLEU score was 0.6571, reflecting a reasonable degree of n-gram precision, although it may also suggest some limitations in grammatical or semantic accuracy.

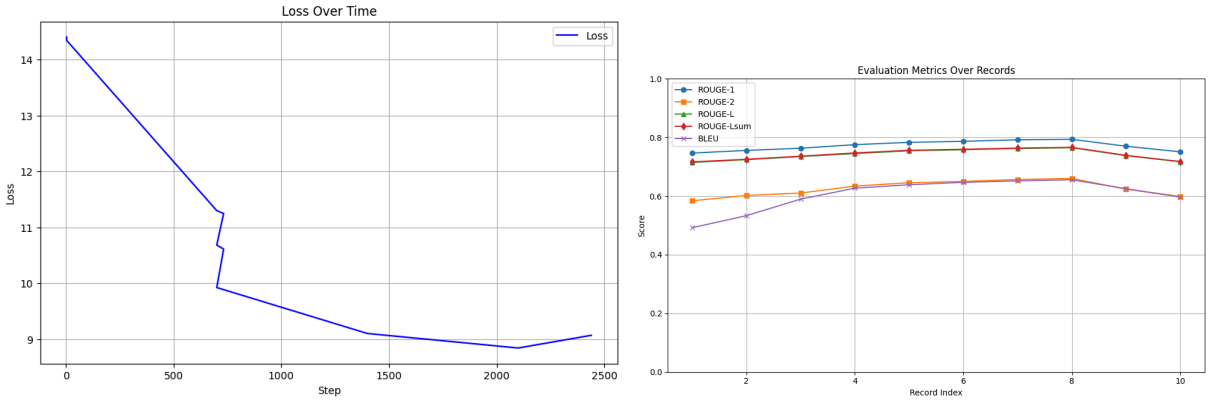
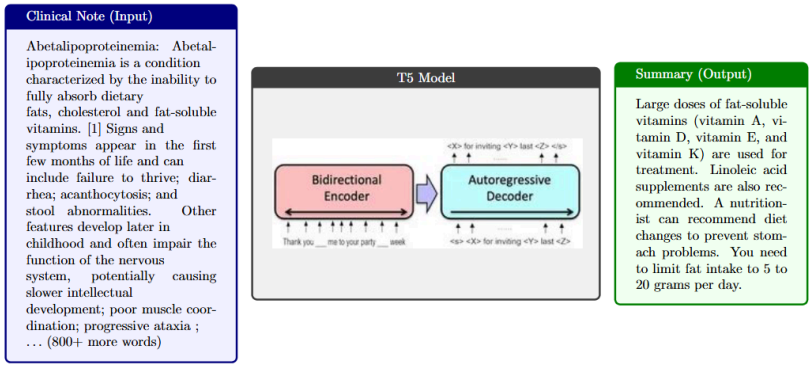


Figure 15. *Evaluation of performance metrics and loss over time.*

The model is achieving good performance, but starts to overfit at epoch 8. This is a clear indication that, while the model has a strong capability of learning the patterns required from this experiment, more data is required.

Compared to the results from the paper "Text Summarization of Medical Documents using Abstractive Techniques" [43] which uses a T5 model and a BART with the SUMPUBMED dataset to train a summarization task, it can be evaluated that our implementation improves on previous implementations.

Performance metrics	ROUGE-1	ROUGE-2	ROUGE-L
T5 small from [43] trained on SUMPUBMED	0.127168	0.027778	0.259259
BART large cnn from [43] on SUMPUBMED	0.306358	0.114583	0.351724
Our T5 trained on [33]	**0.7961**	**0.6556**	**0.7681**



Medical Question and answer

The generative question answering model fine-tuned via supervised instruction tuning (SFT) was evaluated after 10 epochs of training. The metrics for generative sequence to sequence tasks used are perplexity, which roughly corresponds to how diverse the model's text generation is, the lower the better, and ROUGE presented in the section above. The evaluation of perplexity shows a metric of 9131.40, indicating challenges in model calibration and output fluency. In terms of output quality, the model achieved a ROUGE-1 score of 0.269, ROUGE-2 of 0.055, ROUGE-L of 0.198, and ROUGE-Lsum of 0.228, reflecting limited overlap with reference answers at unigram, bigram levels, and the longest common subsequences. Additionally, the BLEU score was 0.028, suggesting low n-gram precision in generated responses. Overall the accuracy on Q&A question and answering ("exact match") was 25% which indicates clear opportunities for improvement on the quality of data.

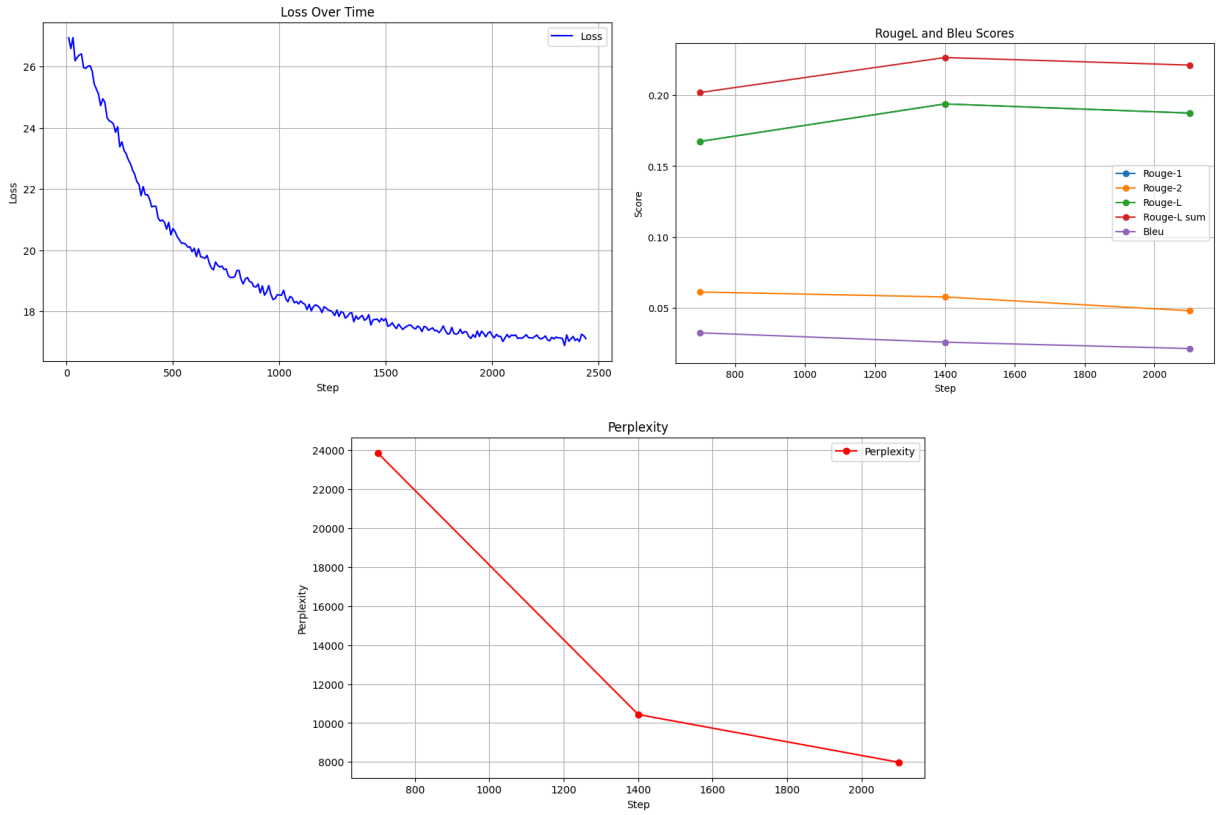


Figure 16. *Evaluation over the steps for the model training. The model shows a clear convergence over minima for loss over time, while achieving reasonable results in ROUGE-L score (increasing). The model is adjusting to the desired data distribution as shown by the perplexity metric*

In contrast with the measures shown above, a State of the art model for Medical question and answering, Google's Med-PALM [44] shows a better performance with a model that has 540 billion parameters. This also indicates that the performance could be improved by using a larger architecture as our experiment is based on the GPT2 base model, a 117 Million parameter architecture, 0.02% the size of the State of the art model. A ***example case for a Q&A*** can be found in ***appendix D.3, figure A.3.2.***

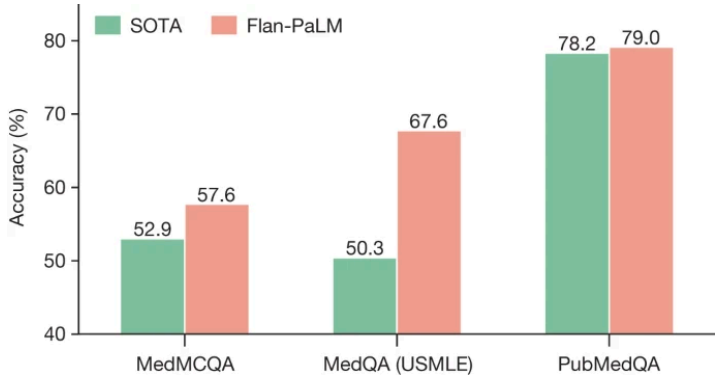


Figure 6. Results from Flan-PALM, Google's Medical Q&A generative model. Figure extracted from Link

Ethical Considerations

The main ethical concern of our project is the situation in which it is used as the sole decider in patients diagnosis, as opposed to a pre-screening aid as was originally designed. There are certain limitations of our dataset that may lead to inaccurate results and threats to external validity. Additionally, the potential biases in the dataset might disproportionately affect certain demographic groups, leading to disparities in diagnostic outcomes. Our dataset being relatively small and the gathering methods of the data not being explicitly stated lead to unconfidence in the data being representative of the broader population, possibly leading to bias. Incorrect pre-screen labeling can lead to wasted time and frustration for both patients and physicians. Beyond wasted time, incorrect predictions could delay critical medical interventions, potentially leading to worsened health outcomes or increased anxiety for patients.

We explicitly acknowledge that due to the delicate and high-risk nature of medical diagnosis, our model should be used to aid in physician decision-making as opposed to replacing it. The data we plan to use has been completely anonymized from patient personal information, mitigating privacy concerns. Although the model we plan to create is a black box model, raising concerns around explainability, our experiments involving dimensionality reduction and feature importance are intended to help make the model's decision-making process easier to understand. To enhance interpretability, we have provided necessary visualizations, such as heatmaps and feature importance rankings, that could assist physicians in understanding why a certain prediction was made.

Conclusions

This study evaluated transformer-based models for medical symptom classification across diverse settings. RoBERTa achieved the highest accuracy, but BERT-base was selected for further experiments due to its comparable performance and much greater computational efficiency. Robustness testing revealed that even strong models are sensitive to input noise. Cross-lingual evaluation showed that multilingual pretraining is essential for transferring to non-English inputs, with Chinese-BERT fine-tuning achieving the best results. Embedding analyses confirmed meaningful symptom clustering, while

training only on simple examples reduced generalization. Overall, the findings emphasize the need for smart architectural choices, multilingual adaptation, and diverse training data to build clinically robust NLP systems.

Roles

- **Sizhe Chen:** I was primarily responsible for **Experiment 3**, which investigated zero-shot cross-lingual generalization using multilingual transformer models. This involved translating the test set from English to Chinese, fine-tuning both monolingual and multilingual models on English data, evaluating their performance on Chinese inputs, and visualizing the results using confusion matrices and per-class metrics. I also contributed to the overall performance analysis by interpreting the multilingual model's ability to generalize across languages. In addition, I took the lead in writing the **Conclusion** section of the report, synthesizing key findings from all five experiments into a cohesive summary that highlighted the broader implications of our work and suggested future research directions.
- **Kayla Haeussler:** I cleaned and consolidated our primary dataset used in all experiments, which consisted of preprocessing and manually mapping the datasets' original labels to broader categories to create the final dataset (based on my own web research on medical conditions and my own discretion on picking categories based on what was currently listed in the data). I was responsible for Experiment 1, including training the three models, evaluating results through confusion matrices and accuracy metrics, and performing manual hyperparameter selection. Additionally, I organized and packaged the BERT-base implementation into a clean pipeline in a separate Jupyter notebook for my teammates. I structured and designed our slideshow presentation. In the report, I contributed to the writing of the data section, ethical considerations as well as edits throughout the report and any section relating to experiment 1.
- **Ramil Mammadov:** I spearheaded **Experiment 4: Embedding Analysis and Interpretability,** where I designed and implemented a four-stage methodological framework to probe our fine-tuned BERT-base model's internal representations. My responsibilities included enabling and visualizing multi-layer attention patterns with BertViz; extracting and projecting 768-dimensional [CLS] embeddings via t-SNE; developing local feature-attribution analyses using SHAP's KernelExplainer and Captum's Integrated Gradients; and constructing the end-to-end inference pipeline within Hugging Face's pipeline("text-classification"). I produced all related figures (4.1-4.6), authored the Methods and Results write-ups, and interpreted the clustering, attention dynamics, and attribution findings to demonstrate the model's explainable performance on unseen clinical statements.
- **Alejandro Paredes La Torre:** Experiment 5, Summarization model using a T5 encode-decoder model and GPT2 model for Medical Q&A on its entirety. I wrote the introduction part, the Related work and partially the datasets part. All mathematical formulations found in the appendix sections were written in latex on overleaf and pasted in this document. They were extracted from the different papers that they reference to.
- **Zihan Xiao:** I was responsible for designing and implementing the robustness testing experiment. I fine-tuned the BERT-base model using PyTorch, generated adversarial inputs with insertion and spelling noise, and evaluated prediction consistency across perturbed samples. I also created visualizations (e.g., bar plots, confusion matrices) and wrote the corresponding report section with results interpretation.

References

[1] Jiang, F., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243.

[2] Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731.

[3] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5, 8869–8879.

[4] Manogaran, G., & Lopez, D. (2018). Health data analytics using scalable logistic regression with stochastic gradient descent. *International Journal of Advanced Intelligence Paradigms*, 10(1), 118–132.

[5] Keerthika, T., & Premalatha, K. (2019). An effective feature selection for heart disease prediction with aid of hybrid kernel SVM. *International Journal of Business Intelligence and Data Mining*, 15(3), 306–326.

[6] Sadek, R. M., et al. (2019). Parkinson's disease prediction using artificial neural network. *International Journal of Academic Health and Medical Research*, 3(1), 1–8.

[7] Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*.

[8] Abdulmohsin, H. A., Al-Khateeb, B., Hasan, S. S., & Dwivedi, R. (2022). Automatic illness prediction system through speech. *Computers and Electrical Engineering*, 102, 108224.

[9] Habib, M., Faris, M., Qaddoura, R., Alomari, A., & Faris, H. (2021). A predictive text system for medical recommendations in telemedicine: a deep learning approach in the Arabic context. *IEEE Access*, 9, 85690–85708.

[10] Weng, W. H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, 17(1), 1–13.

[12] Al-Shawwa, B., Glynn, E., Hoffman, M. A., Ehsan, Z., & Ingram, D. G. (2021). Outpatient health care utilization for sleep disorders in the Cerner Health Facts database. *Journal of Clinical Sleep Medicine*, 17(2), 203–209.

[13] Kulaylat, A. S., Schaefer, E. W., Messaris, E., & Hollenbeak, C. S. (2019). Truven health analytics MarketScan databases for clinical research in colon and rectal surgery. *Clinics in Colon and Rectal Surgery*, 32(1), 54–60.

[14] Schriml, L. M., et al. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1), D940–D946.

[15] Yagnik, N., Jhaveri, J., Sharma, V., & Pila, G. (2024). MedLM: Exploring language models for medical question answering systems. *arXiv preprint arXiv:2401.11389*.

[19] Zhang, Y., Warstadt, A., Li, H. S., & Bowman, S. R. (2020). When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.

[20] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.

[21] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186).

[22] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[23] Lee, J., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.

[24] Alsentzer, E., et al. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 72–78).

[25] Liu, Y., Ge, T., Mathews, K. S., Ji, H., & McGuinness, D. L. (2018). Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. *arXiv preprint arXiv:1804.04225*.

[26] Minarro-Giménez, J. A., Marin-Alonso, O., & Samwald, M. (2014). Exploring the application of deep learning techniques on medical text corpora. In *e-Health – For Continuity of Care* (pp. 584–588). IOS Press.

[29] Mooney, P. T. (2018). Medical Speech, Transcription, and Intent [Data set]. *Kaggle*. https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent

[30] Barman, N. R., & Collaborators. (2023). Symptom2Disease [Data set]. *Kaggle*. https://www.kaggle.com/datasets/niyarrbarman/symptom2disease

[31] QuyenAnhDE. (2023). Diseases_Symptoms [Data set]. *Hugging Face*. https://huggingface.co/datasets/QuyenAnhDE/Diseases_Symptoms

[33] Savery, M., Abacha, A. B., Gayen, S., & Demner-Fushman, D. (2020). Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1), 322.

[34] Han, T., et al. (2023). MedAlpaca—An open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.

[35] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? In *Proceedings of ACL*.

[36] Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *ACL*.

[37] Van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), 2579–2605..

[38] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[39] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

[42] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45).

[43] Lalitha, E., Ramani, K., Shahida, D., Deepak, E. V. S., Bindu, M. H., & Shaikshavali, D. (2023, May). Text summarization of medical documents using abstractive techniques. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 939-943). IEEE.

[44] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

[45] Deep Translator Developers. (2024). *deep-translator* (Version x.x) [Python Package]. https://pypi.org/project/deep-translator/

[46] Belinkov, Y., & Bisk, Y. (2018). *Synthetic and natural noise both break neural machine translation*. Proceedings of the 6th International Conference on Learning Representations (ICLR). https://aclanthology.org/N18-1170/

[47] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). *Beyond accuracy: Behavioral testing of NLP models with CheckList*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 4902–4912. https://aclanthology.org/2020.acl-main.740/

[48] Michel, P., Neubig, G., & Levy, O. (2019). *Are sixteen heads really better than one?* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 1406–1410. https://aclanthology.org/W19-4104/

Appendices

Appendix A: Data Manipulation and Consolidation

The merging process required careful handling due to differences in label specificity across datasets. While the symptom description formats were relatively consistent, the associated condition labels varied significantly in granularity. To address this, the original labels were manually mapped by a team member into eight broader, high-level categories. An example of this consolidation is shown in the table below. This step was motivated by the understanding that model performance often improves when the label space is simplified and better balanced.

Table 1. *Overview of source datasets and example label mapping.*

	Dataset 1	Dataset 2	Dataset 3
Name	Medical Speech Transcription and Intent from Kaggle	Symptom 2 Disease from Kaggle	Disease Symptoms from Hugging Face
# of Labels	25 labels	24 labels	392 labels
# of Rows	706 rows Initially 6,661 rows, however only 706 of those rows contained unique text values	1200 rows	400 rows
Example Symptom Text/Label Pairs			
Example Symptom Text/Label Pairs	“I have a cut that has become red and oozes puss.”	“I have raised lumps, a rash that looks red and inflamed, discoloured areas of skin that are different colours from the rest of my skin, and itching on my skin.”	“Small, painful blisters on the lips or around the mouth”
Original Label	Infected Wound	Fungal Infection	Cold Sore
Final Label	**Infections**		

After the manual mapping to broader labels, the final dataset had the following value counts for each label, exhibiting mild variation in class sizes but no severe imbalance. The largest category ("Infections") contained 368 examples, while the smallest category ("Allergic/Immunologic Reactions") contained 101 examples. Although some variation in sample size existed, all categories had sufficient representation to allow effective model training without applying oversampling or class weighting techniques.
For additional detail on which categories from the original dataset fell under each of the labels in our final dataset, please consult our project GitHub repository.

Table 2. *Final Dataset Label Value Counts*

Label	Count in Final Dataset
Infections	368
Dermatological & Skin Conditions	346
Chronic Conditions	310
Pain & Injuries	298
Respiratory & Sensory Issues	242

Gastrointestinal & Hepatobiliary Conditions	180
Neurological & General Symptoms	131
Allergic/Immunologic Reactions	101

Additionally, the distribution of token length across the 3 datasets was considered. As shown below, Dataset 1 primarily contains short symptom descriptions, with most samples between 5 and 20 tokens. In contrast, Dataset 3 has considerably longer descriptions, typically between 30 and 60 tokens. Dataset 2 was in the middle but had a much smaller sample size. While the datasets varied in the level of detail in their symptom descriptions, all followed a natural language style reflective of how a patient might communicate symptoms to a doctor. Because BERT-based models are designed to handle variable-length inputs, and padding was applied consistently during tokenization, differences in average token length were not expected to adversely impact model performance.

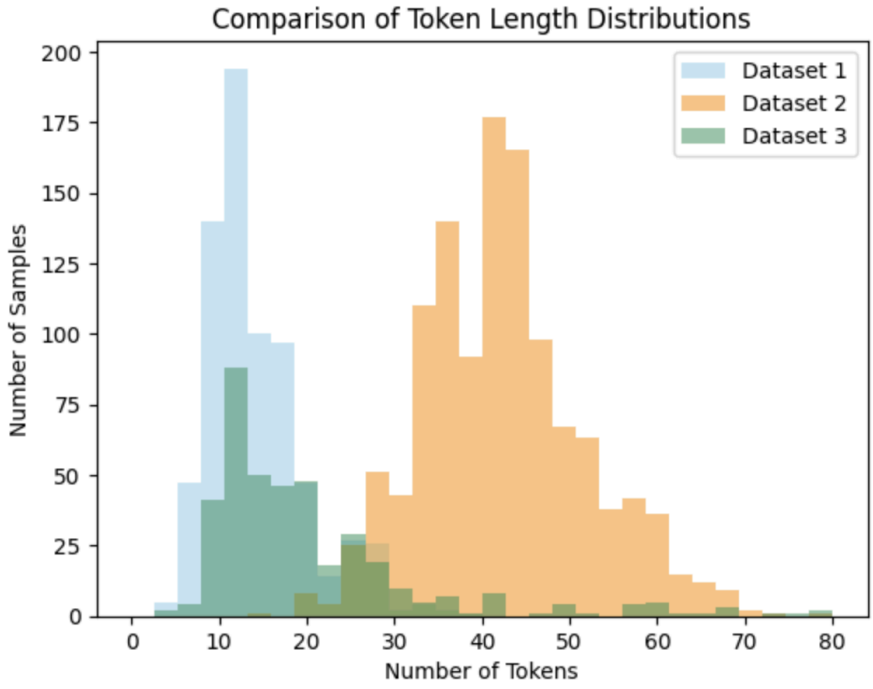


Figure 3. *Distribution of Token Length for Patient Phrases in Each Original Dataset*

Appendix B: Experiment 1, Pre-Trained Model Comparison

Hyperparameter Selection

Each model was fine-tuned using the Hugging Face Trainer API with consistent weight decay of 0.01, and model-specific numbers of epochs selected based on observed validation performance. BERT-base and RoBERTa both achieved early convergence, stabilizing after approximately 5 epochs without signs of overfitting. DistilBERT showed slower improvement and was allowed to train for 8 epochs to maximize performance before plateauing. For additional information on hyperparameter selection, see Appendix B.

Weight decay was applied to introduce L2 regularization, helping prevent overfitting given the relatively small size of our dataset. A value of 0.01 was selected following Hugging Face recommendations and

common practice in transformer fine-tuning, providing moderate regularization without impairing the model's ability to adapt to the classification task. Performance was assessed on the test set using accuracy, precision, recall, and F1-score, with additional insight from confusion matrices to highlight misclassification patterns. This setup allowed for a fair comparison between models with different capacities and training schemes—BERT as a strong baseline, RoBERTa for improved contextual representation, and DistilBERT for efficient inference with minimal performance loss.

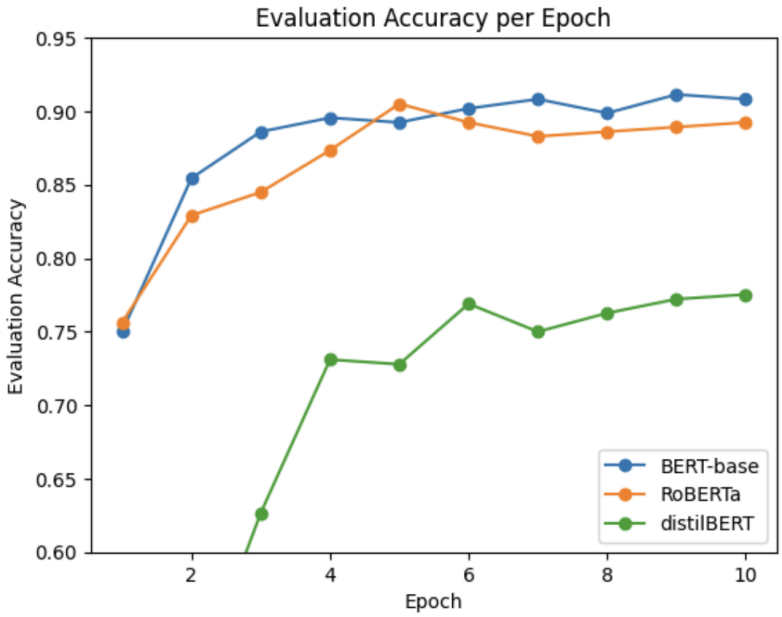


Figure 1. *Evaluation accuracy per epoch for BERT-base, RoBERTa, and DistilBERT. Each model was fine-tuned on the same dataset with consistent hyperparameters. This plot was used to select the optimal number of training epochs for each model by observing where validation accuracy plateaued.*

Confusion Matrices

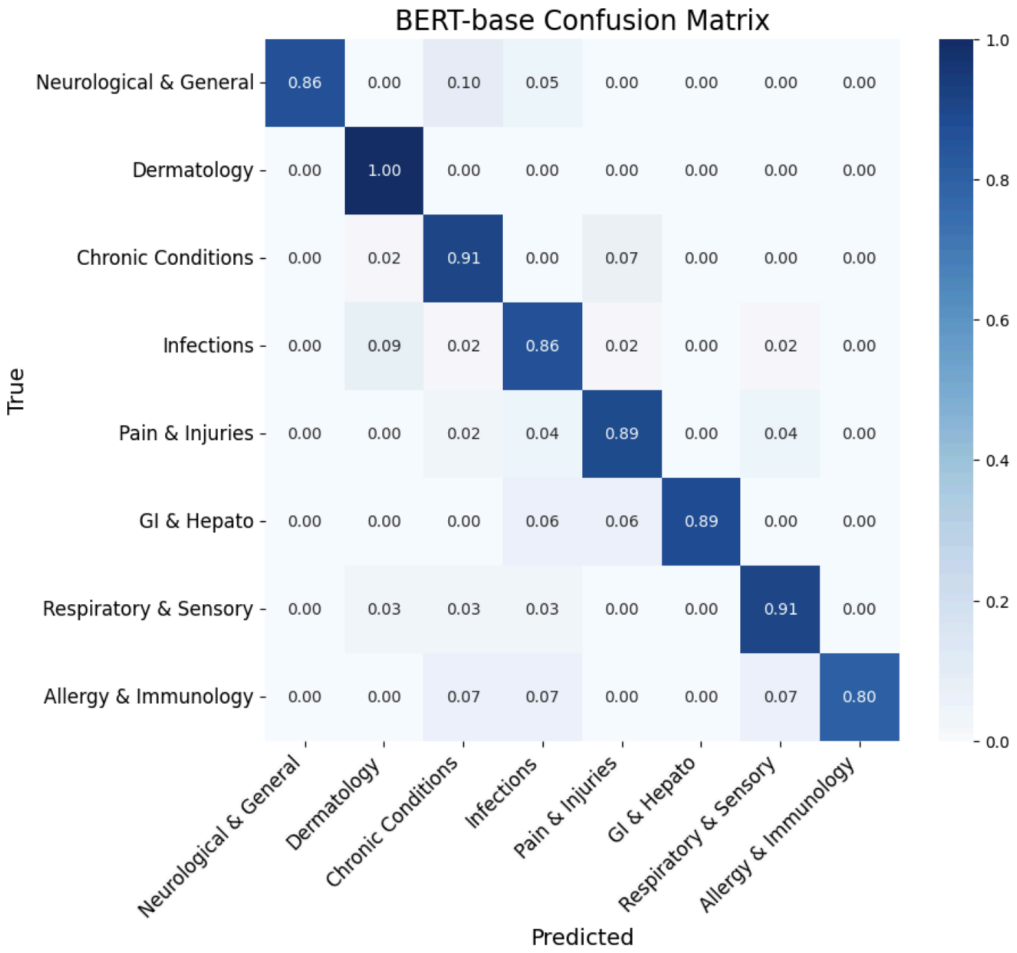


Figure 1. *BERT-base Confusion Matrix*

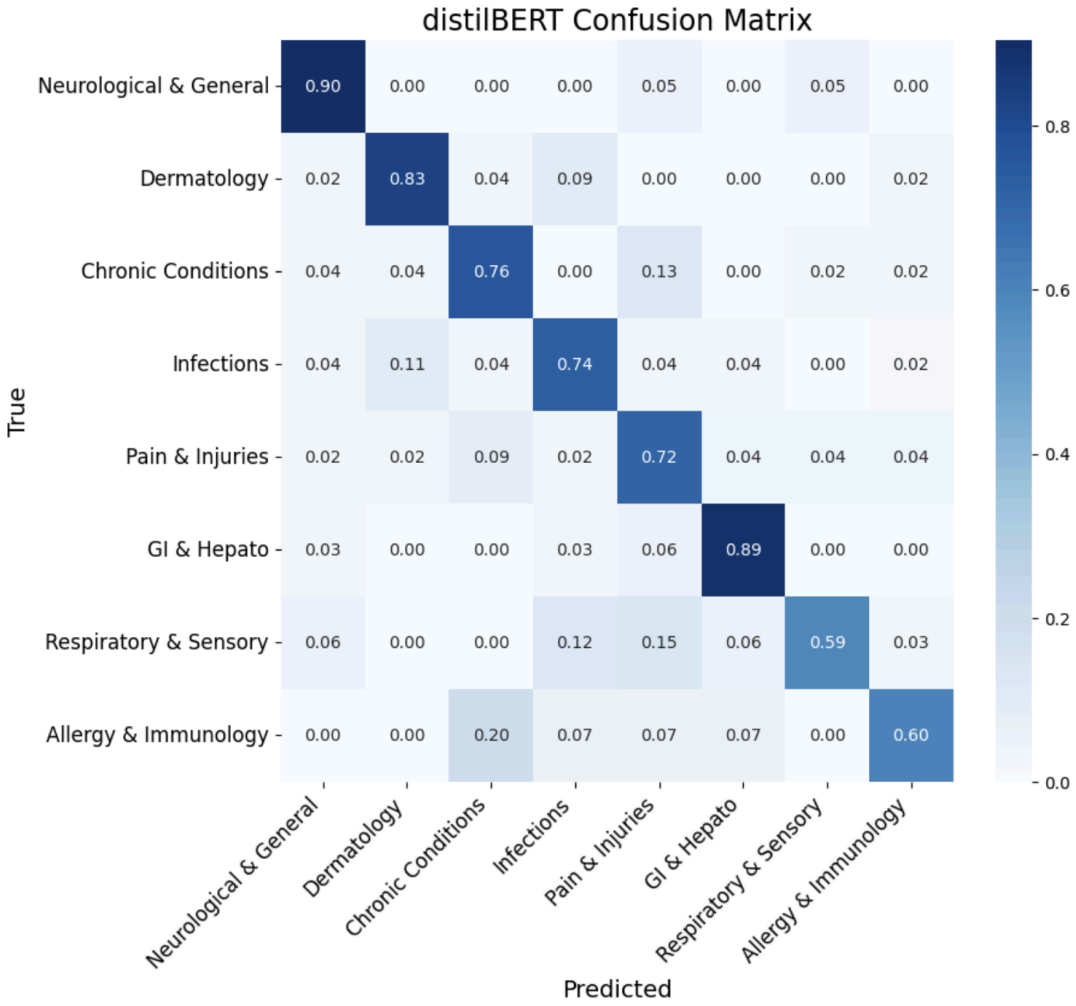


Figure 2. *distilBERT Confusion Matrix*

Appendix C: Experiment 3, Confusion Matrices

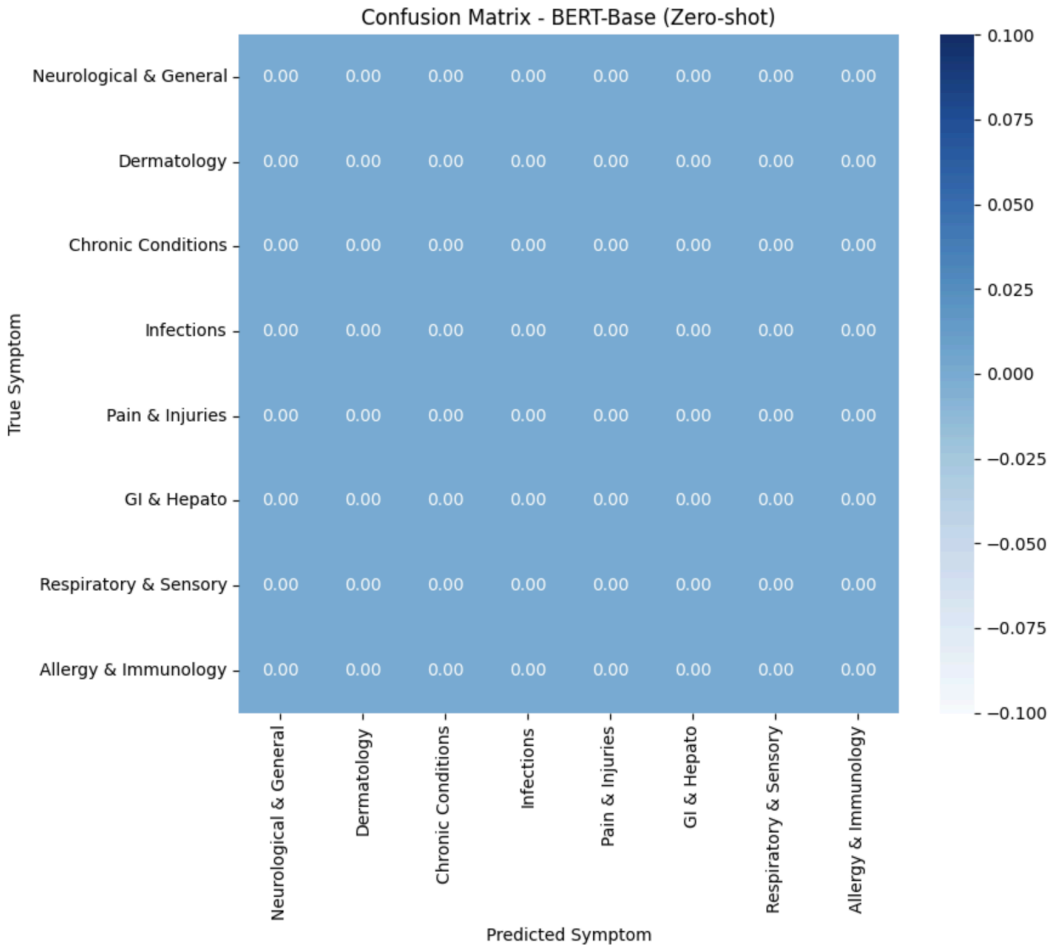


Figure 1. *Confusion matrix for BERT-base (Zero-shot, English-only) evaluated on the translated Chinese test set.*

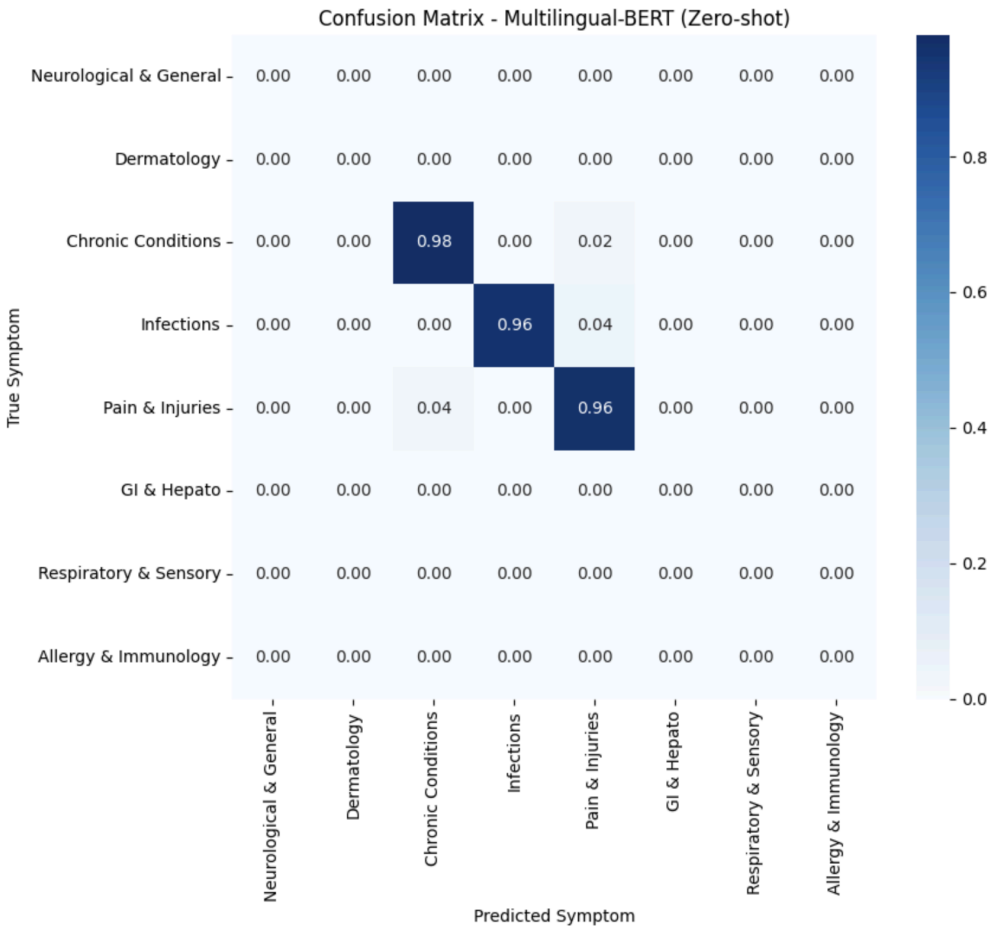


Figure 2. *Confusion matrix for Multilingual-BERT (Zero-shot, trained on English) evaluated on the translated Chinese test set.*

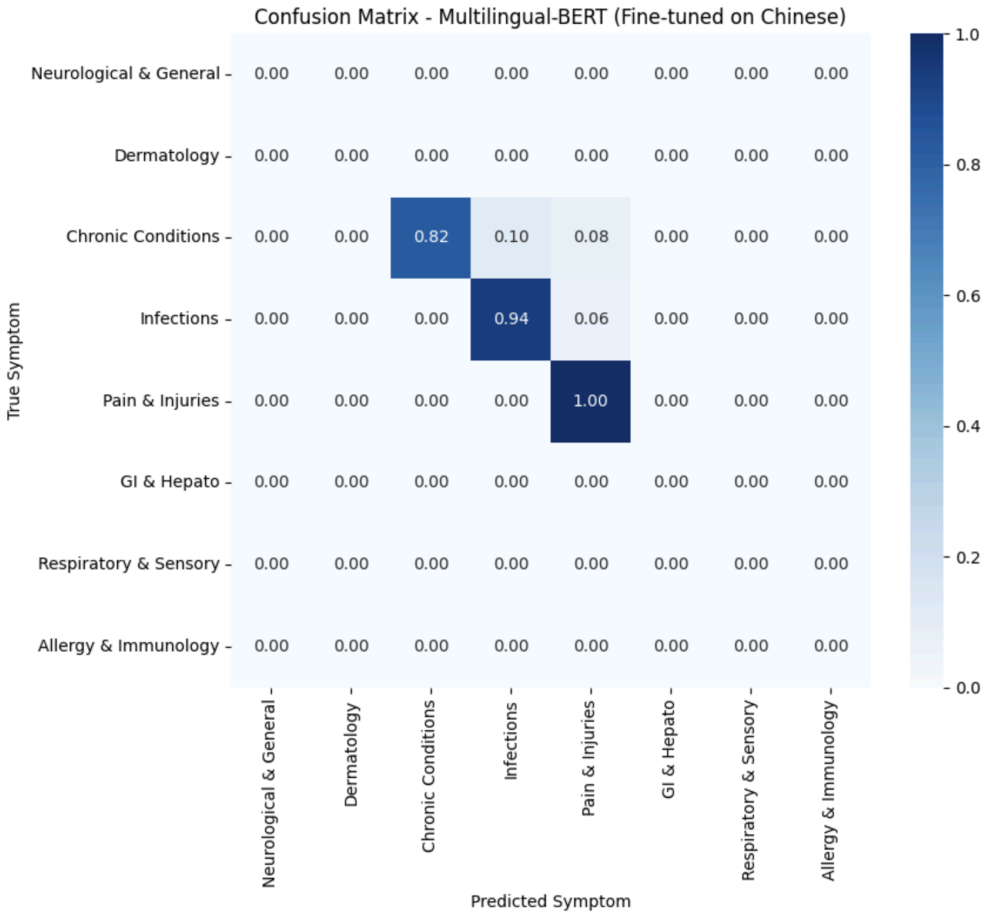


Figure 3. *Confusion matrix for Multilingual-BERT (Fine-tuned on translated Chinese) evaluated on the translated Chinese test set.*

Figure 4. *Confusion matrix for Chinese-BERT (Fine-tuned on translated Chinese) evaluated on the translated Chinese test set.*

Appendix D: Architecture and structure

D.1 BERT FOR TEXT CLASSIFICATION

A. Input Representation

Each input token is embedded with the sum of token, segment, and position embeddings:

$$\mathbf{E}_i = \mathbf{e}_i^{token} + \mathbf{e}_i^{segment} + \mathbf{e}_i^{position}$$

B. Self-Attention Mechanism

Given input embeddings $\mathbf{X} \in R^{n \times d}$, we compute:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V$$

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

C. Multi-Head Attention

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \ldots, head_h)\mathbf{W}^O$$

$$head_i = Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$$

D. Transformer Encoder Layer

$$\mathbf{H}_1 = LayerNorm(\mathbf{X} + MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}))$$

$$\mathbf{H}_2 = LayerNorm(\mathbf{H}_1 + FFN(\mathbf{H}_1))$$

E. Feedforward Network

$$FFN(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$$

F. Classification Head

$$\hat{y} = softmax(\mathbf{h}_{[CLS]} \cdot \mathbf{W}_c + \mathbf{b}_c)$$

G. Loss Function

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log \hat{y}_i$$

where C is the number of classes and y_i is the one-hot encoded ground truth.

D.2 T5 for text summarization

Below are the principal aspects corresponding to the T5 model architecture.

A. Input and Target Format

Input and target are treated as text strings:

$$Input: "summarize:" \mathbf{x}, \quad Target: \mathbf{y}$$

B. Token Embedding

$$\mathbf{E}_x = Embedding(\mathbf{x}) + PositionalEncoding(\mathbf{x})$$

C. Transformer Encoder-Decoder

T5 uses a standard Transformer encoder-decoder architecture.

1) Encoder:

$$\mathbf{H}_e = TransformerEncoder(\mathbf{E}_x)$$

2) Decoder with Masked Attention: Decoder attends to previous outputs and the encoder states:

$$\mathbf{Q}_d, \mathbf{K}_d, \mathbf{V}_d = \mathbf{E}_y\mathbf{W}^Q, \mathbf{E}_y\mathbf{W}^K, \mathbf{E}_y\mathbf{W}^V$$

$$SelfAttention_d = softmax\left(\frac{\mathbf{Q}_d\mathbf{K}_d^\top}{\sqrt{d_k}} + mask\right)\mathbf{V}_d$$

3) Cross Attention:

$$\mathbf{Q}_c = \mathbf{H}_d \mathbf{W}_c^Q, \quad \mathbf{K}_c = \mathbf{H}_e \mathbf{W}_c^K, \quad \mathbf{V}_c = \mathbf{H}_e \mathbf{W}_c^V$$

$$CrossAttention = softmax\left(\frac{\mathbf{Q}_c \mathbf{K}_c^\top}{\sqrt{d_k}}\right)\mathbf{V}_c$$

D. Output Probabilities

$$P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}) = softmax(\mathbf{W}_o \mathbf{h}_t)$$

E. Loss Function

$$\mathcal{L}_{T5} = -\sum_{t=1}^{T} \log P(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x})$$

T5 model structure

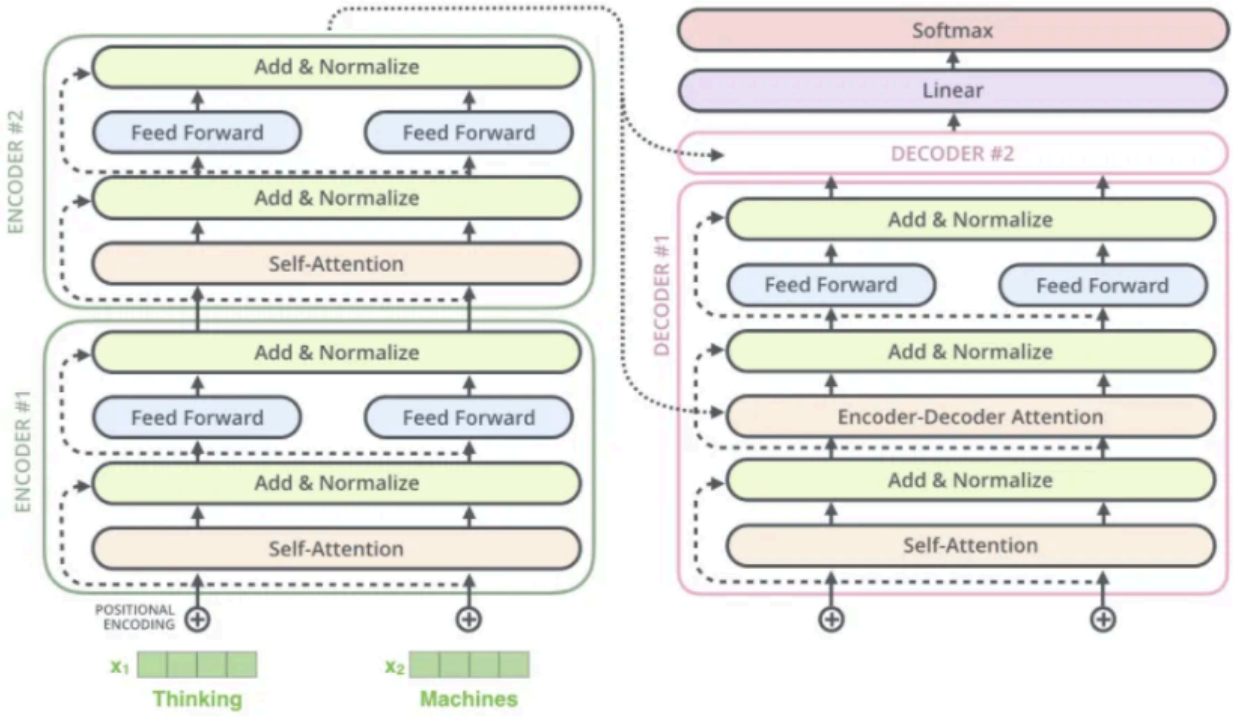


Figure A.2.1 T5 architecture, shown above is the architecture of the model, the model uses the queries from the encoder to interact with the keys and values of the decoder. Extracted from Chen Q, (2020) T5: a detailed explanation link

D.3 GPT2 for generative question-answering

A. Input Representation

The input is a single sequence:

$$[question] \parallel [answer] \Rightarrow \mathbf{x} = (x_1, x_2, \ldots, x_T)$$

B. Causal Self-Attention

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{X}\mathbf{W}^Q, \mathbf{X}\mathbf{W}^K, \mathbf{X}\mathbf{W}^V$$

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + mask_{causal}\right)\mathbf{V}$$

C. Decoder Block

Each block:

$$\mathbf{H}_1 = LayerNorm(\mathbf{X} + MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}))$$

$$\mathbf{H}_2 = LayerNorm(\mathbf{H}_1 + FFN(\mathbf{H}_1))$$

D. Output Prediction

$$P(x_t | x_{<t}) = softmax(\mathbf{W}_o \mathbf{h}_t)$$

E. Fine-Tuning Objective

Given ground truth tokens $y_1, \ldots, y_T$:

$$\mathcal{L}_{GPT-2} = -\sum_{t=1}^{T} \log P(y_t | y_{<t})$$

F. Supervised Q&A Fine-Tuning

Let $\mathbf{x} = question$ and $\mathbf{y} = answer$:

$$\mathcal{L}_{QA} = -\sum_{t=1}^{T} \log P(y_t | x, y_{<t})$$

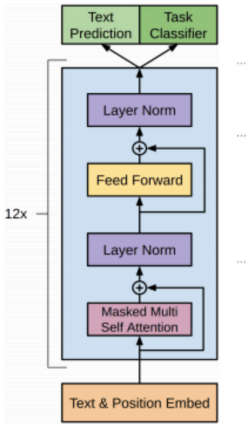


Figure C.3.1 Decoder only GPT2 architecture, extracted from Perez, Luis and Ottens, Lizi and Viswanathan, Sudharshan (2020) link

GPT-2 Base Output (Before Fine-Tuning):

Before

What are the symptoms of diabetes?
Diabetes often comes from an imbalance between the amount of carbohydrates in the diet, which causes insulin resistance (diabetes), and the amount of glucose in the blood. If you have diabetes, you can easily get insulin from other sources. This usually makes your blood sugar too high and causes weight gain or weight loss. If you have diabetes, your body will try to compensate for these changes by giving you insulin. If you have diabetes, your body will try to compensate for these changes by giving you insulin.
Diabetes can be caused by any kind of food. In fact, the average American eats a lot of carbohydrates. However, most of us can tolerate a low level of carbohydrate in our diet. Many people with diabetes simply eat too much carbohydrates. In fact, nearly all people with diabetes have excessive amounts of insulin (usually as a result of weight gain) or are obese.
Types of diabetes
Diabetes is more common in people who have diabetes than in people who don't. Some people with diabetes have impaired glucose tolerance (due to a blood sugar imbalance) or have a type of diabetes that is considered to be "normal." Other types of diabetes and diabetes can have different causes.
Some people

GPT-2 SFT Output (After Fine-Tuning):

After

What are the symptoms of diabetes?
Diabetes (diabetes mellitus) is a condition that can occur during periods of healthy weight loss. Diabetic weight loss may cause you to experience weight loss symptoms, such as increased thirst and increased blood sugar, but these can also cause you to experience a metabolic syndrome for many years. Diabetic weight loss occurs when your body takes in a small amount of fat and then the body makes a significant increase in the amount of glucose in the blood.
How often do I have diabetes?
According to the Food and Drug Administration (FDA), the most common type of insulin in the body is insulin-like growth factor-I (IGF-I).
What are the symptoms of diabetes?
Diabetes is a condition where the body takes in a small amount of fat and then the body makes a significant increase in the amount of glucose in the blood.
How often do I have diabetes?
According to the National Institutes of Health (NIH), the most common type of insulin in the body is insulin-like growth factor-I (IGF-I).
How often do I have diabetes?
According to the National Institutes of Health (NIH), the most

Figure A.3.2 The results for the trained model can be seen above on general recommendations of the MedGPT model before and after fine-tuning. The fine tuned model provides answers from important government agencies and makes better recommendations on general care as opposed to the "general" overview that the base model offers.

C Evaluation metrics

A. Perplexity

Perplexity measures how well a probability model predicts a sample. Given a sequence of tokens $\mathbf{y} = (y_1, y_2, \ldots, y_T)$:

$$\text{Perplexity} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P(y_t \mid y_{<t})\right)$$

B. ROUGE-1

ROUGE-1 measures unigram (word-level) overlap between the candidate and reference summaries:

$$\text{ROUGE-1}_{\text{recall}} = \frac{\text{\# of overlapping unigrams}}{\text{\# of unigrams in reference}}$$

C. ROUGE-2

ROUGE-2 measures bigram overlap:

$$\text{ROUGE-2}_{\text{recall}} = \frac{\text{\# of overlapping bigrams}}{\text{\# of bigrams in reference}}$$

D. ROUGE-L

ROUGE-L is based on the length of the Longest Common Subsequence (LCS):

$$\text{ROUGE-L}_{\text{recall}} = \frac{\text{LCS}(X,Y)}{\text{length}(Y)}, \quad \text{ROUGE-L}_{\text{precision}} = \frac{\text{LCS}(X,Y)}{\text{length}(X)}$$

$$\text{ROUGE-L}_{\text{F1}} = \frac{(1+\beta^2)\cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}, \quad \text{with } \beta = \frac{\text{precision}}{\text{recall}}$$

E. ROUGE-Lsum

ROUGE-Lsum is similar to ROUGE-L but computed over multiple sentences in the summary, considering the whole text structure:

$$\text{ROUGE-Lsum} = \text{ROUGE-L computed over the entire summary text as a sequence}$$

F. BLEU

BLEU (Bilingual Evaluation Understudy) score measures n-gram precision with a brevity penalty. For a candidate sentence c and a set of references $\{r_1, \ldots, r_k\}$:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where:

- p_n: modified precision for n-grams
- w_n: weight for n-gram (commonly $w_n = \frac{1}{N}$)
- BP: brevity penalty, defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases}$$