Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Natural Language Processing to Improve Provider-Patient Relationships through Improved Empathy, Knowledge, Attitudes and Behaviors around Psycho-Oncology.
Principal Investigator:	Arwa Aburizik, MD, MS[1], Martin Kivlighan, PhD[2] [1]Department of Internal Medicine, Division of Hematology-Oncology, Carver College of Medicine [2]Department of Psychological and Quantitative Foundations
Prepared by (IIAI):	Avinash Reddy Mudireddy
Other investigators:	
Date:	November 25, 2025

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	No Significant challenges arose in data analysis that prevented our team from moving forward with the original aim. While the mechanisms underlying this project generated ideas for alternate research paths, the project did not yield results that would drive a proposal for extramural funding.

If yes … Planned submission date	NA
Funding agency	NA
Grant mechanism	N/A
If no … Why not? What went wrong?	N/A

Brief summary of accomplished results:

Healthcare providers face significant operational inefficiencies and revenue loss due to patient appointment "no-shows." While clinical notes contain valuable structured clues about a patient's resources or potential barriers (e.g., transportation issues, anxiety level), effectively leveraging complex, unstructured text to accurately predict attendance remains a significant technical challenge.

We successfully developed and fine-tuned a specialized medical Large Language Model (MedGemma-4b) that predicts encounter status with a strong Test AUROC of 0.731. By implementing a robust pipeline with Parameter-Efficient Fine-Tuning (LoRA) and class-imbalance handling, we established a stable and reliable baseline that effectively extracts predictive signals from clinical text, proving that a targeted, efficient model can perform well without needing excessive complexity.

Research report:

Aims (provided by PI):

1. Train models using Natural Language Processing to predict oncology provider psycho-oncology knowledge, attitudes and behaviors (KABp), and empathy based on their clinical notes.
2. Improve provider-patient communication and relationship (PCR) based on findings through training and feedback systems.

Provider-Patient Communication and Relationship (PCR) is a burgeoning area of research aimed at improving healthcare services [1, 2]. Beyond improving outcomes through evidence-based medicine, patient-centered care recognizes that interpersonal dynamics in medicine are influenced by cultural, societal and institutional factors that are reflected in communication and relationships between medical providers and patients [3]. Our team has conducted research on whole-person care including access to care related to provider and patient factors [4, 5]. Additionally, our research has identified provider factors related to emotional health and cognitive load which are equally influential on access to whole person care. This current research aims to develop innovative and novel methods for detecting provider PCR and provider KABp to automate the process of measuring these critical aspects of whole person care and improve health service delivery through enhanced trainings and feedback systems for providers.

Research has demonstrated that interventions aimed at enhancing PCR have a direct and positive effect on patient and provider outcomes [6, 7]. These interventions have largely focused on rectifying inadequacies, deficits and biases manifested by medical providers to enhance patient outcomes through better communication [8, 9] , but have not necessarily focused on improving provider attitudes, knowledge, skills, and empathy as an underlying and critical aspect of PCR. Researchers have indeed established that empathy and KABp impact PCR in a bidirectional fashion, but these studies have utilized self-report measures of these constructs which limit the clinical utility of assessing provider empathy, knowledge, skills, and attitude as a means to provide feedback to providers about these constructs. Therefore, there is a critical need to develop methods for automating the detection of provider attitudes, knowledge, and empathy to eventually improve PCR through enhanced trainings and ongoing feedback of these constructs in clinical practice.

This project will use NLP to develop models to predict provider knowledge, attitudes, and empathy based on the documentation from actual clinical encounters. In this research study, the clinical note is used to connect the data from empathy questionnaires and KABp questionnaires with the writing styles of oncology providers. Using these advancements in machine learning will allow our team to develop models to detect the variables of interest from actual clinical encounter documentation.

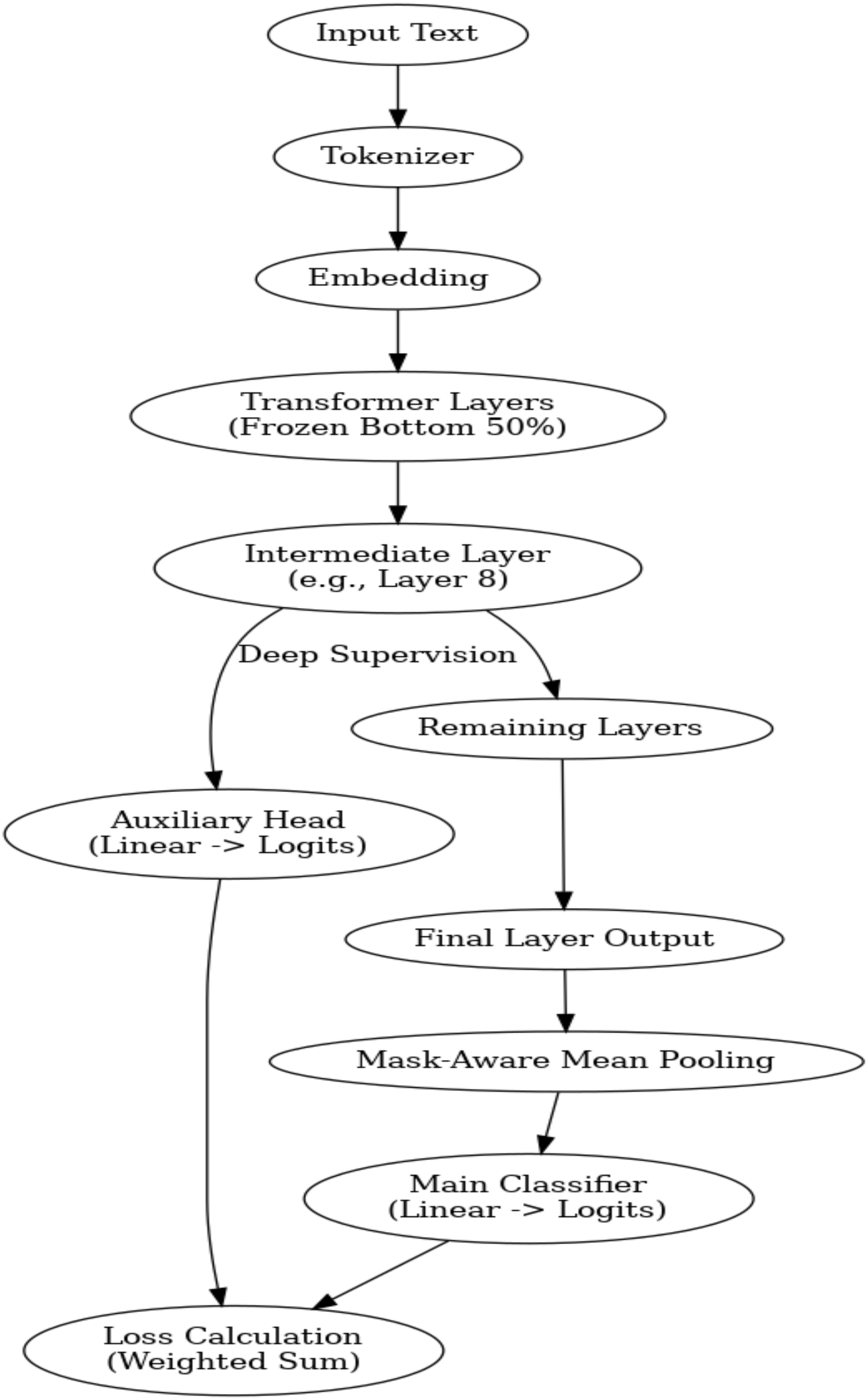
Modified Aims at IIAI due to data issue(changed the project goals midway):

Predicting if a patient will show up for their scheduled appointment is an essential aspect of healthcare operations, relying on patient clinical notes prior to referrals.

Data: we pulled the completed visits in medical and surgical oncology from which a referral to behavioral oncology originated and then divided them into those that resulted in a fulfilled Behavioral Oncology appointment versus those that resulted in a missed or canceled Behavioral Oncology appointment. Subsequently, the language in the clinical notes associated with the encounters within the group from which behavioral oncology appointments were completed were analyzed and contrasted with the narrative text in the notes of encounters from which originated referrals without a completed behavior oncology appointment.

AI/ML Approach:

Predicting whether a patient will attend their scheduled appointment is a critical task in healthcare operations. Accurate predictions can help reduce revenue loss from no-shows and improve patient access to care. Clinical notes contain a wealth of information about the patient's situation, intentions, and potential barriers to attending their appointments. This project aims to leverage this information using state-of-the-art Natural Language Processing (NLP) models to build an accurate prediction model.


Input Text
Tokenizer
Embedding
Transformer Layers
(Frozen Bottom 50%)
Intermediate Layer
(e.g., Layer 8)
Deep Supervision
Remaining Layers
Auxiliary Head
(Linear -> Logits)
Final Layer Output
Mask-Aware Mean Pooling
Main Classifier
(Linear -> Logits)
Loss Calculation
(Weighted Sum)

Baseline Model

Our baseline model is a unsloth/medgemma-4b-it-bnb-4bit model, a large language model optimized for medical text. The selection and implementation of this baseline were guided by the principles outlined in the paper “A Practical Guide to Fine-tuning Large Language Models” (2408.13296v1.pdf). The following techniques from the paper are incorporated into our baseline script:

- Parameter-Efficient Fine-Tuning (PEFT): We use LoRA (Low-Rank Adaptation) to efficiently fine-tune the model. This involves freezing the base model’s weights and learning low-rank updates for the query, key, value, output, and MLP projections. This is controlled by the --lora_r, --lora_alpha, and --lora_dropout arguments.
- Freezing Backbone: The script includes a --freeze_layers option to freeze the lower 50% of the transformer layers, preserving pre-trained knowledge.
- Prompt Engineering: We use deterministic instruction headers and support optional few-shot exemplars to align the model to the classification task. This is enabled with the --use_prompts argument.
- Robust Classifier Head: A custom classifier head (UnslothSeqClassifier) performs mask-aware mean pooling over the last hidden state for a stable and robust classification output.
- Quantization and Mixed Precision: We use 4-bit quantization (load_in_4bit=True) and bfloat16 mixed-precision training to reduce the model’s memory footprint and improve training speed.

Technique 1: Deep Supervision with Auxiliary Head

To improve the model’s learning process, we introduced a deep supervision mechanism inspired by the “Tiny Recursion Model” (TRM) paper. We added an auxiliary classification head to an intermediate layer of the transformer model. The total loss during training is a weighted sum of the loss from the final classification head and the auxiliary head. This encourages the model to learn meaningful representations in its earlier layers.

Implementation: - We modified the UnslothSeqClassifier class to include an optional auxiliary classifier. The *train_loop* was updated to calculate a combined loss: *total_loss = (1 - w) * final_loss + w * aux_loss*, where *w* is the *aux_loss_weight*.

Technique 2: Data Augmentation

We implemented data augmentation using back-translation (English -> French -> English) to improve generalization, though it did not yield performance gains in initial experiments.

Experimental methods, validation approach:

Imbalance-aware Training: To handle the imbalanced nature of the dataset, we use class-weighted cross-entropy and stratified sampling for the train/validation/test splits. We also track metrics like AUROC, balanced accuracy, and PR curves.

Validation and Monitoring: The training process includes epoch-wise validation, the AdamW optimizer with a linear warmup scheduler, and the generation of training history and learning curves for monitoring.

Dual-stage Adapters: The script supports a two-stage training regimen with the --dual_lora flag, allowing for a “merge and refine” workflow.

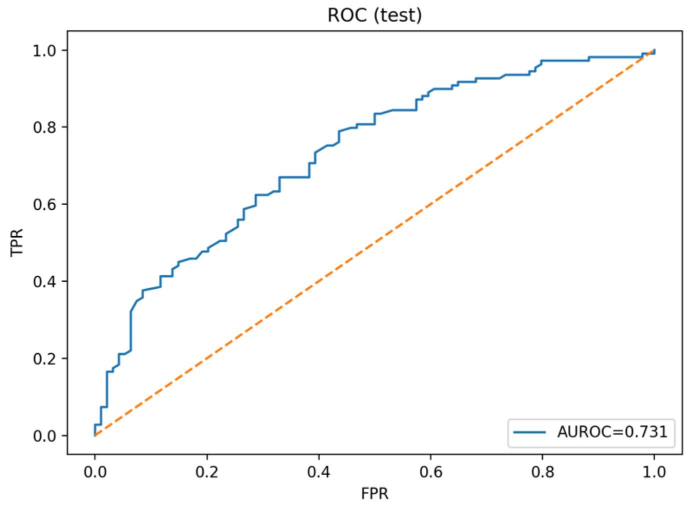
Interpretability: The script includes a function for Integrated Gradient (integrated_gradients_word_importance) to provide token-level attributions for model interpretability.

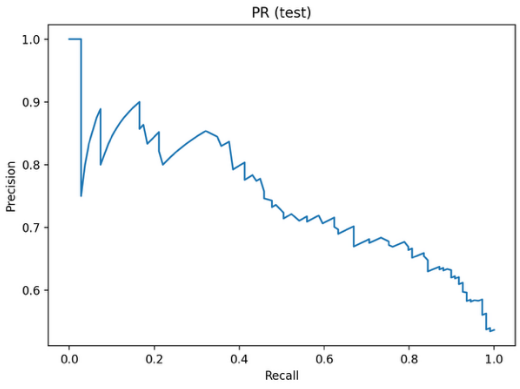
Key parameters: - Model: unsloth/medgemma-4b-it-bnb-4bit - Finetuning: LoRA with r=16, alpha=16 - Epochs: 3 - Learning Rate: 1e-4

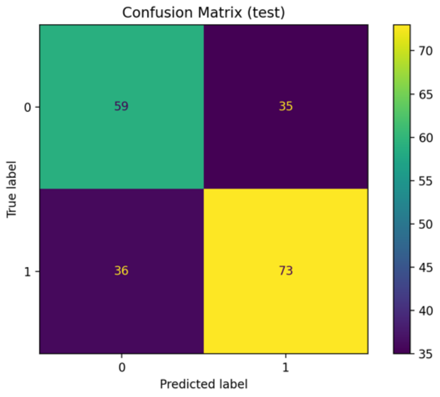
Results:

Baseline Performance:
The initial baseline model achieved a Test AUROC of 0.731 for 1 epoch.

	precision	recall	f1-score	support
0	0.6211	0.6277	0.6243	94
1	0.6759	0.6697	0.6728	109
accuracy			0.6502	203
macro avg	0.6485	0.6487	0.6486	203
weighted avg	0.6505	0.6502	0.6504	203

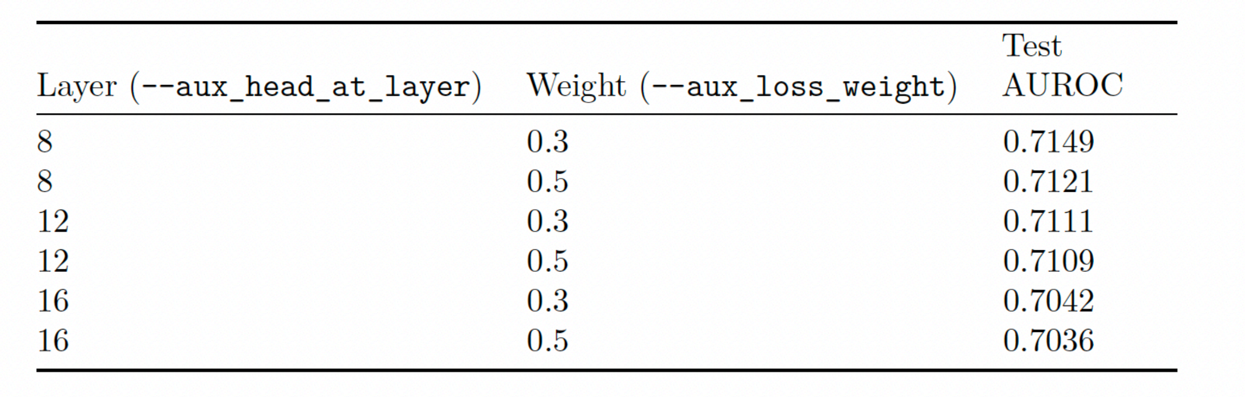






Deep Supervision Tuning:

We ran a series of experiments to find the optimal layer and weight for the auxiliary head. The best performance with the auxiliary head was achieved with Layer 8 and Weight 0.3 (Test AUROC ~0.715). Increasing epochs to 8 led to overfitting.



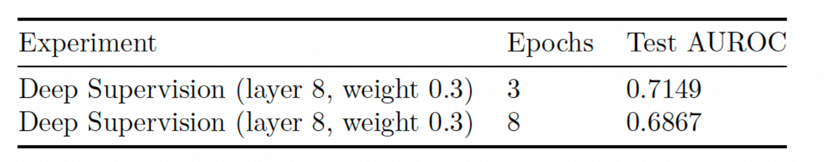
Layer (--aux_head_at_layer)	Weight (--aux_loss_weight)	Test AUROC
8	0.3	0.7149
8	0.5	0.7121
12	0.3	0.7111
12	0.5	0.7109
16	0.3	0.7042
16	0.5	0.7036

Data Augmentation:

Data augmentation (10% fraction) combined with deep supervision did not improve performance over the baseline (AUROC 0.7116).

Longer Training

To see if more training would improve performance, we ran our best-performing experiment so far (--aux_head_at_layer 8, --aux_loss_weight 0.3) for 8 epochs.



Experiment	Epochs	Test AUROC
Deep Supervision (layer 8, weight 0.3)	3	0.7149
Deep Supervision (layer 8, weight 0.3)	8	0.6867

Increasing the number of epochs to 8 resulted in a lower test AUROC, which suggests that the model is overfitting.

In conclusion, fine-tuning the LLM offers predictable results, but the baseline model remains adequate for now. Standard improvement techniques were ineffective, likely due to the limited sample size in this research.

We are trying to teach a computer to read doctor's notes and predict whether a patient will show up for their appointment. We started with a very powerful AI model that already had a lot of smart features. Our goal is to make it even smarter. Our first new idea was to give the AI a "helper" to check its work halfway through its "thinking" process. We hoped this would make the AI smarter, but it didn't improve the results in our first try. We then tried to show the AI more examples by translating the notes to another language and back, but that didn't help either. We also found that training the model for too long can make it worse, not better.

Ideas/aims for future extramural project: Further projects are ongoing and will utilize artificial intelligence and natural language processing to detect biased linguistic patterns in evaluations of medical faculty by patients and students.

Publications resulting from project: Our team with our research specialist to write a manuscript based on the modified aim. That is, the planned manuscript will focus on how our team trained the NLM and the predictors for appointment attendance that we have identified.