UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

HEALTHCARE INNOVATIONS EMPOWERED BY DISCRIMINATIVE AND GENERATIVE ARTIFICIAL INTELLIGENCE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

PAUL CALLE CONTRERAS

Norman, Oklahoma

2025

HEALTHCARE INNOVATIONS EMPOWERED BY DISCRIMINATIVE AND GENERATIVE ARTIFICIAL INTELLIGENCE

A DISSERTATION APPROVED FOR THE
SCHOOL OF COMPUTER SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Chongle Pan (Chair)

Dr. Dean Hougen

Dr. Chao Lan

Dr. Qinggong Tang

Acknowledgments

I am deeply grateful to Dr. Chongle Pan—my committee chair, advisor, co-author, and mentor. His unwavering support, guidance, and high standards have shaped this work and my growth as a researcher. Although we now part ways professionally, I look forward to staying in touch.

Heartfelt thanks to my father, Paul, for his constant encouragement, and to my mother, whose memory continues to inspire me. I am also thankful to my friend and sister, Paola, for her love and patience. I appreciate my labmates—Adrien, Dongyu, Justin, Li, Yunlong, Sinaro, Haoyang, and Zuyuan—for their camaraderie, feedback, and countless conversations that improved this work. To the friends I met during graduate school: Alejandra, Alexander, Angie, Cristian, Helene, and Luis, thank you for your support and for making this journey lighter.

Finally, I thank my dissertation committee for their time, thoughtful feedback, and guidance.

Table of Contents

List Of Tables

List Of Figures

Abstract

Artificial Intelligence (AI) is reshaping healthcare through discriminative and generative models, among other paradigms. Discriminative models classify data by learning boundaries, while generative models capture underlying distributions to produce new data. This dissertation advances both directions with novel computational frameworks.

For discriminative models, I developed NACHOS (Nested and Automated Cross-validation with Hyperparameter Optimization on Supercomputers), an algorithm that integrates nested cross-validation, automated hyperparameter optimization, and high-performance computing (HPC) to more reliably estimate test performance and quantify uncertainty at scale. A companion algorithm, DACHOS (Deployment with Automated Cross-validation and Hyperparameter Optimization using Supercomputing) supports reproducible deployment by retraining the best-selected configuration on full datasets.

To validate the NACHOS framework, I applied it to real-time medical imaging with optical coherence tomography (OCT). In the percutaneous nephrostomy study, ResNet50 achieved a mean accuracy of 82.6% $\pm$ 3.0% under NACHOS evaluation, highlighting that deeper architectures more effectively capture fine-grained tissue differences. In the epidural anesthesia study, a biologically informed sequential binary pipeline outperformed a conventional multi-class approach, reaching 96.7% $\pm$ 1.3% overall accuracy and $>$ 99% precision for detecting the epidural space—a clinically critical milestone. Together, these case studies demonstrate that rigorous evaluation strategies and thoughtful task reformulation can substantially improve robustness and safety in time-critical decision support.

For generative models, I applied large language models (LLMs) to enhance a smoking cessation app whose limited corpus of $\sim$ 900 messages risked repetitiveness. I systematically compared five open-source LLMs and ChatGPT, optimizing prompts and decoding strategies using perplexity and LIWC analysis, and validated outcomes through expert counselor evaluations. The results indicated that ChatGPT and the larger language models consistently generated the most credible and persuasive content.

Overall, this dissertation contributes computational frameworks that advance discriminative and generative AI in healthcare, demonstrating novel pipelines for performance estimation with uncertainty quantification on HPC systems in medical imaging and message generation for patient support systems.

Chapter 1

Introduction

This chapter introduces the fundamental concepts of artificial intelligence (AI) models that underpin the work presented in this dissertation. It begins by outlining the broad categories of models commonly used in AI research and applications. Particular emphasis is placed on discriminative and generative approaches, as these paradigms represent two complementary ways of learning from data and form the foundation for many of the methods applied in later chapters.

1.1 Background and Motivation

Artificial Intelligence (AI) is rapidly transforming the landscape of healthcare, offering new paradigms for disease diagnosis, treatment planning, patient engagement, and health system optimization. Fueled by the convergence of large-scale biomedical data, advances in computing infrastructure, and breakthroughs in machine learning algorithms, AI has emerged as a powerful tool to augment clinical decision-making, reduce diagnostic error, personalize interventions, and streamline healthcare (Topol, 2019; Esteva et al., 2019). From imaging diagnostics and surgical robotics to drug discovery and population health management, AI is beginning to permeate nearly every aspect of modern medicine (Rajpurkar et al., 2017; Beam and Kohane, 2018; Davenport and Kalakota, 2019).

Although enthusiasm for AI in medicine has surged in recent years, its conceptual roots date back several decades. Early expert systems such as MYCIN (Shortliffe

and Buchanan, 1975) and INTERNIST-1 (Miller et al., 1985) sought to codify clinical expertise into rule-based frameworks for infectious disease and internal medicine, respectively. While these pioneering systems demonstrated the potential of computational reasoning in healthcare, their reliance on handcrafted rules and knowledge engineering limited their scalability and adaptability across diverse clinical contexts. The transition to statistical learning and, later, deep learning marked a paradigm shift, enabling algorithms to learn directly from large-scale biomedical datasets rather than manually encoded knowledge (Jordan and Mitchell, 2015).

At its core, AI in healthcare is driven by the ability to learn patterns from diverse data—whether structured (e.g., electronic health records, laboratory values) or unstructured (e.g., medical images, clinical notes, or patient-reported symptoms)—and to make predictions or generate insights that can support both providers and patients. Increasingly, the promise of AI lies not in single-modality models but in multimodal integration, where information from imaging, EHRs, genomics, and wearable sensors can be combined to provide richer and more holistic decision support (Esteva et al., 2019).

Within this landscape, two broad families of models dominate: discriminative models, which focus on predicting clinical outcomes or classifying medical images, and generative models, which synthesize new data such as text, images, or simulations for training, communication, or augmentation purposes. In recent years, discriminative deep learning models have demonstrated performance comparable to—or in some cases exceeding—human experts in tasks such as radiographic interpretation (Ardila et al., 2019), dermatological screening (Esteva et al., 2017), and retinal disease classification (Gulshan et al., 2016). These models are increasingly integrated into imaging workflows and electronic health record systems to serve as clinical decision support tools. However, their reliability and generalizability remain concerns, particularly when deployed

in heterogeneous real-world settings where data distribution shifts or rare edge cases occur. Rigorous evaluation strategies, including uncertainty quantification and robust cross-validation protocols, are therefore essential to ensure trustworthy deployment.

Simultaneously, the advent of large-scale generative models—particularly large language models (LLMs) and generative adversarial networks (GANs)—has opened new frontiers in healthcare innovation. These models are capable of generating synthetic medical data (Frid-Adar et al., 2018), producing detailed clinical documentation (Giorgi et al., 2023), and engaging in natural language interactions with patients (Singhal et al., 2023). Their potential to scale personalized communication, improve health literacy, and support behavioral change interventions is of increasing interest, especially in fields such as mental health, primary care, and public health campaigns (Giorgi et al., 2023; Sezgin and McKay, 2024).

Despite the growing promise of AI in healthcare, the translation from proof-of-concept to clinical utility requires overcoming several challenges. These include data privacy concerns, interpretability and transparency of AI models, integration with existing clinical workflows, regulatory approval, and acceptance by both clinicians and patients (Giorgi et al., 2023; Price and Cohen, 2019). Moreover, computational cost and the need for scalable infrastructure pose barriers to widespread adoption, particularly for high-resolution imaging or large model training (Shen et al., 2024; Bakhtiarnia et al., 2024).

This dissertation is situated at the intersection of these developments. It explores how both discriminative and generative AI models can be harnessed to address critical needs in healthcare, spanning domains such as medical imaging, procedural guidance, and language-based behavioral support. By combining rigorous algorithmic evaluation, high-performance computing, and domain-specific adaptation, the work aims to contribute scalable and trustworthy AI solutions that can empower clinical innovation.

1.2 Discriminative vs Generative Models

AI models are often grouped into discriminative and generative paradigms, which represent two complementary strategies for learning from data. While not the only way to classify models, this distinction provides a useful foundation for understanding many of the approaches discussed in this dissertation.

1.2.1 Definitions

Discriminative models learn the mapping from inputs x to outputs y directly—typically by estimating the conditional distribution $p(y \mid x)$ or learning a decision function $f(x)$ that separates classes. In contrast, generative models aim to capture the underlying data-generating process—by estimating the joint $p(x, y)$ or the marginal $p(x)$. This distinction provides generative models with the ability to sample new data, impute missing values, or compute $p(y \mid x)$ indirectly through Bayes' rule (Jebara, 2012; Banh and Strobel, 2023). Conceptually, discriminative approaches can be viewed as learning decision boundaries in feature space, while generative approaches attempt to model the structure of the data manifold itself.

1.2.2 Training objectives

The training objectives of the two families reflect their goals. Discriminative models optimize predictive performance on labeled tasks (e.g., minimizing cross-entropy for classification). When large annotated datasets are available, discriminative methods tend to achieve superior predictive accuracy. Generative training, by contrast, seeks to maximize (explicitly or implicitly) the likelihood of the observed data. In practice, this can be accomplished via exact likelihood estimation, approximate objectives such as the evidence lower bound (ELBO), or adversarial schemes that pit generators against

discriminators. As a result, generative models can exploit unlabeled or partially labeled data and enable semi-supervised or unsupervised learning—an important capability in clinical settings where annotation is costly.

1.2.3 Representative Families

Discriminative models include logistic models, support vector machines (SVMs), boosted decision trees, and a wide range of deep discriminative networks such as convolutional neural networks (CNNs) (He et al., 2016), and transformers trained for classification. Generative models, on the other hand, comprise variational autoencoders (VAEs) (Kingma and Welling, 2013); which explicitly optimize latent-variable likelihoods; generative adversarial networks (GANs) (Goodfellow et al., 2014) which employ adversarial optimization to learn implicit densities; autoregressive models that factorize p(x) sequentially; diffusion and score-based models that iteratively denoise samples from noise to data (Ho et al., 2020); and large language models (LLMs) (Brown et al., 2014), which generate coherent text and increasingly multimodal outputs.

1.2.4 Advantages, Limitations, and Trade-offs

Discriminative models are attractive for their predictive accuracy, scalability, and relative training simplicity when sufficient labeled data are available (Bishop and Nasrabadi, 2006). Their primary limitation lies in their dependence on labeled datasets and their vulnerability to distribution shifts: when applied to new scanners, populations, or institutions, performance can degrade sharply (Recht et al., 2019). Generative models offer complementary strengths: they can generate synthetic samples for rare conditions (Frid-Adar et al., 2018), exploit large unlabeled datasets(Radford et al., 2015), and learn transferable representations that benefit downstream discriminative

tasks (Raghu et al., 2019). However, they are often more computationally expensive, prone to instability during training (e.g., GAN mode collapse), and susceptible to producing unrealistic or hallucinated content (e.g., in LLMs)(Ji et al., 2023). Increasingly, modern healthcare AI leverages both paradigms in tandem—for example, using generative pretraining to capture broad medical knowledge, followed by discriminative fine-tuning for specific diagnostic tasks (Singhal et al., 2023).

1.2.5 Hybrid and Emerging Paradigms

The traditional boundary between discriminative and generative models is becoming less rigid. Semi-supervised approaches combine generative modeling with discriminative objectives, as in VAEs with classification heads (Zhu and Zhang, 2019)(Kingma et al., 2014). Self-supervised learning and masked autoencoding pretrain generative representations that later support discriminative tasks (He et al., 2022)(Chen et al., 2020). In healthcare, foundation models such as Med-PaLM (Tu et al., 2024), BioGPT (Luo et al., 2022), and multimodal transformers (Liu et al., 2021b) exemplify this convergence by integrating generative training with discriminative evaluation.

1.2.6 Healthcare Use Cases

Discriminative models drive many clinical decision support systems. Examples include disease detection from medical images (e.g., diabetic retinopathy screening in Gulshan et al. (2016)), triage in radiology workflows, and risk prediction from structured electronic health records. Generative models support data augmentation, simulation,

de-identification, clinical documentation drafting, and patient-facing language generation; they can also improve downstream discriminative tasks via pretraining or synthetic data (e.g., GAN-based augmentation for lesion classification in Frid-Adar et al. (2018), LLMs for drafting notes or patient messages (Singhal et al., 2023)).

1.2.7 Evaluation and Reliability

Discriminative systems are commonly evaluated using task-specific predictive metrics—such as the area under the receiver operating characteristic curve (AUROC) and the F1-score—and must be examined for probability calibration and robustness to dataset shift. By contrast, generative systems require distinct evaluation criteria (e.g., marginal or joint likelihoods or their surrogates such as the evidence lower bound, reconstruction error, Fréchet Inception Distance, structured human evaluation, and factuality/faithfulness audits), because high-fidelity samples do not necessarily imply calibrated or clinically safe behavior. Both model families are vulnerable to distribution shift; notably, deep generative models may assign spuriously high likelihood to out-of-distribution inputs, underscoring the necessity of rigorous validation prior to clinical deployment (Nalisnick et al., 2018; Ovadia et al., 2019).

1.3 Unmet Needs in AI-enabled Healthcare

Advances in AI have accelerated markedly, yet several cross-cutting gaps continue to limit clinical impact. While proof-of-concept studies demonstrate high accuracy in retrospective datasets, the translation of AI into routine clinical care remains constrained by concerns around trust, scalability, and integration. Four needs are especially salient: interpretability (making model decisions understandable to clinicians), generalization

(ensuring robust performance across diverse patient populations and settings), real-time imaging decision support, scalable evaluation/training workflows, and personalized patient communication using generative models.

1.3.1 Interpretable and Generalizable Deep Learning Models

Despite strong benchmark performance, many clinical AI systems remain vulnerable to distribution shift, spurious correlations, and limited explainability, all of which can undermine safety and clinician trust. Models may exploit "shortcuts" rather than causal signals, degrading performance across sites, scanners, or populations (Zech et al., 2018; Recht et al., 2019; Geirhos et al., 2020). Robust deployment therefore requires a combination of uncertainty estimation, out-of-distribution detection, and interpretable reasoning mechanisms (Ovadia et al., 2019). Ongoing debate also concerns when intrinsically interpretable models should be preferred to post-hoc explanations in high-stakes settings (Doshi-Velez and Kim, 2017; Rudin, 2019).

Methodologically, rigorous evaluation protocols—such as nested cross-validation, external validation cohorts, and multi-site benchmarking—are essential to guard against selection bias and overly optimistic performance estimates (Varoquaux, 2018). At the same time, there is ongoing debate about when intrinsically interpretable models (e.g., generalized additive models, decision lists) should be preferred to post-hoc explanation techniques (e.g., saliency maps, SHAP, LIME) in high-stakes medical decision-making (Ovadia et al., 2019). Addressing this interpretability–performance trade-off remains central to building trustworthy AI systems for healthcare .

1.3.2 Real-Time Clinical Decision Support from Imaging

For imaging-driven care—triage, guidance, and intra-procedural decisions—AI must deliver low-latency, well-calibrated predictions that integrate seamlessly with clinical workflows and user interfaces (e.g., PACS, operating rooms) (Kelly et al., 2019; Sendak et al., 2020). Meeting this bar entails (i) fast, memory-efficient inference for 2D/3D data streams; (ii) robustness to motion, artifacts, and device variability; and (iii) human-factors design that mitigates automation bias and alarm fatigue. Although deep learning can match or exceed expert performance retrospectively, translating these gains to prospective, real-time use remains challenging—particularly for modalities such as optical coherence tomography (OCT) (Fujimoto and Swanson, 2016). The development of clinically integrated AI tools thus requires advances not only in algorithms but also in hardware acceleration, deployment pipelines, and regulatory alignment.

1.3.3 Scalable AI Workflows (NCV + HPC)

Training and evaluating modern AI models is computationally intensive, especially when rigorous protocols such as nested cross-validation (NCV) are applied to reduce selection bias in performance estimation. This burden is further compounded by hyperparameter optimization (HPO) (Bischl et al., 2023). As datasets grow in size and models in complexity, the demand for scalable workflows becomes acute.

High-performance computing (HPC) resources—combining multi-GPU/CPU parallelism, data/model parallel training, mixed precision arithmetic, and efficient communication backends—are essential to make such workflows feasible within realistic timelines (Dean et al., 2012; Sergeev and Del Balso, 2018). Scalable infrastructures not only enable large-batch or multi-node training but also support reproducibility, a

key requirement for regulatory approval and scientific trust. Without these infrastructures, rigorous evaluation practices like NCV risk being underutilized in healthcare AI research due to prohibitive computational costs.

1.4 Personalized Patient Communication Using Generative Models

Clinical outcomes often hinge on behavioral engagement and patient communication, particularly in chronic disease management and preventive care. Behavior change requires tailored messaging aligned with an individual's stage, preferences, and cultural context. Generative models—including large language models (LLMs)—offer a scalable mechanism to create and adapt such messages, draft documentation, and generate patient education materials (Thirunavukarasu et al., 2023; Liu et al., 2023).

However, outputs from these models must be carefully assessed for factual accuracy, tone, cultural competence, and persuasiveness, using both automated metrics and expert human review (Hawkins et al., 2008; Kreuter and Wray, 2003). Early studies suggest that LLMs encode substantial clinical knowledge and can assist in drafting clinically relevant text (Singhal et al., 2023), but they are also prone to hallucination, bias, and overconfidence (Ji et al., 2023). Accordingly, guardrails—including prompting strategies, retrieval-augmented generation, and human-in-the-loop review—are required to ensure safe deployment. Beyond technical safeguards, frameworks for equity, accountability, and patient-centered design are needed to ensure that generative systems enhance, rather than undermine, health communication (Rajpurkar et al., 2017).

1.5 Overview

This dissertation is organized to move from foundational methodology to clinical application in imaging, and finally to language-based interventions, with contributions articulated within each chapter.

Chapter 2: Integration of nested cross-validation, automated hyperparameter optimization, and high-performance computing to reduce and quantify the variance of test performance estimation of deep learning models

This chapter introduces the NACHOS and DACHOS pipelines, combining nested cross-validation, automated hyperparameter optimization, and high-performance computing for trustworthy performance estimation of deep learning models.

Chapter 3: Deep Learning for OCT-Guided Surgical Imaging

This chapter applies CNN-based discriminative modeling to endoscopic OCT for percutaneous nephrostomy (PCN) and epidural anesthesia, targeting real-time tissue awareness at the needle tip

Chapter 4: Towards AI-Driven Healthcare: Systematic Optimization, Linguistic Analysis, and Clinicians' Evaluation of Large Language Models for Smoking Cessation Interventions

This chapter explores LLM-based message generation for smoking cessation, combining linguistic analysis, prompt optimization, and clinician evaluation for message quality.

Chapter 5: Conclusion and Future Directions

This chapter synthesizes from all chapters and outlines future paths for clinical translation, model explainability, and integration into health systems.

Chapter 2

Integration of nested cross-validation, automated hyperparameter optimization, and high-performance computing to reduce and quantify the variance of test performance estimation of deep learning models

This chapter is reproduced from P. Calle, A. Bates, J. C. Reynolds, Y. Liu, H. Cui, S. Ly, C. Wang, Q. Zhang, A. J. de Armendi, S. S. Shettar, K.-M. Fu, Q. Tang, and C. Pan, "Integration of nested cross-validation, automated hyperparameter optimization, high-performance computing to reduce and quantify the variance of test performance estimation of deep learning models," Computer Methods and Programs in Biomedicine 272, 109063 (2025).

2.1 Introduction

Deep learning (DL) has achieved expert-level performance in numerous medical applications, including diabetic retinopathy detection from fundus photographs (Gulshan et al., 2016), skin cancer classification from dermoscopy images (Esteva et al., 2017), and pneumonia detection in chest X-rays (Rajpurkar et al., 2017). Despite these advances, the clinical deployment of DL systems remains limited due to concerns about generalizability, transparency, and regulatory readiness (Kelly et al., 2019; He et al.,

2020; Topol, 2019). One central concern is the reliability of performance estimation: many DL studies report high accuracy using a single test split, but the variance and bias of these estimates are rarely assessed (Recht et al., 2019; Liu et al., 2019a; Stanley et al., 2024; Tejani et al., 2024; Vrudhula et al., 2024; Yang et al., 2020). For example, Roberts et al. (2021) found widespread methodological issues in COVID-19 diagnostic models, where overoptimistic performance claims stemmed from inadequate evaluation protocols.

In a typical deep learning study, a small fraction (i.e., 10%-20%) of the labeled data is held out in the test set for benchmarking the performance of the final model in unseen data, while the majority of the labeled data is used in the training and validation sets for model development. Relying on a single test split is akin to evaluating a school's academic performance based on one randomly chosen class—any performance estimate derived from it may be unrepresentative. In particular, the variance of the test performance metrics is typically unknown (because there is only one test set) and large (because only a small fraction of the labeled data is allocated to the test set). To make deep learning models trustworthy for medical decision-making, it is essential to estimate their performance metrics with low variance using more test data and measure the variance of the obtained estimates (Berrar, 2019; Bates et al., 2024).

The key objectives of this study are:

1. To demonstrate how nested cross-validation (NCV) provides reduced-variance and uncertainty-quantified estimates of deep learning model performance.

2. To integrate automated hyperparameter optimization (AHPO) within NCV using high-performance computing (HPC).

3. To introduce the NACHOS (**N**ested and **A**utomated **C**ross-validation and **H**yperparameter **O**ptimization using **S**upercomputing) pipeline, which unifies NCV, AHPO, and HPC into a scalable framework for benchmarking.

4. To compare different data partitioning strategies in NCV and show their effects on test performance metrics.

5. To present DACHOS (**D**eployment with **A**utomated **C**ross-validation and **H**yperparameter **O**ptimization using **S**upercomputing), a deployment-oriented algorithm that produces a final model optimized and trained on the full dataset.

NCV is an effective procedure to meet this requirement. Briefly, the entire dataset is partitioned into k folds that are rotated through a cross-testing loop. A model development procedure with a (k-1)-fold cross-validation loop is nested within the cross-testing loop. The output of NCV is k estimates of the test performance metrics of k models. The average and variance of these k estimates reflect the expected performance and variability of this model development procedure across the entire dataset.

NCV has been used in a few medical machine learning studies. Nawabi et al. (2021) employed NCV to benchmark the performance of a random forest classifier for prediction of survival for acute intracerebral hemorrhage using extracted radiomic features obtained from non-enhanced computed tomography images. Their random forest classifier achieved an average test accuracy of 72% with a 95% confidence interval between 70% and 74%. We utilized NCV to benchmark the performance of convolutional neural networks (CNNs) for analysis of Optical Coherence Tomography (OCT) images in multiple endoscopic applications (Wang et al., 2024, 2022b, 2021, 2022a). For example, the average test classification accuracy of CNN was found to be 82.6% with 3.0% standard error for detecting three different renal tissues from their OCT images. However, NCV is still under-utilized in the medical field. Roberts et al. (2021) conducted an analysis

of COVID-19 research papers published between January and October 2020 and found a notable lack of NCV utilization, highlighting a gap in methodological rigor in the field.

A challenge of using NCV in a study is the need to implement AHPO between the cross-testing loop and the cross-validation loop. Most medical deep learning studies perform manual hyperparameter optimization (MHPO). Practitioners can manually evaluate various model architectures, learning rates, regularization methods, and other hyperparameters based on cross-validation performance and select the configurations with the best validation performance to build the final model. However, it is impractical to perform MHPO independently and consistently in every test fold of NCV. Instead, AHPO needs to be performed during each testing iteration of the k-fold cross-testing loop in NCV to automatically identify the model configuration with the best cross-validation performance. AHPO within NCV provides reproducible model optimization and prevents inadvertent information leakage from the test set to the validation set during MHPO.

A second challenge of using NCV with AHPO is the need for significantly more computing than cross-validation with MHPO. Fortunately, the computation in NCV and AHPO can be readily partitioned by data folds for the cross-testing loop or the cross-validation loop and by model configurations for the AHPO loop. The folds can be distributed across many GPUs to compute in parallel. Thus, high-performance computing (HPC) can be used to complete NCV and AHPO within a reasonable amount of wall-clock time.

While many deep learning pipelines, NiftyNet (Gibson et al., 2018), TorchIO (Pérez-García et al., 2021), DeepNeuro (Beers et al., 2021), and GaNDLF (Pati et al., 2023)

support model training and evaluation for medical imaging, they lack integrated support for NCV, AHPO, and HPC—limiting their utility for generating robust, reproducible benchmarks needed in clinical AI. To address these limitations, we developed NACHOS (Nested and Automated Cross-validation and Hyperparameter Optimization using Supercomputing) to integrate NCV and AHPO into a parallelized computational workflow on HPC. A repository of chest X-ray datasets from the TorchXRayVision library (Cohen et al., 2022), along with an kidney OCT dataset, derived from Wang et al. (2021), were used to demonstrate NACHOS. We compared different strategies for partitioning the two datasets into k folds in NCV. The results showed the significance of partitioning in benchmarking the test performance of deep learning models (Yagis et al., 2021; Tampu et al., 2022).

The outcome of the NACHOS algorithm is a reduced-variance and uncertainty-quantified estimation of test performance of the models generated by this computational procedure. To build the final model for production use, we developed an algorithm named **D**eployment with **A**utomated **C**ross-validation and **H**yperparameter **O**ptimization using **S**upercomputing (DACHOS). DACHOS identifies the overall best model configuration using all the data for the AHPO and cross-validation and then uses this configuration to train a model using all the data. Because the final model for deployment was hyperparameter-optimized and trained using more data than the k models generated in the k-fold NCV, its test performance, although unknown, is expected to be better than the average test performance of the k NCV models. We also demonstrated DACHOS using the chest X-ray repository and kidney OCT dataset. While NACHOS is designed for rigorous performance evaluation through NCV and AHPO, DACHOS focuses on training a final model for deployment using all available.

2.2 Methods

2.2.1 Overview of the NACHOS and DACHOS Frameworks

The proposed methodology introduces two complementary frameworks: NACHOS and DACHOS. Both are designed to automate and standardize performance benchmarking and model generation in medical imaging using deep learning. NACHOS focuses on performance evaluation through nested cross-validation combined with AHPO. It rotates all data folds through the test set to reduce variance and provides uncertainty-aware model performance estimates. DACHOS, in turn, uses the same automated pipeline to produce a final production model, trained on the full dataset with the best hyperparameters identified during cross-validation. Both frameworks share a similar structure: they iterate over hyperparameter configurations and data folds in a controlled, distributed computing environment. The primary difference lies in their goals: NACHOS emphasizes performance estimation across folds, while DACHOS is designed to maximize model performance for real-world use by training on the entire dataset. The next sections detail the nested loop structures that implement these workflows.

2.2.2 NACHOS algorithm

The NACHOS algorithm (Algorithm 1) is comprisef of three nested loops: the cross-testing (CT) loop, the AHPO loop, and the cross-validation (CV) loop. First, the dataset D is divided into k folds: $F_0, F_1, \ldots, F_{k-1}$. The CT loop iterates over $i \in I = \{0, 1, 2, \ldots, k-1\}$, where the fold F_i is held out as the test set, and the remaining folds are used for training and validation. The AHPO loop then iterates over $j \in J = \{0, 1, 2, \ldots, n-1\}$, where n is the number of hyperparameter configurations to be tried, and each h_j denotes the j^{th} hyperparameter configuration. Within the CV

loop, the index $m \in I \setminus \{i\}$ is used to reserve the fold F_m for validation while the model is trained on the remaining $k-2$ folds. The model's performance on the validation fold F_m is recorded as v_m^j. After completing cross-validation, the average validation performance—i.e., cross-validation performance—for each hyperparameter h_j, denoted as $\bar{v}^j$, is calculated. Once the AHPO loop is completed, the best-performing hyperparameter h_{j^*} is selected based on the highest cross-validation performance. The model is then trained using h_{j^*} on all folds except the test fold F_i and evaluated on the test fold F_i, with the result recorded as t_i. After completing all k iterations of the cross-testing loop, the test performance values $t_0, t_1, \ldots, t_{k-1}$ are aggregated. The nested loop structure of the NACHOS algorithm is illustrated in Figure 2.1A, showing the cross-testing loop, AHPO loop, and cross-validation loop with their respective reserved folds. The average test performance $\bar{t}$ and its standard error (SE) are computed as:

$$\bar{t} = \frac{1}{k}\sum_{i=0}^{k-1} t_i, \qquad SE = \frac{\sigma_t}{\sqrt{k}}, \qquad \sigma_t = \sqrt{\frac{1}{k-1}\sum_{i=0}^{k-1}(t_i - \bar{t})^2}.$$

For all experiments in this study, the performance metric used was classification accuracy, defined as:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \\ &= \frac{TP + TN}{TP + TN + FP + FN}. \end{aligned}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives.

Using this metric, three types of accuracy were distinguished:

- **Validation accuracy:** The accuracy obtained on a single validation fold F_m during the cross-validation loop. It measures how well a model trained on $k-2$ folds generalizes to the reserved validation fold.

- **Cross-validation accuracy:** For each hyperparameter configuration h_j, the cross-validation accuracy is the average of its validation accuracies across all validation folds, v_m^j. This value is used to identify the best-performing configuration, h_{j^*}.

- **Test accuracy:** After selecting h_{j^*}, the model is retrained on all folds except the held-out test fold F_i and evaluated on F_i. The resulting accuracy t_i represents an unbiased estimate of generalization for that CT iteration.

These metrics provide a reduced-variance and uncertainty-quantified estimate of model performance using nested cross-validation.

In the current implementation of NACHOS, the AHPO loop used a random search algorithm (Bergstra and Bengio, 2012) that randomly samples n combinations of values from a set of predefined hyperparameter choices. The choices of the batch size were powers of 2, ranging from 16 (2^4) to 128 (2^7), which aligns with GPU memory allocation efficiency and is widely adopted in CNN training (Masters and Luschi, 2018). The learning rate and weight decay values were sampled from logarithmic scales (10^{-2}) to 0.0001 (10^{-4}), a common heuristic given the sensitivity of deep networks to orders of magnitude differences in step size (Goodfellow et al., 2016). Momentum values (0.5, 0.9, 0.99) with and without Nesterov acceleration were included based on prior evidence that momentum accelerates convergence and remains the optimizer of choice for large-scale CNNs (Goodfellow et al., 2016).

Finally, three model architectures were chosen: ResNet50 (He et al., 2016), InceptionV3 (Szegedy et al., 2016), and Xception (Chollet, 2017), representing a spectrum of architectural depth, multi-scale feature extraction, and efficiency through depthwise separable convolutions, respectively. Each hyperparameter configuration was created by sampling one value for each parameter, and the AHPO loop iterated over $n = 9$ such

configurations to identify the best-performing model under nested cross-validation, see Table 2.1. The decision to limit the search to 9 configurations reflects a pragmatic trade-off between computational cost and experimental coverage, consistent with prior findings that even small random search samples often yield competitive hyperparameter settings in practice (Bergstra and Bengio, 2012).

2.2.3 DACHOS algorithm

The DACHOS algorithm (Algorithm 2) generates a production model, M, for deployment using AHPO and cross-validation. The dataset D is split into k folds for cross-validation. The AHPO loop iterates through n hyperparameter configurations h_j, $j \in J = \{0, 1, 2, \ldots, n-1\}$, which should be the same as those used by NACHOS. The cross-validation loop iterates through $m \in I = \{0, 1, 2, \ldots, k-1\}$ to select the fold F_m for validation and train the model on the remaining folds. The validation performance is recorded as v_m^j. After the k-fold cross-validation is completed for the hyperparameter configuration h_j, its average validation performance—i.e., cross-validation performance—is calculated as $\bar{v}^j$. Once cross-validation for all hyperparameter configurations is completed, the best-performing hyperparameter h_{j^*} is selected based on its cross-validation performance. Finally, the production model, M, is trained using h_{j^*} with the entire dataset D. The workflow of the DACHOS algorithm is summarized in Figure 2.1B, where the AHPO loop and cross-validation loop are used to select the optimal hyperparameter configuration prior to training the final production model. The DACHOS algorithm maximizes the performance of the production model, M, for deployment by using the entire dataset for AHPO and then using the entire dataset

Algorithm 1: NACHOS

Required

k: number of folds

n: number of set of hyperparameter configurations from random search

D: Dataset

$I = \{0, 1, \ldots k-1\}$

$J = \{0, 1, \ldots n-1\}$

$H = \{h_0, h_1, \ldots h_n\}$

Defined

v_m^j: Validation performance for hyperparameter configuration h_j on the validation fold F_m

$\bar{v}^j$: Average validation performance of hyperparameter configuration h_j across all validation folds

t_i: Test performance on the test fold F_i

Result

average and standard error for performance metric

```
Split D into k folds: F_0, F_1, ... F_{k-1}
/* Cross-testing loop                                          */
for i ∈ I do
    Set F_i as test dataset
    /* AHPO loop                                               */
    for j ∈ J do
        Set h_j as hyperparameter configuration
        /* Cross-validation loop                               */
        for m ∈ I \ {i} do
            Set F_m as validation dataset
            // Distribution of tasks
            Train model on remaining folds (I \ {i, m}) with h_j
            v_m^j ← Evaluate model for validation performance on F_m
        v̄^j ← Average values v_m^j, m ∈ I \ {i})
    j* ← arg max_j {v̄^j : j ∈ J}
    Train model on folds: I \ {i} with h_{j*}
    t_i ← Evaluate model for test performance on F_i
return average and standard error for values t_i, i ∈ I
```

for model training. As in NACHOS, the performance metric used throughout the DACHOS pipeline was classification accuracy, computed using the same formula. The validation performances v_m^j and their average $\bar{v}^j$ for each hyperparameter configuration were based on this metric. The final model M, trained on the entire dataset with

the selected optimal hyperparameter h_{j^*}, inherits the configuration that achieved the highest average cross-validation accuracy (i.e., $\bar{v}^{j^*} = \max_j \bar{v}^j$).

2.2.4 Parallelization strategy

The NACHOS and DACHOS algorithms were parallelized using a Python implementation of the Message Passing Interface (MPI) standard provided in the `mpi4py` (Dalcin and Fang, 2021) library. Both algorithms employed MPI point-to-point communication to enable direct interaction between parallel processes. The NACHOS algorithm distributes a total of $k(k-1)n$ training tasks over g GPUs, where k is the number of folds for NCV, n is the number of hyperparameter configurations for AHPO, and g is the number of GPUs. The DACHOS algorithm parallelizes kn training tasks over g GPUs. When launched, the two algorithms create a manager process along with g worker processes, with each worker process assigned to a separate GPU. The manager process is responsible for assigning the tasks and sending their hyperparameter configurations, test folds, and validation folds to the worker processes for computing on their assigned GPUs. When a worker process completes a training task, it requests a new task from the manager process until all the tasks are completed. The dynamic scheduling provides effective load balancing and ensures linear scalability.

2.2.5 Fault Tolerance Mechanism

To manage unexpected failures of training tasks in a job, NACHOS and DACHOS implement fault tolerance using a checkpointing system that includes two types of checkpoints: a metadata checkpoint and a model checkpoint. During training, the system continuously records the hyperparameter configuration h_j, the test fold F_i, the validation fold F_m, and the epoch number in the metadata checkpoint. After each

epoch, a model checkpoint is saved while the previous one is deleted to conserve space. In the event of a failure, the job is relaunched and the manager process redistributes all training tasks. Each task corresponds to a unique combination (F_i, h_j, F_m). Upon receiving a task, a worker process consults the metadata checkpoint to determine whether it has already been completed.

If completed: the worker skips the task and requests a new one from the manager.

If incomplete: the worker loads the latest model checkpoint (if available) and resumes training from the last saved epoch. Once finished, the worker reports completion and requests another task.

This process continues until the manager confirms that all tasks are finished. By coordinating checkpoints with task redistribution, NACHOS and DACHOS ensure that interrupted jobs resume without redundant computation or manual intervention. A schematic of this recovery mechanism is shown in Figure 2.2

2.2.6 Platform and Dependencies

NACHOS and DACHOS were implemented using Python 3.8, TensorFlow 2.6.2, and and scientific computing libraries. They support both distributed GPU workstations and HPC environments. Details regarding software dependencies, cluster configuration, and hardware specifications are provided in the Appendix Section 6.1 to support reproducibility. The library can also generate class activation maps for instance-wide prediction interpretation or feature importance (Badré and Pan, 2022, 2023) using GradCAM (Selvaraju et al., 2017).

2.2.7 Medical Imaging Datasets and Partitioning schemes

A comprehensive chest X-ray repository was assembled using four publicly available datasets included in the TorchXRayVision library (Cohen et al., 2022) the NIH ChestX-ray14 dataset (`https://www.kaggle.com/datasets/nih-chest-xrays/data`) (Wang et al., 2017), the CheXpert v1.0 dataset (`https://stanfordmlgroup.github.io/competitions/chexpert/`) (Irvin et al., 2019), the MIMIC-CXR dataset (`https://doi.org/10.6084/m9.figshare.10303823`) (Johnson et al., 2019), and the PadChest dataset (`https://bimcv.cipf.es/bimcv-projects/padchest/`) (Bustos et al., 2020). These datasets collectively provide a large and diverse collection of frontal-view chest radiographs labeled for various thoracic pathologies, making them suitable for benchmarking deep learning models in medical image analysis.

The NACHOS and DACHOS algorithms were evaluated on a binary classification task, in which deep learning models were trained to classify Posterior-Anterior (PA) chest X-ray images as either cardiomegaly or no finding. In the MIMIC-CXR and PadChest datasets, some lateral images were mistakenly labeled as PA. These incorrectly labeled images were identified through manual inspection and removed from our chest X-ray repository. All images were resized to a resolution of 224×224 pixels through interpolation.

To create balanced data for benchmarking, we used a reproducible random selection (fixed seed) to choose 620 cardiomegaly images and 620 no-finding images from each of the four datasets. These images were combined to build the chest X-ray repository with a total of 4,960 images (4 datasets $\times$ 2 classes $\times$ 620 images per class per dataset). Consistent with the chest X-ray preprocessing approach, no contrast normalization or data augmentation was applied to the OCT dataset. The chest X-ray repository was partitioned into four folds using three different partitioning levels. In the image-level partitioning, images were randomly distributed across four folds. In the patient-level

partitioning, all images from the same patient were assigned to the same fold. Finally, in the dataset-level partitioning, each dataset was exclusively allocated to a separate fold.

An OCT dataset was derived from our previous study (Wang et al., 2022b) for a renal tissue classification task. The OCT images were originally captured as 3D volumes, each containing multiple 2D cross-sectional B-scan images. These 2D cross-sectional images with a resolution of 185×210 pixels were used as the input data in this study. Similarly to chest X-ray repository, no contrast normalization nor data augmentation was performed. The OCT dataset contains 600 images of the cortex tissue, 600 images of the medulla tissue, and 600 images of the pelvis tissue from each kidney. A total of 10 kidneys were included, yielding 18,000 images (10 kidneys $\times$ 3 tissue types $\times$ 600 images per tissue type per kidney). The OCT dataset was partitioned into 10 folds at three levels: image, volume, and kidney. In the image-level partitioning, all 18,000 images were randomly split into 10 folds. In the volume-level partitioning, images from the same volume were assigned to the same fold. In kidney-level partitioning, the 10 kidneys were divided into 10 folds, with each fold containing all images from a respective kidney. The datasets used can be found in `https://doi.org/10.5281/zenodo.14847200`

Algorithm 2: DACHOS

Required

k: number of folds
D: Dataset
$I = \{0, 1, \dots k-1\}$
$J = \{0, 1, \dots n-1\}$
$H = \{h_0, h_1, \dots h_n\}$

Defined

v_m^j: Validation performance for hyperparameter configuration h_j on the validation fold F_m
$\bar{v}^j$: Average validation performance of hyperparameter configuration h_j across all validation folds

Result

M: Model ready to be deployed

Split D into k folds: $F_0, F_1, \dots F_{k-1}$
`/* AHPO loop */`
for $j \in J$ **do**
    Set h_j as hyperparameter configuration
    `/* Cross-validation loop */`
    for $m \in I$ **do**
        Set F_m as validation dataset
        `// Distribution of tasks`
        Train model on remaining folds ($I \setminus \{m\}$) with h_j
        $v_m^j \leftarrow$ Evaluate model for validation performance on F_m
    $\bar{v}^j \leftarrow$ Average values $v_m^j, m \in I$
$j^* \leftarrow \arg\max_j \{\bar{v}^j : j \in J\}$
$M \leftarrow$ Train model on D with h_{j^*}
return M

Table 2.1: Randomly generated hyperparameter configurations for AHPO

Index	Architecture	Batch size	Learning rate	Decay	Momentum	Nesterov
h_0	ResNet50	128	0.01	0.01	0.9	Enabled
h_1	InceptionV3	16	0.001	0.001	0.9	Disabled
h_2	ResNet50	64	0.01	0.01	0.99	Enabled
h_3	Xception	16	0.001	0.001	0.5	Disabled
h_4	ResNet50	64	0.01	0.01	0.5	Disabled
h_5	ResNet50	32	0.01	0.01	0.99	Enabled
h_6	ResNet50	32	0.0001	0.0001	0.99	Disabled
h_7	ResNet50	32	0.01	0.01	0.9	Enabled
h_8	InceptionV3	64	0.01	0.01	0.5	Disabled

	X-ray	**OCT**
# Images	4,960	18,000
# Classes	2	3
Partitioning available	Image, Patient, Dataset	Image, Volume, Kidney
Image resolution	224×224	185×210
Details	# datasets: 4 # patients: 4,678 # images: 4,960	# kidneys: 10 # volumes p/kidney p/class: 30 # images p/volume: 20

Table 2.2: Summary of dataset characteristics for X-ray and OCT.

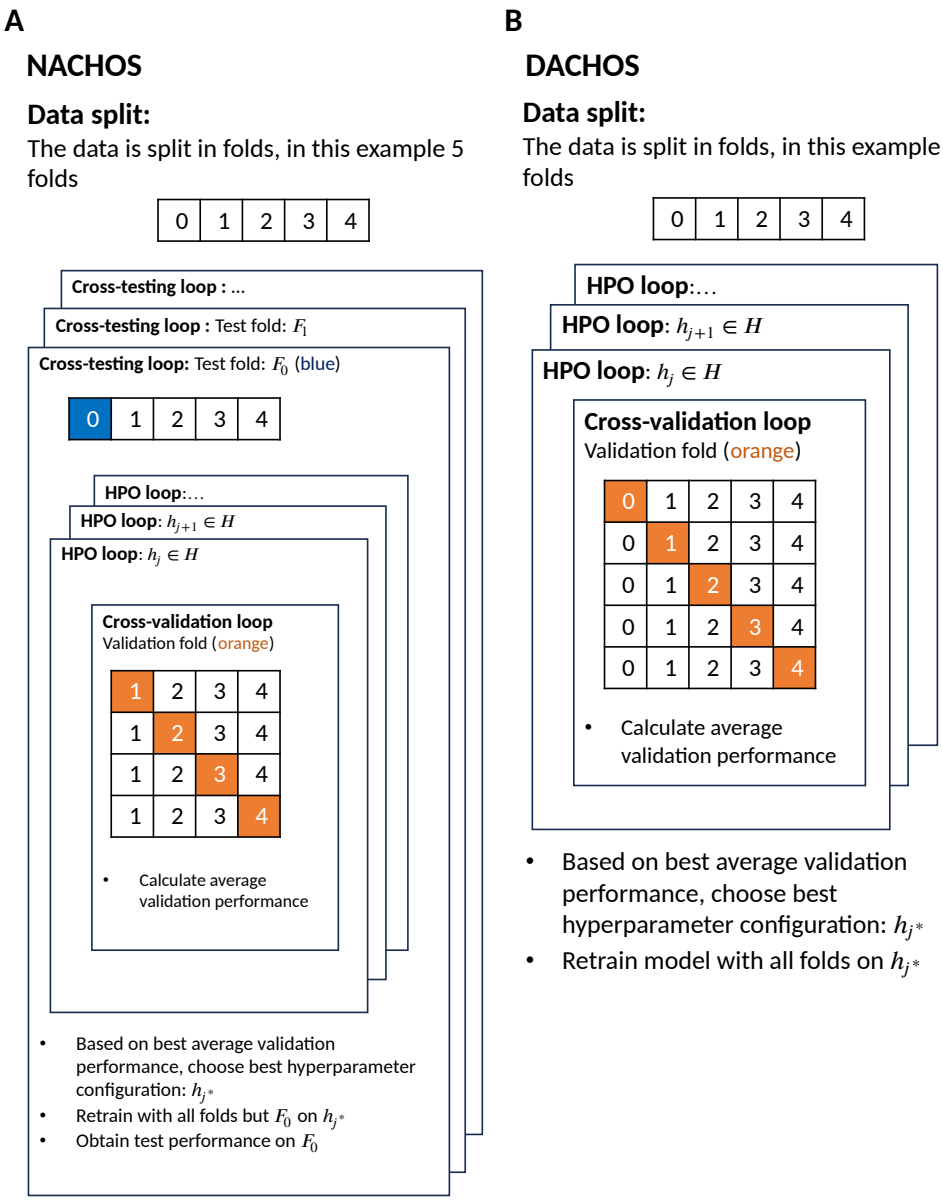

Figure 2.1: **Schematic comparison of the NACHOS and DACHOS algorithms.** [A] NACHOS uses a nested structure with three loops: the cross-testing loop reserves one fold for testing (blue), the AHPO loop explores different hyperparameter configurations, and the cross-validation loop evaluates each configuration by rotating the validation fold (orange). The best configuration is retrained on all folds except the test fold and evaluated on the held-out test fold. [B] DACHOS employs AHPO and cross-validation without an outer cross-testing loop. Hyperparameter configurations are compared based on average cross-validation performance (orange), and the best configuration is used to retrain the final production model on the entire dataset.

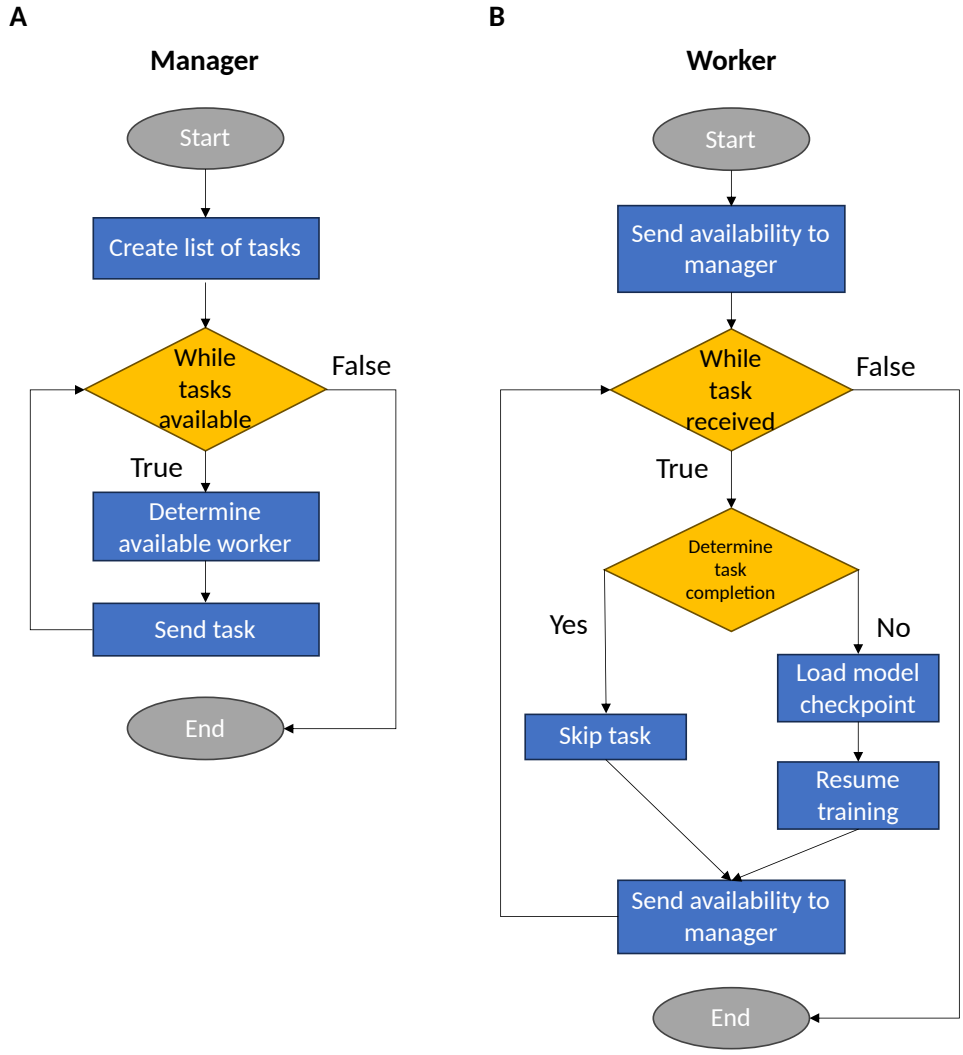

Figure 2.2: **Fault-tolerant workflow for distributed training.** [A] The manager process creates a list of tasks and assigns them iteratively to available workers until all tasks are completed. [B] The worker process reports its availability, receives a task, and checks metadata to determine completion status. If the task is finished, it is skipped; otherwise, the worker loads the latest model checkpoint and resumes training. Afterward, the worker reports availability for reassignment. This coordination ensures efficient scheduling, prevents redundant computation, and enables recovery from failures.

2.3 Results

2.3.1 Reduced-Variance Estimation of the Test Performance with Uncertainty Quantification Using the NACHOS Algorithm

Table 2.3 presents the benchmarking results generated by the NACHOS algorithm on the chest X-ray repository for the cardiomegaly detection task. The data was partitioned into four folds corresponding to the four datasets. NACHOS included three nested loops: the CT loop, the AHPO loop, and the CV loop. For each test fold, the AHPO loop evaluated the cross-validation performance of nine hyperparameter configurations shown in Table 2.1.

When F_0 was reserved as the test fold, configuration h_2 had the highest cross-validation accuracy of $\bar{v}^2 = 0.72$, which was the average of the validation accuracies. After finding h_2 as the configuration with the best cross-validation accuracy, a model was trained on F_1, F_2, and F_3 using h_2, and then tested on F_0 to generate a test accuracy of $t_0 = 0.79$.

This process was repeated for the remaining test folds—F_1, F_2, and F_3. The standard deviation of the test accuracies was 0.06. The variability of the test performance stemmed from the different allocations of data folds for training, validation, and testing, as well as the stochastic nature of stochastic gradient descent in model training and random search in AHPO.

The average test accuracy across all four folds was 0.75 with a standard error of 0.03, corresponding to a 95% t-based confidence interval of [0.67, 0.84]. This average test accuracy, derived from four model instances, reflected the expected performance of model instances that can be generated by our development procedure. The standard

deviation of this average test accuracy (0.03) was half ($\sqrt{4}$) of the standard deviation of the individual test accuracies (0.06) from different test folds, because the average test accuracy represented the test performance from all four folds. From a practical application perspective, this variability implies that when the model is applied to new but similar clinical data, its accuracy would typically be expected to fall within about ±3% of the reported average.

The test accuracies were lower than the cross-validation accuracies in test folds F_1 and F_2, and higher in test folds F_0 and F_3. Although the AHPO was expected to make the cross-validation accuracies higher than the test accuracies, the results were inconsistent, probably due to data variability.

The average confusion matrix across the four test folds is shown in Table 2.5. On average, the models correctly classified 426 ± 82 of the 612 "No-Finding" cases and 505 ± 66 of the 612 "Cardiomegaly" cases. Misclassifications occurred in 194 ± 82 "No-Finding" images predicted as "Cardiomegaly" and 115 ± 66 "Cardiomegaly" images predicted as "No-Finding." These results suggest that false positives for cardiomegaly were more frequent than false negatives, which could have implications for clinical workflows—potentially leading to more follow-up examinations but reducing the likelihood of missed diagnoses. The standard deviations reflect moderate variability across folds, indicating that performance on new datasets from similar imaging sources would likely remain within the reported ranges.

The results from the NACHOS algorithm on the kidney OCT dataset are summarized in Table 2.4. The OCT images were partitioned into 10 folds, with each fold containing all the images from one kidney. When F_0 was reserved as the test set, hyperparameter configuration h_5 yielded the best cross-validation accuracy of 0.93, which was the average of 9 validation accuracies from the 9-fold cross-validation using F_1 to F_9. Then, a model trained on the nine folds from F_1 to F_9 using configuration h_5

achieved a test accuracy of $t_0 = 0.73$ on the reserved test fold F_0. The cross-testing loop in NACHOS repeated this process for the remaining nine folds. The test accuracies had a wide range from 0.71 to 0.96, with a standard deviation of 0.09, reflecting the significant data variability from kidney to kidney in this dataset.

The average test accuracy from the 10 test folds was 0.84 with a standard error of 0.03, corresponding to a 95% t-based confidence interval of [0.77, 0.90]. The cross-testing procedure reduced the standard deviation of the average test accuracy estimation by approximately three folds ($\sqrt{10}$) from 0.09 to 0.03 by averaging across 10 individual test accuracies from separate kidney samples. The standard deviation of the individual folds' test accuracy was higher in the kidney OCT data (0.09) than the chest X-ray data (0.06), but the standard error of the average test accuracy was 0.03 in both datasets due to the higher number of test folds used in the kidney OCT data compared to the chest X-ray data. From a practical standpoint, this indicates that when the model is applied to new kidney samples in a clinical context, performance can be expected to vary by roughly $\pm 3\%$ around the mean accuracy, reflecting moderate but manageable variability in generalization across patients.

The average confusion matrix across the ten test folds is shown in Table 2.6. The models achieved high accuracy but with notable differences in per-class performance and variability across folds. Cortex was correctly identified in 454 $\pm$ 136 of 600 cases, with the most frequent confusion being misclassification as Pelvis (106 $\pm$ 147), indicating that these two tissues share imaging characteristics that can occasionally lead to errors. Medulla achieved a high correct classification rate (512 $\pm$ 137 of 600), with very few false predictions as Pelvis (8 $\pm$ 19) and some misclassification as Cortex (80 $\pm$ 136), suggesting strong discriminative features for this class. Pelvis was correctly classified in 543 $\pm$ 76 of 600 cases, with most of its misclassifications occurring as Cortex (55 $\pm$ 72). The relatively large standard deviations reflect variability in model

performance across different kidneys, likely due to inter-kidney anatomical differences and imaging conditions. Clinically, these results suggest that while Medulla and Pelvis classifications are generally robust, distinguishing Cortex from Pelvis in new kidney data remains more challenging, and further refinement or additional training data may help improve consistency.

Figure 2.3 shows Grad-CAM results for cross-testing, highlighting the regions most influential in the model's decision, for test fold 7 on Pelvis images. Figure 2.3A illustrates a correct classification, while Figure 2.3B shows an incorrect one. The primary difference in the red-highlighted region is that Figure 2.3A contains a greater concentration of bright speckle clusters, whereas Figure 2.3B appears smoother and more uniform, which likely contributed to the misclassification.

2.3.2 Development of Production Model for Deployment Using the DACHOS Algorithm

The DACHOS algorithm was used to produce a final model for deployment. The results of applying DACHOS to the chest X-ray repository are shown in Table 2.7. When hyperparameter configuration h_0 was used, the validation accuracies were $v_0^0 = 0.71$ for fold F_0, $v_1^0 = 0.71$ for fold F_1, $v_2^0 = 0.69$ for fold F_2, and $v_3^0 = 0.70$ for fold F_3, which yielded a cross-validation accuracy of $\bar{v}^0 = 0.71$ for h_0. DACHOS and NACHOS produced different validation accuracies for the same configuration, because DACHOS used 3 data folds for training while NACHOS used only 2. Four-fold cross-validation was performed for all remaining hyperparameter configurations; configuration h_5 achieved the highest cross-validation accuracy of $\bar{v}^5 = 0.75$. This configuration was then used to train a final model on all four data folds.

For the same configuration, 78% of the cross-validation accuracies for NACHOS were lower than the cross-validation accuracy for DACHOS. For example, with hyperparameter configuration h_5, NACHOS achieved cross-validation accuracies of 0.72, 0.73, 0.71, and 0.69 for test folds F_0 to F_3 using 2 data folds for training. In contrast, DACHOS achieved a cross-validation accuracy of 0.75 with h_5 using 3 data folds for training. An additional data fold available for training in DACHOS may have contributed to its increased cross-validation accuracy.

The DACHOS algorithm was also applied to the kidney OCT dataset (Table 2.8). Here, hyperparameter configuration h_0 achieved validation accuracies of 0.63, 0.88, 0.85, 0.76, 0.91, 0.97, 0.92, 0.87, 0.94, and 0.88 for folds F_0 to F_9, which averaged to a cross-validation accuracy of $\bar{v}^0 = 0.86$. This process was repeated for the remaining

Table 2.3: NACHOS accuracy results for chest X-ray repository

		Fold reserved for Test			
	Hyperparameter configuration	F_0	F_1	F_2	F_3
AHPO/Cross-Validation	h_0	v_1^0: 0.53 v_2^0: 0.61 v_3^0: 0.75	v_0^0: 0.50 v_2^0: 0.50 v_3^0: 0.77	v_0^0: 0.50 v_1^0: 0.49 v_3^0: 0.50	v_0^0: 0.54 v_1^0: 0.58 v_2^0: 0.50
	h_1	v_1^1: 0.66 v_2^1: 0.68 v_3^1: 0.80	v_0^1: **0.80** v_2^1: **0.74** v_3^1: **0.84**	v_0^1: 0.81 v_1^1: 0.76 v_3^1: 0.71	v_0^1: 0.80 v_1^1: 0.72 v_2^1: 0.50
	h_2	v_1^2: **0.69** v_2^2: **0.70** v_3^2: **0.76**	v_0^2: 0.60 v_2^2: 0.67 v_3^2: 0.71	v_0^2: 0.70 v_1^2: 0.64 v_3^2: 0.66	v_0^2: 0.67 v_1^2: 0.68 v_2^2: 0.61
	h_3	v_1^3: 0.67 v_2^3: 0.63 v_3^3: 0.82	v_0^3: 0.67 v_2^3: 0.65 v_3^3: 0.79	v_0^3: 0.71 v_1^3: 0.63 v_3^3: 0.76	v_0^3: 0.69 v_1^3: 0.66 v_2^3: 0.56
	h_4	v_1^4: 0.50 v_2^4: 0.50 v_3^4: 0.50	v_0^4: 0.51 v_2^4: 0.50 v_3^4: 0.78	v_0^4: 0.50 v_1^4: 0.50 v_3^4: 0.67	v_0^4: 0.50 v_1^4: 0.50 v_2^4: 0.50
	h_5	v_1^5: 0.69 v_2^6: 0.69 v_3^5: 0.76	v_0^5: 0.76 v_2^5: 0.67 v_3^5: 0.76	v_0^5: 0.70 v_1^5: 0.69 v_3^5: 0.74	v_0^5: **0.73** v_1^5: **0.72** v_2^5: **0.64**
	h_6	v_1^6: 0.50 v_2^6: 0.51 v_3^6: 0.77	v_0^6: 0.77 v_2^6: 0.66 v_3^6: 0.76	v_0^6: 0.75 v_1^6: 0.50 v_3^6: 0.78	v_0^6: 0.73 v_1^6: 0.50 v_2^6: 0.50
	h_7	v_1^7: 0.70 v_2^7: 0.49 v_3^7: 0.78	v_0^7: 0.77 v_2^7: 0.69 v_3^7: 0.80	v_0^7: 0.74 v_1^7: 0.67 v_3^7: 0.74	v_0^7: 0.51 v_1^7: 0.69 v_2^7: 0.56
	h_8	v_1^8: 0.50 v_2^8: 0.61 v_3^8: 0.78	v_0^8: 0.80 v_2^8: 0.71 v_3^8: 0.80	v_0^8: **0.81** v_1^8: **0.74** v_3^8: **0.77**	v_0^8: 0.81 v_1^8: 0.73 v_2^8: 0.50
	best: h_{j^*}	h_2	h_1	h_8	h_5
	Average	$\bar{v}^2$: 0.72	$\bar{v}^1$: 0.79	$\bar{v}^8$: 0.77	$\bar{v}^5$: 0.69
Cross-Testing	Accuracy	t_0: 0.79	t_1: 0.71	t_2: 0.69	t_3: 0.82
	Average and standard error 0.75±0.03				

Note: F_i: represents the test fold i. h_j: represents the hyperparameter configuration j. In the AHPO/CV loop, inside each cell, the validation accuracy v_m^j for hyperparameter configuration j and validation fold m is located. For each test fold, the best hyperparameter configuration is selected by comparing the average validation accuracy and the whole cell is highlighted in blue. For instance, for F_0, the hyperparameter configuration h_2 has the highest average accuracy. $\bar{v}^2$=(0.69+0.70+0.76)/3=0.72. In the cross-testing loop, the selected hyperparameter configuration is used to calculate the test accuracy t_i for test fold i. For test fold F_0, test accuracy is t_0=0.79

Table 2.4: NACHOS accuracy results for kidney OCT dataset

		Fold reserved for Test									
	Hyperparameter configuration	F_0	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
AHPO/ Cross-Validation	h_0	0.91±0.02	0.87±0.04	0.83±0.06	0.85±0.04	0.86±0.03	0.86±0.03	**0.87±0.03**	0.85±0.03	0.85±0.03	0.87±0.03
	h_1	0.92±0.01	0.86±0.04	0.87±0.03	0.83±0.06	0.85±0.06	0.88±0.04	0.83±0.04	0.87±0.03	0.85±0.03	0.87±0.04
	h_2	0.89±0.02	0.87±0.03	**0.91±0.02**	0.88±0.02	0.85±0.04	**0.89±0.02**	0.82±0.04	**0.90±0.02**	0.88±0.02	**0.89±0.03**
	h_3	0.89±0.03	0.84±0.05	0.85±0.05	0.82±0.06	0.83±0.06	0.80±0.06	0.81±0.06	0.83±0.04	0.83±0.06	0.84±0.05
	h_4	0.86±0.03	0.88±0.04	0.85±0.04	0.81±0.04	0.83±0.04	0.85±0.03	0.84±0.03	0.83±0.03	0.80±0.03	0.83±0.04
	h_5	**0.93±0.01**	**0.89±0.03**	0.89±0.02	**0.90±0.02**	**0.88±0.03**	0.89±0.02	0.82±0.03	0.90±0.02	0.87±0.02	0.87±0.03
	h_6	0.88±0.02	0.86±0.04	0.89±0.03	0.85±0.06	0.82±0.05	0.88±0.03	0.85±0.04	0.83±0.03	0.85±0.03	0.87±0.04
	h_7	0.89±0.01	0.87±0.04	0.87±0.03	0.86±0.03	0.87±0.03	0.88±0.03	0.87±0.03	0.83±0.02	0.80±0.03	0.88±0.03
	h_8	0.91±0.02	0.89±0.04	0.89±0.03	0.86±0.04	0.86±0.04	0.88±0.03	0.85±0.05	0.88±0.04	**0.89±0.03**	0.88±0.04
	best: h_{j^*}	h_5	h_5	h_2	h_5	h_5	h_2	h_0	h_2	h_8	h_2
Cross-Testing	Accuracy	t_0: 0.73	t_1: 0.71	t_2: 0.79	t_3: 0.75	t_4: 0.88	t_5: 0.87	t_6: 0.94	t_7: 0.90	t_8: 0.96	t_9: 0.86
	Average and standard error 0.84±0.03										

Note: F_i: represents the test fold i. h_j: represents the hyperparameter configuration j. In the AHPO/CV loop, inside each cell, the average and standard error of the validation accuracies v_m^j for hyperparameter configuration j and validation fold m are located. The cells highlighted in blue have the highest average accuracy for a test fold. For instance, for F_0, the hyperparameter configuration h_5 has the highest average accuracy. $\bar{v}^5$=0.93. In the cross-testing loop, the selected hyperparameter configuration is used to calculate the test accuracy t_i.

Table 2.5: Average confusion matrix for chest X-ray classification Cross-Testing loop NACHOS

		Predicted	
		No-Finding	Cardiomegaly
Truth	No-Finding	426±82	194±82
	Cardiomegaly	115±66	505±66

Note: average and standard deviation. Ground Truth: 612 No-Finding and 612 Cardiomegaly

hyperparameter configurations. Hyperparameter configuration h_2 had the highest average accuracy $\bar{v}^2 = 0.90$, and h_3 had the lowest $\bar{v}^3 = 0.84$. The production model was trained using h_2 on the entire dataset.

Table 2.6: Average confusion matrix for chest OCT classification Cross-Testing loop NACHOS

		Predicted		
		Cortex	Medulla	Pelvis
Truth	Cortex	454±136	40±58	106±147
	Medulla	80±136	512±137	8±19
	Pelvis	55±72	2±6	543±76

Note: average and standard deviation. Ground Truth: 600 Cortex, 600 Medulla, and 600 Pelvis

2.3.3 Evaluation and Interpretation of Model Performance Across Different Partitioning Levels

NACHOS and DACHOS required data to be partitioned into multiple folds for NCV and CV. An appropriate partitioning design was essential for evaluating a model's ability to generalize to unseen data from a new measurement, a new patient, or a new location. The input data in the chest X-ray repository was organized into three partitioning levels, including the image level, the patient level, and the dataset level, for testing to evaluate different partitioning designs (Figure 2.4A). The chest X-ray repository comprised a total of 4,960 images from 4,678 patients across four datasets. There were 1,187 patients in the CheXpert dataset, 1,165 patients in the MIMIC-CXR dataset, 1,105 patients in the ChestX-ray8 dataset, and 1,221 patients in the PadChest dataset. The dataset-level partitioning was used in the previous results sections to evaluate the model performance on unseen data from a different location. Here, we compared the test performance benchmarked using NACHOS across image-level, patient-level, and dataset-level partitioning (Figure 2.4B). Because each patient had approximately 1.06 images on average in the chest X-ray repository, the image-level and patient-level partitioning yielded similar average test accuracies of 0.811 and 0.809, respectively, reflecting the test performance of the models on unseen data from new images or new patients within the four datasets. These accuracies were much higher than the average test accuracy of 0.750 from the dataset-level partitioning. The variability in test accuracies across data folds also increased from 0.008 for the image-level partitioning and 0.009 for the patient-level partitioning to 0.061 for the dataset-level stratification. A one-way ANOVA found no statistically significant difference in accuracy distributions among the three strategies ($F = 3.70$, $p = 0.067$), with variance attributed to partitioning (treatment) estimated at 0.00088 and residual variance at

0.001297. However, pairwise effect size analysis indicated practically meaningful differences: the contrast between image-level and patient-level was small (Cohen's d = 0.24), whereas comparisons involving dataset-level were large (image vs. dataset: d = 1.40; patient vs. dataset: d = 1.35). These results suggest that while statistical significance was not reached at the conventional 0.05 threshold, dataset-level partitioning had a substantial practical impact, markedly reducing test accuracy compared with the other strategies. From a clinical reliability perspective, the low fold-to-fold variability at the image and patient levels indicates that model performance is relatively stable when tested on data similar to the training set, which increases confidence in consistent results within the same clinical setting. In contrast, the larger variability at the dataset level reflects reduced predictability when applying the model to data from new sites, underscoring the need for external validation before deployment in different clinical environments.

The test performances were also benchmarked using NACHOS on the kidney OCT dataset at three partitioning levels: image, volume, and kidney (Figure 2.5A). The kidney-level partitioning was used in the previous results sections to evaluate the model performance on unseen data from a new kidney. Here, we compared the test performance benchmarked by NACHOS using the image-level, volume-level, and kidney-level partitioning (Figure 2.5B). The image-level partitioning resulted in a perfect test accuracy of 1.00 across all data folds, owing to nearly complete data redundancy among contiguous images from the same volume. The volume-level partitioning generated an average test accuracy of 0.97 and a standard deviation of 0.01 across different data folds, suggesting still substantial data redundancy among volumes from the same kidney. In contrast, the kidney-level partitioning yielded an average test accuracy of 0.84

and a standard deviation of 0.09. The reduced accuracy and increased variability reflected the real-world variability of the OCT data from different kidneys and the need for the models to generalize well across kidneys. A one-way ANOVA was conducted to compare mean accuracies across the three splits. The analysis indicated a significant effect of split on accuracy, F(2,27)=27.08, p=3.52×10^{-7}. The between-split variance component was estimated at 0.007043, and the within-split (error) variance component at 0.002700. Post-hoc comparisons using Tukey's HSD revealed that: Split 1 vs. Split 3: Split 1 (M = 0.811) had significantly higher accuracy than Split 3 (M = 0.750), mean difference = 0.162, 95% CI [0.104, 0.219], $p < 0.001$. Split 2 vs. Split 3: Split 2 (M = 0.809) had significantly higher accuracy than Split 3 (M = 0.750), mean difference = 0.130, 95% CI [0.072, 0.187], $p < 0.001$. Split 1 vs. Split 2: No significant difference was found, mean difference = 0.032, 95% CI [–0.026, 0.090], p = 0.367. Effect size analysis showed that these differences were not only statistically significant but also large in magnitude: image-level vs. volume-level (d=3.94), image-level vs. kidney-level (d=2.56), and volume-level vs. kidney-level (d=2.04), underscoring the substantial practical impact of partitioning choice on OCT model performance. From a clinical reliability perspective, the very low fold-to-fold variability at the image and volume levels suggests that the model would produce consistent results when applied to data that is highly similar to its training set. In contrast, the markedly higher variability at the kidney level indicates that performance can fluctuate substantially when encountering entirely new patient samples, highlighting the importance of validating the model across diverse clinical scenarios before broader adoption.

2.3.4 Parallelization of NACHOS and DACHOS Over Multiple GPUs

We compared the execution time of the NACHOS algorithm on various GPU systems using the kidney OCT dataset with kidney-level partitioning and a single hyperparameter configuration. These systems included a Beowulf cluster of GPU workstations with NVIDIA GeForce RTX 4090 or NVIDIA RTX A6000 GPUs on a local Ethernet network, as well as the OSCER supercomputer with NVIDIA A100 GPUs. The execution time was 21.9 hours on a single RTX A6000, 13.8 hours on a single RTX 4090, and 11.7 hours on a single A100 (Figure 2.6A). The RTX A6000 has 336 tensor cores and a memory bandwidth of 112.5 GB/s, the RTX 4090 has 512 tensor cores and a memory bandwidth of approximately 1000 GB/s, and the A100 has 432 tensor cores and a memory bandwidth of approximately 2000 GB/s. The RTX 4090 outperformed the A6000 likely due to its higher number of tensor cores. The A100 outperformed the RTX 4090 likely due to its greater memory bandwidth and optimizations for deep learning applications. The peak memory usage on all these GPUs was approximately 18.5 GB. These results demonstrate the portability of NACHOS and DACHOS across a variety of computing systems.

NACHOS and DACHOS can distribute the training across multiple GPUs with fault tolerance to reduce the wall-clock time. Figure 2.6B presents the speedup ratios by the number of GPUs. NACHOS reached linear scalability on RTX A6000s. The execution time was reduced to 11.1 hours on 2 RTX A6000s with a 2.0× speedup, further to 6.9 hours on 3 RTX A6000s with a 3.2× speedup, and finally to 5.4 hours on 4 RTX A6000s with a 4.1× speedup.

NACHOS also achieved linear speedups on the RTX 4090s. The execution time was 13.8 hours on 1 RTX 4090, 7.0 hours on 2 RTX 4090s (2.0× speedup), 4.7 hours on

3 RTX 4090s (2.9× speedup), and 3.6 hours on 4 RTX 4090s (3.8× speedup). Linear speedups were also achieved on A100s: 2.0× on 2 A100s, 4.0× on 4 A100s, and 7.6× on 8 A100s. The execution time was reduced to only 1.6 hours using 8 A100s. Detailed results are provided in Table 6.1 in the Appendix.

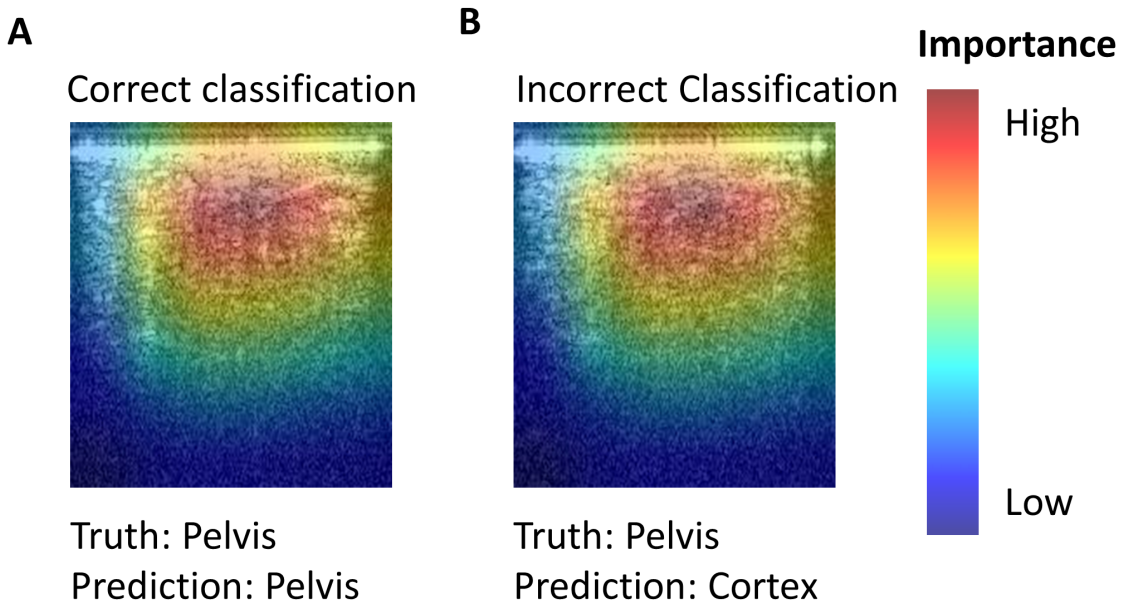

Figure 2.3: **Grad-CAM explainability results for cross-testing on OCT dataset for fold k7.** [A] Correct classification [B] Incorrect Classification

Table 2.7: DACHOS accuracy results for chest X-ray repository

Hyperparameter configuration	Fold reserved for Validation				
	F_0	F_1	F_2	F_3	Average
h_0	v_0^0: 0.71	v_1^0: 0.71	v_2^0: 0.69	v_3^0: 0.70	$\bar{v}^0$: 0.71
h_1	v_0^1: 0.79	v_1^1: 0.63	v_2^1: 0.71	v_3^1: 0.71	$\bar{v}^1$: 0.71
h_2	v_0^2: 0.78	v_1^2: 0.70	v_2^2: 0.69	v_3^2: 0.65	$\bar{v}^2$: 0.71
h_3	v_0^3: 0.75	v_1^3: 0.69	v_2^3: 0.66	v_3^3: 0.80	$\bar{v}^3$: 0.72
h_4	v_0^4: 0.50	v_1^4: 0.50	v_2^4: 0.59	v_3^4: 0.78	$\bar{v}^4$: 0.59
h_5	v_0^5: 0.75	v_1^5: 0.74	v_2^5: 0.71	v_3^5: 0.80	$\bar{v}^5$: **0.75**
h_6	v_0^6: 0.73	v_1^6: 0.71	v_2^6: 0.73	v_3^6: 0.77	$\bar{v}^6$: 0.73
h_7	v_0^7: 0.77	v_1^7: 0.72	v_2^7: 0.63	v_3^6: 0.75	$\bar{v}^7$: 0.72
h_8	v_0^8: 0.81	v_1^8: 0.74	v_2^8: 0.63	v_3^8: 0.81	$\bar{v}^8$: 0.75
best: h_{j^*}	h_5				

Note: F_m: represents the validation fold m. h_j: represents the hyperparameter configuration j. Inside each cell, the validation performance v_m^j for hyperparameter configuration j and validation fold m is located. In order to select the configuration hyperparameter, the highest average accuracy $\bar{v}^j$ is selected. Among them, h_5 has the highest average validation accuracy $\bar{v}^5$=0.75, highlighted in blue. Despite apparent equality due to rounding, h_5's average validation accuracy is higher than h_8.

Table 2.8: DACHOS accuracy results for kidney OCT dataset

	Fold reserved for Validation										
Hyperparameter configuration	F_0	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	Average
h_0	0.63	0.88	0.85	0.76	0.91	0.97	0.92	0.87	0.94	0.88	$\bar{v}^0$: 0.86
h_1	0.51	0.91	0.79	0.93	0.90	0.91	0.98	0.84	0.97	0.90	$\bar{v}^1$: 0.86
h_2	0.77	0.88	0.81	0.91	0.96	0.93	0.92	0.91	0.97	0.92	$\bar{v}^2$: **0.90**
h_3	0.37	0.90	0.82	0.83	0.87	0.91	0.91	0.92	0.98	0.87	$\bar{v}^3$: 0.84
h_4	0.68	0.89	0.86	0.84	0.78	0.99	0.88	0.86	0.93	0.91	$\bar{v}^4$: 0.86
h_5	0.77	0.88	0.89	0.79	0.94	0.97	0.92	0.91	0.96	0.93	$\bar{v}^5$: 0.90
h_6	0.66	0.88	0.81	0.91	0.84	0.95	0.91	0.90	0.97	0.92	$\bar{v}^6$: 0.88
h_7	0.69	0.87	0.84	0.93	0.85	0.79	0.95	0.86	0.97	0.91	$\bar{v}^7$: 0.87
h_8	0.64	0.89	0.69	0.94	0.92	0.92	0.97	0.90	0.97	0.90	$\bar{v}^8$: 0.88
best: h_{j^*}	h_2										

Note: F_m: represents the validation fold m. h_j: represents the hyperparameter configuration j. Inside each cell, the validation accuracy v_m^j for hyperparameter configuration j and validation fold m is located. In order to select the configuration hyperparameter, the highest average accuracy $\bar{v}^j$ is selected. Among them, h_2 has the highest average validation accuracy $\bar{v}^2$=0.90, highlighted in blue. Despite apparent equality due to rounding, h_2's average validation accuracy is higher than h_5.

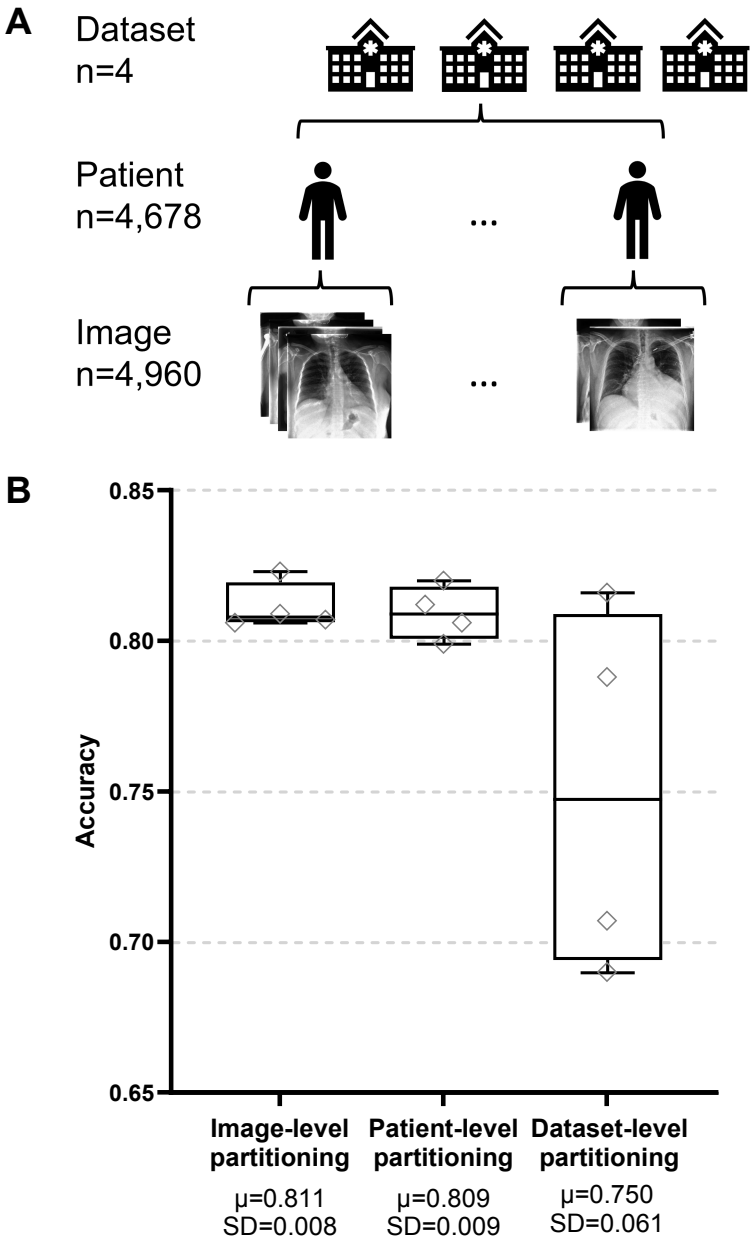

Figure 2.4: **Data partitioning schemes of the chest X-ray repository.** [A] Data structure of the chest X-ray repository. The repository includes four datasets, each originating from a different set of institutions, capturing variations in imaging protocols and patient populations. [B] Mean (μ) and standard deviation (SD) of the test accuracy from data partition on the image level, the patient level, and the dataset level. The four open circles represent the test accuracies of four partitions of mixed images regardless of patients, four partitions of mixed patients regardless of their source datasets, and four partitions corresponding to the four datasets. The dataset-level partitioning had significantly lower mean and higher variability in the test accuracies.

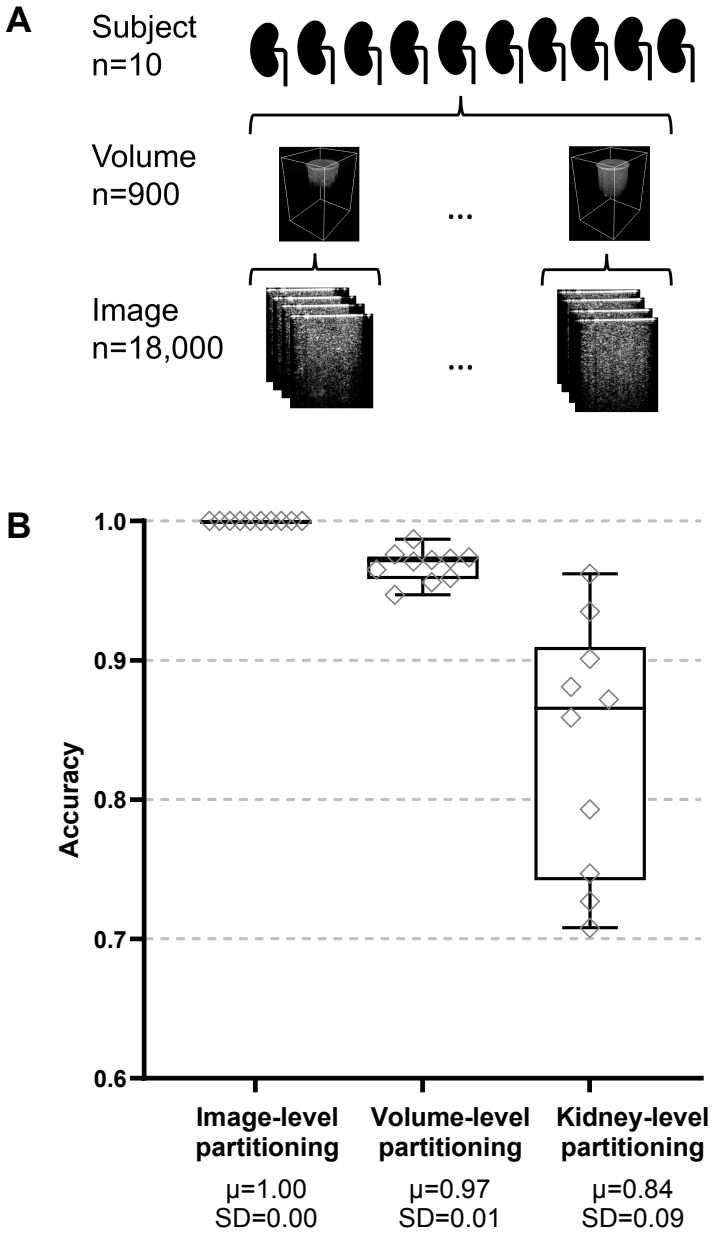

Figure 2.5: **Data partitioning schemes of the kidney OCT dataset.** [A] Data structure of the kidney OCT dataset. The dataset includes 10 kidneys, each generating 90 OCT volumes. 20 B-scan images are extracted from each volume. [B] Mean (μ) and standard deviation (SD) of the test accuracy from data partition on the image level, the volume level, and the kidney level. The 10 open circles represent the test accuracies of 10 partitions of mixed images regardless of volume, 10 partitions of mixed volume regardless of their source kidney, and 10 partitions corresponding to the 10 kidneys. The kidney-level partitioning had significantly lower mean and higher variability in the test accuracies.

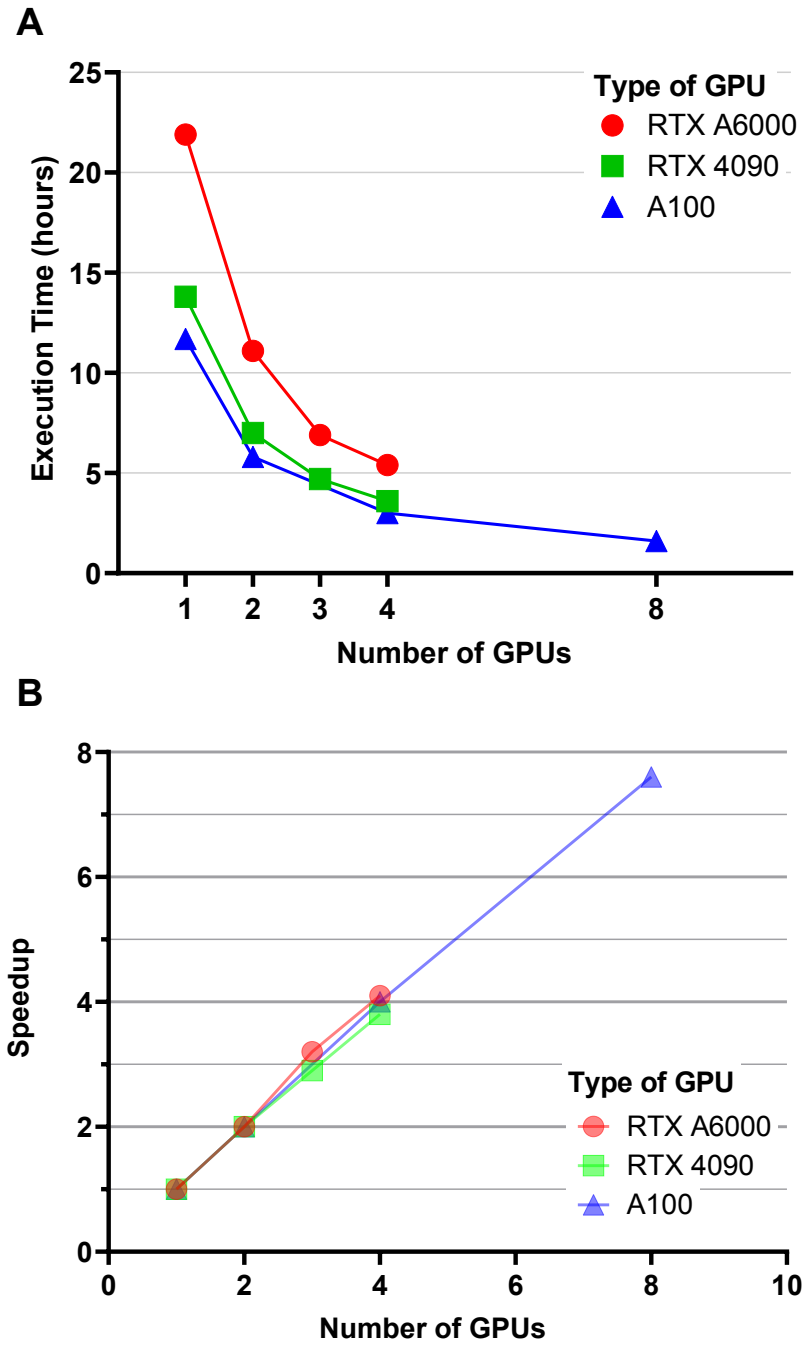

Figure 2.6: **Scalability of NACHOS parallelization.** [A] Execution time as a function of the number of GPUs for three GPU types: RTX A6000 (Beowulf cluster), RTX 4090 (Beowulf cluster), and A100 (supercomputer). [B] Speedup relative to the number of GPUs. NACHOS achieved linear speedup on the A6000 GPUs, RTX 4090 GPUs, and the A100 GPUs.

2.4 Discussion

Accurate and robust benchmarking of the performance of machine learning models has been a challenge in the field of medical imaging (Varoquaux and Cheplygina, 2022). A commonly used approach is to split a labeled dataset into a training set for learning model parameters, a validation set for optimizing model hyperparameters, and a test set for benchmarking the performance of the obtained model. Although cross-validation is often used to rotate data partitions between the training set and the validation set, most machine learning studies split out a single fixed test set for performance benchmarking.

NACHOS features NCV in an automated and user-friendly machine learning workflow for medical imaging applications. k-fold NCV offers two key advantages over a single test split for test performance benchmarking. First, NCV reduces the variance of the test performance estimation by rotating all data partitions through the test set. Specifically, the variance of the average performance score from k-fold NCV should be k times lower than the variance of the point estimate of the performance score from a single test split. Second, and more importantly, the variance of the performance estimation is not only reduced but also quantified during cross-testing over k test partitions.

For example, if a user holds out only the last partition for the test set, they would estimate the classification accuracy to be 0.82 in the chest X-ray repository and 0.86 in the kidney OCT dataset, with large ($\pm$0.06 and $\pm$0.09, respectively) variabilities that are unknown to the user. If the user uses NCV, they would be able to better estimate the accuracy as 0.75 ± 0.03 in the chest X-ray repository and as 0.84 ± 0.03 in the kidney OCT dataset. These use cases demonstrate that NCV reduces and quantifies the variance of performance benchmarking for deep learning models.

The partitioning level used by NCV is also important for accurate and robust performance benchmarking. The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) (Mongan et al., 2020) emphasizes transparent reporting of data partitioning and recommends partitioning at the patient level or higher. The classification accuracy in the chest X-ray repository decreased from 0.809 with the patient-level partitioning to 0.750 with the dataset-level partitioning (Figure 2.4B). This means that it is more difficult for models to generalize to a new dataset acquired by other institutions than to generalize to new patients from a previously seen dataset.

Similarly, Zech et al. (2018) found that pneumonia classification models often perform better on internal test datasets originating from the same institution as the training data, than on external test datasets from institutions different from the ones used for training. The choice of the partitioning level for NCV performance benchmarking should match the intended use scenario of the models. For instance, the patient-level partitioning can be used to evaluate models designed for use within the same hospitals that produced the training data, as it mimics the scenario of encountering new patients within these hospitals. The institution-level partitioning should be used to evaluate models intended to be deployed to new hospitals.

NACHOS benchmarks the average test performance of models generated by a reproducible workflow using a specific dataset. AHPO is needed in NACHOS because it is impractical to perform laborious manual hyperparameter optimization consistently across all test folds in NCV. Random search (Bergstra and Bengio, 2012) is a simple, yet highly effective, AHPO method. In the chest X-ray repository (Table 2.3), the best hyperparameter configuration achieved improvements ranging from 0.07 to 0.10 over the average cross-validation accuracy of all configurations. In the kidney OCT dataset (Table 2.4), AHPO delivered performance gains between 0.02 and 0.05 above the overall average.

NCV and AHPO in NACHOS incur a significant computational cost that needs to be distributed across multiple GPUs to shorten the wall-clock time of model development. The parallelization in NACHOS achieved linear speedups on different kinds of GPUs (Figure 2.6A and Figure 2.6B). For example, distributing training across four RTX A6000 GPUs reduced runtime from 21.9 to 5.4 hours—representing a 75% reduction in runtime to a single GPU. On A100s, runtime decreased from 11.7 hours on a single GPU to just 1.6 hours on eight GPUs—an 86% reduction in runtime. These results underscore not only the portability of NACHOS across heterogeneous GPU architectures but also its strong scalability and efficiency for large-scale medical AI workflows.

After NACHOS measures the test performance of models from a model development workflow, DACHOS is used to generate a production model for deployment using this workflow. The actual test performance of this production model is unknown but should be higher than the test performance of the models benchmarked by NACHOS. This is because the AHPO in DACHOS can use an extra data fold compared to the AHPO in NACHOS, and the final model training in DACHOS can use two additional data folds compared to the training in NACHOS.

Beyond performance benchmarking, NACHOS provides built-in support for interpretability through automated generation of class activation maps and receiver operating characteristic (ROC) curves, offering insight into both how a model makes predictions and where errors may occur. Such interpretability features are increasingly recognized as essential for explainable AI in healthcare, making them a core benefit of adopting NACHOS in addition to its variance-aware benchmarking capabilities.

In addition to variance-aware benchmarking and hyperparameter optimization, NACHOS incorporates engineering features that make it suitable for large-scale medical AI

workflows. The framework demonstrated linear GPU scalability, ensuring that computationally intensive NCV and AHPO processes can be executed efficiently across modern HPC resources. Furthermore, NACHOS includes a fault-tolerant checkpointing system that safeguards progress during long training runs, enabling workflows to resume seamlessly after interruptions. Together, these capabilities underline the framework's robustness and practicality for large-scale, distributed model development in medical imaging.

2.4.1 Limitations

This study has several limitations that warrant consideration. First, we evaluated NACHOS and DACHOS only on chest X-ray and OCT datasets, and focused exclusively on image classification tasks. The framework needs to be further evaluated for other modalities such as CT or MRI, and for additional tasks such as segmentation and detection. While the principles of nested cross-validation, automated hyperparameter optimization, and distributed parallelization are broadly applicable, empirical validation across diverse imaging types and tasks will be necessary to confirm the generalizability of the framework.

Second, although nested cross-validation provides a strong foundation for variance-aware performance estimation, the study did not include evaluation on fully independent, external datasets. For the chest X-ray repository, our use of dataset-level partitioning partially addresses this by rotating test folds across data originating from different institutions, thereby simulating cross-site generalization. However, we acknowledge that this approach does not replace validation on entirely unseen, external datasets. Such validation across independent multi-institutional cohorts will be an important next step to more rigorously establish the robustness and real-world reproducibility of the framework.

Third, we did not conduct a quantitative comparison of NACHOS against existing machine learning pipelines such as NiftyNet or TorchIO. Such comparisons could provide more direct evidence of practical improvements in performance benchmarking, computational efficiency, and reproducibility. Our emphasis here was on introducing the unified integration of nested cross-validation, automated hyperparameter optimization, and high-performance computing within a single workflow, and demonstrating its feasibility across two representative datasets.

Fourth, while our results demonstrate scalability and portability across GPU platforms, the framework's dependence on multi-GPU setups and high-performance computing clusters introduces a barrier to adoption in lower-resource settings. This reflects a tradeoff between the reproducibility and robustness gained from nested cross-validation with automated hyperparameter optimization and the computational burden required to execute these workflows at scale. Broader adoption may require adaptations such as more efficient AHPO strategies, or cloud-based solutions to lower the entry barrier for institutions with limited infrastructure.

Fifth, while the framework is intended to support deployment in clinical workflows, this study does not present clinical validation or comparison against human expert performance. As such, the effectiveness, safety, and practical utility of models developed through NACHOS and DACHOS in real clinical environments remain unproven. Future work should include pilot clinical studies and direct benchmarking against human experts to more fully establish the framework's role in supporting trustworthy clinical AI adoption.

Sixth, the framework assumes that training and evaluation datasets contain accurate, noise-free labels. In practice, medical datasets often exhibit label noise and interobserver variability, especially in complex or ambiguous cases. Such label inconsistencies can influence both model performance and the estimated variance across

folds While the reproducibility and variance quantification benefits of nested cross-validation still hold under moderate label noise, future work should explore strategies for robustness, such as incorporating consensus labeling, or integrating noise-aware training methods (Li et al., 2018).

In conclusion, NACHOS integrates NCV, AHPO, and HPC into an automated workflow. NCV reduces and measures variance in test performance estimation by rotating all data folds through the test set. AHPO enhances model performance by searching for optimal hyperparameters and avoids the impracticality of manual tuning across multiple test folds in NCV. To mitigate the substantial computational costs of NCV and AHPO, NACHOS can distribute computation across multiple GPUs with linear speedup. After NACHOS completes model benchmarking, DACHOS can be used to generate final production models with potentially higher performance.

2.5 Conclusion

NACHOS integrates NCV, AHPO, and HPC into an automated workflow. NCV reduces and measures variance in test performance estimation by rotating all data folds through the test set. This variance reduction applies to the estimation process rather than guaranteeing lower variability when the model is applied to new, external clinical environments. AHPO enhances model performance by searching for optimal hyperparameters and avoids the impracticality of manual tuning across multiple test folds in NCV. To mitigate the substantial computational costs of NCV and AHPO, NACHOS can distribute computation across multiple GPUs with linear speedup. Although validated here on two imaging modalities (chest X-ray and OCT) and classification tasks, its design is general and can be extended to other medical imaging modalities and tasks such as segmentation. To our knowledge, this is the first framework to unify nested

cross-validation, automated hyperparameter optimization, and high-performance computing for medical imaging. While each of these components has been applied individually in prior studies, their integration in NACHOS represents a novel and impactful advancement. This combination not only ensures variance-aware performance benchmarking and reproducible optimization but also makes such workflows computationally feasible at scale, thereby addressing a key gap in current medical AI development pipelines.

Following benchmarking, DACHOS can be used to generate final models that may achieve higher accuracy. However, these models should be viewed as technically optimized research outputs, not as clinically validated tools ready for deployment. While NCV reduces estimation variability, some performance variation is inherent to dataset characteristics and may still be observed when applying models to new clinical settings.

By providing transparent, reproducible, and scalable evaluation, NACHOS and DACHOS help address reproducibility concerns in medical AI and could inform best practices for model validation. Importantly, the entire codebase (v1.0.0) is openly available at `https://github.com/thepanlab/NACHOS`, enabling broad community adoption, extension to other domains, and independent verification of results—thereby maximizing transparency and accelerating progress beyond this study's scope.

Beyond methodological contributions, NACHOS and DACHOS address critical gaps in current development pipelines that directly impact regulatory trust and reproducibility in clinical settings. By standardizing performance benchmarking, quantifying estimation variance, and enabling transparent validation across data partitions, the framework provides a foundation for more reliable evaluation of medical AI systems. These qualities are essential not only for advancing research reproducibility but also for informing best practices and regulatory standards that govern the safe and responsible integration of AI into healthcare.

Looking ahead, several extensions of this work are planned. First, NACHOS will be expanded to support additional AHPO strategies, such as Hyperband (Li et al., 2018) and Bayesian Optimization Hyperband (Falkner et al., 2018), to further improve search efficiency. Second, the framework will be applied to a broader range of medical imaging modalities and tasks, including segmentation and multi-label classification. Finally, integration into pilot clinical workflows will be explored to evaluate practical usability, interoperability with existing hospital systems, and potential barriers to real-world adoption.

Chapter 3

Deep learning for tool guidance using OCT

This chapter presents two studies of using machine learning with OCT images: the first is related to Percutaneous Nephrostomy (PCN) Guidance and the second to epidural Anesthesia Needle Guidance.

Published in:

- C. Wang[#], P. Calle[#], N. B. Tran Ton, Z. Zhang, F. Yan, A. M. Donaldson, N. A. Bradley, Z. Yu, K.-m. Fung, C. Pan, and Q. Tang, "Deep-learning-aided forward optical coherence tomography endoscope for percutaneous nephrostomy guidance," Biomedical Optics Express 12, 2404-2418 (2021).
- C. Wang[#], P. Calle[#], J. C. Reynolds, S. Ton, F. Yan, A. M. Donaldson, A. D. Ladymon, P. R. Roberts, A. J. de Armendi, K.-m. Fung, S. S. Shettar, C. Pan, and Q. Tang, "Epidural anesthesia needle guidance by forward-view endoscopic optical coherence tomography and deep learning," Scientific Reports 12, 9057 (2022).

#: Equal contribution

3.1 Introduction

Optical Coherence Tomography (OCT) is a non-invasive imaging technique that is analogous to ultrasound, but instead of sound waves, it uses low-coherence light to capture high-resolution cross-sectional images of biological tissues. First introduced

by Huang et al. (1991) for imaging the retina and human coronary arteries ex vivo. OCT quickly gained attention for its potential in ophthalmology. During the 1990s, early clinical studies demonstrated its utility in diagnosing various ocular conditions, including macular diseases (Puliafito et al., 1995) and macular edema (Hee et al., 1995), among others. Since then, OCT has become a cornerstone in ophthalmic imaging due to its capability to provide micrometer-scale resolution and real-time visualization.

Andrews et al. (2008) demonstrated the application of high-speed swept-source OCT to visualize the living rat kidney in situ, enabling detailed observation of renal tissues under normal and pathological conditions. Another study (Andrews et al., 2014) demonstrated the use of OCT and Doppler OCT (DOCT) as intra-operative guidance tools for assessing living human donor kidneys during transplantation. Using a handheld, sterile-compatible OCT probe, surgeons were able to visualize renal and real-time blood flow dynamics. This approach offered a promising, non-invasive method to guide clinical decision-making by providing immediate structural and functional assessment of donor organ viability during transplant surgery.

Endoscopic optical coherence tomography (OCT) enables high-resolution, cross-sectional imaging of internal tissues through miniature probes integrated into endoscopic instruments. This technology combines the depth-resolving capability of OCT with the accessibility of endoscopy, allowing real-time visualization of subsurface microstructures in hollow organs such as the gastrointestinal tract, airways, and urinary system. Its minimally invasive nature and ability to provide biopsy-like information without tissue removal make it a powerful guidance tool for early disease detection, procedural planning, and intraoperative assessment (Gora et al., 2017).

Endoscopic OCT has been successfully applied in various clinical settings, including early detection of Barrett's esophagus (Poneros and Nishioka, 2003) and specialized Intestinal Metaplasia (Poneros et al., 2001), and assessment of airway wall structure in

obstructive lung diseases (Coxson et al., 2008), Its minimally invasive nature and ability to provide biopsy-like information without tissue removal make it a powerful guidance tool for diagnostics, procedural navigation, and intraoperative decision-making.

Percutaneous nephrostomy (PCN) is a minimally invasive procedure used for accessing the renal collecting system, especially in complex cases where transurethral access is difficult. Despite being routine, PCN remains technically challenging due to the difficulty of accurately guiding the needle to the renal pelvis while avoiding vascular injury. (Goodwin et al., 1955; Young and Leslie, 2025). Conventional imaging methods like ultrasound, fluoroscopy, and CT have limitations in spatial resolution and soft tissue differentiation, leading to failure rates as high as 18% and complications such as bleeding, infection, and prolonged operative times (Ramchandani et al., 2003; Montvilas et al., 2011). To improve guidance, enhanced ultrasound methods (e.g., contrast-enhanced (Cui et al., 2016; Montvilas et al., 2011) and Doppler (Lu et al., 2010)), cone-beam CT (Hawkins et al., 2016), and magnetic navigation devices (Krombach et al., 2001) have been explored. Endoscopic tools integrated with PCN needles offer better precision but are generally limited to 2D imaging and lack subsurface visualization (Miller and Wickham, 1984; Tang et al., 2016). Optical coherence tomography (OCT), with its high axial resolution ($\sim$ 10 µm) and depth-resolved capability, offers a promising alternative (Huang et al., 1991; Li et al., 2009). OCT can distinguish renal cortex, medulla, and calyx based on tissue microstructure, though its penetration depth is limited to 1–2 mm (Andrews et al., 2014; Wang and Chen, 2014). Our lab developed a forward-imaging endoscopic OCT needle system and demonstrated its use in imaging porcine kidneys. Using data from ten kidneys, we trained convolutional neural networks (ResNet34 (He et al., 2016), ResNet50 (He et al., 2016), MobileNetv2 (Sandler et al., 2018)) to classify renal tissue types. The models were validated using

nested cross-validation to account for biological variability, and interpretability methods were applied to understand the image features used for classification. This work supports OCT's potential for real-time, tissue-specific guidance in PCN procedures.

Epidural anesthesia is widely used in surgeries (Hollmann et al., 2001; Moraca et al., 2003; Svircevic et al., 2011) and pain management (Li et al., 2019), but its success depends on precise needle placement into the narrow epidural space (Scott, 1997). Misplacement can lead to serious complications such as paresthesia , post-dural puncture headache, and even paralysis (Horlocker, 2000). Currently, anesthesiologists rely on the loss-of-resistance (LOR) technique (Hoffmann et al., 1999), which is subjective and prone to failure, especially in complex cases (Hermanides et al., 2012; McLeod et al., 2005; Stojanovic et al., 2002; Tang et al., 2014). Imaging methods like ultrasound and fluoroscopy offer limited resolution or contrast for distinguishing critical soft tissues (Richardson and Groen, 2005; Tang et al., 2014). Optical coherence tomography (OCT), a high-resolution imaging modality, has shown promise for visualizing tissue layers during needle insertion (Huang et al., 1991; Li et al., 2000). Building on prior work integrating OCT with epidural needles , we propose a computer-aided diagnosis (CAD) system using convolutional neural networks (CNNs) to classify epidural tissues in real time from forward-view OCT images. This study is the first to combine forward-view OCT and deep learning for epidural guidance and demonstrates its feasibility in distinguishing key anatomical layers.

This chapter combines two studies to present a unified deep learning framework for interpreting OCT images acquired via endoscopic OCT systems. Specifically, the framework is applied to (1) PCN procedures to classify renal tissues, and (2) epidural anesthesia needle insertion for automatic classification of epidural tissues and regression-based distance estimation of the needle tip to critical structures.

3.2 Methods

3.2.1 OCT Imaging System

The first study used swept-source OCT (SS-OCT); whereas, the second study used forward-imaging endoscopic OCT . Both operate at a central wavelength of 1300 nm and an axial resolution of approximately 10.6 µm. A gradient-index (GRIN) rod lens with a diameter of 1.3 mm was integrated with the OCT probe, enabling forward-view endoscopic imaging. This setup provided a lateral imaging field-of-view of approximately 1.25 mm. The GRIN lens was enclosed within a thin-walled protective steel tube, optimized for minimal dispersion and reflection artifacts, allowing high-quality imaging at high scanning speeds of up to 200 kHz.

3.2.2 Data Acquisition and Preprocessing

The PCN study involved imaging ten ex-vivo porcine kidneys, obtaining 1,000 cross-sectional images each for renal cortex, medulla, and calyx. The epidural study collected data from eight ex-vivo porcine spinal columns, acquiring OCT images from five epidural layers: fat, interspinous ligament, ligamentum flavum, epidural space, and spinal cord. Both datasets underwent histological validation to confirm tissue classifications.

To facilitate deep learning analyses, images were preprocessed by cropping to reduce noise and standardize field-of-view. PCN images were resized to 235×301 pixels, while epidural classification images were resized to 181×241 pixels. PCN data is available at `https://zenodo.org/records/7113948`. Epidural data is available at `http://doi.org/10.5281/zenodo.5018581`. Intensity normalization and channel replication were also performed to meet CNN input requirements.

3.2.3 Nested Cross-Validation

We employed Nested Cross-Validation (NCV) to provide an unbiased estimate of model performance. As summarized in Algorithm 1 (Chapter 2), NCV consists of three hierarchical loops:

- **Cross-Testing (Outer Loop)** : For the PCN study, each ex-vivo kidney was assigned as a distinct test fold, ensuring subject-level separation and preventing data leakage. For the epidural study, folds were defined analogously at the specimen level. The outer loop iterated through these folds, holding one out as the independent test set while using the remainder for training and validation.

- **Hyperparameter Optimization (Middle Loop)**: Within each outer split, an HPO loop explored multiple model architectures and hyperparameter configurations. Candidate models included convolutional neural networks adapted for each study (PCN and epidural), with variations in architecture type for these studies.

- **Cross-Validation (Inner Loop)**: For each hyperparameter configuration, a classical k-fold cross-validation was performed on the training set. This inner loop estimated the generalization performance for each candidate configuration, guiding model selection while avoiding overfitting to the outer test fold.

This design ensured that test accuracy reflected true out-of-sample performance, validation accuracy guided hyperparameter selection, and cross-validation accuracy quantified variance within the training data. By structuring the folds at the specimen level, we minimized the risk of information leakage across images derived from the same organ.

3.2.4 Training

3.2.4.1 PCN Study

Three CNN architectures were evaluated: Residual Network 34 (ResNet34) (He et al., 2016), Residual Network 50 (ResNet50) (He et al., 2016), and MobileNetv2 (Sandler et al., 2018), implemented using TensorFlow 2.3 (Abadi et al., 2016). These models were chosen to span a spectrum of representational capacity and computational complexity. ResNet34 provides a shallower baseline with fewer parameters, ResNet50 offers greater depth and capacity for learning fine-grained tissue patterns, and MobileNetv2 represents a lightweight, efficiency-oriented architecture designed for deployment scenarios where computational resources may be limited.

Pre-trained ResNet50 and MobileNetv2 models from ImageNet (Deng et al., 2009) were fine-tuned via a two-stage training strategy (Géron, 2019). Images were resized to 224 $\times$ 244 pixel resolution. This approach is commonly used to quickly adapt task-specific layers without disrupting generic pretrained features (Shin et al., 2016). Specifically, stochastic gradient descent (SGD) with a learning rate of 0.2, momentum of 0.3, and decay of 0.01 was employed. In the second stage, all layers were unfrozen and optimized with SGD using a lower learning rate (0.01), higher momentum (0.9), and smaller weight decay (0.001), settings that stabilize fine-tuning while preserving the pretrained representations. Early stopping criteria were applied to mitigate overfitting, with patience set to 10 epochs (maximum 50) for ResNet50 and 20 epochs (maximum 100) for MobileNetv2.

ResNet34 and ResNet50 were trained from randomly initialized weights to assess performance without transfer learning. For ResNet50, SGD with Nesterov momentum (learning rate 0.01, momentum 0.9, decay 0.01) and a batch size of 32 were used, following established practice for training deep residual networks from scratch. In

contrast, ResNet34 was optimized using the Adam algorithm (learning rate 0.001, β_1=0.9, β_2=0.9999, ε=1e-7) with a batch size of 512. Adam was chosen for ResNet34 because its adaptive learning rates help stabilize training when models are initialized randomly and trained on moderately sized datasets. Hyperparameters were selected based on prior transfer learning literature (Géron, 2019; Shin et al., 2016) and refined through preliminary experiments. For all models, OCT images were single-channel and mean-centered using the statistics of the training set to normalize intensity distributions and improve optimization stability.

3.2.4.2 Epidural Study

For the epidural guidance study, three CNN architectures were evaluated: Inception (Szegedy et al., 2016), ResNet50 (He et al., 2016), and Xception (Chollet, 2017), implemented using Keras within TensorFlow 2.3. These models were chosen to balance architectural diversity and performance potential. Inception introduces multi-scale feature extraction through parallel convolutional pathways, making it well-suited for capturing variable spatial patterns in OCT tissue images. ResNet50 provides deeper representational capacity with residual connections that mitigate vanishing gradients and enable the learning of fine-grained structural features. Xception, as an extension of Inception, leverages depthwise separable convolutions to improve efficiency while maintaining high accuracy, representing a modern, lightweight alternative. Together, these architectures offer complementary strengths for evaluating the complexity of classifying five distinct epidural tissue layers.

All models were trained using stochastic gradient descent (SGD) with Nesterov momentum (learning rate 0.01, momentum 0.9, decay 0.01), a widely adopted optimizer in computer vision that supports stable convergence and robust generalization. A batch size of 32 was used, reflecting a trade-off between training efficiency and the memory

demands of processing high-resolution OCT images. Early stopping was employed with a patience window based on validation loss to prevent overfitting and to reduce unnecessary computation, a common practice in medical imaging where datasets are limited in size. The choice of these hyperparameters was informed by both established deep learning conventions in medical imaging (Shin et al., 2016) and preliminary experiments conducted on the dataset.

3.2.4.3 Explainability

To enhance interpretability of model decisions, we employed Grad-CAM (Selvaraju et al., 2017) to generate class activation maps. These visualizations highlight the regions within the OCT images that most strongly influenced the network's predictions.

3.2.4.4 Code Availability

The source code used in this study is publicly accessible.

- Percutaneous Nephrostomy (PCN): `https://github.com/thepanlab/FOCT_kidney`
- Epidural Anesthesia: `https://github.com/thepanlab/Endoscopic_OCT_Epidural`

3.3 Results

3.3.1 OCT Imaging Results for Renal and Epidural Tissues

3.3.1.1 Renal Tissue Imaging

Figure 3.1A illustrates the OCT endoscopic imaging setup used for renal tissue identification. Renal cortex appeared as brown tissue at the periphery, medulla featured

distinct red renal pyramids located centrally, and calyx tissue was clearly distinguishable due to its white, fibrous appearance centrally. Representative OCT imaging results (Figure 3.1[B-D]) clearly differentiate these tissues based on brightness, structure, and imaging depth. The calyx presented the shallowest imaging depth, with notably higher brightness and structural density near the surface, characterized by horizontal stripes and layered architecture. Cortex and medulla tissues showed relatively homogeneous textures, with medulla exhibiting a deeper imaging penetration compared to the cortex. These distinctions in imaging characteristics confirm the feasibility of differentiating renal tissue types using OCT.

3.3.1.2 Epidural Tissue Imaging

Figure 3.2[A] depicts the OCT endoscopic imaging setup for epidural tissues, with Figure 3.2[B] showing representative cross-sectional OCT images and corresponding histology. The epidural space was easily distinguishable due to the visible gap between the needle tip and the dura mater. Interspinous ligament exhibited distinct, transverse stripes and the greatest imaging penetration depth, reflective of its fibrous nature. Ligamentum flavum appeared brighter near its surface but with shallower penetration. Fat and spinal cord tissues had similar penetration depths, but fat displayed uneven intensity distribution, contrasting the uniform appearance of the spinal cord. Histological validation aligned well with OCT images except for fat tissue, where adipocyte features observed in histology were less distinct due to applied tissue compression.

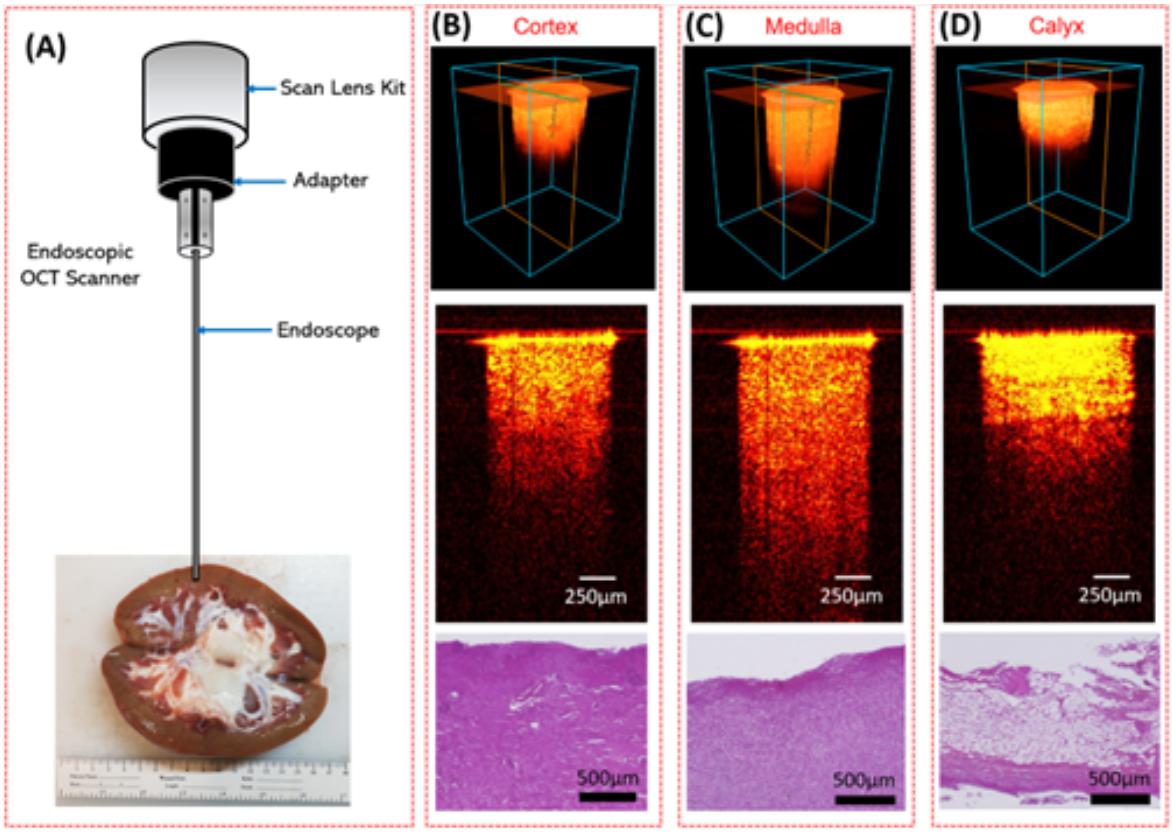

Figure 3.1: **Instrument and sample images for PCN study.** [A] Sample kidney used for endoscopic OCT imaging; [B-D] Representative 3D OCT images, cross-sectional OCT images and the histology results of renal cortex (B), medulla (C) and calyx (D).

3.3.1.3 CNN Model Development and Performance Evaluation

Renal Tissue Classification Results CNN architectures (ResNet34, ResNet50, and MobileNetv2) underwent nested cross-validation to classify renal OCT images. Table 3.1 summarizes cross-validation accuracies, revealing that pretrained (PT) ResNet50 consistently outperformed randomly initialized (RI) versions and other architectures, indicating the benefit of deeper CNN models. Table 3.2 presents cross-testing accuracies, with the average accuracy being 82.6%. Significant variability in accuracy (67.7%-92%) among kidneys highlighted challenges stemming primarily from anatomical diversity

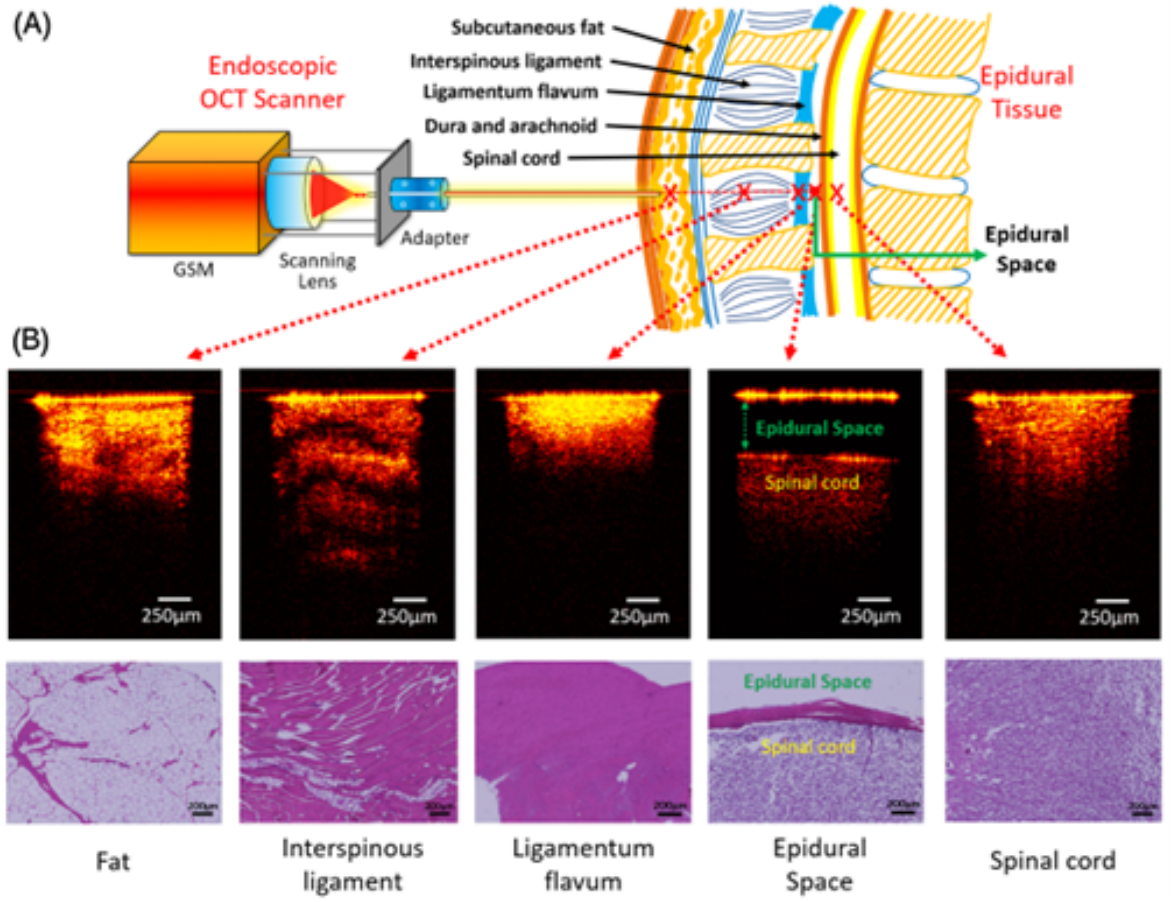

Figure 3.2: **Instrument and sample images for Epidural study.** [A] Endoscopic OCT scanner setup and the representative OCT images of five epidural tissue layer categories. [B] Histology results of different tissue layers.

rather than technical limitations. Average ROC curves from cross-testing loop for cortex (AUC=0.91), medulla (AUC=0.96), and calyx (AUC=0.97) are shown in Figure 3.3. The confusion matrix (Table 3.3) revealed cortex and medulla were commonly misclassified due to overlapping imaging characteristics, particularly when penetration depth differences were minimal.

Grad-CAM visualizations (Figure 3.4) demonstrated distinct classification strategies between randomly initialized (RI) and pretrained (PT) ResNet50 models. PT

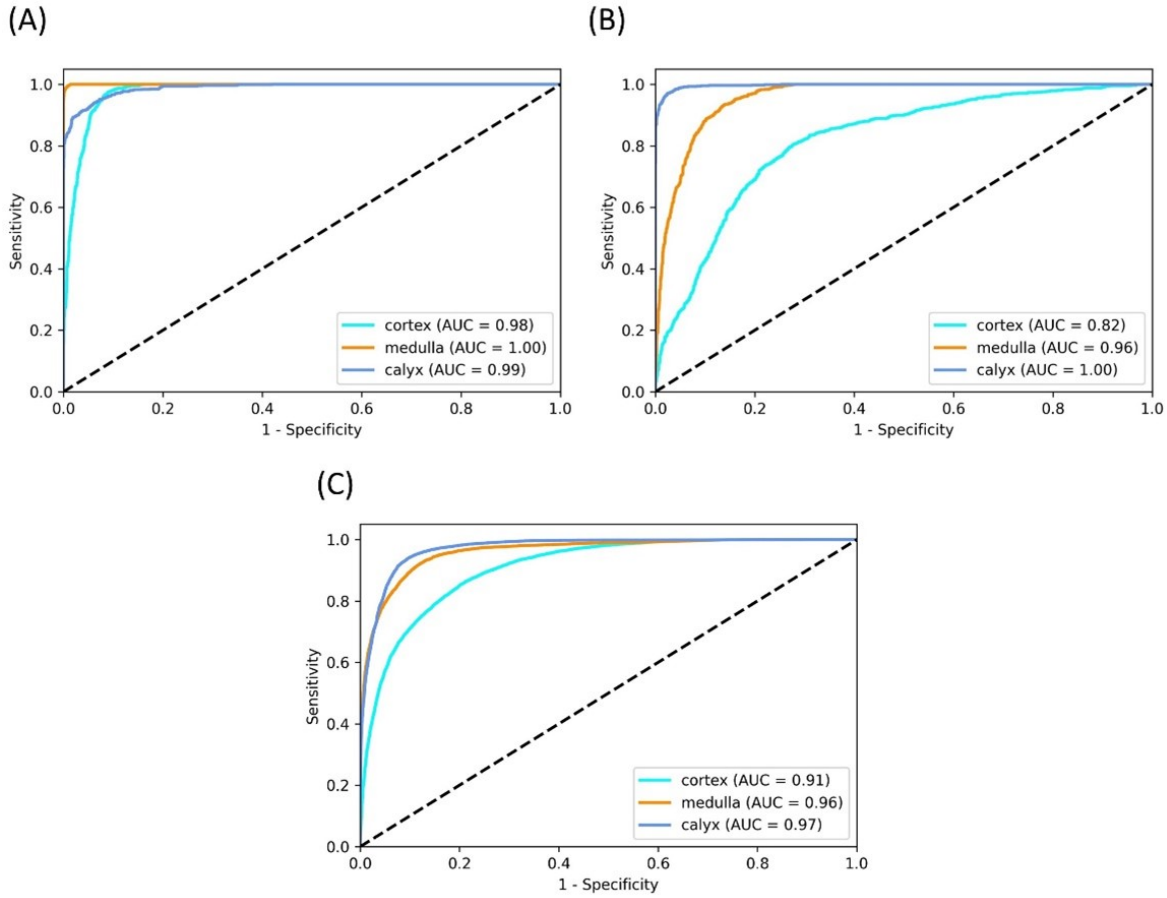

Figure 3.3: **PCN study. ROC multi-class plots for cross-testing loop.** [A] kidney 5, [B] kidney 10 and [C] the average of all kidneys.

models consistently focused on anatomically relevant regions closer to signal-dense areas, suggesting enhanced interpretability and reliability.

Epidural Tissue Classification Results Initial multi-class classification of epidural OCT images yielded moderate accuracy (∼66%), prompting the implementation of a sequential binary classification strategy aligned with needle insertion sequence. CNN models (ResNet50, Xception, and Inception) were evaluated, with ResNet50 delivering the highest overall accuracy (Table 3.4). This approach markedly improved tissue differentiation performance, achieving average test accuracies above 90% across all

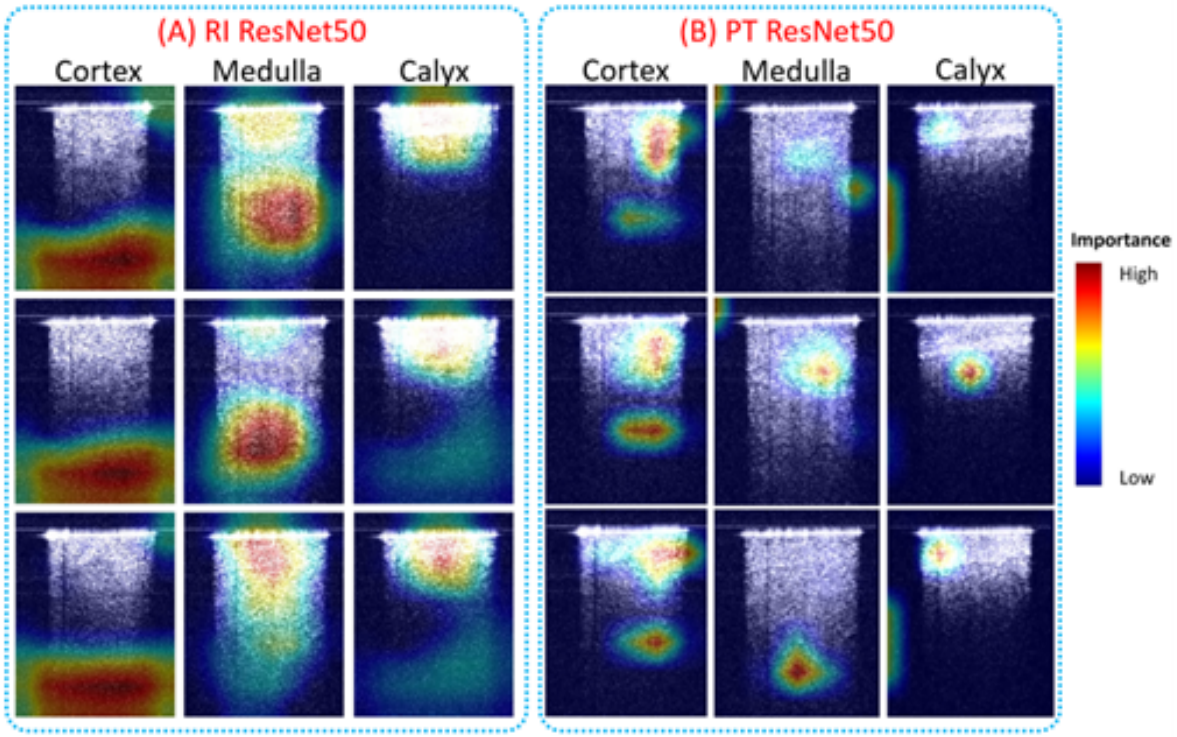

Figure 3.4: **PCN study. GradCAM explanability: Class activation heatmaps** [A] RI ResNet50, and [B] PT ResNet50 for three representative images in each tissue type. Note: The class activation heatmaps used jet colormap and were superimposed on the input images.

binary classifiers, and exceeding 98% accuracy in identifying the epidural space (Table 3.5).

Substantial accuracy variation was observed among subjects, especially for fat versus interspinous ligament classification, reflecting anatomical and compression-induced variability. Grad-CAM visualizations (Figure 3.5A) confirmed logical model behavior by highlighting key anatomical structures within images, with clear distinction strategies adopted by different binary classifiers. Real-time feasibility was demonstrated through simulated needle insertion, effectively illustrating binary classifier transitions (Figure 3.5B).

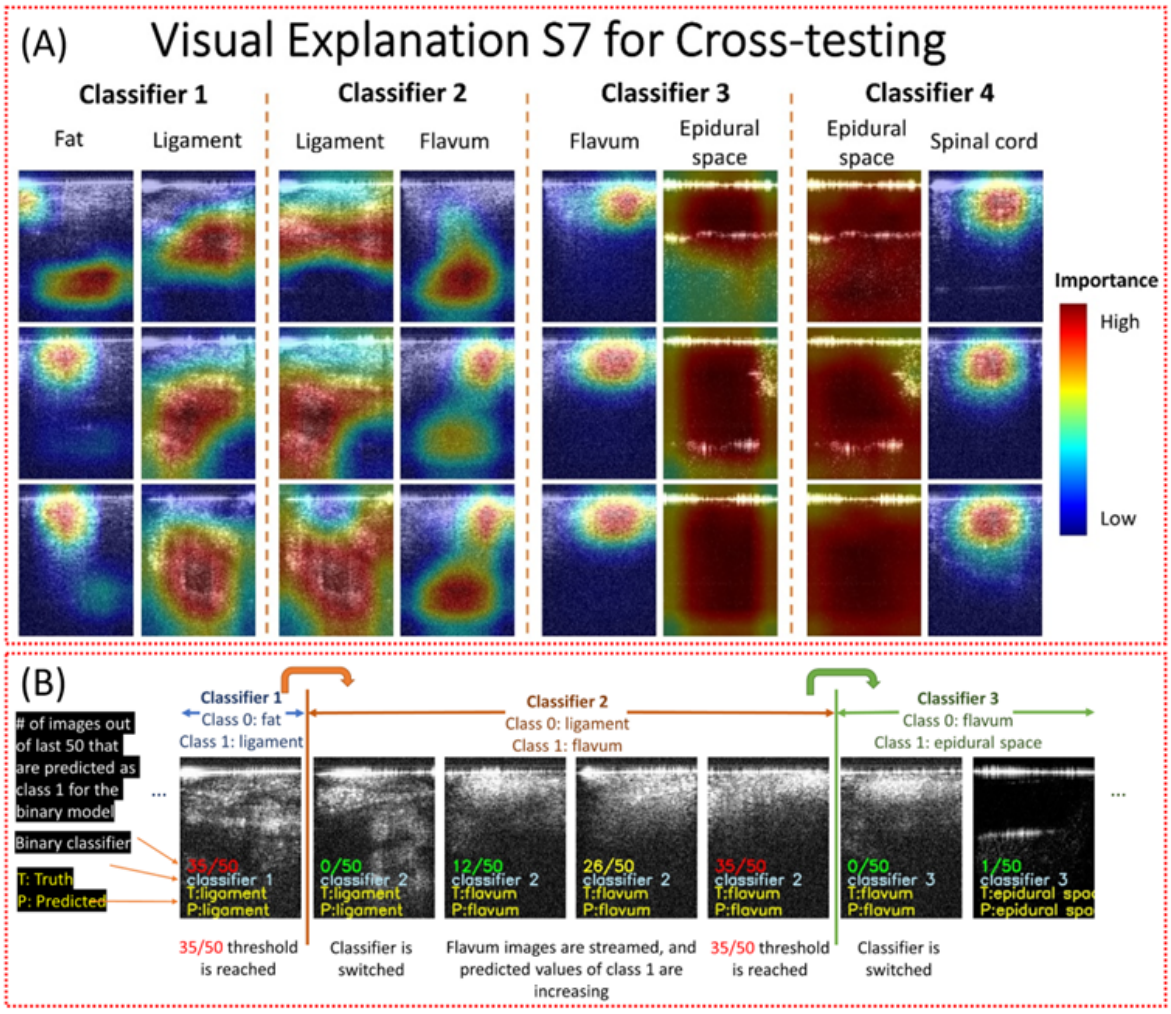

Figure 3.5: **Epidural application**. [A] Class activation heatmaps for Subject 7 using ResNet50 in cross-testing. [B] Video captures of the insertion process

These combined OCT imaging and CNN classification results establish a robust framework for accurate tissue identification and classification during OCT-guided renal and epidural needle procedures, with potential for clinical translation.

Table 3.1: PCN study. The averages and standard error of validation accuracies from cross-validation loop.

Testing folds	**PT MobileNetv2**	**RI ResNet34**	**RI ResNet50**	**PT ResNet50**
K1	$84.1\% \pm 3.5\%$	$83.8\% \pm 3.7\%$	$87.1\% \pm 2.9\%$	$\mathbf{89.2\% \pm 2.4\%}$
K2	$82.1\% \pm 3.9\%$	$84.6\% \pm 2.8\%$	$87.0\% \pm 3.0\%$	$\mathbf{89.8\% \pm 2.4\%}$
K3	$83.9\% \pm 4.2\%$	$84.5\% \pm 3.5\%$	$\mathbf{88.5\% \pm 2.4\%}$	$86.6\% \pm 2.4\%$
K4	$78.0\% \pm 6.3\%$	$83.5\% \pm 2.6\%$	$85.0\% \pm 2.4\%$	$\mathbf{87.8\% \pm 2.3\%}$
K5	$83.3\% \pm 2.7\%$	$82.8\% \pm 4.2\%$	$\mathbf{87.8\% \pm 1.9\%}$	$87.7\% \pm 2.3\%$
K6	$86.2\% \pm 2.7\%$	$84.7\% \pm 3.6\%$	$\mathbf{88.4\% \pm 2.1\%}$	$86.8\% \pm 2.4\%$
K7	$86.0\% \pm 3.8\%$	$82.9\% \pm 3.6\%$	$\mathbf{88.6\% \pm 1.7\%}$	$86.8\% \pm 2.4\%$
K8	$76.1\% \pm 6.0\%$	$82.1\% \pm 3.9\%$	$87.7\% \pm 2.6\%$	$\mathbf{88.0\% \pm 2.5\%}$
K9	$77.7\% \pm 4.8\%$	$79.4\% \pm 4.3\%$	$82.1\% \pm 3.6\%$	$\mathbf{85.1\% \pm 2.3\%}$
K10	$81.8\% \pm 4.6\%$	$86.4\% \pm 2.7\%$	$87.5\% \pm 3.2\%$	$\mathbf{87.7\% \pm 2.1\%}$

Note:The architecture with the highest validation performance in each testing fold is indicated in bold.

Testing folds	**Accuracy**	**Coverage**
K1	75.4%	100%
K2	87.9%	100%
K3	71.4%	100%
K4	92.1%	100%
K5	93.2%	100%
K6	67.7%	100%
K7	83.5%	100%
K8	88.2%	100%
K9	92.2%	100%
K10	74.1%	100%
Average	$\mathbf{82.6\% \pm 3.0\%}$	**100%**

Table 3.2: PCN study. Accuracy for cross-testing loop

True	**Predicted**			**Recall**
	Cortex	**Medulla**	**Calyx**	
Cortex	677 ± 109	204 ± 93	118 ± 93	$\mathbf{68\% \pm 11\%}$
Medulla	75 ± 33	908 ± 41	17 ± 15	$\mathbf{91\% \pm 4\%}$
Calyx	95 ± 30	14 ± 9	892 ± 32	$\mathbf{89\% \pm 3\%}$
Precision	$\mathbf{79\% \pm 4\%}$	$\mathbf{85\% \pm 6\%}$	$\mathbf{91\% \pm 5\%}$	

Table 3.3: PCN study. Confusion matrix with precision and recall for Cortex, Medulla, and Calyx classification for cross-testing loop.

Fat vs Ligament			
Testing Fold	**ResNet50**	**Xception**	**Inception**
S1	$95.1\% \pm 2.2\%$	$82.8\% \pm 8.5\%$	$90.4\% \pm 2.2\%$
S2	$93.6\% \pm 2.4\%$	$88.2\% \pm 6.5\%$	$92.4\% \pm 2.1\%$
S3	$94.5\% \pm 2.5\%$	$84.4\% \pm 6.5\%$	$88.0\% \pm 2.4\%$
S4	$90.9\% \pm 2.9\%$	$84.7\% \pm 6.3\%$	$89.7\% \pm 2.8\%$
S5	$89.9\% \pm 2.3\%$	$86.9\% \pm 4.7\%$	$89.3\% \pm 2.5\%$
S6	$90.6\% \pm 3.7\%$	$82.3\% \pm 8.5\%$	$89.9\% \pm 2.2\%$
S7	$90.7\% \pm 3.5\%$	$86.0\% \pm 3.3\%$	$88.1\% \pm 3.0\%$
S8	$88.7\% \pm 3.8\%$	$82.7\% \pm 6.3\%$	$86.3\% \pm 2.7\%$
Average	$\mathbf{91.8\% \pm 1.0\%}$	$\mathbf{84.7\% \pm 2.2\%}$	$\mathbf{89.3\% \pm 2.2\%}$

Ligament vs Flavum			
Testing Fold	**ResNet50**	**Xception**	**Inception**
S1	$97.9\% \pm 1.7\%$	$98.5\% \pm 1.2\%$	$97.8\% \pm 2.0\%$
S2	$98.7\% \pm 1.2\%$	$98.9\% \pm 1.0\%$	$98.9\% \pm 1.0\%$
S3	$99.4\% \pm 0.5\%$	$99.6\% \pm 0.3\%$	$99.0\% \pm 0.5\%$
S4	$98.5\% \pm 1.3\%$	$98.7\% \pm 0.9\%$	$97.7\% \pm 1.7\%$
S5	$98.8\% \pm 0.9\%$	$98.4\% \pm 0.9\%$	$99.0\% \pm 1.5\%$
S6	$98.8\% \pm 0.9\%$	$98.2\% \pm 1.2\%$	$97.8\% \pm 1.9\%$
S7	$98.4\% \pm 1.4\%$	$99.1\% \pm 0.7\%$	$97.0\% \pm 2.7\%$
S8	$98.7\% \pm 1.1\%$	$98.8\% \pm 0.6\%$	$98.7\% \pm 0.8\%$
Average	$\mathbf{98.6\% \pm 0.4\%}$	$\mathbf{98.8\% \pm 0.3\%}$	$\mathbf{98.1\% \pm 0.6\%}$

Flavum vs Epidural Space			
Testing Fold	**ResNet50**	**Xception**	**Inception**
S1	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S2	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S3	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S4	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S5	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S6	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S7	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S8	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
Average	$\mathbf{100.0\% \pm 0.0\%}$	$\mathbf{100.0\% \pm 0.0\%}$	$\mathbf{100.0\% \pm 0.0\%}$

Epidural Space vs Spinal Cord			
Testing Fold	**ResNet50**	**Xception**	**Inception**
S1	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S2	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S3	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S4	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S5	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S6	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S7	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
S8	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$	$100.0\% \pm 0.0\%$
Average	$\mathbf{100.0\% \pm 0.0\%}$	$\mathbf{100.0\% \pm 0.0\%}$	$\mathbf{100.0\% \pm 0.0\%}$

Table 3.4: Epidural study. Average and standard error for cross-validation loop

Testing Fold	**Fat vs Ligament**	**Ligament vs Flavum**	**Flavum vs Epidural Space**	**Epidural Space vs Spinal Cord**	**Average**
S1	81.3%	98.8%	100.0%	100.0%	95.0% ± 4.6%
S2	67.3%	98.0%	100.0%	100.0%	91.3% ± 8.0%
S3	92.0%	87.7%	98.8%	100.0%	94.6% ± 2.9%
S4	99.9%	99.4%	100.0%	100.0%	99.8% ± 0.1%
S5	98.8%	99.4%	100.0%	100.0%	99.6% ± 0.3%
S6	79.5%	99.0%	100.0%	100.0%	94.6% ± 5.1%
S7	94.2%	99.8%	100.0%	100.0%	98.5% ± 1.4%
S8	98.8%	100.0%	100.0%	100.0%	99.7% ± 0.3%
	89.0% ± 4.2%	**97.8% ± 1.5%**	**99.8% ± 0.2%**	**100.0% ± 0.0%**	**96.65% ± 1.32%**

Table 3.5: Epidural study. Binary Classification accuracies for cross-testing loop.

3.4 Discussion

This dissertation chapter explored the application of endoscopic Optical Coherence Tomography (OCT) combined with deep learning techniques for real-time needle guidance in two clinically significant procedures: percutaneous nephrostomy (PCN) and epidural anesthesia. Both procedures involve delicate navigation through tissue-rich anatomical regions where inaccurate needle placement can lead to significant complications, including bleeding, nerve injury, and failed interventions. Conventional imaging modalities such as fluoroscopy, ultrasound, and computed tomography (CT) offer macroscopic guidance but lack the resolution and contrast needed to differentiate critical tissue structures in real time at the needle tip. To address this limitation, I investigated the integration of high-resolution endoscopic OCT systems and convolutional neural networks (CNNs) to classify tissue microarchitecture and predict needle trajectory more precisely.

3.4.1 Tissue-Level Imaging and Classification in PCN

In the PCN guidance study, the OCT endoscope was used to acquire cross-sectional images from ten porcine kidneys, encompassing three histologically distinct tissue types: renal cortex, medulla, and calyx. These tissues exhibit microstructural differences that are discernible in OCT images due to their varied scattering properties and nephron densities. From a machine learning perspective, the task reduces to a supervised multi-class classification problem.

To assess the discriminative power of deep learning models, I trained and evaluated several CNN architectures, including MobileNetV2, ResNet34, and ResNet50. The results indicate that larger models are better able to capture fine-grained structural differences in the OCT data, albeit with higher computational cost. Specifically,

ResNet34 outperformed MobileNetV2, and ResNet50 achieved the highest performance overall.

Under ResNet50 consistently demonstrated superior performance with a test average accuracy of 82.6% ± 3.0% under nested cross-validation. The classification task was particularly challenging between cortex and medulla, potentially due to biological degradation in the ex vivo samples. These samples were acquired from a local slaughterhouse without controlled preservation, and post-mortem structural collapse (e.g., renal tubules) may have compromised tissue integrity. Nonetheless, our results establish a strong baseline for automatic tissue recognition, and future experiments using well-preserved ex vivo human kidneys or in vivo porcine kidneys are expected to yield higher classification accuracies. Importantly, our CNN pipeline offers a scalable solution to reduce the interpretive burden on clinicians, with inference times suitable for real-time deployment.

3.4.2 Layer-Specific Guidance in Epidural Anesthesia

In parallel, the endoscopic OCT system was validated for epidural anesthesia, a procedure that demands high precision to identify the epidural space while avoiding puncture of the dura mater or damage to adjacent tissues. OCT images were acquired from five epidural tissue layers: fat, interspinous ligament, ligamentum flavum, epidural space, and spinal cord. These tissues differ in morphology and optical characteristics, although certain layers (e.g., fat and spinal cord) can exhibit similar image patterns.

From a machine learning perspective, this task can be formulated either as a single multi-class classification problem or as a sequence of binary decision problems. A multi-class Inception model was trained but achieved only modest accuracy (60–65%), reflecting the difficulty of learning simultaneous decision boundaries across five classes with overlapping features. In contrast, a sequential binary strategy—aligned with the

physical order of tissue traversal by the needle—proved far more effective. Each classifier specialized in distinguishing only between two adjacent layers, thereby reducing class overlap and simplifying the decision boundary at each step.

This design embodies a divide-and-conquer principle, where a complex multi-class problem is decomposed into a series of tractable binary subproblems. The result was a substantial performance gain: >90% accuracy across all layer transitions and >99% precision for detecting entry into the epidural space—a clinically critical event. Beyond this specific application, the study highlights how problem framing in AI (multi-class vs. binary decomposition) can be impactful, illustrating a core computer science insight: that the way a problem is represented often determines the feasibility of its solution.

3.5 Limitations and Future Work

Despite promising results, several limitations must be acknowledged. First, both studies relied primarily on ex vivo tissue, which does not fully replicate the dynamics of living systems, including tissue deformation, perfusion, and blood interference. In future work, I plan to retrain and validate our models using in vivo porcine data. Although prior experience with epidural anesthesia guidance suggests that blood does not significantly occlude the OCT imaging window due to the needle-tissue contact, I will still explore Doppler OCT techniques to enhance blood vessel visibility, especially in the spinal canal and renal cortex, where vascular injury can be catastrophic.

Second, while CNNs have proven effective, their clinical deployment will require interpretability, robustness to domain shifts, and lightweight model architectures. We are currently investigating Grad-CAM-based interpretability, model distillation, and weight pruning to reduce computational overhead and enhance generalizability.

3.6 Conclusion

Together, these studies establish a comprehensive framework that integrates subsurface imaging with machine learning–driven inference to enable minimally invasive surgical guidance. From a computer science perspective, the contributions extend beyond the clinical applications: they demonstrate how problem formulation (multi-class vs. binary decomposition), model capacity (lightweight vs. deep architectures), and validation strategy (nested cross-validation) directly influence performance, and generalization in real-time settings.

The PCN study underscores the importance of model complexity and data quality in learning fine-grained tissue distinctions, while the epidural anesthesia study highlights how reframing a classification task into sequential binary decisions can transform a low-performing multi-class problem into a highly accurate and clinically actionable pipeline. More broadly, the results reinforce a central principle in AI: that how a problem is represented and decomposed can be as influential as the architecture used to solve it.

With these advances, the combination of OCT and deep learning have the potential to evolve into a scalable, trustworthy, and generalizable paradigm for intelligent surgical guidance. Such a system could provide clinicians with real-time feedback at the point of care, reducing the cognitive burden of interpreting complex images while increasing procedural safety.

Chapter 4

Towards AI-Driven Healthcare: Systematic Optimization, Linguistic Analysis, and Clinicians' Evaluation of Large Language Models for Smoking Cessation Interventions

This chapter is reproduced from P. Calle, R. Shao, Y. Liu, E. T. Hébert, D. Kendzor, J. Neil, M. Businelle, and C. Pan, "Towards AI-Driven Healthcare: Systematic Optimization, Linguistic Analysis, and Clinicians' Evaluation of Large Language Models for Smoking Cessation Interventions," in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, Honolulu, HI, USA, 2024, p. Article 436.

4.1 Introduction

Smartphone-delivered message intervention has become an essential part of an effective smoking cessation program (Sequeira et al., 2024; Free et al., 2011). Meta-analytical reviews have demonstrated the efficacy of receiving smartphone-delivered message interventions via short message service (SMS) or smartphone applications in promoting both short-term self-reported quitting behaviors and long-term abstinence (Scott-Sheldon et al., 2016; Spohr et al., 2015; Whittaker et al., 2016). Compared to traditional face-to-face counseling interventions, a short intervention message can offer real-time

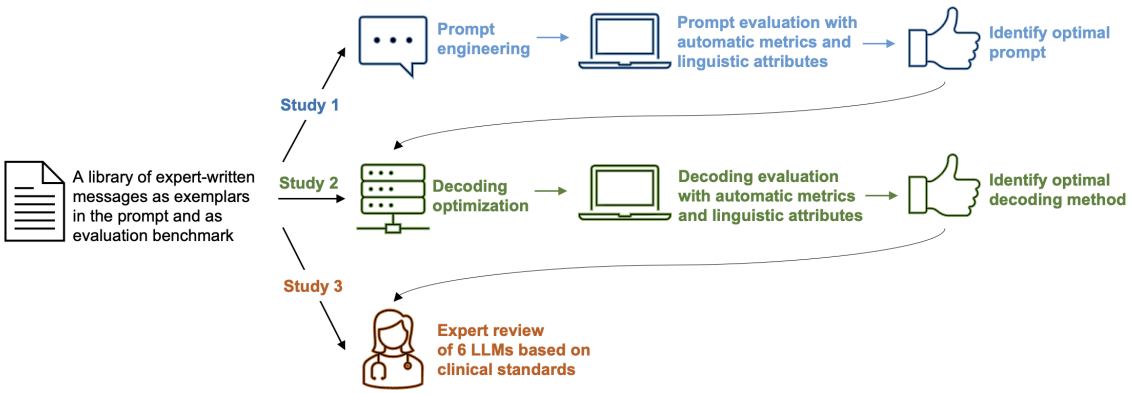

Figure 4.1: An overview of the three studies, including prompt engineering (Study 1), decoding optimization (Study 2) and expert review (Study 3).

and tailored support when individuals are vulnerable to lapse or relapse of smoking, thereby significantly enhancing tobacco abstinence rates (Carreiro et al., 2020; Kong et al., 2014; Perski et al., 2022; Yang et al., 2023). Research in the domain of smoking cessation treatment and behavioral intervention broadly calls for the integration of smartphone-delivered message interventions to improve treatment efficacy (Graham et al., 2020; Mason et al., 2015; Ybarra et al., 2014).

Despite the well-established efficacy of smartphone-delivered message interventions, maintaining user engagement and intervention effectiveness rely on the provision of novel and non-repetitive content (Patrick et al., 2009). Repetition in messages can elicit perceptions of redundancy and a feeling of information overload, which further diminish the perceived utility of the received information and attenuate attitudinal and behavioral changes (Stephens and Rains, 2011). Qualitative feedback echoes this observation; for example, in Brown et al. (2014) participants who received a text messaging intervention designed to provide nutrition education and encourage better dietary choices suggested that future designs of text messaging intervention should avoid content repetition and offer an opportunity to learn new information for each interaction message. Consequently, the development of effective smoking cessation treatment

necessitates the creation of substantial, high-quality intervention messages, a labor-intensive process for experts that poses a significant challenge.

In the field of Natural Language Processing (NLP), the advent of Large Language Models (LLMs) has significantly facilitated the capability of message generation. These LLMs, equipped with vast training datasets, can not only mimic human writing but also draw upon the knowledge embedded in the training data to produce fluent text (Bommasani et al., 2021; Brown et al., 2020). Previous research has examined the feasibility of LLMs for message generation in diverse health-related contexts. For example, Karinshak et al. (2023) compared COVID-19 pro-vaccination messages generated by the GPT-3 model with human-authored messages released by the Centers for Disease Control and Prevention (CDC). Their findings revealed that messages produced by GPT-3 were perceived by crowdworkers as more effective, presenting stronger arguments, and evoking more positive attitudes about vaccination. Similarly, Lim and Schmälzle (2023) undertook a comparison of health awareness messages on folic acid generated by the BLOOM-7B1 model with tweets on the same topic. Their study demonstrated that LLM-generated messages were rated higher in terms of message quality and clarity, compared to human-written tweets. Furthermore, computational text analysis indicated that LLM-generated messages exhibited similar characteristics to those written by humans in terms of sentiment, reading ease, and semantic content.

While prior research has supported the feasibility of LLMs in message generation (Karinshak et al., 2023; Lim and Schmälzle, 2023; Brown et al., 2020), using LLMs to generate intervention messages for smoking cessation treatment necessitates the exploration of two critical questions: 1) How can LLMs be optimized to mimic human expert writing, and 2) Do LLM-generated messages meet clinical standards to be safely implemented in tobacco treatment programs? To address these pivotal questions, the present

study conducted a systematic examination of the message generation and evaluation processes of LLMs within the context of smoking cessation intervention.

In the message generation process of LLMs, both the prompt choice and the decoding method serve as critical determinants that influence the quality of the generated text (Ahmed et al., 2022). To refine LLMs to emulate human expert writing, this study conducted prompt engineering (Study 1) and decoding optimization (Study 2), involving the testing of five prompts and eight decoding methods to generate intervention messages across five LLMs (see Figure 4.1). The evaluation of LLM performance in message generation is a multi-faceted task, requiring a comprehensive assessment of message quality, coherence, relevance, grammaticality, and accuracy. Consequently, the employment of commonly adopted automatic metrics in LLM assessment is often oversimplified and inadequate in capturing the nuanced linguistic properties and overall text quality (Belz and Reiter, 2006; Scott and Moore, 2007; Reiter, 2018; Ananthakrishnan et al., 2007). In recognition of these challenges, our study introduced a comprehensive evaluation framework encompassing diversity, quality, and efficiency. We built upon the computational linguistic approach and utilized the Linguistic Inquiry and Word Count (LIWC) tool to analyze linguistic features of LLM-generated messages (Tausczik and Pennebaker, 2010), using intervention messages written by tobacco treatment experts as the benchmark. In Studies 1 and 2, we identified the prompt and decoding methods that most closely resembled human expert writing. These strategies were subsequently recommended and subjected to evaluation in terms of their clinical utility (Study 3).

While LLMs have demonstrated their capacity in generating effective public health messages (Karinshak et al., 2023; Lim and Schmälzle, 2023) and have exhibited knowledge levels comparable to third-year medical students (Kung et al., 2023; Gilson et al.,

2023; Sedaghat, 2023), their applicability in the high-stakes healthcare context necessitates rigorous evaluations to ensure both reliability and safety. Previous research underscores the importance of expert review in increasing the validity of LLM evaluations, especially for tasks requiring specialized expertise (Liang et al., 2022; Marvin and Linzen, 2018; Wu and Aji, 2023; Jakesch et al., 2023; Clark et al., 2021; Belz and Reiter, 2006). Therefore, to establish the efficacy of LLMs in generating intervention messages for clinical use, we invited certified tobacco treatment specialists (TTS) to conduct an expert review (Study 3). These specialists rigorously evaluated the messages generated by the five LLMs from Studies 1 and 2, and the newly released ChatGPT, on message quality, accuracy, credibility, and persuasiveness, using expert-written messages as the benchmark. Drawing upon their extensive clinical experiences in tobacco treatment counseling, they further evaluated whether the generated messages met the standards in TTS training and were ready to use in clinical practice. Our results indicate that larger LLMs, including ChatGPT, OPT-13B and OPT-30B, can effectively emulate human expert writing to generate well-written, accurate, credible, and persuasive messages, thereby demonstrating the efficacy of LLMs in augmenting smoking cessation interventions in clinical settings.

Our paper contributes to the current research landscape on several key aspects. First, we conducted a systematic examination of message generation and evaluation with LLMs, involving prompt engineering, decoding optimization, and expert review, conducted across six state-of-art LLMs, thereby ensuring the generalizability of our findings. Furthermore, we proposed an innovative and comprehensive evaluation framework for LLM assessment, integrating automatic metrics, linguistic attributes, and expert evaluation, offering a multifaceted assessment that delves into diversity, quality, and efficiency aspects of the LLM performance. To date, this is the first study that

adopted the computational linguistic approach for the assessment of LLMs and established the efficacy of LLMs in specialized healthcare contexts through expert reviews.

This paper is organized as follows: Section II and III address the first research question on prompt engineering and decoding optimization. Section II provides an overview of prompt engineering literature, introduces the LLMs employed for message generation, presents a multi-faceted evaluation framework, and assesses prompts accordingly. Section III reviews literature on decoding parameters and examines how decoding methods influence message generation. Section IV addresses the second research question by incorporating expert reviews of LLM-generated intervention messages. Section V discusses future applications and limitations of LLMs in the healthcare context, and Section VI concludes.

4.2 Study I: Prompt Engineering

4.2.1 Related Work

Large language models (LLMs) are deep learning models trained to understand and generate natural language. Autoregressive LLMs work by using a series of input tokens (words or fragments thereof) to generate subsequent tokens (Brown et al., 2020; Radford et al., 2019). Grounded in the transformer architecture, which is a cutting-edge neural network design characterized by its massive parameter sizes, LLMs can successfully understand patterns within the input tokens through a self-attention mechanism (Vaswani et al., 2017). In recent years, the efficacy of LLMs has been examined and validated for various natural language tasks, including automatic summarization, machine translation, and question answering (Min et al., 2021; Brown et al., 2020).

4.2.1.1 Large Language Models

Currently, there are open-source and closed-source models based on the availability of model weights. Leading open-source models in the field include GPT-J-6B, the BLOOM series, and the OPT series. For the purposes of this research, we employed models compatible with our computational resources, namely GPT-J-6B (Wang, 2021), Bloom-7B1 (Scao et al., 2022), and OPT 6.7B, 13B, and 30B (Zhang et al., 2022). GPT-J-6B is an open-access alternative to GPT-3 and was trained using the Pile dataset (Biderman et al., 2022), a predominantly English-centric text corpus that combines sources such as English Wikipedia and PubMed Central. BLOOM introduces modifications to the conventional Transformer architecture (Vaswani et al., 2017), including the incorporation of ALiBi Positional Embeddings (Press et al., 2022) and the Embedding LayerNormal. BLOOM is trained with the ROOTS dataset (Laurençon et al., 2022), a multi-lingual text corpus. The architecture of OPT is similar to GPT-3 and was trained on a corpus comprised by subdatasets of RoBERTa (Liu et al., 2019b), a subset of the Pile, and a subset of the Pushshift dataset (Baumgartner et al., 2020). The primary language of OPT's corpus is English. We used LLMs of diverse sizes, ranging from 6 billion parameters (as in GPT-J-6B) to 30 billion parameters (as in OPT-30B) to identify the optimal model for message generation in the healthcare context and to enhance the generalizability of our findings.

4.2.1.2 Prompt engineering for LLMs

We employed the "tuning-free prompting" approach (Liu et al., 2021a; Clavié et al., 2023; White et al., 2023; Zhou et al., 2022) to generate intervention messages, a method where the parameters/weights of the model are fixed and in-context learning is used (Brown et al., 2020). This approach leverages the input context to guide the model's responses without the need for task-specific fine-tuning. For example, to instruct LLMs

in generating intervention messages for smoking cessation, we can provide a handful of intervention messages as examples to help models infer the format of message output. The number of examples provided can be used to differentiate the learning process into three categories: zero-shot, one-shot, and few-shot. In zero-shot learning, the model is given a natural language description of the task without any examples, whereas one-shot and few-shot learning provide the model with one or a few context-relevant examples. As concluded by Brown et al. (2020), LLMs are meta-learners that integrate outer-loop gradient descent learning with in-context learning, and few-shot learning, by prepending context-specific examples, can greatly enhance model performance and adaptation to task contexts, especially in areas requiring specialized knowledge and messaging styles, such as smoking cessation intervention.

For message generation tasks using few-shot learning, a prompt is defined as the initial input provided to the LLM for generating message output (Liu et al., 2021a). In our study, this input comprises both the task instruction and the message exemplars authored by human experts. For example, when prompting the BLOOM-7B1 model with this task instruction "Write motivational messages to encourage people to quit smoking:" along with four sample messages written by human experts, the model returned the following message:

"If you know that you are having a strong craving, you can take a deep breath and relax your muscles. When you can relax, you can focus on something else, like reading or watching TV. Do not attempt to fight the urge to smoke. Instead, take the next deep breath and wait for it to pass."

Both the prompt and the accompanying examples can influence the quality and relevance of the model output. Prompting serves multiple purposes: setting the context, clarifying the nature of the task, guiding response format, improving relevance and accuracy, and reducing ambiguity and potential biases (Brown et al., 2020; Liu

et al., 2019b; Radford et al., 2019). As highlighted by the PromptBench benchmark (Zhu et al., 2023), LLMs are sensitive to prompts, underscoring the importance of prompt engineering for optimal model performance. Therefore, enhancing model efficacy necessitates prompt engineering, an iterative refinement process which involves rephasing the prompt for clarity and precision, providing context-relevant examples, and specifying response format.

In this study, we employed manual template engineering to generate prompts (Liu et al., 2021a), (see Table 4.1). The first three prompts were adapted from (Brown et al., 2020), with each subsequent prompt version increased in length and detail. The first version provided a broad instruction with no context ("Message:"), the second version provided a specific theme for the message output ("Messages to help you quit smoking:"), and the third version instructed explicitly on both message theme and desired tone ("Write motivational messages to encourage people to quit smoking:"). Further, the fourth version introduced a variation in placement by positioning the task instruction after the example messages (Example messages + "Write messages like the previous ones:"). By prepending message exemplars before the general instruction 'Write messages like the previous ones,' we were able to test whether providing context-related exemplars and having LLMs infer the task instructions, without explicitly conditioning the tasks, themes, or tones, would lead to more relevant and accurate output. Lastly, the fifth version adopted a structured format (Schick and Schütze, 2021; Liu et al., 2022), using tags to label the task and examples ("Task: Write messages that are on the same topic" + Message 1: ... + Message 2: ...). Structured prompting was introduced to help LLMs better distinguish between task instructions and exemplars and its effectiveness in increasing the accuracy, relevance, and context appropriateness of the model's output was further examined. The message example

dataset contains intervention messages from existing research on smoking cessation intervention written by tobacco treatment experts (Hébert et al., 2018, 2020; Businelle et al., 2014, 2016).

Table 4.1: Five version of prompts used in Study 1

	Task		
Version	Position	Instruction	Type of list
v1	start	Messages:	Numbered
v2	start	Messages to help you quit smoking:	Numbered
v3	start	Write motivational messages to encourage people to quit smoking:	Numbered
v4	end	Write messages like the previous ones:	Numbered
v5	start	Task: Write messages that are on the same topic.	Labeled

4.2.1.3 Perplexity and Computational Linguistic Analysis

This study first evaluated prompt performance with the perplexity measure (Bahl et al., 1983). Perplexity is a standard metric for evaluating the quality of language models and quantifies how "surprised" the model is when seeing a passage of text (Clarkson and Robinson, 1999). A passage of text with a perplexity too high may contain language errors or nonsensical content (Gonen et al., 2022), while a perplexity too low may signify repetitive and uninteresting text (Holtzman et al., 2020).

The words we use in communication not only reflect our identity but also the social and relational context in which communication takes place. Linguistic features, such as emotional tones or the use of complex words, allow individuals to infer the medical expertise of their healthcare provider, subsequently influencing their trust and adherence to health-related messages (Toma and D'Angelo, 2015). To explore the extent to which LLMs emulate human expert writing, we examined the psychometric properties and linguistic features in LLM-generated messages, using human expert writing as the benchmark. We adopted the computational linguistic approach, which

involves quantifying linguistic patterns and psychometric properties in a given text by investigating word usage within predefined psychometric categories (Tausczik and Pennebaker, 2010). In this study, five critical linguistic features, including word count, clout, emotional tone, authenticity, and use of complex words, were chosen to evaluate the properties of the intervention messages (Boyd et al., 2022). The clout score is a proxy for relative social status, confidence, or leadership. A high clout score indicates greater expertise, certainty, and confidence in communication, whereas a low score suggests a tentative style of expression (Kacewicz et al., 2014; Meier et al., 2019). Authenticity reflects the varying degrees of personal and disclosing styles in discourse (Newman et al., 2003). Higher authenticity score indicates honest, personal, and disclosing communication, whereas lower score implies a distanced and impersonal style. Emotional tone ranges from negative (values <50) to positive (values >50), with a midpoint of 50 on the 100-point scale representing a neutral emotional tone (Cohn et al., 2004). Scores for clout, authenticity, and emotional tone are standardized composite variables transformed to a scale ranging from 1 to 100. Complex words refers to the proportion of words that are 7 letters or longer (Boyd et al., 2022).

The five selected linguistic features signify the quality, credibility, and trustworthiness of writings in the context of substance use intervention and health communication. Partch and Dykeman (2019) analyzed the linguistic attributes of text messages used in substance use disorder treatments and found that, in comparison to everyday Twitter posts, intervention messages exhibited greater clout, maintained neutral emotional tones, and displayed lower authenticity. Additionally, Toma and D'Angelo (2015) posited that medical advice messages with a higher word count and frequent utilization of complex words were perceived as more authoritative. Together, previous literature has identified linguistic markers that bolster the perception of credibility and trustworthiness in health intervention messages: a higher word count and greater clout

reduce uncertainties and provide comprehensive and assertive explanations; neutral emotional tones and impersonal writing foster a psychologically detached and objective stance; and the employment of complex words signifies cognitive complexity and thereby enhances the perceptions of formality, expertise, and professionalism.

4.2.2 Method

A library of 899 intervention messages written by tobacco treatment experts and validated in clinical trials served as the training and validation data (Hébert et al., 2020; Businelle et al., 2014, 2016). Drawing upon research in few-shot learning, which has demonstrated significant improvements in text generation performance when increasing the number of examples from 1 to 4, with diminishing returns beyond 4 examples (Zhao et al., 2021), we opted to use 4 message exemplars in combination with each of the 5 prompt versions under examination (refer to Table 4.1 for 5 prompt versions and refer to Table 4.2 for an example of the prompt input with message exemplars). These exemplars were randomly selected from the human-written message library and used as input for 5 state-or-art LLMs, namely GPT-J-6B, Bloom 7B1, and OPT 6.7B, 13B, and 30B. The message generation process was repeated 100 times for each LLM, utilizing the Transformer library from HuggingFace. Computational tasks were performed on a server with an AMD Ryzen Threadripper PRO 3955Wx (16-Cores, 32-Threads) and two Nvidia RTX A6000 48GB GPUs with the Ubuntu 20.04.4 LTS OS. The initial process yielded a total of 21,638 messages.

To ensure relevancy of LLM-generated messages, this study employed a two-step filtering process (see Figure 4.2). In the first step, BLEU-4 scores were computed to measure repetition between each message and the rest of the generated messages from the same LLM (Zhu et al., 2018; Papineni et al., 2002). The BLEU-4 metric evaluates the

Table 4.2: Example of a prompt with four sample messages used in Study 1

Prompt	Messages to help you quit smoking:
Sample messages	1. Remember your progress! How long has it been since you last smoked? Calculate the number of hours and give yourself credit for your hard work!
	2. When you feel stressed, a call to a family member or friend could be a good distraction.
	3. Even after you are smoke free for a while, you may still feel an urge to smoke. The important thing to remember is that having an urge to smoke does not mean you have to smoke!
	4. Nicotine lozenges will help you cope with urges to smoke, but you may still feel urges at times. These urges are only temporary, though. They will pass. Don't let them spoil your hard work.

similarity up to 4 consecutive words between pairs of intervention messages, producing scores that range from 0 to 1, with values nearing 1 indicating high similarity between messages. We set the threshold at 0.5 to discard redundant messages with BLEU-4 scores equal to or larger than 0.5. In the second step, messages were filtered out based on three criteria tailored to the context of smoking cessation and the characteristics of the example dataset: (1) the presence of words like "app", "apps", or "applications". This criterion was applied because the example data contained messages designed for smartphone-based smoking cessation interventions, and a small portion of which aimed to promote app engagement rather than smoking cessation itself. (2) the inclusion of underscores ("_"), which indicated placeholders for information to be inserted. And (3) messages consisting of fewer than six words. This threshold was chosen because expert-written messages contain at least six words, and messages shorter than six words tend to lack the necessary informativeness for effective intervention. Subsequent message filtering based on repetition (BLEU-4 ≥ 0.5) and three context-specific criteria resulted in a final count of 11,558 messages. We combined the filtered LLM-generated

messages (N=11,558) with the original human-written messages (N=899) to create the final sample for evaluation (N=12,457).

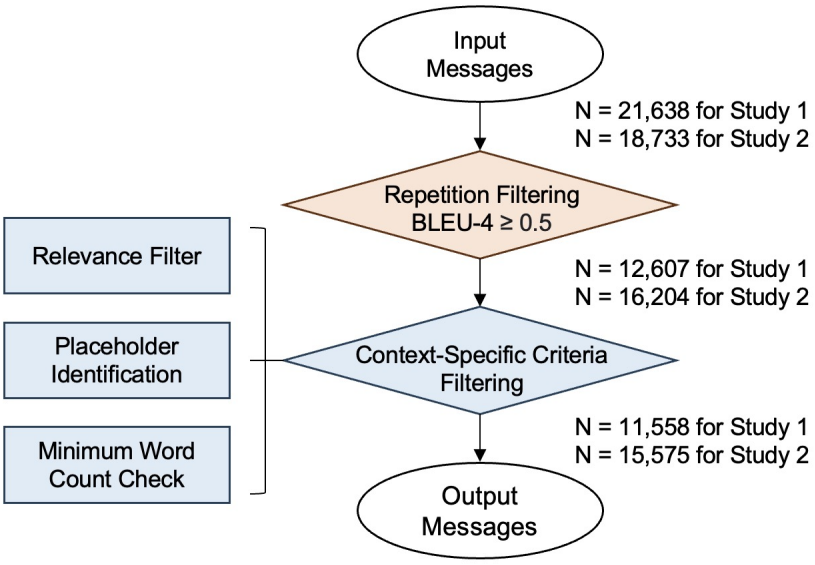

Figure 4.2: Two-step message filtering with sample sizes at each step for studies 1 and 2

Five prompts were rigorously assessed based on three key dimensions: diversity, quality, and efficiency of the generated messages. *Diversity* was evaluated using two primary indicators: perplexity scores and the pass rate of repetition filtering. Perplexity scores for each prompt version and each LLM were computed, utilizing the perplexity score from the expert-writing condition as the gold standard. The selection of the optimal prompt was predicated on its performance in close proximity to this reference perplexity. Importantly, this study refrained from the pursuit of lower perplexity scores, as such metrics could potentially signify a lack of diversity in the generated messages (Papineni et al., 2002; Holtzman et al., 2020). *Quality* was ascertained through a comprehensive analysis of critical linguistic features, using expert-written messages as the benchmark. Linguistic attributes, including word count, clout, emotional tone,

authenticity, and the use of complex words, were quantified for each generated message using the Linguistic Inquiry and Word Count (LIWC) software (LIWC-22, version 1.3.0) (Boyd et al., 2022). Efficiency was gauged by calculating the average number of messages retained after both repetition and criteria-based filtering applied to each iteration. This measure served as a practical indicator of the prompt's efficiency in generating messages that not only meet quality standards but also do so in an efficient manner.

4.2.3 Results

In the initial message generation phase, each of the five prompt versions yielded an average of 6.7 to 13.8 messages per iteration. The initial repetition filtering step discarded between 23.5% and 72.8% of these messages. Subsequent criteria-based filtering removed an additional 3.2% to 7.0% of these messages. As a result, the proportion of messages retained, relative to the initial count generated by each prompt version, ranged from 20.2% to 73.3%. On average, the human-authored messages comprised 26.16 words (SD=10.80). They exhibited high clout (M=76.62, SD=29.90), a low degree of authenticity (M=43.74, SD=37.24), and a neutral emotional tone (M=48.79, SD=38.97). These messages contained approximately 22% complex words (SD=10.74).

Five one-way analysis of variance (ANOVA) tests with post-hoc multiple comparison analyses were conducted to compare linguistic features of messages generated with the five different prompts with expert generated messages. Results indicated that messages generated with the prompts significantly differed from human-written content in terms of word count ($F(5, 12451) = 172.42$, $p < .001$, partial $\eta^2 = .07$), clout ($F(5, 12451) = 293.29$, $p < .001$, partial $\eta^2 = .11$), authenticity ($F(5, 12451) = 102.95$, $p < .001$, partial $\eta^2 = .04$), emotional tone ($F(5, 12451) = 2.97$, $p = .011$, partial $\eta^2 = .00$),

and the use of complex words ($F(5, 12451) = 48.20$, $p < .001$, partial $\eta^2 = .02$). Subsequent comparisons using Dunnett's test were conducted between each prompt and the expert-writing condition (Figure 4.3). Results demonstrated that compared with human-written messages, messages generated with five different prompts consistently had fewer words and used less complex vocabulary (all p values $< .001$). Compared with expert generated messages, messages generated with prompt version 4 exhibited higher authenticity scores, whereas messages generated with the other four prompts had lower authenticity scores (all p values $< .001$). In addition, both the generated messages and the expert generated messages used a relatively neutral emotional tone (all p values $> .05$). Messages generated with prompt versions 1 and 5 displayed clout similar to human writing, whereas messages generated with prompt versions 2 and 3 had significantly higher clout and those generated with prompt version 4 had significantly lower clout. The ANOVA test with post-hoc multiple comparison analysis revealed significant differences in perplexity between each prompt with the expert-writing condition ($F(5, 99) = 20.87$, $p < .001$, partial $\eta^2 = .51$). Post-hoc Dunnett's test further indicated that LLMs using all five prompts had lower perplexity ($M \sim 4.29$ to 4.79; $SD \sim .39$ to .66) compared to messages written by human experts ($M=6.54$, $SD=.63$, all p values $< .001$).

4.2.4 Discussion

Based on a comprehensive evaluation encompassing diversity, quality, and efficiency, our findings highlight the significant impact of different prompts on the ability of LLMs to generate intervention messages, assessed against the expert-written benchmark. Despite advances in LLMs, a discernible gap remains between machine-generated and human-written messages. Across all five prompt versions, LLM-generated messages

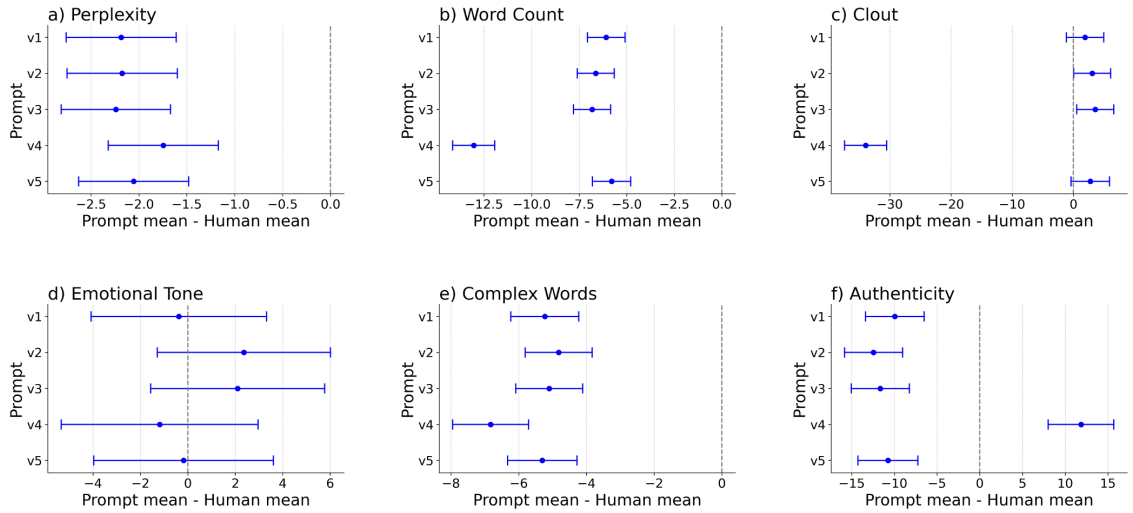

Figure 4.3: Confidence intervals for Dunnett test for prompt mean - human mean

exhibited significant deviations from the expert-writing condition in terms of word count, authenticity, the use of complex vocabulary, and model perplexity. Specifically, messages generated by LLMs were less detailed (evidenced by a reduced word count), more formal and monotonous (with lower authenticity, except for those generated with prompt version 4), less professional (utilized less complex vocabulary), and less diverse (indicated by lower perplexity scores). Notably, our findings indicate that the sequencing of task instructions after message examples, as observed in prompt version 4, adversely impacted the performance of all evaluated Large Language Models (LLMs), extending the ordering effects of training examples on model performance to the relative placement of instructions and examples within prompts (Zhao et al., 2021). In particular, the "example + instruction" arrangement led to highly repetitive outputs, with a repetition rate of 72.8% compared to rates ranging from 23.5% to 33% for other prompts. Additionally, the quality of messages generated using this prompt was suboptimal, as evidenced by a 7% message discard rate due to criteria-based filtering, compared to discard rates of 3.2% to 3.8% for other prompts. Furthermore, linguistic

features of these messages diverged significantly from human writing across all critical dimensions. Finally, the production efficiency of messages generated using this prompt was markedly lower, with a pass rate of 20%, in contrast to pass rates ranging from 62.8% to 73.3% for other prompts.

Prompt versions 1, 2, and 3 in our study featured task-specific instructions that ranged from a general directive ("Message:") to more detailed guidelines with a constrained tone ("motivational") and theme ("smoking cessation"). Existing literature posits that detailed and specific prompts act as semantically meaningful task instructions, thereby facilitating more efficient model learning — akin to how specific task instructions enhance human learning efficiency (Mishra et al., 2022; Schick and Schütze, 2021; Brown et al., 2020). This aligns with a common assumption in few-shot learning research, which suggests that optimal model performance necessitates expertly crafted, clear, and accurate task descriptions (Logan IV et al., 2022; Schick and Schütze, 2021). Contrary to these expectations, our findings indicated that the general instruction (prompt version 1) outperformed its more detailed and specific counterparts (prompt versions 2 and 3) across metrics of message diversity, quality, and efficiency. Specifically, messages generated using prompt version 1 exhibited performance closely aligned with human expert-generated messages in terms of perplexity and key linguistic features such as clout and emotional tone. Moreover, after applying a two-step filtering process, prompt version 1 yielded an average of 7.1 messages per iteration, achieving higher efficiency for message generation compared to 6.6 messages for prompt version 2 and 6.8 messages for prompt version 3.

The finding that a general prompt outperformed detailed and specific ones, although counterintuitive, in fact echoes emerging research that questions the necessity for models to receive semantically meaningful instructions (Mishra et al., 2022; Prasad et al., 2023; Webson and Pavlick, 2022). Empirical studies have shown that, in a few-shot

learning context, models perform comparably well when given irrelevant or misleading instructions as opposed to clear and specific directives. This suggests that prompts may serve more to help models learn the distribution of input text rather than to provide explicit task instructions (Min et al., 2021; Webson and Pavlick, 2022; Lampinen et al., 2022). Furthermore, our findings align with research by Yang et al. (2023), which demonstrated that prompts with a structured format (referred as schema-based prompts, e.g., "Title: ....; Author...") consistently outperform natural language (NL)-based prompts (referred as template prompts, e.g., "The title is ....; The author is ....") in few-shot learning across various NLP tasks. In our study, both prompt version 1 ("Messages:") and version 5 ("Task: Write messages that are on the same topic" + Message1: ... + Message2: ...) adopted a structured format and outperformed the natural language sentences employed for task-specific instructions in prompt versions 2 and 3. Despite comparable quality in the messages generated by prompt versions 1 and 5, the more succinct version 1 exhibited higher efficiency, yielding an average of 7.1 messages per iteration, as compared to an average of 3.5 messages generated by prompt version 5. These findings suggest a general pattern within the context of few-shot learning: model performance benefits from both the brevity and the structured nature of the prompt.

4.3 Study II: Decoding Optimization

4.3.1 Related Work

A decoding method interprets LLM's output probabilities for the subsequent token (word, sub-word, or character) in a sequence and selects the most suitable one for text generation (Abroms et al., 2013). Various decoding methods exist, each with its own set of parameters. The selection of decoding methods and parameter values needs to

balance the quality and diversity of the message output. In this study, we employed three decoding methods, namely temperature sampling, top-k sampling, and nucleus sampling, due to their extensive application in text generation tasks (Abroms et al., 2013; Businelle et al., 2016; Hébert et al., 2018).

Temperature sampling modulates the softmax output probabilities using a temperature parameter T. As the temperature value approaches 0, the distribution becomes more concentrated, thereby favoring the most probable tokens and leading to more deterministic but potentially less diverse output. Conversely, a temperature close to 1 keep the same distribution, increasing the likelihood of sampling less probable tokens. Hashimoto et al. (2019) explored the balance between diversity and quality in temperature annealing across single-sentence natural language generation tasks, including summarization, story generation, and chit-chat dialogue. Their findings revealed a quality-diversity trade-off: while lowering the temperature (T=0.7, as compared to T=1) improved the quality of generated text, it also reduced diversity and introduced repetition issues across all three tasks. Similarly, Holtzman et al. (2020) observed that temperatures exceeding 0.9 yielded messages diversity akin to human writing, measured by self-BLEU scores, and temperatures above 0.7 effectively mitigated repetition issues. In the healthcare context, Schmälzle and Wilcox (2022) conducted a pilot test employing temperature settings of 0.3, 0.5, 0.7, and 1 with a fine-tuned 355M GPT-2 model, aiming to create messages about folic acid. Their results indicated that T=0.7 produced the most balanced outputs in terms of both quality and diversity, whereas T=1 led to incoherent outputs and $T < 0.5$ produced text that was highly homogeneous with the training text. In light of these observations, previous studies generally recommend an optimal temperature threshold within the range of [0.7, 1] for text generation tasks.

Temperature sampling is frequently used in conjunction with other techniques such as top-k or top-p sampling. In top-k sampling, the k most probable subsequent tokens are filtered, and the probability mass is then redistributed exclusively among these k options. This approach introduces a degree of stochasticity into the decoding process, thereby enriching the diversity of the generated text relative to other methods like greedy decoding or beam search. Previous research commonly recommended combining top-k sampling with a temperature setting of T=0.7, (Fan et al., 2018; Radford et al., 2019). The optimal value of k may vary depending on the specific task and requirements (Ippolito et al., 2019). For example, Fan et al. (2018) employed k=2 for summarization tasks and k=40 for text generation. Similarly, in Ippolito et al. (2019), k=10 was recommended for story generation, aiming to produce text that is coherent, contextually relevant, and diverse. Nucleus sampling, also known as top-p sampling, deviates from the fixed-number approach inherent in top-k sampling. Instead, it selects tokens from the smallest subset that has a cumulative probability exceeding a predefined threshold p (Holtzman et al., 2020). This method allows for dynamic adjustments in the number of tokens considered at each decoding step, thereby enhancing the diversity and creativity of the generated text. Holtzman et al. (2020) observed that nucleus sampling with p=0.95 closely matched human performance in terms of both perplexity and self-BLEU scores. Additionally, DeLucia et al. (2021) recommended an optimal p-value range of [0.7, 0.9] for narrative generation using GPT-2, based on four evaluative criteria: interestingness, coherence, fluency, and relevance.

In message generation tasks, the selection of decoding methods is critical for achieving the optimal balance between diversity and quality (Holtzman et al., 2019; Hashimoto et al., 2019; Fan et al., 2018; Caccia et al., 2020; Ippolito et al., 2019). A systematic comparison of temperature sampling, top-k sampling, and nucleus sampling has indicated that these sampling strategies can yield comparable performance with tuned

hyperparameters (Nadeem et al., 2020; Zhang et al., 2021). For example, a configuration with k=500 and T=0.8 was found to perform closely to k=30. However, when the emphasis is on quality over diversity, nucleus sampling has been shown to outperform other decoding methods, as recommended in Zhang et al. (2021). Based on previous findings, this study investigated the efficacy of eight decoding methods (refer to Table 4.3), encompassing temperature sampling, top-k sampling, and nucleus sampling, in the generation of intervention messages for smoking cessation. Specifically, in version 3, we adopted the top-p sampling value recommended by Holtzman et al. (2019) and reduced the value from 0.95 (as in version 3) to 0.9 (in version 2) to examine potential improvements in model performance. Version 5 employed a temperature setting of 0.7, as suggested by Schmälzle and Wilcox (2022) for message generation tasks, and we increased the value to 0.9 in version 4 to explore the impact of a temperature variation on model performance. For top-k sampling, we followed Radford et al. (2019)'s advice, using k=40 in version 6 and adjusting it to k=30 in version 7 to study its effects. Furthermore, we adopted a hybrid approach to versions 1 and 8: version 8 was a combination of versions 5 and 6; and version 1 adjusted values of p, k, and temperature, aiming to achieve a balance between restriction and flexibility to ensure both diversity and coherency for message output.

4.3.2 Method

We replicated the message generation and evaluation process from Study 1. The recommended prompt version 1 from Study 1 was adopted. Eight decoding methods were applied to each LLM, as detailed in Table 4.3. In alignment with Study 1, each LLM produced messages across 100 iterations. This initial phase resulted in a total of 18,731 messages, which subsequently underwent the same filtering process (see Figure 4.2). The refined LLM-generated messages (N=15,575) were amalgamated with the

original human-authored messages (N=899) to constitute the final evaluation sample (N=16,474). Furthermore, eight decoding methods were systematically assessed on the same dimensions in Study 1 including diversity, quality, and efficiency.

4.3.3 Results

We replicated the analysis process from Study 1, including descriptive analysis on message filtering and multiple comparisons of linguistic features and perplexity. In the initial message generation phase, each of the eight decoding methods yielded an average of 3.4 to 7.1 messages per iteration. The initial repetition filtering step discarded between 4.3% and 23.8% of these messages. Subsequent criteria-based filtering removed an additional 1.4% to 4.6% of these messages. As a result, the proportion of messages retained, relative to the initial count generated by each prompt version, ranged from 72.3% to 93.8%.

Five one-way analysis of variance (ANOVA) tests with post-hoc multiple comparison analyses were conducted comparing linguistic features of messages generated with 8 different decoding methods with human expert writing. Results indicated that messages generated with 8 decoding methods significantly differed from expert-written messages in terms of word count ($F(8, 16148) = 137.61$, $p < .001$, partial $\eta^2 = .06$), clout ($F(8, 16148) = 2.84$, $p < .01$, partial $\eta^2 = .00$), authenticity ($F(8, 16148) = 14.82$, $p < .001$, partial $\eta^2 = .01$), emotional tone ($F(8, 16148) = 4.18$, $p < .001$, partial $\eta^2 = .00$), and the use of complex words ($F(8, 16148) = 37.95$, $p < .001$, partial $\eta^2 = .02$). Subsequent comparisons using Dunnett's test were conducted between each decoding strategy and the expert-writing condition (Figure 4.4). Results indicated that, compared to expert-written messages, messages generated with decoding methods 1, 5, and 8 contained significantly fewer words (p values $< .001$), and messages generated with decoding method 3 had higher word count (p values $< .001$). Meanwhile, decoding methods 2, 4, 6, and 7 yielded messages of lengths comparable to those written by human experts. All messages generated by LLMs, regardless of the decoding strategy employed, exhibited a level of clout similar to that of expert-written messages. Messages generated using decoding methods 3 and 7 aligned with human writing in terms

of authenticity, whereas messages generated with decoding methods 1, 2, 4, 5, 6, and 8 appeared more distanced and impersonal in style (p values $< .05$). Messages generated with decoding methods 1, 3, 4, 5, 6, 7, and 8 used neutral emotional tones like human writing, and decoding methods 2 and 3 produced more positive messages (p values $< .05$). All LLM-created messages used fewer complex words than human expert writing (all p values $< .001$). Further, the ANOVA test with post-hoc multiple comparison analysis revealed significant differences in perplexity between each decoding strategy with the expert-writing condition ($F(8, 155) = 40.92$, $p < .001$, partial $\eta^2 = .68$). Post-hoc Dunnett's test further indicated that LLMs using decoding versions 1, 5, and 8 had similar perplexity ($M \sim 5.87$ to 6.18; $SD \sim .54$ to .61) to human experts writing (M=6.54, SD=.63), whereas LLMs using decoding versions 2, 3, 4, 6, and 7 exhibited significantly higher perplexity ($M \sim 7.66$ to 8.50; $SD \sim .70$ to .87, all p values $< .01$).

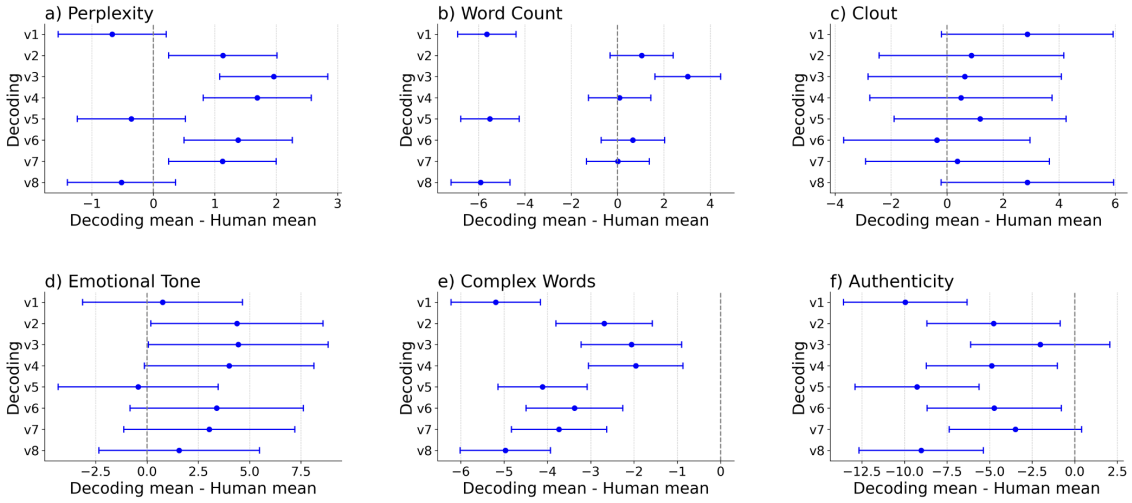

Figure 4.4: Confidence intervals for Dunnett test for decoding mean – human mean.

4.3.4 Discussion

Results revealed that temperature sampling (version 5, T=0.7) and a hybrid approach (version 1, p=0.9, k=50, T=0.8; version 8, k=40, T=0.7) exhibited perplexity scores closely to that of the expert-writing condition. Consistent with prior studies (Holtzman et al., 2019; Schmälzle and Wilcox, 2022; Fan et al., 2018; Radford et al., 2019), decoding methods with higher p values (version 2, p=0.9; version 3, p=0.95) or higher temperature settings (version 4, T=0.9) were found to sample too many unlikely tokens, resulting in messages that were diverse yet incoherent. The perplexity findings also echoed the repetition filtering process: decoding methods (version 1, 5, 8) that exhibited lower and approximating-reference perplexity scores had a higher proportion of messages (7.2% to 7.7%) discarded due to repetition; conversely, decoding methods with higher perplexity scores (version 2, 3, 4) demonstrated lower rates of repetition (3.7% to 4.5%). These coherent findings between perplexity and repetition filtering reconfirm the coherence-diversity trade-off that decoding methods favoring the most probable tokens tend to produce more deterministic and less diverse message outputs, thereby raising the likelihood of repetition issues (Holtzman et al., 2020; Hashimoto et al., 2019).

Moreover, decoding versions 6 (k=40) and 7 (k=30), which were designed to produce more deterministic and coherent outputs through lower k-values, still displayed significantly higher perplexity scores compared to the expert-writing condition. On the other hand, prompt versions 1 and 8, despite employing equal or higher k-values (version 1, p=0.9, k=50, T=0.8; version 8, k=40, T=0.7), achieved perplexity scores closely approximating human performance. This suggests that a hybrid approach, incorporating nucleus sampling with top-k and temperature adjustments (as in version 1), or top-k sampling with temperature adjustments (as in version 8), may offer a more

balanced model performance in terms of message quality and diversity (Holtzman et al., 2020).

Table 4.3: Parameters for the decoding methods used in Study 2

Version	top p[1]	top k[2]	temperature[3]
v1	0.90	50	0.8
v2	0.90	NU	NU
v3	0.95	NU	NU
v4	NU	NU	0.9
v5	NU	NU	0.7
v6	NU	40	NU
v7	NU	30	NU
v8	NU	40	0.7

Notes: NU is short for Not Used, [1]meaning p=1 in top-p sampling,[2]T=1 in temperature sampling, and [3]no capping in top-k sampling

4.4 Study III: Expert Review

4.4.1 Related Work

The development of LLMs has revolutionized the NLP field and offered new opportunities for augmenting clinical practices (Karinshak et al., 2023; Lim and Schmälzle, 2023; Agrawal et al., 2022). However, their application in clinical contexts has raised ethical concerns related to misinformation and message quality (Liebrenz et al., 2023). Consequently, the evaluation of LLMs' applicability in clinical practices necessitates a thorough examination of their safety and quality, considering subdimensions such as credibility/trustworthiness, contextual relevance, and accuracy (Reddy, 2023). While automated metrics like perplexity and BLEU scores are commonly used due to their cost-effectiveness, speed, and repeatability, they have been criticized for their limitations in assessing linguistic properties and overall text quality (Reiter and Belz, 2009; Scott and Moore, 2007; Reiter, 2018; Ananthakrishnan et al., 2007). Automated metrics can be under-informative and may not provide an accurate representation of text quality. For instance, low BLEU scores, often interpreted as indicative of poor text quality, may actually result from correct but unconventional phrasing (Ananthakrishnan et al., 2007). Moreover, incremental improvements in BLEU scores, typically in the range of 1-2 points as observed in most experimental studies, corresponded to true improvements only about half of the time when subjected to human evaluations (Mathur et al., 2020). In addition, previous studies noted a lack of correlation between automated metrics and human evaluations, highlighting the limitations of relying solely on automated metrics for comprehensive LLM evaluation (Reiter and Belz, 2009; Novikova et al., 2017; Ma et al., 2019). When applying LLMs in healthcare, automatic metrics fall short in assessing model performance for safety and quality concerns. Therefore,

human evaluation continues to be a critical element in the assessment of LLMs, frequently serving as the gold standard against which automated metrics are compared (Ouyang et al., 2022; Liang et al., 2022; Novikova et al., 2017). This underscores the importance of incorporating human judgment into the evaluation framework to capture the nuanced aspects of text quality and safety that automated metrics may not fully encapsulate.

4.4.1.1 Human Evaluations of Message Generation with LLMs

Human evaluation typically involves recruiting annotators to manually assess the quality of text generated by LLMs. This approach closely approximates real-world application scenarios and provides a more comprehensive and nuanced understanding of a model's performance. Evaluation criteria often include fluency, informativeness, relevance, coherence, accuracy, clarity, grammaticality, and appropriateness (Lee et al., 2022; Chang et al., 2015; Marvin and Linzen, 2018)

However, due to budgetary constraints, human evaluations often rely on crowd-sourced workers from platforms such as Amazon Mechanical Turk rather than on specialized experts (Liang et al., 2022; Marvin and Linzen, 2018; Wu and Aji, 2023; Jakesch et al., 2023; Clark et al., 2021). Although this approach is cost-effective, it has received criticisms for potentially compromising the validity of evaluations. Research indicates that non-expert evaluations may not consistently align well with expert assessments, especially for tasks that require specialized expertise, such as disinformation detection (Belz and Reiter, 2006). Crowd-sourced annotators have been observed to focus on superficial textual features, such as text length and grammatical precision, over more substantive criteria like content accuracy and consistency (Wu and Aji, 2023; Clark et al., 2021). Further analysis has suggested that heuristics employed by non-expert annotators, including features like nonsensical and repetitive text, grammatical issues,

rare bigrams, and long words, can be flawed and misleading in evaluation of LLM-generated messages (Jakesch et al., 2023). This raises significant concerns in contexts that are sensitive to safety and reliability, such as messages that provide health behavior recommendations, where the deployment of LLMs necessitates rigorous evaluation to ensure both reliability and safety. The lack of specialized expertise in the evaluation process could compromise its validity, thereby posing the risk of disseminating inaccurate or unsafe information through LLM-generated outputs. Therefore, while human evaluation remains a critical component of LLM assessment, the qualifications and expertise of the evaluators should be carefully considered, especially in high-stakes contexts.

4.4.1.2 ChatGPT

During this study, ChatGPT, a state-of-the-art LLM, attracted considerable attention for its exceptional performance in natural language tasks (OpenAI, 2023). Operating on a closed-source paradigm, ChatGPT is powered by GPT-3.5, an LLM trained on OpenAI's 175-billion parameter foundation model. The training process utilized a vast corpus of text data from the internet and employed a combination of reinforcement and supervised learning methods. ChatGPT has outperformed its GPT-based predecessors in linguistic capabilities and has been evaluated for its applicability in various sectors, including healthcare education, research, and practice (Kung et al., 2023; Sallam, 2023). For example, a study by Gilson et al. (2023) assessed the efficacy of ChatGPT in a medical setting and found that ChatGPT was capable of utilizing logical reasoning and external information to provide accurate, coherent, and contextually relevant answers in question-answering scenarios, demonstrating a level of competency comparable to that of third-year medical students. Given ChatGPT's proven utility in

medical contexts, this study aims to further include ChatGPT as an additional LLM and examine its feasibility in generating intervention messages for smoking cessation.

4.4.2 Method

Distinct from other LLMs, ChatGPT incorporates an additional fine-tuning process utilizing reinforcement learning from human feedback (Ouyang et al., 2022). This methodology involves ranking a broad spectrum of responses to diverse prompts based on human-labeled feedback, thereby enabling the model to discern between high-quality and suboptimal outputs. Given this unique fine-tuning process, it is not advisable to assume that the optimal prompt identified in Study 1, which was empirically validated with other LLMs, would yield similar results when applied to ChatGPT. To address this issue, a separate evaluation was conducted to identify the most effective prompt for ChatGPT.

In this evaluation, five prompts from Study 1 were initially tested on ChatGPT each across 20 iterations, to compare their effectiveness in message generation. The initial generation phase returned an average of 3 to 11 messages per iteration, except for prompt version 5, which returned only 1 message per iteration. Repetition filtering removed 17% of the messages generated using prompt version 1, but had no impact on messages generated with other prompt versions. Criteria-based filtering further discarded 1% of the messages generated with prompt version 2. Prompt versions 3 and 4 demonstrated superior performance, yielding multiple messages per iteration, all of which passed both repetition and criteria-based filtering, thereby achieving a 100% pass rate. Compared with version 3, version 4 returned more messages per iteration, thereby was recommended and applied to ChatGPT for subsequent message generation, taking into account both quality and efficacy considerations.

Following the message generation process in Study 1, prompt version 4 combined with four message exemplars randomly selected from the expert-written message library were used as input for the ChatGPT interface for 100 iterations, using its default decoding method. The initial process yielded a total of 493 messages. Subsequent message filtering discarded 0 messages for repetition (BLEU-4 ≥ 0.5) and 4 messages for context-specific criteria, resulting in a final count of 489 messages.

Founded in 2008, the Council for Tobacco Treatment Training Programs `www.ctttp.org` has developed an interdisciplinary approach to implementing training standards (Sheffer et al., 2016) for Tobacco Treatment Specialists (TTS). The TTS training ensures that TTS are well-prepared to address the dynamic and complex needs of tobacco users in diverse settings. Seven certified tobacco treatment specialists (TTS) were invited to participate in an expert review of messages generated by LLMs in Study 3, using expert-written messages as the benchmark (See Table 4.4 for examples of intervention messages and scores from expert evaluation). Messages generated using prompt version 1 and decoding version 1 from Study 2 were chosen for review due to their superior performance and efficiency. Each specialist was randomly assigned to review a set of 100 messages, with approximately 16 messages stratified and randomly selected from each of the five LLM message banks (N=624 for GPT-J-6B, N=300 for Bloom 7B1, N=498 for OPT 6.7B, N=537 for OPT 13B, and N=599 for OPT 30B), ChatGPT message bank (N=489), and the expert-written message library (N=899).

Drawing on the widely validated message evaluation protocol for health interventions (Cappella, 2018; Kim and Cappella, 2019), each message was evaluated using 10-point scales for quality, accuracy, credibility, and persuasiveness. Message quality was assessed using a single item: "Considering both content and style, how well-written is the message?" with scores ranging from 1 (poorly written) to 10 (very well-written)

adopted from Cacioppo et al. (1983). Accuracy was assessed by a single item: "How accurate is this message? (Accurate refers to no misinformation and no factual errors)" (Jung et al., 2016). Credibility was assessed by a single item asking "how credible does this message seem to you?" (Appelman and Sundar, 2016). Persuasiveness was assessed by a single item asking "To what extent do you feel this message can help smokers avoid smoking? (Baig et al., 2019). Additionally, a binary measure was employed asking whether the message met the standards in TTS training, making it ready to use in smoking cessation interventions.

Table 4.4: Examples of Intervention Messages and Scores from Expert Evaluation in Study 3

Creator	Examples of Intervention Messages	MQ[1]	AC[2]	CR[3]	PE[4]	TTS[5]
Human experts	Nicotine doesn't REALLY relax you. Your body is used to nicotine, and it seems more relaxed when you feed the addiction. Ex-smokers report feeling much less nervous just a few weeks after quitting!	7	8	8	7	Yes
	Ride out the desire to smoke. It WILL go away! Don't fool yourself into thinking you can have "just one cigarette."	2	2	2	1	No
GPT-J-6B	If you are currently in a cigarette-smoking episode, you have a few options: a. If you are not in the mood to talk to someone, just talk to yourself. b. Try to distract yourself with a phone call or text message. c. Get outside and move your body. d. If you are too uncomfortable to talk to yourself, talk to a friend, family member, or counselor. e. Call or text a friend and say you are feeling down. Let them know you are ready to quit smoking. f. If you are feeling too down, call your counselor.	10	10	10	10	Yes

	Nicotine replacement therapy (such as patches, gums, and lozenges) may help reduce the amount of nicotine in your blood and help you feel less anxious during a quit attempt.	1	1	1	1	No
BLOOM-7B1	Smoking has a direct relationship to heart disease and stroke.	9	9	9	7	Yes
	Quit smoking in the morning. If you start smoking in the morning, you have more chances of quitting.	1	2	1	2	No
OPT-6.7B	If you're going to quit smoking, try to quit by doing something fun. You're less likely to smoke when you're having fun!	10	8	9	8	Yes
	If you need help, you can call a quit line. If you are in a pinch, you can also call a quit line. If you have a quit partner, ask them to call a quit line and check in on you.	1	2	1	2	No
OPT-13B	If you smoke, you are not alone. There are millions of people who smoke and are trying to quit.	8	8	8	7	Yes
	The sooner you quit smoking, the better your chances of quitting for good.	3	1	1	1	No
OPT-30B	If you are feeling anxious, depressed, or irritable, you may have withdrawal symptoms. Talk to your health care provider or a counselor about how to cope with these feelings.	9	9	9	9	Yes
	Do not worry about having cravings. It is normal.	3	2	2	1	No
ChatGPT	Remember that quitting smoking is a journey, not a destination. Take it one day at a time and celebrate every small victory along the way.	10	10	10	10	Yes
	Consider the financial benefits of quitting smoking. Not only will you save money on cigarettes, but you'll also save money on healthcare costs in the long run.	10	1	3	3	No

Notes: [1]MQ: Message Quality, [2]AC: Accuracy, [3]CR: Credibility, [4]PE: Persuasiveness, [5]TTS: Meet TTS training standards to use in smoking cessation intervention

4.4.3 Results

Four one-way analysis of variance (ANOVA) tests with post-hoc multiple comparison analyses were conducted to compare the message quality, accuracy, credibility, and persuasiveness of LLM-generated messages with expert-written messages. Results revealed significant differences between the LLM-generated and expert-written messages in terms of message quality ($F(6, 693) = 12.93$, $p < .001$, partial $\eta^2 = .10$), accuracy ($F(6, 693) = 12.75$, $p < .001$, partial $\eta^2 = .10$), credibility ($F(6, 693) = 13.62$, $p < .001$, partial $\eta^2 = .11$), and persuasiveness ($F(6, 693) = 13.36$, $p < .001$, partial $\eta^2 = .10$). Subsequent Dunnett's test were conducted to compare each LLM with the expert-writing condition. Results indicated that, messages generated by ChatGPT achieved the highest mean scores across all four evaluation metrics. Specifically, ChatGPT significantly outperformed human experts in message quality (M=7.46, SD =2.22 vs. M=6.47, SD=2.59, $p = .04$) and demonstrated comparable performance in credibility, accuracy, and persuasiveness. Additionally, OPT-13B and OPT-30B exhibited equal performance to human experts across four evaluation metrics. Conversely, BLOOM-7B1, GPT-J-6B, and OPT-6.7B constantly underperformed human experts on four evaluations criteria. Furthermore, a Chi-Square test of independence was conducted to examine the pass rates across different LLMs in comparison to the expert-writing condition. The results indicated significant differences in pass rates, $\chi^2(1, N=700) = 10.98$, $p < .001$. Post-hoc analyses with Bonferroni adjustments revealed that ChatGPT achieved a significantly higher pass rate (71%) compared to the expert-writing condition (48%). OPT-30B and OPT-13B exhibited similar pass rates, with 40% for OPT-30B and 33% for OPT-13B. In contrast, OPT-6.7B, GPT-J-6B, and BLOOM-7B1 demonstrated significantly lower pass rates, registering at 25%, 22%, and 19%, respectively.

4.4.4 Discussion

Our findings offer compelling evidence for the capabilities of larger LLMs like ChatGPT, OPT-13B, and OPT-30B in generating high-quality intervention messages for smoking cessation. Notably, ChatGPT not only met but exceeded human performance in critical evaluation metrics. Specifically, it outperformed human experts in terms of message quality and generated a significantly higher rate of messages that met the TTS standards, as assessed by certified TTS. These results demonstrate the efficacy of LLMs in generating intervention messages for clinical practices. The superior performance of ChatGPT in generating ready-to-use , high-quality intervention messages implies its potential as a valuable tool for healthcare professionals in the field of tobacco treatment and suggests possibilities for broader adoption in various healthcare contexts.

4.5 General Discussion and Limitation

Collectively, these three studies present a thorough exploration of message generation and evaluation with LLMs in the specialized context of smoking cessation interventions. Based upon a comprehensive evaluation framework that integrates automated metrics, linguistic attributes, and expert assessments, we have demonstrated that succinct prompt with general instructions (Study 1) and a hybrid decoding approach, which incorporates nucleus sampling with top-k and temperature adjustments (Study 2), achieved optimal performance on message diversity, quality, and generation efficiency, closely approaching the level of human expert writing. Furthermore, expert review (Study 3) revealed that larger LLMs, including ChatGPT, OPT-13B and OPT-30B, can effectively emulate human expert writing to generate well-written, accurate,

credible, and persuasive messages, with about half of the generated messages meeting the clinical standards to be directly adopted in smoking cessation intervention.

This research makes several substantial contributions to the field: 1) It identifies optimal prompt and decoding strategies for message generation in specialized healthcare contexts, offering generalizable insights across five state-of-art LLMs, and thereby establishing a roadmap for their effective deployment in healthcare; 2) It adopts the computational linguistic approach and introduces a comprehensive evaluation framework that assesses LLM performance in terms of message diversity, quality, and efficiency, offering a robust framework for LLM evaluations for future studies; 3) It underscores the potential of LLMs to augment health interventions by generating a substantial volume of high-quality, accurate, and persuasive messages, thereby offer the opportunity to enhance the scalability and efficacy of future health interventions; and 4) It offers practical guidelines for healthcare professionals on the adoption of LLMs in clinical practice, addressing key considerations such as model optimization, message filtering, comprehensive evaluation, and expert review. Importantly, our findings demonstrate that advanced LLMs like ChatGPT, OPT-13B, and OPT-30B can serve as valuable adjuncts in clinical settings for content generation. Moreover, this research addresses, for the first time, the safety and quality concerns associated with the use of LLM-generated messages in healthcare contexts through the rigorous review from clinicians.

4.5.1 Optimizing LLMs for Message Generation in Healthcare

Optimizing LLMs for message generation necessitates a nuanced understanding of both the nature of the messages to be generated and the specific characteristics of the LLMs employed. Results from Study 1 indicate that a succinct prompt, featuring general instructions placed prior to message exemplars, yielded optimal performance in the generation of healthcare-related messages. This observation is consistent with existing

literature (Mishra et al., 2022; Prasad et al., 2023; Webson and Pavlick, 2022; Min et al., 2021; Lampinen et al., 2022), which posits that, in the context of few-shot learning, LLMs derive greater benefit from message exemplars for learning the distribution of input text, as opposed to relying heavily on explicit task instructions. This finding is particular true for the generation of specialized content, such as healthcare messages, which exhibit unique linguistic patterns that distinguish them from general, everyday discourse (Toma and D'Angelo, 2015; Partch and Dykeman, 2019; Tausczik and Pennebaker, 2010). In a similar vein, we hypothesize that the suboptimal performance observed with prompt version 4, which positioned the task instruction after the message exemplars, may be due to a disruption of in-context learning of the latent concept about smoking cessation. A Bayesian interpretation of the in-context learning posits that a list of independent and identically distributed (IDD) training examples provided in the prompt enables the LLM to locate a hidden concept shared among these examples (Xie et al., 2021; Min et al., 2022). The task instruction after these examples may have shifted the posterior probability distribution of the learned concept, leading to inferior performance.

Intriguingly, despite its suboptimal performance across five other LLMs, prompt version 4 yielded the most effective results when applied to ChatGPT. This underscores the notion that prompt selection should be tailored to the specific attributes of the LLM in use. ChatGPT's unique fine-tuning process (Ouyang et al., 2022), which enables it to better interpret and act upon task instructions within the prompt. In our task, the instruction after the examples may have allowed ChatGPT to better act on the latent concept learned from the IID training examples. This capability distinguishes ChatGPT from other LLMs and aligns with recent findings in the literature (Webson and Pavlick, 2022).

Moreover, Study 2 identified the optimal decoding method for generating intervention messages, specifically recommending a hybrid decoding approach that combines nucleus sampling with top-k and temperature adjustments (decoding version 1, p=0.9, k=50, T=0.8). This approach outperformed the restricted top-k sampling methods (decoding versions 6, k=40 and version 7, k=30) by offering a balanced performance in terms of both message quality and diversity, consistent with existing literature (Holtzman et al., 2020). Given the critical importance of safety and quality in healthcare messaging, our study aimed for LLMs to closely emulate human expert writing. The decoding method recommended in Study 2, tailored to this specific requirement, utilizes a restrictive strategy to prioritize accuracy over creativity. Consistent with prior research, our study suggests that different tasks necessitate different decoding methods. For instance, in tasks where accuracy is paramount, such as summarization, translation, or message generation for a healthcare context, more restrictive decoding methods are often employed (e.g., k=2, as in Fan et al. (2018)). Conversely, tasks prioritizing diversity, like creative writing, advertising copywriting, or storytelling, may benefit from more flexible decoding methods (e.g., k=100, as cited in Bhandari and Brennan (2023)). Therefore, we suggest customizing decoding methods to align with the specific requirements of the NLP task in future studies.

4.5.2 Evaluating LLMs in Healthcare Contexts

The evaluation of LLMs is important for ensuring their safe and effective deployment in healthcare settings. Traditional automatic metrics like perplexity or BLEU often prove inadequate for comprehensively assessing model performance in the context of automatic message generation (Reiter and Belz, 2009; Novikova et al., 2017; Ma et al.,

2019). Moreover, these metrics can yield results that are difficult to interpret (Ananthakrishnan et al., 2007). To address these limitations, our study proposes the augmentation of automatic metrics with computational linguistic analysis. This multi-faceted approach allows for a more nuanced and interpretable understanding of model performance. Specifically, results indicate the incompetence of perplexity to correctly reflect subtle messaging style of linguistic features. This means that LLMs can produce messages that differ significantly from those written by experts, despite having similar perplexity. Conversely, a significant discrepancy in perplexity scores between LLMs and human experts does not necessarily indicate differences in linguistic patterns either. Our work integrates computational linguistic analysis to contextualize model performance in message generation tasks, focusing on critical linguistic features. This approach provides a more holistic and interpretable evaluation of Large Language Models' (LLMs) capabilities.

Furthermore, it is worth noting that even in the best-performing case of ChatGPT, approximately 30% of generated messages did not meet the TTS standards for clinical use. Therefore, while LLMs can augment smoking cessation interventions, our results emphasize the importance of subjecting LLM-generated messages to systematic evaluation before presenting to patients. In particular, we propose a practical guideline for healthcare professionals on the integration of LLMs in clinical practice. We advocate that the process of LLM-based automated message generation should encompass model optimization, message filtering, comprehensive evaluation, and expert review to ensure its safety, quality, and effectiveness for clinical application.

4.5.3 Limitations

The current study is subject to several limitations. First, this study primarily focused on prompt engineering and decoding optimization in the message generation process.

Subsequent research may improve the model performance through two aspects:1) incorporating model fine-tuning as a third strategy, and 2) pre-processing the human-written messages before prompting. In particular, expert-written intervention messages adopted in this study were originally designed for a smartphone-based smoking cessation intervention, which covered a wide range of topics, such as coping with smoking urge, motivation to quit smoking, mood change and anxiety during withdrawal. Therefore, to improve model performance, future studies could pre-categorize expert-written messages by topics and increase the coherency of message examples in prompting. Second, despite the engagement of certified tobacco treatment specialists for the expert assessment of messages generated by LLMs, the sample size of these specialists was relatively small (N=7, reviewing a total of 700 messages). This limitation was contingent on the availability of TTS specialists within our tobacco treatment group, which could introduce bias and potentially compromise the robustness of the evaluations. Consequently, the assessments may not comprehensively represent the spectrum of opinions and clinical practices within the field. Third, the cross-sectional nature of the expert review, combined with the volume of messages (N=100) assigned to each specialist for evaluation, resulted in an out-of-context assessment of the messages. For instance, certain messages may have been crafted to address specific scenarios, such as experiencing the social pressure to smoke or coping with stress eating during nicotine withdrawal. Consequently, the validity of the evaluation concerning message persuasiveness was compromised when devoid of this contextual information. Future research could better balance between the quantity of messages and the need for in-context review. By providing contextual information specific to each message, evaluators could assess the messages' effectiveness within the scenarios they are intended to address. This approach would likely yield a more nuanced and accurate understanding of the messages' persuasive effectiveness, thereby improving the validity of future research.

4.6 Conclusion

We conducted a systematic examination of message generation and evaluation process with LLMs to address two critical questions regarding the applicability of LLMs in healthcare: 1) How to optimize LLMs to mimic human expert writing and 2) Do LLM-generated messages meet clinical standards to be safely implemented in tobacco treatments? Through three studies on prompt engineering, decoding optimization, and expert review, we identified optimal prompt and decoding strategies across state-of-art LLMs for message generation in healthcare, using human expert writing as the benchmark. We proposed a comprehensive evaluation framework encompassing automatic metrics, linguistic attributes, and expert review to assess LLMs in healthcare context on diversity, quality, and efficiency. Further, LLM-generated messages were evaluated by certified TTS on message quality, accuracy, credibility, and persuasiveness. Drawing upon their extensive clinical experiences in tobacco treatment counseling, the TTS concluded that larger LLMs, including ChatGPT, OPT-13B and OPT-30B, can effectively emulate human expert writing to generate well-written, accurate, credible, and persuasive messages, thereby demonstrating their applicability in clinical practices.

Chapter 5

Summary and Conclusions

5.1 Summary of Key Contributions

This dissertation advances both the methodological foundations of AI evaluation and the application of discriminative and generative models in healthcare. Together, these contributions illustrate how scalability, and clinical alignment can be achieved across diverse problem domains.

5.1.1 Chapter 2: Methodology (NACHOS/DACHOS)

First, I developed NACHOS, a reusable pipeline that couples nested cross-validation (NCV) with automated hyperparameter optimization (AHPO) and high-performance computing (HPC) to produce unbiased test-set estimates with quantified uncertainty. By separating outer test folds from inner model selection, NACHOS prevents data leakage and optimistic bias, enabling variance-aware evaluation at scale. The pipeline supports distributed, fault-tolerant execution across multi-GPU and multi-node environments, using checkpoints coordination to parallelize NCV folds and HPO trials. Reproducibility is ensured by design through configuration-driven experiments, with comprehensive archiving of models and checkpoints. A companion stage, DACHOS, retrains the selected model on the full training set with optimized hyperparameters to yield a deployment-ready model aligned with the evaluation protocol.

5.1.2 Chapter 3: Imaging (Discriminative)

Second, I applied discriminative deep learning to real-time procedural guidance using optical coherence tomography (OCT). From a machine learning perspective, this work illustrates how task formulation, architectural capacity, and validation strategy shape outcomes in safety-critical imaging.

For percutaneous nephrostomy (PCN), I framed tissue recognition as a supervised multi-class classification problem and developed CNN-based models capable of distinguishing cortex, medulla, and calyx. Using nested cross-validation to avoid data leakage at the specimen level, ResNet50 achieved $\sim 82.6\% \pm 3.0\%$ mean test accuracy, highlighting how deeper networks capture fine-grained structural differences while underscoring the computational trade-offs inherent in scaling model complexity.

For epidural anesthesia guidance, I advanced beyond standard multi-class framing and implemented a sequential binary pipeline aligned with the biological order of needle traversal. This divide-and-conquer decomposition reduced class overlap and simplified decision boundaries, yielding $\sim 96.65\% \pm 1.32\%$ overall accuracy and $>99\%$ precision in detecting the epidural space—a clinically critical milestone. These results demonstrate that problem representation and decomposition can be as critical as architecture, showing that the way a task is framed often governs the feasibility and reliability of real-time clinical solutions.

5.1.3 Chapter 4: Messaging (Generative)

Third, I explored generative modeling for behavioral health interventions. I designed a systematic LLM generation-and-selection framework for smoking-cessation messaging, which evaluates prompt and decoding configurations across multiple large language

models. Candidate outputs were screened using quantitative metrics such as perplexity, lexical diversity, and LIWC-based psycholinguistic profiles. To ground the system in clinical practice, seven certified counselors rated message quality, accuracy, credibility, and persuasiveness, producing a unique labeled corpus linking automated metrics to expert judgment. ChatGPT and large open-source models (e.g., OPT-13B/30B) achieved the strongest ratings. The framework further integrated guardrails including repetition controls and policy-based filters to mitigate unsafe or low-quality outputs.

5.2 Future work

5.2.1 Methodology (NACHOS/DACHOS)

Architectural capabilities will be expanded to include spatiotemporal and 3D network models, such as transformers and 3D CNNs. Beyond classification, additional tasks such as tissue segmentation will be incorporated to broaden clinical applicability. Interpretability efforts will progress beyond saliency-based methods like Grad-CAM toward concept-based (Yeh et al., 2020) and counterfactual explanations (Wachter et al., 2017), providing deeper insights that can strengthen clinician trust and adoption. In parallel, advances in containerization and workflow management will enhance portability across HPC clusters and cloud platforms, making the pipeline more accessible to both academic researchers and clinical partners

5.2.2 Discriminative models (OCT)

The next stage of imaging research will emphasize clinical realism and robustness. Data collection will expand from ex-vivo porcine experiments to in-vivo porcine and,

ultimately, human datasets. Multi-site external validation will be essential to quantify generalization across institutions and devices.

5.2.3 Generative models (LLM)

Future directions focus on developing a patient-facing conversational agent that delivers just-in-time adaptive interventions (JITAIs) for smoking cessation and beyond. Such an agent will use LLMs to provide context-aware, stage-tailored messages while maintaining a privacy-preserving longitudinal memory of prior interactions and engagement patterns. Retrieval-augmented prompting from vetted clinical guidelines will enhance factuality, while adaptive mechanisms will personalize tone, cadence, and content in real time. Ultimately, this line of research aims to move from one-shot message generation toward continuous, personalized, and safe patient engagement systems.

5.2.4 Concluding Remarks

Collectively, these contributions highlight a unifying theme: AI in healthcare must be more than accurate—it must be rigorous, scalable, interpretable, and aligned with human expertise. The methodological framework (NACHOS/DACHOS) ensures robust evaluation and reproducibility; the discriminative imaging studies demonstrate how deep learning can support real-time, high-stakes clinical procedures; and the generative messaging work illustrates the promise of LLMs in enhancing patient communication and behavioral interventions.

By bridging discriminative and generative approaches under a common philosophy of trustworthy AI, this dissertation lays the groundwork for next-generation clinical systems that combine technical sophistication with practical utility. As healthcare increasingly embraces AI, progress will depend not only on algorithmic advances but also

on addressing interpretability, scalability, real-time integration, and patient-centered communication—the very gaps this dissertation has sought to address.

Chapter 6

Appendix

6.1 Implementation Details and Hardware Configuration

NACHOS and DACHOS were implemented using Python 3.8 and TensorFlow 2.6.2. They use `dill` for saving and loading configuration checkpoints, `mpi4py` for parallelization, `fasteners` for process locking and unlocking, `NumPy` for working with arrays, `scikit-learn` for computing performance metrics, `SciPy` for statistical analysis of the results, and `termcolor` for color-coded standard output messages. NACHOS and DACHOS are capable of generating learning curves, confusion matrices, and Receiver Operating Characteristic (ROC) curves for result visualization. They can also generate class activation maps for instance-wide prediction interpretation or feature importance , using GradCAM, which requires `Matplotlib` and `seaborn`.

NACHOS and DACHOS were designed to operate on both a supercomputer with GPU nodes and a Beowulf cluster of GPU workstations connected via a local Ethernet network. The algorithms were tested on the Schooner supercomputer, utilizing GPU nodes equipped with NVIDIA A100 GPUs. Jobs on the supercomputer were managed through the SLURM system, with the GPU count per node specified. Additional experiments were performed on a Beowulf cluster comprising GPU workstations with NVIDIA RTX A6000 and NVIDIA RTX 4090 GPUs, running Ubuntu 20.04.6 LTS. In the Beowulf cluster, data were distributed across all workstations, and jobs were

executed using a configuration file that specified the GPU count and the IP address of each workstation. The repository for this paper can be fount at `https://github.com/thepanlab/NACHOS`.

6.2 Tables

Table 6.1: Performance comparison across different GPU types and configurations.

GPU type	Memory bandwidth	# tensor cores	# GPUs	Time (hrs)
RTXA6000	112.5 GB/s	336	1	21.9
RTXA6000	112.5 GB/s	336	2	11.1
RTXA6000	112.5 GB/s	336	3	6.9
RTXA6000	112.5 GB/s	336	4	5.4
RTX4090	1000.0 GB/s	512	1	13.8
RTX4090	1000.0 GB/s	512	2	7.0
RTX4090	1000.0 GB/s	512	3	4.7
RTX4090	1000.0 GB/s	512	4	3.6
RTXA6000 + RTX4090	-	-	8	2.2
A100	2000.0 GB/s	432	1	11.7
A100	2000.0 GB/s	432	2	5.8
A100	2000.0 GB/s	432	4	3.0
A100	2000.0 GB/s	432	8	1.6

Reference List

Abadi, M., and Coauthors, 2016: Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.

Abroms, L. C., J. Lee Westmaas, J. Bontemps-Jones, R. Ramani, and J. Mellerson, 2013: A content analysis of popular smartphone apps for smoking cessation. *American Journal of Preventive Medicine*, **45 (6)**, 732–736, https://doi.org/https://doi.org/10.1016/j.amepre.2013.07.008, URL https://www.sciencedirect.com/science/article/pii/S0749379713004790.

Agrawal, M., S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, 2022: Large language models are few-shot clinical information extractors. Association for Computational Linguistics, 1998–2022, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.18653/v1/2022.emnlp-main.130, URL https://aclanthology.org/2022.emnlp-main.130https://doi.org/10.18653/v1/2022.emnlp-main.130.

Ahmed, I., J. Rose, E. Keilty, C. Cooper, and P. Selby, 2022: Generation and classification of motivational-interviewing-style reflections for smoking behaviour change using few-shot learning with transformers. *Preprint posted online June*, **13**, 10–36 227.

Ananthakrishnan, R., P. Bhattacharyya, M. Sasikumar, and R. M. Shah, 2007: Some issues in automatic evaluation of english-hindi mt: more blues for bleu.

Andrews, P. M., H.-W. Wang, J. Wierwille, W. Gong, J. Verbesey, M. Cooper, and Y. Chen, 2014: Optical coherence tomography of the living human kidney. *Journal of Innovative Optical Health Sciences*, **07 (02)**, 1350 064, https://doi.org/10.1142/s1793545813500648, URL https://www.worldscientific.com/doi/abs/10.1142/S1793545813500648.

Andrews, P. M., and Coauthors, 2008: High-resolution optical coherence tomography imaging of the living kidney. *Laboratory Investigation*, **88 (4)**, 441–449, https://doi.org/10.1038/labinvest.2008.4, URL https://doi.org/10.1038/labinvest.2008.4.

Appelman, A., and S. S. Sundar, 2016: Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, **93 (1)**, 59–79, https://doi.org/10.1177/1077699015606057, URL http://journals.sagepub.com/doi/10.1177/1077699015606057.

Ardila, D., and Coauthors, 2019: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, **25 (6)**, 954–961, https://doi.org/10.1038/s41591-019-0447-x, URL https://doi.org/10.1038/s41591-019-0447-x.

Badré, A., and C. Pan, 2022: Lina: A linearizing neural network architecture for accurate first-order and second-order interpretations. *IEEE Access*, **10**, 36 166–36 176, https://doi.org/10.1109/ACCESS.2022.3163257.

Badré, A., and C. Pan, 2023: Explainable multi-task learning improves the parallel estimation of polygenic risk scores for many diseases through shared genetic basis. *PLOS Computational Biology*, **19 (7)**, e1011 211, https://doi.org/10.1371/journal.pcbi.1011211, URL https://doi.org/10.1371/journal.pcbi.1011211.

Bahl, L. R., F. Jelinek, and R. L. Mercer, 1983: A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-5 (2)**, 179–190, https://doi.org/10.1109/TPAMI.1983.4767370, URL http://ieeexplore.ieee.org/document/4767370/.

Baig, S. A., S. M. Noar, N. C. Gottfredson, M. H. Boynton, K. M. Ribisl, and N. T. Brewer, 2019: UNC perceived message effectiveness: Validation of a brief scale. *Annals of Behavioral Medicine*, **53 (8)**, 732–742, https://doi.org/10.1093/abm/kay080, URL https://academic.oup.com/abm/article/53/8/732/5131526.

Bakhtiarnia, A., Q. Zhang, and A. Iosifidis, 2024: Efficient high-resolution deep learning: A survey. *ACM Computing Surveys*, **56 (7)**, 1–35.

Banh, L., and G. Strobel, 2023: Generative artificial intelligence. *Electronic Markets*, **33 (1)**, 63, https://doi.org/10.1007/s12525-023-00680-1, URL https://doi.org/10.1007/s12525-023-00680-1.

Bates, S., T. Hastie, and R. Tibshirani, 2024: Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, **119 (546)**, 1434–1445, https://doi.org/10.1080/01621459.2023.2197686, URL https://doi.org/10.1080/01621459.2023.2197686, doi: 10.1080/01621459.2023.2197686.

Baumgartner, J., S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, 2020: The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, **14 (1)**, 830–839, https://doi.org/10.1609/icwsm.v14i1.7347, URL https://ojs.aaai.org/index.php/ICWSM/article/view/7347.

Beam, A. L., and I. S. Kohane, 2018: Big data and machine learning in health care. *Jama*, **319 (13)**, 1317–1318.

Beers, A., and Coauthors, 2021: Deepneuro: an open-source deep learning toolbox for neuroimaging. *Neuroinformatics*, **19 (1)**, 127–140, https://doi.org/10.1007/s12021-020-09477-5, URL https://doi.org/10.1007/s12021-020-09477-5.

Belz, A., and E. Reiter, 2006: Comparing automatic and human evaluation of nlg systems. Association for Computational Linguistics, 313–320, 11th Conference of the European Chapter of the Association for Computational Linguistics, URL https://aclanthology.org/E06-1040.

Bergstra, J., and Y. Bengio, 2012: Random search for hyper-parameter optimization. *Journal of machine learning research*, **13 (2)**.

Berrar, D., 2019: Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, **1 (April)**, 542–545.

Bhandari, P., and H. M. Brennan, 2023: Trustworthiness of children stories generated by large language models. *arXiv preprint arXiv:2308.00073*.

Biderman, S., K. Bicheno, and L. Gao, 2022: Datasheet for the pile. *arXiv preprint arXiv:2201.07311*.

Bischl, B., and Coauthors, 2023: Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, **13 (2)**, e1484, https://doi.org/https://doi.org/10.1002/widm.1484, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1484.

Bishop, C. M., and N. M. Nasrabadi, 2006: *Pattern recognition and machine learning*, Vol. 4. Springer.

Bommasani, R., and Coauthors, 2021: On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boyd, R. L., A. Ashokkumar, S. Seraj, and J. W. Pennebaker, 2022: The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 1–47.

Brown, O. N., L. E. O'Connor, and D. Savaiano, 2014: Mobile MyPlate: A Pilot Study Using Text Messaging to Provide Nutrition Education and Promote Better Dietary Choices in College Students. *Journal of American College Health*, **62 (5)**, 320–327, https://doi.org/10.1080/07448481.2014.899233, URL http://www.tandfonline.com/doi/abs/10.1080/07448481.2014.899233.

Brown, T. B., and Coauthors, 2020: Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Businelle, M. S., P. Ma, D. E. Kendzor, S. G. Frank, D. J. Vidrine, and D. W. Wetter, 2016: An ecological momentary intervention for smoking cessation: evaluation of feasibility and effectiveness. *Journal of Medical Internet Research*, **18 (12)**, e6058.

Businelle, M. S., P. Ma, D. E. Kendzor, L. R. Reitzel, M. Chen, C. Y. Lam, I. Bernstein, and D. W. Wetter, 2014: Predicting quit attempts among homeless smokers seeking cessation treatment: An ecological momentary assessment study. *Nicotine & Tobacco Research*, **16 (10)**, 1371–1378, https://doi.org/10.1093/ntr/ntu088, URL https://doi.org/10.1093/ntr/ntu088.

Bustos, A., A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, 2020: Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, **66**, 101 797, https://doi.org/https://doi.org/10.1016/j.media.2020.101797, URL https://www.sciencedirect.com/science/article/pii/S1361841520301614.

Caccia, M., L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin, 2020: Language GANS falling short.

Cacioppo, J. T., R. E. Petty, and K. J. Morris, 1983: Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of personality and social psychology*, **45 (4)**, 805.

Cappella, J. N., 2018: Perceived message effectiveness meets the requirements of a reliable, valid, and efficient measure of persuasiveness. *Journal of Communication*, **68 (5)**, 994–997, https://doi.org/10.1093/joc/jqy044, URL https://doi.org/10.1093/joc/jqy044.

Carreiro, S., M. Newcomb, R. Leach, S. Ostrowski, E. D. Boudreaux, and D. Amante, 2020: Current reporting of usability and impact of mHealth interventions for substance use disorder: A systematic review. *Drug and Alcohol Dependence*, **215**, 108 201, https://doi.org/10.1016/j.drugalcdep.2020.108201, URL https://linkinghub.elsevier.com/retrieve/pii/S0376871620303665.

Chang, S., P. Dai, J. Chen, and E. H. Chi, 2015: Got many labels? deriving topic labels from multiple sources for social media posts using crowdsourcing and ensemble learning. *Proceedings of the 24th International Conference on World Wide Web*, 397–406.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton, 2020: A simple framework for contrastive learning of visual representations. *International conference on machine learning*, PmLR, 1597–1607.

Chollet, F., 2017: Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Clark, E., T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith, 2021: All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

Clarkson, P., and T. Robinson, 1999: Towards improved language model evaluation measures. *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, ISCA, 1927–1930, https://doi.org/10.21437/Eurospeech.1999-423, URL https://www.isca-speech.org/archive/eurospeech_1999/clarkson99_eurospeech.html.

Clavié, B., A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, 2023: Large language models in the workplace: A case study on prompt engineering for job type classification. *International Conference on Applications of Natural Language to Information Systems*, Springer, 3–17.

Cohen, J. P., and Coauthors, 2022: Torchxrayvision: A library of chest x-ray datasets and models. *International Conference on Medical Imaging with Deep Learning*, PMLR, 231–249.

Cohn, M. A., M. R. Mehl, and J. W. Pennebaker, 2004: Linguistic Markers of Psychological Change Surrounding September 11, 2001. *Psychological Science*, **15 (10)**, 687–693, https://doi.org/10.1111/j.0956-7976.2004.00741.x, URL http://journals.sagepub.com/doi/10.1111/j.0956-7976.2004.00741.x.

Coxson, H. O., B. Quiney, D. D. Sin, L. Xing, A. M. McWilliams, J. R. Mayo, and S. Lam, 2008: Airway wall thickness assessed using computed tomography and optical coherence tomography. *American journal of respiratory and critical care medicine*, **177 (11)**, 1201–1206.

Cui, X.-W., A. Ignee, T. Maros, B. Straub, J.-G. Wen, and C. F. Dietrich, 2016: Feasibility and usefulness of intra-cavitary contrast-enhanced ultrasound in percutaneous nephrostomy. *Ultrasound in Medicine & Biology*, **42 (9)**, 2180–2188, https://doi.org/https://doi.org/10.1016/j.ultrasmedbio.2016.04.015, URL https://www.sciencedirect.com/science/article/pii/S0301562916300527.

Dalcin, L., and Y.-L. L. Fang, 2021: mpi4py: Status update after 12 years of development. *Computing in Science & Engineering*, **23 (4)**, 47–54.

Davenport, T., and R. Kalakota, 2019: The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, **6 (2)**, 94–98, https://doi.org/https://doi.org/10.7861/futurehosp.6-2-94, URL https://www.sciencedirect.com/science/article/pii/S2514664524010592.

Dean, J., and Coauthors, 2012: Large scale distributed deep networks. *Advances in neural information processing systems*, **25**.

DeLucia, A., A. Mueller, X. L. Li, and J. Sedoc, 2021: Decoding methods for neural narrative generation. Association for Computational Linguistics, 166–185, Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and

Metrics (GEM 2021), https://doi.org/10.18653/v1/2021.gem-1.16, URL https://aclanthology.org/2021.gem-1.16https://doi.org/10.18653/v1/2021.gem-1.16.

Deng, J., W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li, 2009: Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

Doshi-Velez, F., and B. Kim, 2017: Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, 2017: Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, **542 (7639)**, 115–118, https://doi.org/10.1038/nature21056, URL https://doi.org/10.1038/nature21056.

Esteva, A., and Coauthors, 2019: A guide to deep learning in healthcare. *Nature Medicine*, **25 (1)**, 24–29, https://doi.org/10.1038/s41591-018-0316-z, URL https://doi.org/10.1038/s41591-018-0316-z.

Falkner, S., A. Klein, and F. Hutter, 2018: BOHB: Robust and efficient hyperparameter optimization at scale. PMLR, URL http://proceedings.mlr.press/v80/falkner18a.html, 1437–1446 pp.

Fan, A., M. Lewis, and Y. Dauphin, 2018: Hierarchical neural story generation. Association for Computational Linguistics, 889–898, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), https://doi.org/10.18653/v1/P18-1082, URL https://aclanthology.org/P18-1082https://doi.org/10.18653/v1/P18-1082.

Free, C., and Coauthors, 2011: Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. *The Lancet*, **378 (9785)**, 49–55, https://doi.org/10.1016/S0140-6736(11)60701-0, URL https://doi.org/10.1016/S0140-6736(11)60701-0, doi: 10.1016/S0140-6736(11)60701-0.

Frid-Adar, M., I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, 2018: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, **321**, 321–331, https://doi.org/https://doi.org/10.1016/j.neucom.2018.09.013, URL https://www.sciencedirect.com/science/article/pii/S0925231218310749.

Fujimoto, J., and E. Swanson, 2016: The development, commercialization, and impact of optical coherence tomography. *Investigative ophthalmology & visual science*, **57 (9)**, OCT1–OCT13.

Geirhos, R., J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, 2020: Shortcut learning in deep neural networks. *Nature Machine*

Intelligence, **2 (11)**, 665–673, https://doi.org/10.1038/s42256-020-00257-z, URL https://doi.org/10.1038/s42256-020-00257-z.

Gibson, E., and Coauthors, 2018: Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, **158**, 113–122, https://doi.org/https://doi.org/10.1016/j.cmpb.2018.01.025, URL https://www.sciencedirect.com/science/article/pii/S0169260717311823.

Gilson, A., C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, 2023: How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, **9**, e45 312, https://doi.org/10.2196/45312, URL https://mededu.jmir.org/2023/1/e45312.

Giorgi, J., A. Toma, R. Xie, S. Chen, K. An, G. Zheng, and B. Wang, 2023: Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 323–334.

Gonen, H., S. Iyer, T. Blevins, N. A. Smith, and L. Zettlemoyer, 2022: Demystifying Prompts in Language Models via Perplexity Estimation. arXiv, URL http://arxiv.org/abs/2212.04037, arXiv:2212.04037 [cs].

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio, 2016: *Deep learning*, Vol. 1. MIT press Cambridge.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014: Generative adversarial nets. *Advances in neural information processing systems*, **27**.

Goodwin, W. E., W. C. Casey, and W. Woolf, 1955: Percutaneous trocar (needle) nephrostomy in hydronephrosis. *Journal of the American Medical Association*, **157 (11)**, 891–894, https://doi.org/10.1001/jama.1955.02950280015005, URL https://doi.org/10.1001/jama.1955.02950280015005.

Gora, M. J., M. J. Suter, G. J. Tearney, and X. Li, 2017: Endoscopic optical coherence tomography: technologies and clinical applications [invited]. *Biomedical Optics Express*, **8 (5)**, 2405–2444, https://doi.org/10.1364/BOE.8.002405, URL https://opg.optica.org/boe/abstract.cfm?URI=boe-8-5-2405.

Graham, A. L., G. D. Papandonatos, M. A. Jacobs, M. S. Amato, S. Cha, A. M. Cohn, L. C. Abroms, and R. Whittaker, 2020: Optimizing Text Messages to Promote Engagement With Internet Smoking Cessation Treatment: Results From a Factorial Screening Experiment. *Journal of Medical Internet Research*, **22 (4)**, e17 734, https://doi.org/10.2196/17734, URL https://www.jmir.org/2020/4/e17734.

Gulshan, V., and Coauthors, 2016: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, **316 (22)**, 2402–2410, https://doi.org/10.1001/jama.2016.17216, URL https://doi.org/10.1001/jama.2016.17216.

Géron, A., 2019: *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Hashimoto, T. B., H. Zhang, and P. Liang, 2019: Unifying human and statistical evaluation for natural language generation. Association for Computational Linguistics, 1689–1701, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), https://doi.org/10.18653/v1/N19-1169, URL https://aclanthology.org/N19-1169https://doi.org/10.18653/v1/N19-1169.

Hawkins, C. M., K. Kukreja, T. Singewald, E. Minevich, N. D. Johnson, P. Reddy, and J. M. Racadio, 2016: Use of cone-beam ct and live 3-d needle guidance to facilitate percutaneous nephrostomy and nephrolithotripsy access in children and adolescents. *Pediatric Radiology*, **46 (4)**, 570–574, https://doi.org/10.1007/s00247-015-3499-1, URL https://doi.org/10.1007/s00247-015-3499-1.

Hawkins, R. P., M. Kreuter, K. Resnicow, M. Fishbein, and A. Dijkstra, 2008: Understanding tailoring in communicating about health. *Health Education Research*, **23 (3)**, 454–466, https://doi.org/10.1093/her/cyn004, URL https://doi.org/10.1093/her/cyn004.

He, K., X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, 2022: Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16 000–16 009.

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, M., Z. Li, C. Liu, D. Shi, and Z. Tan, 2020: Deployment of artificial intelligence in real-world practice: Opportunity and challenge. *Asia-Pacific Journal of Ophthalmology*, **9 (4)**, 299–307, https://doi.org/https://doi.org/10.1097/APO.0000000000000301, URL https://www.sciencedirect.com/science/article/pii/S2162098923001858.

Hee, M. R., and Coauthors, 1995: Quantitative assessment of macular edema with optical coherence tomography. *Archives of Ophthalmology*, **113 (8)**, 1019–1029, https://doi.org/10.1001/archopht.1995.01100080071031, URL https://doi.org/10.1001/archopht.1995.01100080071031.

Hermanides, J., M. W. Hollmann, M. F. Stevens, and P. Lirk, 2012: Failed epidural: causes and management. *BJA: British Journal of Anaesthesia*, **109 (2)**, 144–154, https://doi.org/10.1093/bja/aes214, URL https://doi.org/10.1093/bja/aes214.

Ho, J., A. Jain, and P. Abbeel, 2020: Denoising diffusion probabilistic models. *Advances in neural information processing systems*, **33**, 6840–6851.

Hoffmann, V. L., M. P. Vercauteren, J. P. Vreugde, G. H. Hans, H. C. Coppejans, and H. A. Adriaensen, 1999: Posterior epidural space depth: safety of the loss of resistance and hanging drop techniques. *British Journal of Anaesthesia*, **83 (5)**, 807–809, https://doi.org/https://doi.org/10.1093/bja/83.5.807, URL https://www.sciencedirect.com/science/article/pii/S0007091217382776.

Hollmann, M. W., K. S. Wieczorek, M. Smart, and M. E. Durieux, 2001: Epidural anesthesia prevents hypercoagulation in patients undergoing major orthopedic surgery. *Regional Anesthesia & Pain Medicine*, **26 (3)**, 215–222, https://doi.org/10.1053/rapm.2001.23209, URL https://rapm.bmj.com/content/rapm/26/3/215.full.pdf.

Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi, 2019: The curious case of neural text degeneration. *International Conference on Learning Representations*.

Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi, 2020: The Curious Case of Neural Text Degeneration. arXiv, URL http://arxiv.org/abs/1904.09751, arXiv:1904.09751 [cs].

Horlocker, T. T., 2000: Complications of spinal and epidural anesthesia. *Anesthesiology Clinics of North America*, **18 (2)**, 461–485, https://doi.org/https://doi.org/10.1016/S0889-8537(05)70172-3, URL https://www.sciencedirect.com/science/article/pii/S0889853705701723.

Huang, D., and Coauthors, 1991: Optical coherence tomography. *science*, **254 (5035)**, 1178–1181.

Hébert, E. T., E. M. Stevens, S. G. Frank, D. E. Kendzor, D. W. Wetter, M. J. Zvolensky, J. D. Buckner, and M. S. Businelle, 2018: An ecological momentary intervention for smoking cessation: the associations of just-in-time, tailored messages with lapse risk factors. *Addictive behaviors*, **78**, 30–35.

Hébert, E. T., and Coauthors, 2020: A mobile just-in-time adaptive intervention for smoking cessation: pilot randomized controlled trial. *Journal of medical Internet research*, **22 (3)**.

Ippolito, D., R. Kriz, J. Sedoc, M. Kustikova, and C. Callison-Burch, 2019: Comparison of Diverse Decoding Methods from Conditional Language Models. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 3752–3762, https://doi.org/10.18653/v1/P19-1365, URL https://www.aclweb.org/anthology/P19-1365.

Irvin, J., and Coauthors, 2019: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on*

Artificial Intelligence, **33 (01)**, 590–597, https://doi.org/10.1609/aaai.v33i01.3301590, URL https://ojs.aaai.org/index.php/AAAI/article/view/3834.

Jakesch, M., J. T. Hancock, and M. Naaman, 2023: Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, **120 (11)**, e2208839 120, https://doi.org/doi:10.1073/pnas.2208839120, URL https://www.pnas.org/doi/abs/10.1073/pnas.2208839120.

Jebara, T., 2012: *Machine learning: discriminative and generative*, Vol. 755. Springer Science & Business Media.

Ji, Z., and Coauthors, 2023: Survey of hallucination in natural language generation. *ACM computing surveys*, **55 (12)**, 1–38.

Johnson, A. E. W., T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, 2019: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, **6 (1)**, 317, https://doi.org/10.1038/s41597-019-0322-0, URL https://doi.org/10.1038/s41597-019-0322-0.

Jordan, M. I., and T. M. Mitchell, 2015: Machine learning: Trends, perspectives, and prospects. *Science*, **349 (6245)**, 255–260.

Jung, E. H., K. Walsh-Childers, and H.-S. Kim, 2016: Factors influencing the perceived credibility of diet-nutrition information web sites. *Computers in Human Behavior*, **58**, 37–47.

Kacewicz, E., J. W. Pennebaker, M. Davis, M. Jeon, and A. C. Graesser, 2014: Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology*, **33 (2)**, 125–143, https://doi.org/10.1177/0261927X13502654, URL http://journals.sagepub.com/doi/10.1177/0261927X13502654.

Karinshak, E., S. X. Liu, J. S. Park, and J. T. Hancock, 2023: Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, **7 (CSCW1)**, 1–29.

Kelly, C. J., A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, 2019: Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, **17 (1)**, 195.

Kim, M., and J. N. Cappella, 2019: Reliable, valid and efficient evaluation of media messages: Developing a message testing protocol. *Journal of Communication Management*, **23 (3)**, 179–197.

Kingma, D. P., D. J. Rezende, S. Mohamed, and M. Welling, 2014: Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, **27**.

Kingma, D. P., and M. Welling, 2013: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, G., D. M. Ells, D. R. Camenga, and S. Krishnan-Sarin, 2014: Text messaging-based smoking cessation intervention: A narrative review. *Addictive Behaviors*, **39 (5)**, 907–917, https://doi.org/10.1016/j.addbeh.2013.11.024, URL https://linkinghub.elsevier.com/retrieve/pii/S0306460313004164.

Kreuter, M. W., and R. J. Wray, 2003: Tailored and targeted health communication: strategies for enhancing information relevance. *American journal of health behavior*, **27 (1)**, S227–S232.

Krombach, G. A., and Coauthors, 2001: Us-guided nephrostomy with the aid of a magnetic field–based navigation device in the porcine pelvicaliceal system. *Journal of Vascular and Interventional Radiology*, **12 (5)**, 623–628, https://doi.org/https://doi.org/10.1016/S1051-0443(07)61488-2, URL https://www.sciencedirect.com/science/article/pii/S1051044307614882.

Kung, T. H., and Coauthors, 2023: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, **2 (2)**, e0000 198, https://doi.org/10.1371/journal.pdig.0000198, URL https://dx.plos.org/10.1371/journal.pdig.0000198.

Lampinen, A., and Coauthors, 2022: Can language models learn from explanations in context? Association for Computational Linguistics, 537–563, Findings of the Association for Computational Linguistics: EMNLP 2022, https://doi.org/10.18653/v1/2022.findings-emnlp.38, URL https://aclanthology.org/2022.findings-emnlp.38https://doi.org/10.18653/v1/2022.findings-emnlp.38.

Laurençon, H., and Coauthors, 2022: The bigscience roots corpus: A 1.6tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, **35**, 31 809–31 826.

Lee, K., Y. Y. Lee, M. Suh, J. K. Jun, B. Park, Y. Kim, and K. S. Choi, 2022: Impact of COVID-19 on cancer screening in South Korea. *Scientific Reports*, **12 (1)**, 11 380, https://doi.org/10.1038/s41598-022-15778-3, URL https://www.nature.com/articles/s41598-022-15778-3.

Li, L., K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, 2018: Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, **18 (1)**, 6765–6816.

Li, Q., and Coauthors, 2009: Automated quantification of microstructural dimensions of the human kidney using optical coherence tomography (oct). *Optics Express*, **17 (18)**, 16 000–16 016, https://doi.org/10.1364/OE.17.016000, URL https://opg.optica.org/oe/abstract.cfm?URI=oe-17-18-16000.

Li, X., and Coauthors, 2000: Optical coherence tomography: advanced technology for the endoscopic imaging of barrett's esophagus. *Endoscopy*, **32 (12)**, 921–930.

Li, Y., H. Dong, S. Tan, Y. Qian, and W. Jin, 2019: Effects of thoracic epidural anesthesia/analgesia on the stress response, pain relief, hospital stay, and treatment costs of patients with esophageal carcinoma undergoing thoracic surgery: A single-center, randomized controlled trial. *Medicine*, **98 (7)**, e14 362, https://doi.org/10.1097/md.0000000000014362, URL https://journals.lww.com/md-journal/fulltext/2019/02150/effects_of_thoracic_epidural_anesthesia_analgesia.12.aspx.

Liang, P., and Coauthors, 2022: Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liebrenz, M., R. Schleifer, A. Buadze, D. Bhugra, and A. Smith, 2023: Generating scholarly content with chatgpt: ethical challenges for medical publishing. *The Lancet Digital Health*, **5 (3)**, e105–e106.

Lim, S., and R. Schmälzle, 2023: Artificial intelligence for health message generation: an empirical study using a large language model (llm) and prompt engineering. *Frontiers in Communication*, **8**, 1129 082.

Liu, C.-L., H.-y. Lee, and W.-t. Yih, 2022: Structured prompt tuning. *arXiv preprint arXiv:2205.12309*.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, 2021a: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, X., and Coauthors, 2019a: A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, **1 (6)**, e271–e297, https://doi.org/10.1016/S2589-7500(19)30123-2, URL https://doi.org/10.1016/S2589-7500(19)30123-2, doi: 10.1016/S2589-7500(19)30123-2.

Liu, Y., and Coauthors, 2019b: Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., and Coauthors, 2023: Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-radiology*, **1 (2)**, 100 017.

Liu, Z., Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, 2021b: Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10 012–10 022.

Logan IV, R., I. Balazevic, E. Wallace, F. Petroni, S. Singh, and S. Riedel, 2022: Cutting down on prompts and parameters: Simple few-shot learning with language models. Association for Computational Linguistics, 2824–2835, Findings of the Association for Computational Linguistics: ACL 2022, https://doi.org/10.18653/v1/2022.findings-acl.222, URL https://aclanthology.org/2022.findings-acl.222https://doi.org/10.18653/v1/2022.findings-acl.222.

Lu, M.-H., X.-Y. Pu, X. Gao, X.-F. Zhou, J.-G. Qiu, and J. Si-Tu, 2010: A comparative study of clinical value of single b-mode ultrasound guidance and b-mode combined with color doppler ultrasound guidance in mini-invasive percutaneous nephrolithotomy to decrease hemorrhagic complications. *Urology*, **76 (4)**, 815–820, https://doi.org/https://doi.org/10.1016/j.urology.2009.08.091, URL https://www.sciencedirect.com/science/article/pii/S0090429510004966.

Luo, R., L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, 2022: Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, **23 (6)**, bbac409.

Ma, Q., J. Wei, O. Bojar, and Y. Graham, 2019: Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. Association for Computational Linguistics, 62–90, Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), https://doi.org/10.18653/v1/W19-5302, URL https://aclanthology.org/W19-5302https://doi.org/10.18653/v1/W19-5302.

Marvin, R., and T. Linzen, 2018: Targeted syntactic evaluation of language models. Association for Computational Linguistics, 1192–1202, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.18653/v1/D18-1151, URL https://aclanthology.org/D18-1151https://doi.org/10.18653/v1/D18-1151.

Mason, M. J., and Coauthors, 2015: Development and Outcomes of A Text Messaging Tobacco Cessation Intervention with Urban Adolescents. *Substance Abuse*, **36 (4)**, 500–506, https://doi.org/10.1080/08897077.2014.987946, URL http://journals.sagepub.com/doi/10.1080/08897077.2014.987946.

Masters, D., and C. Luschi, 2018: Revisiting small batch training for deep neural networks. arxiv 2018. *arXiv preprint arXiv:1804.07612*.

Mathur, N., T. Baldwin, and T. Cohn, 2020: Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4984–4997.

McLeod, A., A. Roche, and M. Fennelly, 2005: Case series: ultrasonography may assist epidural insertion in scoliosis patients. *Canadian Journal of Anesthesia*, **52 (7)**, 717–720, https://doi.org/10.1007/BF03016559, URL https://doi.org/10.1007/BF03016559.

Meier, T., R. L. Boyd, J. W. Pennebaker, M. R. Mehl, M. Martin, M. Wolf, and A. B. Horn, 2019: "LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE- LIWC2015. preprint, PsyArXiv. https://doi.org/10.31234/osf.io/uq8zt, URL https://osf.io/uq8zt.

Miller, R. A., H. E. Pople Jr, and J. D. Myers, 1985: *Internist-I, an experimental computer-based diagnostic consultant for general internal medicine*, 139–158. Springer.

Miller, R. A., and J. E. A. Wickham, 1984: Percutaneous nephrolithotomy: Advances in equipment and endoscopic techniques. *Urology*, **23 (5, Supplement)**, 2–6, https://doi.org/https://doi.org/10.1016/0090-4295(84)90234-6, URL https://www.sciencedirect.com/science/article/pii/0090429584902346.

Min, B., and Coauthors, 2021: Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. arXiv, URL http://arxiv.org/abs/2111.01243, arXiv:2111.01243 [cs].

Min, S., X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, 2022: Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Mishra, S., D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi, 2022: Reframing instructional prompts to gptk's language. Association for Computational Linguistics, 589–612, Findings of the Association for Computational Linguistics: ACL 2022, https://doi.org/10.18653/v1/2022.findings-acl.50, URL https://aclanthology.org/2022.findings-acl.50https://doi.org/10.18653/v1/2022.findings-acl.50.

Mongan, J., L. Moy, and C. E. Kahn Jr, 2020: Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. Radiological Society of North America.

Montvilas, P., J. Solvig, and T. E. B. Johansen, 2011: Single-centre review of radiologically guided percutaneous nephrostomy using "mixed" technique: Success and complication rates. *European Journal of Radiology*, **80 (2)**, 553–558, https://doi.org/https://doi.org/10.1016/j.ejrad.2011.01.109, URL https://www.sciencedirect.com/science/article/pii/S0720048X11001689.

Moraca, R. J., D. G. Sheldon, and R. C. Thirlby, 2003: The role of epidural anesthesia and analgesia in surgical practice. *Annals of Surgery*, **238 (5)**, 663–673, https://doi.org/10.1097/01.sla.0000094300.36689.ad, URL https://journals.lww

.com/annalsofsurgery/fulltext/2003/11000/the_role_of_epidural_anesthesia_and_analgesia_in.5.aspx.

Nadeem, M., T. He, K. Cho, and J. Glass, 2020: A systematic characterization of sampling algorithms for open-ended language generation. Association for Computational Linguistics, 334–346, Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, URL https://aclanthology.org/2020.aacl-main.36.

Nalisnick, E., A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, 2018: Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*.

Nawabi, J., and Coauthors, 2021: Imaging-based outcome prediction of acute intracerebral hemorrhage. *Translational Stroke Research*, **12 (6)**, 958–967, https://doi.org/10.1007/s12975-021-00891-8, URL https://doi.org/10.1007/s12975-021-00891-8.

Newman, M. L., J. W. Pennebaker, D. S. Berry, and J. M. Richards, 2003: Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, **29 (5)**, 665–675, https://doi.org/10.1177/0146167203029005010, URL http://journals.sagepub.com/doi/10.1177/0146167203029005010.

Novikova, J., O. Dušek, A. Cercas Curry, and V. Rieser, 2017: Why we need new evaluation metrics for nlg. Association for Computational Linguistics, 2241–2252, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.18653/v1/D17-1238, URL https://aclanthology.org/D17-1238https://doi.org/10.18653/v1/D17-1238.

OpenAI, 2023: Chatgpt.

Ouyang, L., and Coauthors, 2022: Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., Vol. 35, 27 730–27 744, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Ovadia, Y., and Coauthors, 2019: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, **32**.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu, 2002: Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Partch, M., and C. Dykeman, 2019: A corpus linguistics study of text message interventions in substance use disorder treatment.

Pati, S., and Coauthors, 2023: Gandlf: the generally nuanced deep learning framework for scalable end-to-end clinical workflows. *Communications Engineering*, **2 (1)**, 23, https://doi.org/10.1038/s44172-023-00066-3, URL https://doi.org/10.1038/s44172-023-00066-3.

Patrick, K., F. Raab, M. A. Adams, L. Dillon, M. Zabinski, C. L. Rock, W. G. Griswold, and G. J. Norman, 2009: A Text Message–Based Intervention for Weight Loss: Randomized Controlled Trial. *Journal of Medical Internet Research*, **11 (1)**, e1, https://doi.org/10.2196/jmir.1100, URL http://www.jmir.org/2009/1/e1/.

Perski, O., E. T. Hébert, F. Naughton, E. B. Hekler, J. Brown, and M. S. Businelle, 2022: Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. *Addiction*, **117 (5)**, 1220–1241, https://doi.org/10.1111/add.15687, URL https://onlinelibrary.wiley.com/doi/10.1111/add.15687.

Poneros, J. M., S. Brand, B. E. Bouma, G. J. Tearney, C. C. Compton, and N. S. Nishioka, 2001: Diagnosis of specialized intestinal metaplasia by optical coherence tomography. *Gastroenterology*, **120 (1)**, 7–12, https://doi.org/https://doi.org/10.1053/gast.2001.20911, URL https://www.sciencedirect.com/science/article/pii/S0016508501291 15X.

Poneros, J. M., and N. S. Nishioka, 2003: Diagnosis of barrett's esophagus using optical coherence tomography. *Gastrointestinal Endoscopy Clinics*, **13 (2)**, 309–323, https://doi.org/10.1016/S1052-5157(03)00012-6, URL https://doi.org/10.1016/S1052-5157(03)00012-6, doi: 10.1016/S1052-5157(03)00012-6.

Prasad, A., P. Hase, X. Zhou, and M. Bansal, 2023: Grips: Gradient-free, edit-based instruction search for prompting large language models. Association for Computational Linguistics, 3845–3864, Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, https://doi.org/10.18653/v1/2023.eacl-main.277, URL https://aclanthology.org/2023.eacl-main.277https://doi.org/10.18653/v1/2023.eacl-main.277.

Press, O., N. Smith, and M. Lewis, 2022: Train short, test long: Attention with linear biases enables input length extrapolation. *International Conference on Learning Representations*.

Price, W. N., and I. G. Cohen, 2019: Privacy in the age of medical big data. *Nature Medicine*, **25 (1)**, 37–43, https://doi.org/10.1038/s41591-018-0272-7, URL https://doi.org/10.1038/s41591-018-0272-7.

Puliafito, C. A., and Coauthors, 1995: Imaging of macular diseases with optical coherence tomography. *Ophthalmology*, **102 (2)**, 217–229, https://doi.org/https://doi.org/10.1016/S0161-6420(95)31032-9, URL https://www.sciencedirect.com/science/article/pii/S0161642095310329.

Pérez-García, F., R. Sparks, and S. Ourselin, 2021: Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, **208**, 106 236, https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106236, URL https://www.sciencedirect.com/science/article/pii/S0169260721003102.

Radford, A., L. Metz, and S. Chintala, 2015: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, 2019: Language models are unsupervised multitask learners. *OpenAI blog*, **1 (8)**, 9.

Raghu, M., C. Zhang, J. Kleinberg, and S. Bengio, 2019: Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, **32**.

Rajpurkar, P., and Coauthors, 2017: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ramchandani, P., J. F. Cardella, C. J. Grassi, A. C. Roberts, D. Sacks, M. S. Schwartzberg, C. A. Lewis, and S. o. I. R. S. o. P. Committee, 2003: Quality improvement guidelines for percutaneous nephrostomy. *Journal of vascular and interventional radiology*, **14 (9)**, S277–S281.

Recht, B., R. Roelofs, L. Schmidt, and V. Shankar, 2019: Do imagenet classifiers generalize to imagenet? *International conference on machine learning*, PMLR, 5389–5400.

Reddy, S., 2023: Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, **41**, 101 304, https://doi.org/https://doi.org/10.1016/j.imu.2023.101304, URL https://www.sciencedirect.com/science/article/pii/S2352914823001508.

Reiter, E., 2018: A structured review of the validity of bleu. *Computational Linguistics*, **44 (3)**, 393–401, https://doi.org/10.1162/coli_a_00322, URL https://doi.org/10.1162/coli_a_00322.

Reiter, E., and A. Belz, 2009: An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, **35 (4)**, 529–558, https://doi.org/10.1162/coli.2009.35.4.35405, URL https://doi.org/10.1162/coli.2009.35.4.35405.

Richardson, J., and G. J. Groen, 2005: Applied epidural anatomy. *Continuing Education in Anaesthesia Critical Care & Pain*, **5 (3)**, 98–100, https://doi.org/10.1093/bjaceaccp/mki026, URL https://doi.org/10.1093/bjaceaccp/mki026.

Roberts, M., and Coauthors, 2021: Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, **3 (3)**, 199–217, https://doi.org/10.1038/s42256-021-00307-0, URL https://doi.org/10.1038/s42256-021-00307-0.

Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1 (5)**, 206–215, https://doi.org/10.1038/s42256-019-0048-x, URL https://doi.org/10.1038/s42256-019-0048-x.

Sallam, M., 2023: ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, **11 (6)**, 887, https://doi.org/10.3390/healthcare11060887, URL https://www.mdpi.com/2227-9032/11/6/887.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 2018: Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Scao, T. L., and Coauthors, 2022: Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schick, T., and H. Schütze, 2021: Generating datasets with pretrained language models. Association for Computational Linguistics, 6943–6951, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, https://doi.org/10.18653/v1/2021.emnlp-main.555, URL https://doi.org/10.18653/v1/2021.emnlp-main.555.

Schmälzle, R., and S. Wilcox, 2022: Harnessing artificial intelligence for health message generation: The folic acid message engine. *Journal of Medical Internet Research*, **24 (1)**, e28 858.

Scott, D., 1997: Identification of the epidural space: Loss of resistance to air or saline? *Regional Anesthesia: The Journal of Neural Blockade in Obstetrics, Surgery & Pain Control*, **22 (1)**, 1–2, https://doi.org/10.1136/rapm-00115550-199722010-00001, URL https://rapm.bmj.com/content/rapm/22/1/1.full.pdf.

Scott, D., and J. Moore, 2007: An NLG evaluation competition? Eight reasons to be cautious.

Scott-Sheldon, L. A. J., R. Lantini, E. G. Jennings, H. Thind, R. K. Rosen, E. Salmoirago-Blotcher, and B. C. Bock, 2016: Text Messaging-Based Interventions for Smoking Cessation: A Systematic Review and Meta-Analysis. *JMIR mHealth and uHealth*, **4 (2)**, e49, https://doi.org/10.2196/mhealth.5436, URL http://mhealth.jmir.org/2016/2/e49/.

Sedaghat, S., 2023: Early applications of chatgpt in medical practice, education and research. *Clinical Medicine*, **23 (3)**, 278–279.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017: Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sendak, M. P., J. D'Arcy, S. Kashyap, M. Gao, M. Nichols, K. Corey, W. Ratliff, and S. Balu, 2020: A path for translation of machine learning products into healthcare delivery. *EMJ Innov*, **10**, 19–00 172.

Sequeira, M., and Coauthors, 2024: The systematic development of a mobile phone delivered text-messaging tobacco cessation intervention in india. *Nicotine & Tobacco Research*, **27 (9)**, 1616–1625, https://doi.org/10.1093/ntr/ntae306, URL https://doi.org/10.1093/ntr/ntae306.

Sergeev, A., and M. Del Balso, 2018: Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint arXiv:1802.05799*.

Sezgin, E., and I. McKay, 2024: Behavioral health and generative ai: a perspective on future of therapies and patient care. *npj Mental Health Research*, **3 (1)**, 25, https://doi.org/10.1038/s44184-024-00067-w, URL https://doi.org/10.1038/s44184-024-00067-w.

Sheffer, C. E., and Coauthors, 2016: Increasing the quality and availability of evidence-based treatment for tobacco dependence through unified certification of tobacco treatment specialists. *Journal of Smoking Cessation*, **11 (4)**, 229–235.

Shen, L., Y. Sun, Z. Yu, L. Ding, X. Tian, and D. Tao, 2024: On efficient training of large-scale deep learning models. *ACM Computing Surveys*, **57 (3)**, 1–36.

Shin, H.-C., and Coauthors, 2016: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, **35 (5)**, 1285–1298.

Shortliffe, E. H., and B. G. Buchanan, 1975: A model of inexact reasoning in medicine. *Mathematical biosciences*, **23 (3-4)**, 351–379.

Singhal, K., and Coauthors, 2023: Large language models encode clinical knowledge. *Nature*, **620 (7972)**, 172–180, https://doi.org/10.1038/s41586-023-06291-2, URL https://doi.org/10.1038/s41586-023-06291-2.

Spohr, S. A., R. Nandy, D. Gandhiraj, A. Vemulapalli, S. Anne, and S. T. Walters, 2015: Efficacy of SMS Text Message Interventions for Smoking Cessation: A Meta-Analysis. *Journal of Substance Abuse Treatment*, **56**, 1–10, https://doi.org/10.1016/j.jsat.2015.01.011, URL https://linkinghub.elsevier.com/retrieve/pii/S0740547215000264.

Stanley, E. A. M., R. Souza, A. J. Winder, V. Gulve, K. Amador, M. Wilms, and N. D. Forkert, 2024: Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. *Journal of the American Medical Informatics Association*, https://doi.org/10.1093/jamia/ocae165, URL https://doi.org/10.1093/jamia/ocae165.

Stephens, K. K., and S. A. Rains, 2011: Information and Communication Technology Sequences and Message Repetition in Interpersonal Interaction. *Communication Research*, **38 (1)**, 101–122, https://doi.org/10.1177/0093650210362679, URL http://journals.sagepub.com/doi/10.1177/0093650210362679.

Stojanovic, M. P., T.-N. Vu, O. Caneris, J. Slezak, S. P. Cohen, and C. N. Sang, 2002: The role of fluoroscopy in cervical epidural steroid injections: An analysis of contrast dispersal patterns. *Spine*, **27 (5)**, 509–514, URL https://journals.lww.com/spinejournal/fulltext/2002/03010/the_role_of_fluoroscopy_in_cervical_epidural.11.aspx.

Svircevic, V., D. van Dijk, A. P. Nierich, M. P. Passier, C. J. Kalkman, G. Van Der Heijden, and L. Bax, 2011: Meta-analysis of thoracic epidural anesthesia versus general anesthesia for cardiac surgery. *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]*.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, 2016: Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tampu, I. E., A. Eklund, and N. Haj-Hosseini, 2022: Inflation of test accuracy due to data leakage in deep learning-based classification of oct images. *Scientific Data*, **9 (1)**, 580, https://doi.org/10.1038/s41597-022-01618-6, URL https://doi.org/10.1038/s41597-022-01618-6.

Tang, Q., C.-P. Liang, K. Wu, A. Sandler, and Y. Chen, 2014: Real-time epidural anesthesia guidance using optical coherence tomography needle probe. *CLEO: Applications and Technology*, Optical Society of America, AM2O. 3.

Tang, Q., and Coauthors, 2016: Depth-resolved imaging of colon tumor using optical coherence tomography and fluorescence laminar optical tomography. *Biomedical Optics Express*, **7 (12)**, 5218–5232, https://doi.org/10.1364/BOE.7.005218, URL https://opg.optica.org/boe/abstract.cfm?URI=boe-7-12-5218.

Tausczik, Y. R., and J. W. Pennebaker, 2010: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, **29 (1)**, 24–54, https://doi.org/10.1177/0261927X09351676, URL http://journals.sagepub.com/doi/10.1177/0261927X09351676.

Tejani, A. S., Y. S. Ng, Y. Xi, and J. C. Rayan, 2024: Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics*, **44 (5)**, e230 067,

https://doi.org/10.1148/rg.230067, URL https://pubs.rsna.org/doi/abs/10.1148/rg.230067.

Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, 2023: Large language models in medicine. *Nature medicine*, **29 (8)**, 1930–1940.

Toma, C. L., and J. D. D'Angelo, 2015: Tell-tale words:linguistic cues used to infer the expertise of online medical advice. *Journal of Language and Social Psychology*, **34 (1)**, 25–45, https://doi.org/10.1177/0261927x14554484, URL https://journals.sagepub.com/doi/abs/10.1177/0261927X14554484.

Topol, E. J., 2019: High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, **25 (1)**, 44–56, https://doi.org/10.1038/s41591-018-0300-7, URL https://doi.org/10.1038/s41591-018-0300-7.

Tu, T., and Coauthors, 2024: Towards generalist biomedical ai. *NEJM AI*, **1 (3)**, AIoa2300 138, https://doi.org/doi:10.1056/AIoa2300138, URL https://ai.nejm.org/doi/full/10.1056/AIoa2300138.

Varoquaux, G., 2018: Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, **180**, 68–77.

Varoquaux, G., and V. Cheplygina, 2022: Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, **5 (1)**, 48, https://doi.org/10.1038/s41746-022-00592-y, URL https://doi.org/10.1038/s41746-022-00592-y.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, 2017: Attention is all you need. *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., Vol. 30, URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vrudhula, A., A. C. Kwan, D. Ouyang, and S. Cheng, 2024: Machine learning and bias in medical imaging: Opportunities and challenges. *Circulation: Cardiovascular Imaging*, **17 (2)**, e015 495, https://doi.org/doi:10.1161/CIRCIMAGING.123.015495, URL https://www.ahajournals.org/doi/abs/10.1161/CIRCIMAGING.123.015495.

Wachter, S., B. Mittelstadt, and C. Russell, 2017: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, **31**, 841.

Wang, B., 2021: Mesh-transformer-jax: Model-parallel implementation of transformer language model with jax. URL https://github.com/kingoflolz/mesh-transformer-jax.

Wang, C., and Coauthors, 2021: Deep-learning-aided forward optical coherence tomography endoscope for percutaneous nephrostomy guidance. *Biomedical Optics Express*, **12 (4)**, 2404–2418, https://doi.org/10.1364/BOE.421299, URL http://opg.optica.org/boe/abstract.cfm?URI=boe-12-4-2404.

Wang, C., and Coauthors, 2022a: Computer-aided veress needle guidance using endoscopic optical coherence tomography and convolutional neural networks. *Journal of Biophotonics*, **15 (5)**, e202100 347.

Wang, C., and Coauthors, 2022b: Epidural anesthesia needle guidance by forward-view endoscopic optical coherence tomography and deep learning. *Scientific Reports*, **12 (1)**, 9057, https://doi.org/10.1038/s41598-022-12950-7, URL https://doi.org/10.1038/s41598-022-12950-7.

Wang, C., and Coauthors, 2024: Enhancing epidural needle guidance using a polarization-sensitive optical coherence tomography probe with convolutional neural networks. *Journal of biophotonics*, **17 (2)**, e202300 330.

Wang, H.-W., and Y. Chen, 2014: Clinical applications of optical coherence tomography in urology. *Intravital*, **3 (1)**, e28 770.

Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, 2017: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106.

Webson, A., and E. Pavlick, 2022: Do prompt-based models really understand the meaning of their prompts? Association for Computational Linguistics, 2300–2344, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, https://doi.org/10.18653/v1/2022.naacl-main.167, URL https://aclanthology.org/2022.naacl-main.167https://doi.org/10.18653/v1/2022.naacl-main.167.

White, J., and Coauthors, 2023: A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Whittaker, R., H. McRobbie, C. Bullen, A. Rodgers, and Y. Gu, 2016: Mobile phone-based interventions for smoking cessation. *Cochrane Database of Systematic Reviews*, https://doi.org/10.1002/14651858.CD006611.pub4, URL https://doi.wiley.com/10.1002/14651858.CD006611.pub4.

Wu, M., and A. F. Aji, 2023: Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Xie, S. M., A. Raghunathan, P. Liang, and T. Ma, 2021: An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Yagis, E., and Coauthors, 2021: Effect of data leakage in brain mri classification using 2d convolutional neural networks. *Scientific Reports*, **11 (1)**, 22 544, https://doi.org/10.1038/s41598-021-01681-w, URL https://doi.org/10.1038/s41598-021-01681-w.

Yang, M.-J., S. K. Sutton, L. M. Hernandez, S. R. Jones, D. W. Wetter, S. Kumar, and C. Vinci, 2023: A Just-In-Time Adaptive intervention (JITAI) for smoking cessation: Feasibility and acceptability findings. *Addictive Behaviors*, **136**, 107 467, https://doi.org/10.1016/j.addbeh.2022.107467, URL https://linkinghub.elsevier.com/retrieve/pii/S0306460322002337.

Yang, Z., Y. Yu, C. You, J. Steinhardt, and Y. Ma, 2020: Rethinking bias-variance trade-off for generalization of neural networks. *International Conference on Machine Learning*, PMLR, 10 767–10 777.

Ybarra, M. L., J. S. Holtrop, T. L. Prescott, and D. Strong, 2014: Process evaluation of a mHealth program: Lessons learned from Stop My Smoking USA, a text messaging-based smoking cessation program for young adults. *Patient Education and Counseling*, **97 (2)**, 239–243, https://doi.org/10.1016/j.pec.2014.07.009, URL https://linkinghub.elsevier.com/retrieve/pii/S0738399114002833.

Yeh, C.-K., B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, 2020: On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, **33**, 20 554–20 565.

Young, M., and S. W. Leslie, 2025: *Percutaneous Nephrostomy*. StatPearls Publishing, Treasure Island (FL), URL http://europepmc.org/abstract/MED/29630257http://europepmc.org/books/NBK493205https://www.ncbi.nlm.nih.gov/books/NBK493205.

Zech, J. R., M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, 2018: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, **15 (11)**, e1002 683, https://doi.org/10.1371/journal.pmed.1002683, URL https://doi.org/10.1371/journal.pmed.1002683.

Zhang, H., D. Duckworth, D. Ippolito, and A. Neelakantan, 2021: Trading off diversity and quality in natural language generation. Association for Computational Linguistics, 25–33, Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), URL https://aclanthology.org/2021.humeval-1.3.

Zhang, S., and Coauthors, 2022: Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh, 2021: Calibrate before use: Improving few-shot performance of language models. PMLR, URL https://proceedings.mlr.press/v139/zhao21c.html, 12697–12706 pp.

Zhou, Y., A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, 2022: Large language models are human-level prompt engineers. *The Eleventh International Conference on Learning Representations*.

Zhu, K., and Coauthors, 2023: Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Zhu, Q., and R. Zhang, 2019: A classification supervised auto-encoder based on predefined evenly-distributed class centroids. *arXiv preprint arXiv:1902.00220*.

Zhu, Y., S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu, 2018: Texygen: A benchmarking platform for text generation models. Association for Computing Machinery, URL https://doi.org/10.1145/3209978.3210080, 1097–1100 pp., https://doi.org/10.1145/3209978.3210080.