Determining the Effects of Domain-Specific Pretraining of Long-Context Transformer Encoder Models for Automated CPT Code Assignment in Cedars-Sinai Pathology Reports

Introduction

dAutomated CPT coding systems are needed

- Accurate CPT coding is vital to the financial health of both pathology departments and patients.

Emergent methods in NLP remain untested

- Using long-context encoders, such that pathology reports need not be truncated, is underexplored.
- Pretraining on additional reports has not been explored.

Access to large report corpora

- Cedars-Sinai has a large private corpus of unused pathology reports, while DHMC's DPLM corpus (Levy et al., 2022) remains useful for pretraining.

Objectives

Objective Develop a *reliable predictor* of primary CPT codes from pathology reports capable of reliably accelerating coding in real-world implementation.

Secondary objectives

- **Compare** the effectiveness of long-context transformer encoders to Naive Bayes, random forests, and XGBoost (eXtreme Gradient Boosting).
- **Investigate** transfer learning for CPT code prediction tasks by pretraining all models on the DPLM dataset before fine-tuning on Cedars.
- **Assess** our trained model's ability to use clinically relevant features for its predictions via SHAP explanations.

Corpora Analysis

Pathology report corpora contained 5 primary CPT: 88302, 88304, 88305, 83307, and 88309.

Corpus	Report Sections	Reports
Cedars-Sinai	46	174,045
DPLM	11	59,923

Table 1: Corpora sections and report counts.

Both corpora are **dominated by CPT 88305 cases**, with few low-complexity (88302) and high-complexity (88309) reports.

14.05% of Cedars-Sinai and 3.78% of DPLM reports exceed 512 tokens, highlighting the **need for long-context encoders** beyond BERT's 512-token limit.

Latent representations (Figure 2 reveals overlap between our pretraining (DPLM) and finetuning (Cedars-Sinai) datasets.

Methods

Transformers

Pretrained long-context encoders Clinical ModernBERT, BioClinical ModernBERT, Clinical Longformer, SciBERT Longformer, and Clinical-BigBird on DPLM corpus ⟶ finetuned to Cedars-Sinai corpus.

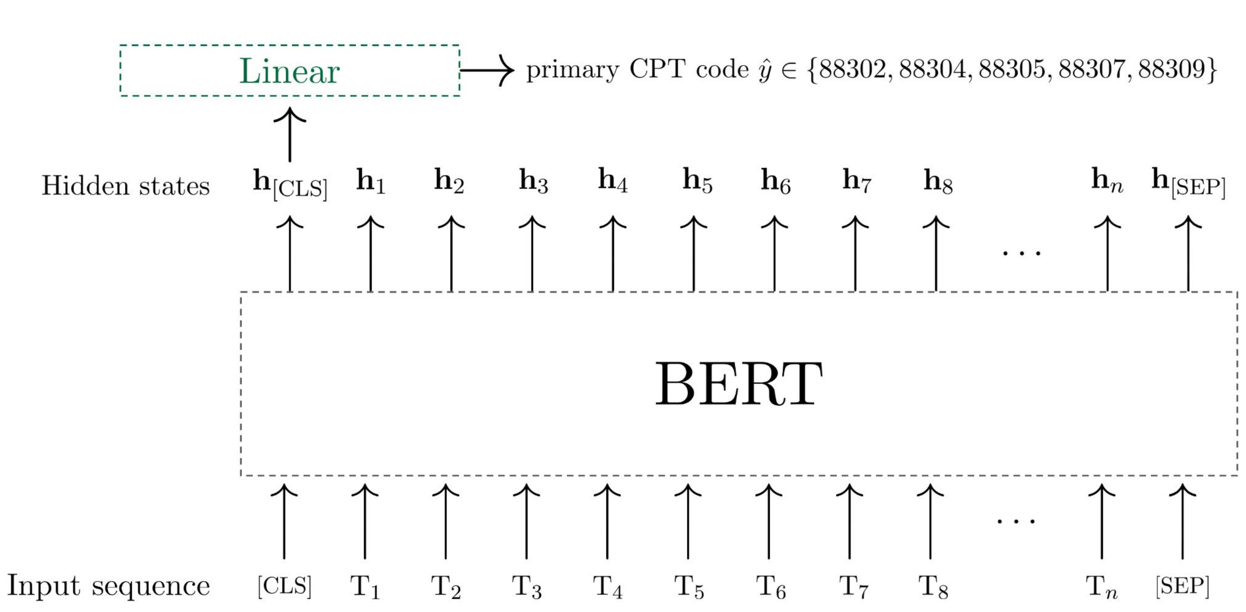

Figure 1: High-level BERT CPT code sequence classification pipeline.

Focal loss

We observed that focal loss (Lin et al., 2017) was beneficial to training our transformer encoder models, boosting F1 subtly. It was then used in all downstream experiments

$$\mathcal{L}_{focal} = -\alpha(1-p_t)^{\gamma}\log(p_t) \qquad \mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{focal}(y_i, \hat{y}_i)$$

Baseline approaches

We built Bag-of-Words and TF-IDF embeddings from the Cedars-Sinai corpus and applied Multinomial Naive Bayes, Random Forests, and XGBoost classifiers for comparison.

Results

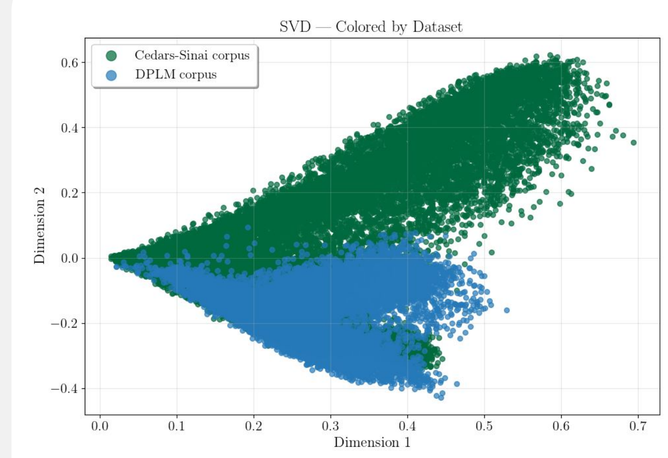

Figure 2: 2D SVD representations of all TF-IDF rows, broken down by billing code and by corpus.

Model	Accuracy	F1-Score	Precision	Recall	ROC-AUC	Cohen's κ	Inference Time (ms) ↓
SciBERT Longformer*	0.957483	**0.891239**	0.879124	**0.904974**	0.993567	**0.907560**	23.497
BioClinical ModernBERT*	0.957886	0.884051	0.877607	0.893444	0.995095	0.906741	151.503
Clinical ModernBERT*	0.957139	0.879889	0.909039	0.856634	0.987255	0.896817	29.462
Clinical-BigBird*	0.952772	0.869078	0.875185	0.867616	0.991923	0.887508	30.974
Clinical Longformer*	0.957828	0.880499	0.873080	0.888816	0.986401	0.899354	39.018
SciBERT Longformer	0.955817	0.877449	0.865417	0.895203	0.994384	0.898172	
BioClinical ModernBERT	0.957886	0.877089	0.853830	0.904636	**0.995568**	0.901980	
Clinical ModernBERT	0.955358	0.873371	0.864374	0.883508	0.991785	0.897605	
Clinical-BigBird	0.956794	0.882305	0.882253	0.882722	0.988368	0.900548	
Clinical Longformer	**0.959609**	0.879815	0.876555	0.883367	0.993308	0.903246	
BOW NB	0.808400	0.653900	0.614500	0.811300	0.954200	0.629900	0.006000
BOW RF	0.938600	0.831800	0.909800	0.772900	0.992200	0.847000	0.217000
BOW XGB	0.948200	0.855500	**0.917900**	0.805600	0.993300	0.875100	0.016000
TF-IDF NB	0.917300	0.700200	0.905200	0.628900	0.963400	0.780200	**0.004000**
TF-IDF RF	0.933600	0.818400	0.903300	0.755400	0.991100	0.833100	0.185000
TF-IDF XGB	0.946700	0.850200	0.915000	0.799000	0.993000	0.872700	0.016000

*Table 2: Performance of encoders and baseline approaches on the Cedars-Sinai test set with standard training vs. DPLM pretraining. Bold indicates best, * denotes pretraining.*

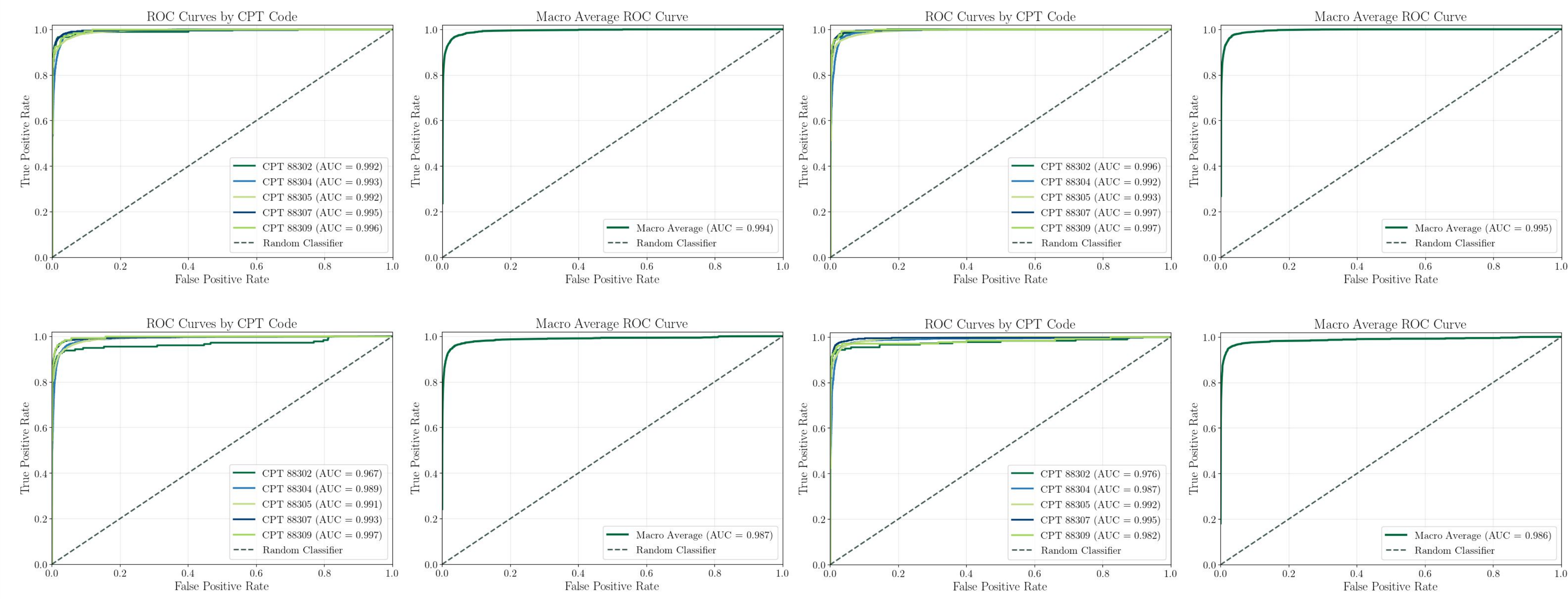

Figure 3: ROC curves calculated per-CPT code alongside macro-averaged ROC curves for DPLM pretrained models SciBERT Longformer (top left), BioClinical ModernBERT (top right), Clinical ModernBERT (bottom left), and Clinical Longformer (bottom right). DPLM-pretrained models SciBERT Longformer and BioClinical ModernBERT show notably higher ROC-AUC scores for CPT 88302 and 88309 in particular.

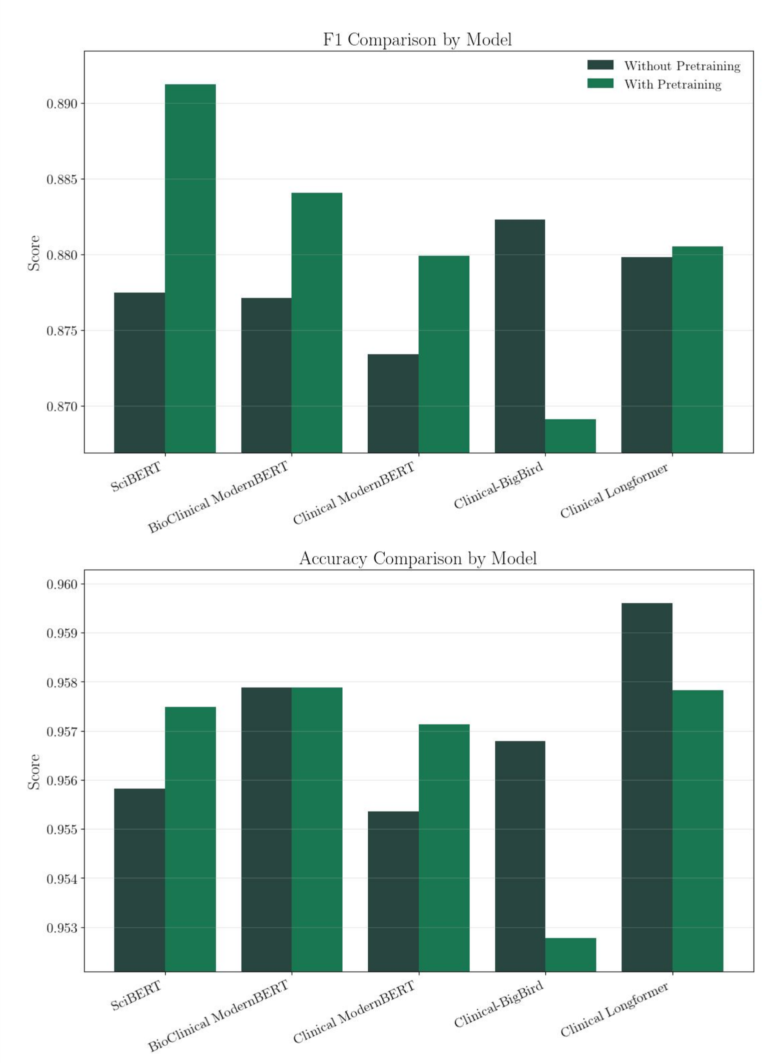

Figure 4: F1-score and accuracy comparison by transformer encoder model with and without domain-specific pretraining. Pretraining improved both metrics for most models.

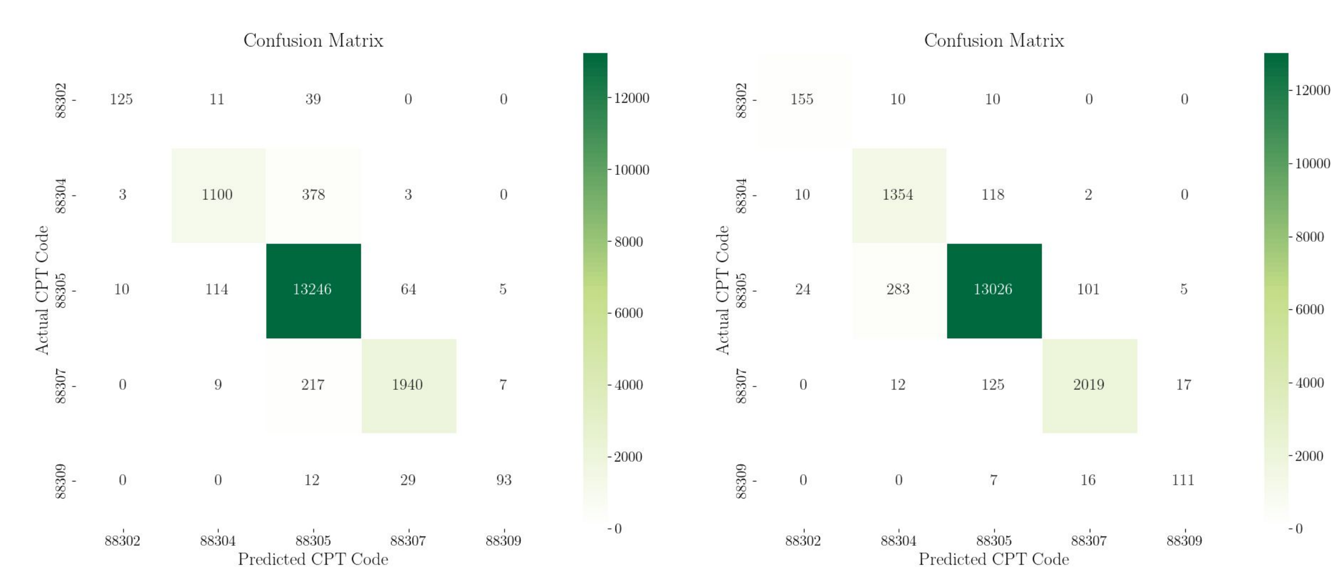

Figure 5: Confusion matrices on the Cedars-Sinai test set. XGBoost with BoW embeddings (left) struggles with rare codes such as 88302 and 88309, while SciBERT Longformer (right) demonstrates stronger performance across both common and rare codes.

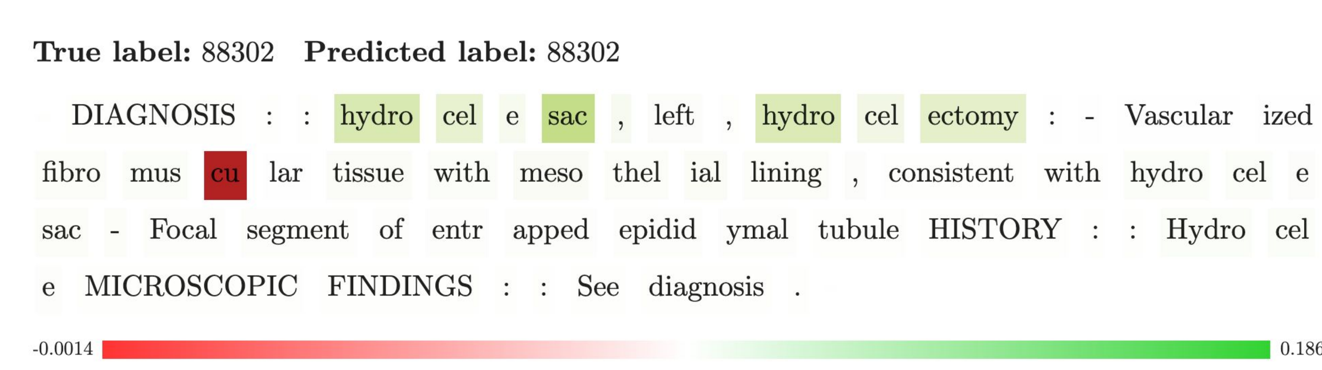

Figure 6: SHAP explanation for SciBERT Longformer's correct CPT 88302 prediction showing high importance tokens for "hydrocele sac," a keyword linked to CPT 88302.

Key Research Findings

- Best model (SciBERT Longformer + DPLM pretraining): 95.8% accuracy, 0.891 F1
- Outperformed baselines like XGBoost (BoW embeddings)
- Pretraining on a domain-specific corpus yielded improvements, especially for rare CPT codes
- SHAP analysis showed models highlight meaningful medical terms (e.g., "hydrocele sac")

Discussion & Conclusions

We **leveraged long-context transformer models** (e.g., SciBERT Longformer) to improve automated CPT code assignment from pathology reports. Unlike previous models limited to 512 tokens, our approach uses whole-report embeddings, **capturing subtle details** other approaches missed.

Implications

- More accurate CPT coding reduces financial and clinical errors
- Demonstrates the value of domain-specific pretraining, even with modest data sizes
- Learned representations may transfer to other coding tasks (e.g., ICD-10)

Limitations

- Relies on past human coding (may contain errors)
- Class imbalance (CPT 88305 dominant)
- Generalizability across institutions needs validation

We plan to extend our proposed framework to **multi-code prediction for complex cases, implement an uncertainty-based case review system for coders, and incorporate subspecialty metadata for further analysis.**

Acknowledgements

We thank Dr. Joshua Levy for mentorship and help orchestrating this research, Sean McOsker for technical advice and guidance, and all supporters of the EDIT AI Summer Research Internship for funding this research.

edit. Cedars Sinai Dartmouth Health