DON'T PAY ATTENTION, PLANT IT: PRETRAINING ATTENTION VIA LEARNING-TO-RANK

Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art Extreme Multi-Label Text Classification models rely on multi-label attention to focus on key tokens in input text, but learning good attention weights is challenging. We introduce PLANT—**P**retrained and **L**everaged **A**tte**NT**ion—a plug-and-play strategy for initializing attention. PLANT works by *planting* label-specific attention using a pretrained Learning-to-Rank model guided by mutual information gain. This architecture-agnostic approach integrates seamlessly with LLM backbones (e.g., we consider `Mistral-7B`, `LLaMA3-8B`, `DeepSeek-V3`, and `Phi-3`). PLANT outperforms SOTA methods across tasks like ICD coding, legal topic classification, and content recommendation. Gains are especially pronounced in few-shot settings, with substantial improvements on rare labels. Ablation studies confirm that attention initialization is a key driver of these gains. We make our code and trained models available.

1 INTRODUCTION

Extreme Multi-Label Text Classification (XMTC) entails assigning the most relevant subset of labels to a given instance from a (very) large label set. This setting emerges naturally in domains featuring vast, structured taxonomies such as e-commerce, legal categorization, and healthcare. In such settings, manual labeling is both costly and error-prone. For example, in clinical settings (Table 1), ICD coding—the task of assigning standardized codes for diagnoses and procedures based on clinical notes (Moons et al., 2020; WHO, 2025)—may be viewed as an instance of XMTC.

428.0: *Congestive heart failure*	202.8: *Other malignant lymphomas*	770.6: *Transitory tachypnea of newborn*
··· DIAGNOSES: **Acute congestive heart failure**, Diabetes mellitus, Pulmonary edema ···	··· 55 year-old female with **non Hodgkin's lymphoma** and C1 esterase inhibitor deficiency ···	··· Chest x-ray: **transient tachypnea of the newborn** with respiratory distress ···

Table 1: Examples of clinical text with ICD codes (Wang et al., 2024d; Zhang et al., 2025). Blue: code/label; red bold: disease mentions; teal: other relevant clinical findings.

Building XMTC models is challenging due to the high-dimensional label space and heavily skewed label distributions Bhatia et al. (2016). For example, in ICD coding there can be 170000 unique codes (CDC, 2024). Many are rare: In the MIMIC-III dataset Johnson et al. (2016) approximately 5411 out of 8929 codes appear <10 times. The task is further exacerbated by the often lengthy narratives in clinical texts. For example, in the MIMIC-III dataset, discharge summaries frequently contain detailed clinical histories comprising an average of 709.3 tokens, and often exceeding 1500 tokens (Johnson et al., 2016; 2023; Mullenbach et al., 2021; Nguyen et al., 2023). However, only a small fraction of these tokens are informative for assigning relevant ICD codes.

LLMs can be used zero-shot for XMTC tasks, but this poses challenges. For instance, prompts for such tasks tend to include long and flat label lists, resulting in *attention dilution*: The fixed attention budget is spread thin across thousands of tokens, weakening focus on rare tail labels (Peysakhovich & Lerer, 2023; Vandemoortele et al., 2025). This limitation is similarly evident in long-context retrieval tasks (Kamradt, 2023; Hsieh et al., 2024; Liu et al., 2024a), where LLMs struggle to locate relevant items. Task-specific fine-tuning may address such issues by embedding knowledge of the labels directly into model parameters during training, obviating the need for attention over long label

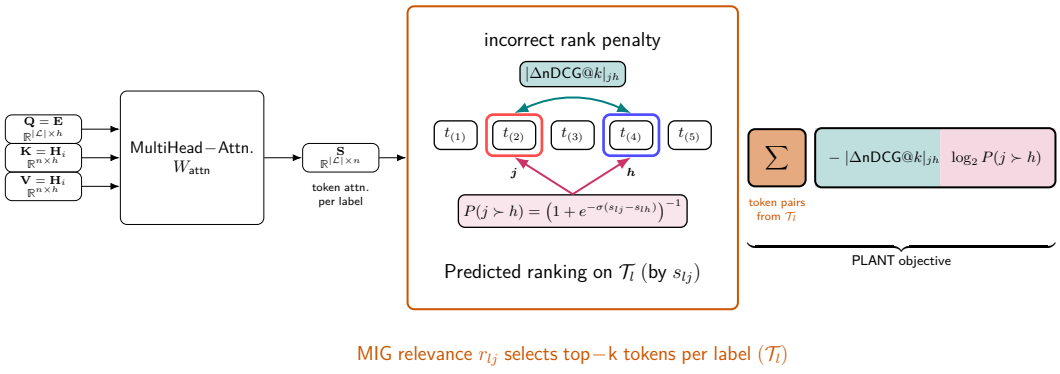


Figure 1: **PLANT** Attention. On the left, the MultiHead – Attention module (Vaswani et al., 2017), parameterized by $\boldsymbol{W}_{\text{attn}}$, takes as input queries $\mathbf{Q} = \mathbf{E}$ (label embeddings), keys $\mathbf{K} = \mathbf{H}_i$, and values $\mathbf{V} = \mathbf{H}_i$, and produces $\mathbf{S} \in \mathbb{R}^{|\mathcal{L}| \times n}$, representing the token-level attention distribution for each label. The **orange box** highlights the set of top-k tokens per label, $\mathcal{T}_l$, selected via Mutual Information Gain r_{lj} between labels and tokens. Within this set, two tokens j (**red**) and h (**blue**) are compared, with j being more relevant than h. The MultiHead – Attention module is trained to maximize the probability of correctly ranking tokens j and h ($P(j \succ h)$), while penalizing incorrect rankings in proportion to their impact on the nDCG@k metric if j and h were swapped ($|\Delta\text{nDCG@}k|_{jh}$). Finally, the **summation box** aggregates over all token pairs in $\mathcal{T}_l$, yielding the PLANT objective—(**nDCG term** $\times$ **probability term**)—that is optimized to initialize $\boldsymbol{W}_{\text{attn}}$.

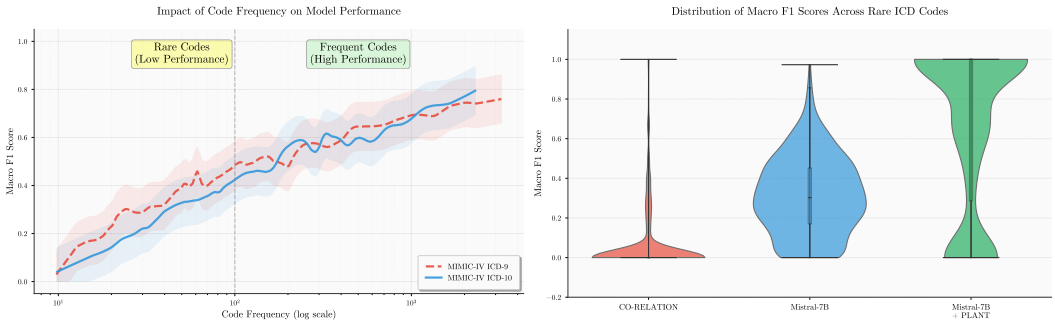
lists in the prompt and thus mitigating attention dilution (Yang et al., 2023a; Boukhers et al., 2024; Zhang et al., 2025; Barreiros et al., 2025).

In current approaches to XMTC, *attention mechanisms* Bahdanau et al. (2014) help address the challenges of high-dimensional, skewed label spaces. Existing XMTC models (Lu et al., 2023; Li et al., 2023; Nguyen et al., 2023; Yang et al., 2023b; Chen et al., 2023a; Zhang & Wang, 2024; Luo et al., 2024) almost always include a multi-label attention layer that allocates per-label attention weights to the input tokens (Wang et al., 2023a; Xiong et al., 2023; Yuan et al., 2024; Liu et al., 2025b). Intuitively, this is akin to a dedicated "spotlight" for each label: in high-dimensional spaces, it avoids the inefficiency of a single global focus by creating tailored text representations that highlight most relevant tokens per label. For skewed distributions, this ensures subtle cues for tail labels are not overshadowed by head labels, enabling better prediction of sparse classes.

Regardless of the specific encoder architecture, removing this attention layer significantly harms performance. A recent study by Xiong et al. (2023) highlights the importance of label-specific attention for *product-to-tag matching* by showing that removing this component leads to a sharp drop in P@1 (-15.69 points). Elsewhere, results on *scientific paper classification* show that stacking attention layers further boosts performance: Micro-F1 improves by a few points, showing that deeper attention enhances the model's capacity to represent label-specific features Liu et al. (2025b).

The premise of this work is that we can be smarter about how we initialize attention module weights. SOTA XMTC models begin with random label attention weights, requiring ranking all tokens for each label from scratch. This is data-intensive due to the high-dimensional label space. Skewed label distributions exacerbate this issue, as rare labels require even more data. Insufficient data, however, causes models to require more training epochs, often leading to overfitting rather than meaningful generalization—ultimately hurting rare label performance. Studies like Edin et al. (2023) show that SOTA models struggle to predict rare ICD diagnosis codes (Figure 2a). Models perform similarly across codes with comparable frequencies, indicating that the high proportion of rare codes impacts performance. Correlations between code frequency and F1 score are moderately high, showing that rare codes are predicted less accurately than common ones. This underscores the need for efficient attention mechanisms, as starting with random weights may be suboptimal.

Building on evidence that label-specific attention is pivotal in XMTC—its removal leads to sharp performance drops—we argue that how this attention is initialized is also crucial. By PLANT ing attention (i.e., seeding it with mutual-information signals and pretrained Learningto-Rank activations), we inject label-specific priors that boosts rare label performance, similar to how fine-tuning pretrained models outperforms training from scratch.



(a) (b)

Figure 2: (a) Rare codes have near-zero macro-F1. (b) Macro-F1 distribution on `MIMIC-III-few` for rare codes across CO-RELATION (Luo et al., 2024) (mean=0.054), `Mistral-7B` (0.309), and `Mistral-7B` +PLANT (0.663). `Mistral-7B` +PLANT yields far more rare codes with higher F1. See Section 3 (RQ4).

Our **main contributions** are as follows: (1) We introduce PLANT (**P**retrained and **L**everaged **A**tte**NT**ion), a plug-and-play strategy for initializing attention. PLANT replaces random initialization with relevance-guided attention weights via a two-stage framework: Stage 1 pretrains the attention layer as a *Learning-to-Rank* (L2R) module using *mutual information*; Stage 2 leverages these weights to train the full model end-to-end, improving rare-label performance. PLANT is architecture-agnostic and can be seamlessly integrated with LLM backbones — such as `Mistral-7B`, `LLaMA3-8B`, `DeepSeek-V3`, or `Phi-3`— without any modification; (2) In extensive experiments across ICD coding, legal topic classification, and content recommendation, we report consistent gains using PLANT across backbones and datasets, and we analyze through careful ablations which aspects of PLANT are responsible for these.

2 PLANT

In Extreme Multilabel Classification (XMTC) tasks, the goal is to assign to an input text multiple relevant labels from a very large label set. Formally, denote the dataset by $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid \boldsymbol{y}_i \in \{0,1\}^{|\mathcal{L}|}, i = 1, \ldots, N\}$, where: $\boldsymbol{x}_i$ is an input instance (e.g., a text document), and $\boldsymbol{y}_i$ is a binary vector indicating the presence ($y_{il} = 1$) or absence ($y_{il} = 0$) of each label $l \in \mathcal{L}$, where $\mathcal{L}$ denotes the label set (which may contain tens of thousands of unique labels). The objective is to learn a prediction function $f_\theta : \boldsymbol{x}_i \mapsto \mathbb{R}^{|\mathcal{L}|}$ parameterized by θ that outputs labels for each input $\boldsymbol{x}_i$. For each label $l \in \mathcal{L}$, the output $f_\theta(\boldsymbol{x}_i)_l \in \mathbb{R}$ is the score assigned for the l-th label.

Model Architecture. We start with a pretrained transformer-based LLM $\mathcal{M}_{\text{base}}$, selected from widely used models such as `Mistral-7B`, `LLaMA3-8B`, `DeepSeek-V3`, and `Phi-3`, known for strong general (Team et al., 2023; Grattafiori et al., 2024; Jiang et al., 2024; Abdin et al., 2024) and domain-specific performance in ICD coding (Yang et al., 2022a; 2023c; Falis et al., 2024; Madan et al., 2024; Nerella et al., 2024; Asensio Blasco et al., 2025; He et al., 2025; Liu et al., 2025a; Yuan et al., 2025). Due to computational constraints, we use both Low-Rank Adaptation (LoRA) and 4-bit quantization in all experiments (Frantar et al., 2022; Hu et al., 2022; Dettmers et al., 2023; Liu et al., 2024b; Aidouni, 2024). We adapt $\mathcal{M}_{\text{base}}$ into $\mathcal{M}_{\text{adapt}}$; see Appendix A for details.

Given an input $\boldsymbol{x}_i$, we tokenize it into $\boldsymbol{t}_i$ and pass it through the adapted model $\mathcal{M}_{\text{adapt}}$ to obtain hidden states: $\boldsymbol{H}_i = \mathcal{M}_{\text{adapt}}(\boldsymbol{t}_i) \in \mathbb{R}^{n \times h}$, where n is the sequence length and h the hidden size. We extract the final token's representation $\boldsymbol{h}_n \in \mathbb{R}^h$ for label prediction.

Specifically, we define trainable label embeddings $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{L}| \times h}$, one per label. These embeddings serve as query vectors in a multi-head attention module. The module, denoted as MultiHead, defines learnable parameters $\boldsymbol{W}_{\text{attn}}$:

$$\boldsymbol{Q} = \boldsymbol{E},\ \boldsymbol{K}\,\&\,\boldsymbol{V} = \boldsymbol{H}_i, \quad \boldsymbol{A}, \boldsymbol{S} = \text{MultiHead}^1(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}; \boldsymbol{W}_{\text{attn}}), \quad \boldsymbol{A} \in \mathbb{R}^{|\mathcal{L}| \times h},\ \boldsymbol{S} \in \mathbb{R}^{|\mathcal{L}| \times n}, \tag{1}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ represent the query, key, and value inputs, respectively. During training, the parameters $\boldsymbol{W}_{\text{attn}}$ are optimized to learn label-specific attention weights $\boldsymbol{S}$, which determine how each label query attends to tokens in the sequence, and the output $\boldsymbol{A}$, which represents the attended representations for each label query.

The resulting attention output is boosted by learnable matrices $\mathbf{B}_a, \mathbf{B}_m \in \mathbb{R}^{|\mathcal{L}|\times h}$ as: $\mathbf{A}_{\text{boost}} = \mathbf{B}_a + \mathbf{A} \cdot \mathbf{B}_m$, where $\cdot$ denotes element-wise multiplication. This "boosts" label-specific signals in a learned, differentiable manner. This is motivated by the need to enhance task-specific signals in recent Mixture-of-Experts (MoE) frameworks Cai et al. (2024); Yu et al. (2024); Chen et al. (2023b). An adaptive average pooling layer reduces the dimensionality of $\boldsymbol{A}_{\text{boost}}$ to: $\boldsymbol{P}_i = \mathsf{Pool}(\boldsymbol{A}_{\text{boost}}) \in \mathbb{R}^{|\mathcal{L}|\times p}$. [1] A shared linear projection $\boldsymbol{W}_c \in \mathbb{R}^{p\times 1}$ then computes the final logits as $\hat{\boldsymbol{y}}_i = \boldsymbol{P}_i \boldsymbol{W}_c \in \mathbb{R}^{|\mathcal{L}|}$, each entry in y_i is the predicted (raw) relevance score for the corresponding label.

Two-Stage Training

PLANT entails a two stage optimization strategy. The first focuses on pretraining MultiHead (Equation 1) label-wise attention weights; the second entails fine-tuning the model end-to-end.

Stage 1: Pretraining Attention as L2R (Figure 1) In Stage 1 we train the multi-head attention module parameters ($\mathcal{M}_{\text{adapt}}$, $\mathbf{E}$ and MultiHead). The MultiHead module outputs label-specific attention scores $\boldsymbol{S} = [s_{lj}] \in \mathbb{R}^{|\mathcal{L}|\times n}$, where s_{lj} is the attention score for token j with respect to label l.

We train this following a learning-to-ranking objective focused on the top-k tokens per label selected by **Mutual Information Gain** (MIG) computed from the training set.[2]

$$\mathcal{L}^{(i)}_{\text{rank}}(\boldsymbol{S}) = -\sum_{l=1}^{|\mathcal{L}|} \sum_{\substack{j,h\in\mathcal{T}_l \\ r_{lj}>r_{lh}}} |\Delta\text{nDCG@k}|_{jh} \times \log_2 \left(1 + e^{-\sigma(s_{lj}-s_{lh})}\right)^{-1} \tag{2}$$

wt. for pairwise swap impact

MIG relevance

(differentiable) approx ranking prob.

where $\mathbf{r}_l = [r_{lj}] \in \mathbb{R}^n$ represents the ground-truth relevance of tokens for label l, derived from the MIG between tokens and labels; $\mathcal{T}_l$ is the set of top-k tokens for label l, selected based on r_{lj}; $|\Delta\text{nDCG@k}|_{jh}$ is the change in nDCG@k after swapping j and h in the predicted ranking, where the predicted ranking is determined by sorting tokens in $\mathcal{T}_l$ by their predicted scores s_{lj} in descending order; and σ is the sigmoid scaling factor. We test $k = \{500, 1000, 2000\}$ to evaluate sensitivity.

PLANTed Attention via MIG Ranking MIG scores, denoted by r_{lj}, quantifies how informative token t_j is for predicting label l. Higher r_{lj} indicates stronger relevance of the token to the label. Note that we empirically determined these relevance scores by computing MIG between token occurrences and label assignments across the training corpus (see Appendix B.1). The ranking loss in Equation 2 encourages the MultiHead module to assign higher attention scores (s_{lj}) to tokens with greater relevance. It considers token pairs (j, h) where token j is more relevant than token h for a given label l, i.e., $r_{lj} > r_{lh}$. The term $\left(1 + e^{-\sigma(s_{lj}-s_{lh})}\right)^{-1}$ approximates the probability that token j is ranked above token h. Each pair is weighted by $|\Delta\text{nDCG@k}|_{jh}$, which penalizes incorrect rankings in proportion to their impact on the nDCG@k metric. This loss formulation encourages attention scores to align with the MIG.

Stage 2: Leveraging Attention – Full Training In Stage 2 we train the entire model (Section 2) end-to-end. We start with the finetuned $\mathcal{M}_{\text{adapt}}$ and the initialized weights $\mathbf{W}_{\text{attn}}$ and $\mathbf{E}$ from Stage 1. We optimize the model under focal loss with label smoothing and hard negative mining to address label imbalance (Ben-Baruch et al., 2020; Xiong et al., 2023). The detailed formulation of the focal loss is provided in Appendix A (Eq. 3).

[1] See Appendix A for the number of attention heads in Equation 1 and output size p used in pooling.

[2] See Appendix B.1 for details on pre-computing MIG for a corpus.

To address the challenge of imbalanced labels, where negative labels often dominate, hard negative mining is applied to focus the loss on the most informative examples. This selects all positive labels ($y_{il} = 1$) and the top-m negative labels ($y_{il} = 0$) with the highest predicted probabilities $\sigma(\hat{y}_{il})$, where $m = 1000$. The focal loss is then computed over this selected subset $\mathcal{S}_i \subseteq \mathcal{L}$ by restricting the summation in Eq. 3 to $\mathcal{S}_i$. We refer to this as the *HNM-augmented focal loss*. This approach ensures the model prioritizes learning from difficult negative examples, improving performance on challenging cases. (as shown in Ablation Section 4)

3 EXPERIMENTS

Datasets, Baselines & Implementation: We evaluated PLANT against SOTA models on the `MIMIC-IV/III-full` datasets, which comprise discharge summaries annotated with ICD-9 and ICD-10 codes, respectively. For few-shot learning, we used `MIMIC-III-rare50` and `MIMIC-III-few` subsets to focus on rare codes. To assess generalizability, we also evaluated PLANT on publicly available legal topic classification (`EURLEX-4K`, over long legal documents) and content recommendation (`WIKI10-31K`, tag prediction for Wikipedia-style texts). Dataset descriptions, implementation, baselines & evaluation metrics are in App. C, D and E, respectively.

Notation. ▲/▼ mark significant gains/drops (α=0.05, Wilcoxon Demšar 2006; see App. F), shown only if the test passes and the 95% CI excludes 0. Gains/drops are followed by the CI in plum. **Bold** = best per metric.

RQ1: How effective is PLANT 's two-stage across LLM backbones? Table 2 compares the performance of four LLM backbones (`Mistral-7B`, `LLaMA3-8B`, `DeepSeek-V3`, `Phi-3`)[3]trained end-to-end with *HNM-augmented focal loss* versus PLANT 's two-stage training (Section 2) which, in stage 2, adopts the same *HNM-augmented focal loss*, on `MIMIC-III-full` and `MIMIC-IV-full`. We report the *average absolute gains* obtained by computing the mean difference between each LLM and its PLANT-enhanced counterpart for a given metric. Table 2 highlights the gains from integrating PLANT: green rows show PLANT-enhanced results, with consistent improvements across all metrics and average gains summarized in the last row.

Notably, much smaller models integrated with PLANT outperform significantly larger LLMs used alone. For instance, on `MIMIC-III-full`, `LLaMA3-8B` +PLANT (8B) outperforms `DeepSeek-V3` (336B) by **+1.8** in F1 (Macro) and **+3.0** in P@15. Similarly, `Phi-3` +PLANT (3.8B) surpasses `DeepSeek-V3` by **+2.0** in F1 (Micro) and a substantial **+7.1** in AUC (Macro). This trend persists across `MIMIC-IV-full` as well: `LLaMA3-8B` +PLANT achieves gains of **+3.7** in F1 (Macro) and **+3.5** in P@15 over `DeepSeek-V3`. These results highlight the efficiency of PLANT, which permits smaller models to surpass much larger LLMs across key metrics.

RQ2: Does PLANT[4] outperform SOTA models on ICD-10 code classification? Table 3 compares PLANT with SOTA models on `MIMIC-IV-full`. Performance comparison across `MIMIC-III-full` and `MIMIC-III-top50` is provided in Table 10 in Appendix G. On `MIMIC-IV-full`, which exhibits a more skewed label distribution (see Table 7), PLANT demonstrates average gains of **+0.2–1.4**, including a **+0.7** (95% CI: **0.5–1.0**) gain in F1 (Macro) and a **+1.4** (95% CI: **1.0–1.9**) improvement in Precision@8 over SOTA baselines. PLANT 's larger performance gains on `MIMIC-III-full` and `MIMIC-IV-full` for the macro-averaged metrics highlight its effectiveness in addressing label imbalance.

RQ3: How effective is PLANT on rare labels? Table 4 evaluates PLANT against SOTA models on `MIMIC-III-few` (labels appearing in fewer than 5 samples) and `MIMIC-III-rare50` (50 most rare labels) subsets of `MIMIC-III-full`. PLANT significantly outperforms all baselines. On `MIMIC-III-few`, PLANT achieves substantial aggregate gains of **+30–49** across F1, Precision, and Recall, including a **+36.1** (95% CI: **30.5–41.7**) gain in F1 (Macro) and a **+48.1** (95% CI: **42.6–54.0**) gain in Recall (Macro). For `MIMIC-III-rare50`, PLANT demonstrates even larger improvements, with average gains of **+9–49** across metrics, notably a **+48.6** (95% CI: **41.2–56.4**) gain in F1 (Macro).

[3]See Appendix A for LLM details.

[4]In this setting the base LLM $\mathcal{M}_{\text{base}}$ for PLANT was `Mistral-7B`.

Model	AUC		F1		P@15
	Macro	Micro	Macro	Micro	
Mistral-7B	90.2	98.7	20.0	57.0	53.8
Mistral-7B + PLANT	97.4▲ *(+7.2)*	99.5▲ *(+0.8)*	23.0▲ *(+3.0)*	59.2▲ *(+2.2)*	56.9▲ *(+3.1)*
LLaMA3-8B	90.5	98.8	20.5	57.5	54.0
LLaMA3-8B + PLANT	**97.6**▲ *(+7.1)*	**99.6**▲ *(+0.8)*	**23.5**▲ *(+3.0)*	**59.5**▲ *(+2.0)*	**57.0**▲ *(+3.0)*
DeepSeek-V3	90.0	98.6	19.8	56.8	53.5
DeepSeek-V3 + PLANT	97.2▲ *(+7.2)*	99.4▲ *(+0.8)*	22.8▲ *(+3.0)*	59.0▲ *(+2.2)*	56.5▲ *(+3.0)*
Phi-3	89.8	98.5	19.5	56.5	53.2
Phi-3 + PLANT	97.0▲ *(+7.2)*	99.3▲ *(+0.8)*	22.5▲ *(+3.0)*	58.8▲ *(+2.3)*	56.3▲ *(+3.1)*
Avg. gain with PLANT	Δ+7.2	Δ+0.8	Δ+3.0	Δ+2.2	Δ+3.1

Table 2: **PLANT consistently boosts all LLM backbones on MIMIC-IV-full.** The full table with both MIMIC-III-full and MIMIC-IV-full results is provided in Table 9 in Appendix G.

RQ4: Why PLANT Is Superior to Few-Shot SOTA Models? Figure 2a shows that codes with frequencies <10 have near-zero macro-F1 scores, highlighting the challenge of predicting *rare codes*—a problem PLANT aims to address. To evaluate this, we used the MIMIC-III-few dataset, which contains 685 codes, each appearing in <5 instances. Figure 2b focuses on these rare codes, effectively zooming in on the leftmost part of Figure 2a. We present violin plots (with embedded box plots) of macro-F1 distributions for rare codes across three models: CO-RELATION (Luo et al., 2024) (mean = 0.054), Mistral-7B (mean = 0.308), and Mistral-7B +PLANT (mean = **0.663**). Notably, **54.8%** of rare codes achieve macro-F1 > 0.7 with PLANT, compared to only 2.0% for the base Mistral-7B, and 0.6% for CO-RELATION. These results demonstrate that integrating PLANT with a base LLM not only surpasses specialized few-shot approaches but also markedly enhances the LLM's capacity to model rare labels.

RQ5: How generalizable is PLANT to other imbalanced classification tasks? Table 5 evaluates PLANT on two diverse tasks: legal topic classification (EURLEX-4K) and content recommendation (WIKI10-31K), both characterized by extreme label spaces and imbalanced distributions. On EURLEX-4K, PLANT achieves aggregate gains of **+0.9–2.5** across P@1, P@3, and P@5, including a a **+2.5** (95% CI: **1.7–3.5**) gain in P@3. For WIKI10-31K, PLANT shows a **+2.2** (95% CI: **1.6–2.8**) gain in P@3, though it exhibits a negligible dip in P@5.

Model	AUC		F1		Precision	
	Macro	Micro	Macro	Micro	P@8	P@15
CoRelation (Luo et al., 2024)	97.2	99.6	6.3	57.8	70.0	55.3
PLM-CA (Edin et al., 2024)	91.8	99.1	22.3	58.9	70.5	55.8
GKI-ICD (Zhang et al., 2025)	97.1	99.3	20.6	58.5	70.7	55.8
GPT-4 Zero-Shot (Yuan et al., 2025)	90.5	98.8	5.0	56.0	68.0	53.5
PLANT (Ours)	**97.4**	99.5	**23.0**	**59.2**	**72.1**	**56.9**
	▲*+0.2*	*-0.1*	▲*+0.7*	▲*+0.3*	▲*+1.4*	▲*+1.1*
	[0.08, 0.31]	*[-0.27, 0.06]*	*[0.48, 0.95]*	*[0.12, 0.44]*	*[1.01, 1.88]*	*[0.72, 1.55]*

Table 3: **PLANT sets a new SOTA on MIMIC-IV-full.**

4 ABLATION ANALYSIS

ABLATING COMPONENTS IN PLANT (TABLE 6) To assess the contribution of individual components in PLANT's two-stage training pipeline (Section 2), we perform ablations on two dataset/LLM combinations: MIMIC-III-full with Mistral-7B and MIMIC-IV-full with LLaMA3-8B, where Mistral-7B and LLaMA3-8B are the base LLMs described in Section 2. **Each ablation configuration is compared against the full PLANT setup** (bottom row of Table 6), intentionally isolating the effect of a component. We evaluate using (1) macro-AUC for per-label classification, (2) macro-F1 for rare-label accuracy, and (3) P@15 for top-k prediction quality. For each ablation, we report the average decrease relative to the full PLANT setup ($\nabla -x$).

Model	F1		Precision		Recall	
	Ma.	Mi.	Ma.	Mi.	Ma.	Mi.
MSMN + Contrastive (Lu et al., 2023)	4.3	8.5	4.5	**70.9**	4.2	4.5
GP (Yang et al., 2023b)	30.2	35.3	27.9	38.5	32.9	32.6
Tr-EHR (Yang et al., 2023c)	22.0	32.5	20.5	52.0	23.5	24.0
CoRelation (Luo et al., 2024)	25.0	34.0	23.5	50.5	26.5	27.0
PLM-CA (Edin et al., 2024)	26.5	35.0	24.5	51.5	28.0	28.5
GKI-ICD (Zhang et al., 2025)	24.0	33.5	22.5	49.0	25.5	26.0
PLANT (Ours)	**66.3**	**71.0**	**65.1**	68.6	**81.0**	**81.7**
	▲ *+36.1*	▲ *+35.7*	▲ *+37.2*	▼ *-2.3*	▲ *+48.1*	▲ *+49.1*
	[30.5, 41.7]	*[29.8, 41.2]*	*[31.0, 43.5]*	*[-3.7, -1.0]*	*[42.6, 54.0]*	*[43.3, 54.8]*

Table 4: **Performance on rare labels (`MIMIC-III-few`).** PLANT offers large gains on most metrics. The full table including `MIMIC-III-rare50` appears in Table 11 in Appendix G.

Model	Legal Topic Classification (`EURLEX-4K`)			Content Recommendation (`WIKI10-31K`)		
	P@1	P@3	P@5	P@1	P@3	P@5
ICXML (Zhu & Zamani, 2023)	87.10	74.32	62.70	87.61	79.11	69.78
XRR (Xiong et al., 2023)	87.96	78.88	68.52	89.54	85.38	81.34
X-Transformer w/ RDE (Shi et al., 2024)	84.60	72.61	61.35	86.15	76.99	68.75
MatchXML (Ye et al., 2024)	88.12	75.00	62.22	89.30	80.45	70.89
DE (Gupta et al.)	87.60	74.39	67.80	88.21	80.29	69.91
InceptionXML (Kharbanda et al., 2023) + GANDALF (Kharbanda et al., 2024)	86.98	75.89	68.78	88.76	80.32	69.89
CG (Chai et al., 2024)	87.82	76.71	68.42	87.29	79.81	68.45
PLANT (Ours)	**90.61**	**81.35**	**70.24**	**90.91**	**87.61**	81.33
	▲ *+2.20*	▲ *+2.47*	▲ *+0.90*	▲ *+1.37*	▲ *+2.23*	*-0.01*
	[1.47, 3.25]	*[1.73, 3.49]*	*[0.38, 1.55]*	*[0.76, 1.93]*	*[1.60, 2.84]*	*[-0.27, 0.19]*

Table 5: **PLANT performs strongly across domains**—legal topic classification & tag prediction.

(1)+(2) Benefit of Stage 1 attention initialization. End-to-end training from scratch without Stage 1 degrades performance: using BCE loss yields avg dips (macro-AUC/macro-F1/P@15) of ▽−6.2/▽−4.8/▽−5.3, while Focal+HNM ($\gamma = 2, \epsilon = 0.1, m = 1000$) reduces the dips to ▽−4.8/▽−3.4/▽−3.8. *Takeaway:* Random attention initialization and label imbalance severely harm rare-label accuracy and top-k retrieval; Focal+HNM helps, but pretrained attention still recovers substantial headroom.

(3) Removing focal loss and HNM in Stage 2. Training with vanilla BCE after Stage 1 ($\gamma = 0$, no label smoothing, no HNM), yields changes of (macro-AUC/macro-F1/P@15): ▽−2.5/▽−2.0/▽−2.5; *Takeaway:* While not competitive with the full PLANT, it still outperforms both single-stage BCE and single-stage Focal Loss, demonstrating *Stage 1 attention pretraining alone provides meaningful gains* even with simple BCE.

(4) Effect of label smoothing in Stage 2. Two-stage training without label smoothing (Stage 2 with ϵ=0) results in (macro-AUC/macro-F1/P@15): ▽−0.3/▽−1.1/▽−1.2; *Takeaway:* Mild but consistent loss—overconfidence slightly hurts macro-F1.

(5) Effect of hard negative mining (HNM) in Stage 2. Two-stage training but without HNM (in Stage 2 computing loss over all labels instead of top-m negatives.) changes performance (macro-AUC/macro-F1/P@15): ▽−0.6/▽−1.9/▽−2.8; *Takeaway:* Noticeable drops in both precision and rare-label accuracy—HNM focuses learning on informative negatives.

(6) Importance of MIG vs. naive frequency in Stage 1. Replacing MIG relevance scores with normalized token frequency per label yields (macro-AUC/macro-F1/P@15): ▽−1.0/▽−2.4/▽−3.3; *Takeaway:* Coarser relevance signals harms rare-label and top-k accuracy.

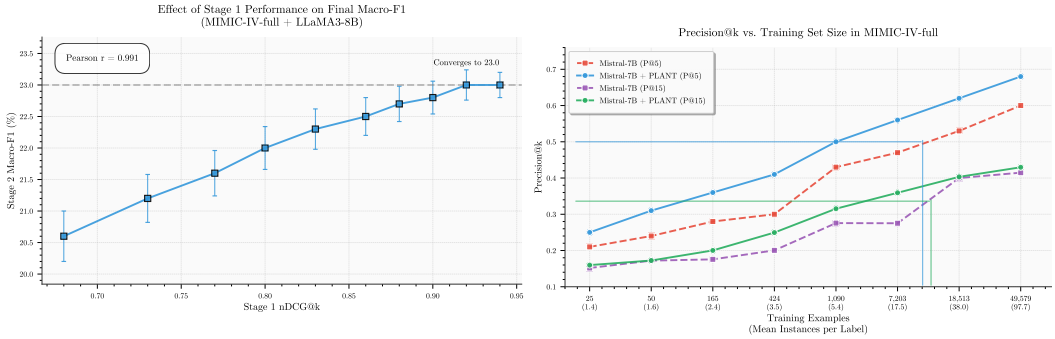


Figure 3: *(Left)* On `MIMIC-IV-full` with `LLaMA3-8B`, better Stage 1 nDCG@k (attn. init. quality) leads to higher Stage 2 (downstream) macro-F1. Extended results (`MIMIC-III-full` +`Mistral-7B`, `MIMIC-IV-full` +`LLaMA3-8B`; macro-F1, P@15) appear in App. G, Fig. 4. *(Right)* PLANT consistently boosts `Mistral-7B` on `MIMIC-IV-full` across training sizes: solid lines (`Mistral-7B` +PLANT) beat dashed baselines on P@5/P@15, with largest gains in low-data regimes. Paired `MIMIC-III-full` +`MIMIC-IV-full` results are in App. G, Fig. 5.

(7) Importance of ranking objective in Stage 1. Compared to full PLANT, replacing Stage 1 pairwise ranking loss (Eq. 2) with MSE, yields *avg dips* (macro-AUC/macro-F1/P@15): $\nabla -1.5$/$\nabla -2.9$/$\nabla -3.5$; *Takeaway:* The pairwise ranking loss is central to "planting" attention weights that reflect MIG-derived token relevance. Replacing it with MSE removes the ranking signal, leading to weaker alignment between learned attention scores and true relevance, which in turn degrades rare-label accuracy & top-k precision.

(8) Effect of attention initialization quality of Stage 1. We vary the number of Stage 1 epochs (1–10) before switching to Stage 2, and measure attention initialization quality at the end of Stage 1 via nDCG@k (computed against MIG relevance as ground truth), alongside final macro-F1 and P@15 after Stage 2 (Figure 3, *left*). Compared to the full PLANT setup (10 epochs), training for only 1 epoch yields *avg dips* (nDCG@k/macro-F1/P@15): $\nabla -0.24$/$\nabla -2.4$/$\nabla -2.6$. As Stage 1 training length increases, nDCG@k steadily improves (e.g., 0.68→0.94 on `MIMIC-IV-full`/`LLaMA3-8B`), and final metrics rise accordingly, saturating at the full PLANT performance. *Takeaway:* Higher nDCG@k at the end of Stage 1 correlates with better rare-label accuracy and top-k precision.

Key Takeaways (1–8): Illustrated in Table 6, Stage 1 attention pretraining is the single most impactful component of PLANT: removing it and training end-to-end from scratch with BCE or HNM-augmented focal loss yields average dips of $\nabla -(4.8 \text{ to } 6.2)$ in macro-AUC, $\nabla -(3.4 \text{ to } 4.8)$ in macro-F1, and $\nabla -(3.8 \text{ to } 5.3)$ in P@15. Analysis of attention initialization quality (Fig. 4) further shows that stronger token-ranking quality at the end of Stage 1 correlates with better downstream macro-F1 and P@15. Within Stage 1, both the MIG relevance signal and the ranking objective are essential for effectively "planting" attention weights: replacing MIG with token frequency causes average dips up to $\nabla -3.3$ (P@15), and replacing the ranking loss with MSE causes up to $\nabla -3.5$ (P@15). Moreover, when Stage 2 is trained with vanilla BCE, PLANT still achieves average gains of $\Delta +(1.4 \text{ to } 2.3)$ (macro-AUC/macro-F1/P@15) over end-to-end training with HNM-augmented focal loss, underscoring that *planted attention* alone contributes substantial improvements.

PLANT UNDER VARYING TRAINING SIZE Annotated data is scarce and costly, especially for rare labels. So we ask: *can PLANT's pretrained attention improve sample efficiency over standard end-to-end training by reducing the labeled examples needed for competitive performance?*

To evaluate this, we compare PLANT 's two-stage training (Section 2) with single-stage end-to-end training using the same architecture (Section 2) and base LLM `Mistral-7B`, on `MIMIC-III-full` and `MIMIC-IV-full` under varying training sizes (Figure 3, right). Both methods are trained on different fractions of balanced training splits, with fixed test sets, up to 5 epochs, and evaluated on P@5 and P@15. The sole difference is in the attention mechanism $\mathbf{A}$ from the MultiHead module (Equation 1): PLANT uses Stage 1 pretrained attention from the

[4] computed as normalized token occurrences in documents for each label

Ablation Config	macro-AUC	macro-F1	P@15
[1]Single-Stage BCE	91.0▼ *(-6.4, [-7.2, -5.4])*	18.0▼ *(-5.0, [-5.9, -3.8])*	52.0▼ *(-4.9, [-5.7, -3.9])*
[1]Single-Stage Focal Loss	92.5▼ *(-4.9, [-5.8, -4.0])*	19.5▼ *(-3.5, [-4.2, -2.7])*	53.5▼ *(-3.4, [-4.0, -2.7])*
[1]PLANT w/ Vanilla BCE	95.0▼ *(-2.4, [-3.0, -1.7])*	21.0▼ *(-2.0, [-2.6, -1.4])*	54.8▼ *(-2.1, [-2.7, -1.5])*
[2]PLANT w/o Label Smoothing	97.0▼ *(-0.4, [-0.7, -0.2])*	21.8▼ *(-1.2, [-1.8, -0.7])*	55.8▼ *(-1.1, [-1.7, -0.6])*
[2]PLANT w/o Hard Neg Mining	96.8▼ *(-0.6, [-1.0, -0.3])*	21.0▼ *(-2.0, [-2.7, -1.3])*	54.5▼ *(-2.4, [-3.1, -1.7])*
[1]PLANT w/ Term Frequency	96.5▼ *(-0.9, [-1.4, -0.5])*	20.5▼ *(-2.5, [-3.2, -1.8])*	54.0▼ *(-2.9, [-3.6, -2.1])*
[1]PLANT w/ MSE	96.0▼ *(-1.4, [-2.1, -0.8])*	20.0▼ *(-3.0, [-3.9, -2.1])*	53.8▼ *(-3.1, [-3.9, -2.3])*
PLANT (full setup)	**97.4**	**23.0**	**56.9**

Table 6: Ablation on `MIMIC-IV-full` with base `LLaMA3-8B`. Full results (`MIMIC-III-full` w/ `Mistral-7B`, `MIMIC-IV-full` w/ `LLaMA3-8B`) appear in Table 12, App. G.

L2R model (Equation 2), emphasizing MIG-ranked tokens, while the baseline learns attention from scratch. **Takeaway:** As shown in Figure 3 (right), PLANT consistently matches or exceeds end-to-end `Mistral-7B` across all training sizes, often with an order of magnitude fewer labels. On `MIMIC-IV-full`, PLANT achieves P@5=0.50 and P@15=0.37 with only **1090** and **2743** instances—matching baselines trained on 10,337 and 12,902. On `MIMIC-III-full`, it reaches P@5=0.47 and P@15=0.30 using just **136** and **235** instances—vs. 1342 and 1578 for the baseline.

5 Conclusion

This work proposed PLANT—a plug-and-play strategy for initializing attention. By pretraining attention as a L2R module with mutual information and then leveraging it in full end-to-end training, PLANT turns attention into a pretrainable component. PLANT is architecture-agnostic, integrates seamlessly with diverse LLM backbones, and boosts performance across tasks. Strikingly, smaller LLMs enhanced with PLANT outperform much larger models used alone, and PLANT is substantially better at predicting rare labels. It also improves sample efficiency, matching the performance of baselines trained on 10× more data. In sum, PLANT shifts attention from something merely learned during training to something we can *plant* and leverage. Looking ahead, its pretraining principle could extend naturally to multimodal tasks, where cross-signal attention is critical.

6 Related Work

Attention has long been used to capture label–text interactions in XMTC. Transformer-based models introduced multi-resolution self-attention for large label spaces (Zhang et al., 2021; Kharbanda et al., 2022), while multi-head attention across text granularities improved weak supervision (Kargupta et al., 2023). Graph and label-centric methods proposed task-specific attention modules to better capture label relevance (Goyal et al., 2023), also leveraging contrastive or knowledge-enhanced attention (Lu et al., 2023; Li et al., 2023). Dynamic pipelines filtered candidate labels using structured signals like diagnoses, procedures, and medications, relying on attention to prioritize relevant labels (Wang et al., 2024b;c). Other studies show that label-guided, dictionary, or bi-attention mechanisms improve alignment between labels and text (Wang et al., 2023b; Wu et al., 2024; Wang et al., 2024a). Meta-learning and label tree structures further advance attention-driven few-shot generalization (Teng et al., 2024; Wang et al., 2024d). Recent work includes attention-based co-ranking (Yan et al., 2025), contrastive dual-attention for rare labels (Huang et al., 2025), and knowledge-integrated attention for medical coding (Zhang et al., 2025). Attention design is central across tasks, but PLANT is the first to *treat it as pretrainable*.

LLMs are increasingly applied to XMTC problems (Asensio Blasco et al., 2025; Yuan et al., 2025; Nerella et al., 2024). However, massive LLMs applied zero-shot may be less accurate than smaller fine-tuned models (Boyle et al., 2023; Zhang et al., 2025). Extensive finetuning, in turn, raises concerns about compute requirements and overfitting (Huang et al., 2022; Michalopoulos et al., 2022; Ng et al., 2023; Kang et al., 2023). Sakai & Lam (2025) highlight that reliance on heavy finetuning fails to improve rare-label performance in high-dimensional, label-skewed spaces. As PLANT is *architecture-agnostic and effective in skewed label settings*, it integrates seamlessly with LLMs to boost rare-label performance without heavy finetuning.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Manal El Aidouni. Mastering qlora: A deep dive into 4-bit quantization and lora parameter efficient fine-tuning. https://manalelaidouni.github.io/4Bit-Quantization-Models-QLoRa.html, June 8 2024.

Elisa Asensio Blasco, Xavier Borrat Frigola, Xavier Pastor Duran, Artur Conesa González, Narcís Macià, David Sánchez Barcenilla, Ricardo Garrido Bejar, and Santiago Frid. From admission to discharge: Leveraging nlp for upstream primary coding with snomed ct. *Journal of Medical Systems*, 49(1):1–7, 2025.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Leonor Barreiros, Isabel Coutinho, Gonçalo M Correia, and Bruno Martins. Explainable icd coding via entity linking. *arXiv preprint arXiv:2503.20508*, 2025.

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

Kush Bhatia, Kunal Dahiya, Himanshu Jain, Purushottam Kar, Anshul Mittal, Yashoteja Prabhu, and Manik Varma. The extreme classification repository: Multi-label datasets and code. *URL http://manikvarma. org/downloads/XC/XMLRepository. html*, 2016.

Zeyd Boukhers, AmeerAli Khan, Qusai Ramadan, and Cong Yang. Large language model in medical informatics: Direct classification and enhanced text representations for automatic icd coding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3066–3069. IEEE, 2024.

Joseph S Boyle, Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q O'Neil. Automated clinical coding using off-the-shelf large language models. *arXiv preprint arXiv:2310.06552*, 2023.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. HyperCore: Hyperbolic and co-graph representation for automatic ICD coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3105–3114, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.282. URL https://aclanthology.org/2020.acl-main.282.

CDC. International classification of diseases, tenth revision, clinical modification (icd-10-cm), 2024. URL https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm. Accessed: 2024-10-14.

Yuyang Chai, Zhuang Li, Jiahui Liu, Lei Chen, Fei Li, Donghong Ji, and Chong Teng. Compositional generalization for multi-label text classification: A data-augmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17727–17735, 2024.

Jiamin Chen, Xuhong Li, Junting Xi, Lei Yu, and Haoyi Xiong. Rare codes count: Mining inter-code relations for long-tail clinical text classification. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 403–413, 2023a.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11828–11837, 2023b.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*, 2022.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Joakim Edin, Alexander Junge, Jakob D Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. Automated medical coding on mimic-iii and mimic-iv: A critical review and replicability study. *arXiv preprint arXiv:2304.10909*, 2023.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. An unsupervised approach to achieve supervised-level explainability in healthcare records. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4869–4890, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.280. URL https://aclanthology.org/2024.emnlp-main.280/.

Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. Can gpt-3.5 generate and code discharge summaries? *Journal of the American Medical Informatics Association*, 31(10):2284–2293, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Nidhi Goyal, Jushaan Kalra, Charu Sharma, Raghava Mutharaju, Niharika Sachdeva, and Ponnurangam Kumaraguru. Jobxmlc: Extreme multi-label classification of job skills with graph neural networks. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2181–2191, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nilesh Gupta, Fnu Devvrit, Ankit Singh Rawat, Srinadh Bhojanapalli, Prateek Jain, and Inderjit S Dhillon. Dual-encoders for extreme multi-label classification. In *The Twelfth International Conference on Learning Representations*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, pp. 102963, 2025.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. Plm-icd: automatic icd coding with pretrained language models. *arXiv preprint arXiv:2207.05289*, 2022.

Hui Huang, Mingfeng Yu, Shuai Yu, Yongbin Qin, and Chuan Lin. Contrastive learning-enhanced dual attention network for multi-label text classification. *Journal of King Saud University Computer and Information Sciences*, 37(6):136, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Greg Kamradt. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023. Accessed: 2025-07-22.

Beichen Kang, Xiaosu Wang, Yun Xiong, Yao Zhang, Chaofan Zhou, Yangyong Zhu, Jiawei Zhang, and Chunlei Tang. Automatic icd coding based on segmented clinicalbert with hierarchical tree structure learning. In *International Conference on Database Systems for Advanced Applications*, pp. 250–265. Springer, 2023.

Priyanka Kargupta, Tanay Komarlu, Susik Yoon, Xuan Wang, and Jiawei Han. Megclass: extremely weakly supervised text classification via mutually-enhancing text granularities. *arXiv preprint arXiv:2304.01969*, 2023.

Siddhant Kharbanda, Atmadeep Banerjee, Erik Schultheis, and Rohit Babbar. Cascadexml: Rethinking transformers for end-to-end multi-resolution training in extreme multi-label classification. *Advances in neural information processing systems*, 35:2074–2087, 2022.

Siddhant Kharbanda, Atmadeep Banerjee, Devaansh Gupta, Akash Palrecha, and Rohit Babbar. Inceptionxml: A lightweight framework with synchronized negative sampling for short text extreme classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 760–769, 2023.

Siddhant Kharbanda, Devaansh Gupta, Erik Schultheis, Atmadeep Banerjee, Cho-Jui Hsieh, and Rohit Babbar. Gandalf: Learning label-label correlations in extreme multi-label classification via label features. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 1360–1371, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672063. URL https://doi.org/10.1145/3637528.3672063.

Fei Li and Hong Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8180–8187, 2020.

Xinhang Li, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. Towards automatic icd coding via knowledge enhanced multi-task learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1238–1248, 2023.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, et al. Application of large language models in medicine. *Nature Reviews Bioengineering*, pp. 1–20, 2025a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024b.

yang liu, hua cheng, russell klopfer, matthew r. gormley, and thomas schaaf. effective convolutional attention network for multi-label clinical document classification. In *proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5941–5953, online and punta cana, dominican republic, November 2021. association for computational linguistics. doi: 10.18653/v1/2021.emnlp-main.481. URL https://aclanthology.org/2021.emnlp-main.481.

Zhi Liu, Yunjie Huang, Xincheng Xia, and Yihao Zhang. All is attention for multi-label text classification. *Knowledge and Information Systems*, 67(2):1249–1270, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chang Lu, Chandan Reddy, Ping Wang, and Yue Ning. Towards semi-structured automatic icd coding via tree-based contrastive learning. *Advances in Neural Information Processing Systems*, 36:68300–68315, 2023.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. Corelation: Boosting automatic icd coding through contextualized code relation learning. *arXiv preprint arXiv:2402.15700*, 2024.

Sumit Madan, Manuel Lentzen, Johannes Brandt, Daniel Rueckert, Martin Hofmann-Apitius, and Holger Fröhlich. Transformer models in biomedicine. *BMC Medical Informatics and Decision Making*, 24(1):214, 2024.

George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. Icdbigbird: a contextual embedding model for icd code classification. *arXiv preprint arXiv:2204.10408*, 2022.

Elias Moons, Aditya Khanna, Abbas Akkasi, and Marie-Francine Moens. A comparison of deep learning methods for icd coding of clinical records. *Applied Sciences*, 10(15):5262, 2020.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, T Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. Clip: A dataset for extracting action items for physicians from hospital discharge notes. *arXiv preprint arXiv:2106.02524*, 2021.

Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, et al. Transformers and large language models in healthcare: A review. *Artificial intelligence in medicine*, pp. 102900, 2024.

Clarence Boon Liang Ng, Diogo Santos, and Marek Rei. Modelling temporal document sequences for clinical icd coding. *arXiv preprint arXiv:2302.12666*, 2023.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. Mimic-iv-icd: A new benchmark for extreme multilabel classification. *arXiv preprint arXiv:2304.13998*, 2023.

Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

Hajar Sakai and Sarah S Lam. Large language models for healthcare text classification: A systematic review. *arXiv preprint arXiv:2503.01159*, 2025.

Jiang-Xin Shi, Tong Wei, and Yu-Feng Li. Residual diverse ensemble for long-tailed multi-label text classification. *Science CHINA Information Science*, 2024.

Congzheng Song, Shanghang Zhang, Najmeh Sadoughi, Pengtao Xie, and Eric Xing. Generalized zero-shot text classification for icd coding. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 4018–4024, 2021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Fei Teng, Quanmei Zhang, Xiaomin Zhou, Jie Hu, and Tianrui Li. Few-shot icd coding with knowledge transfer and evidence representation. *Expert Systems with Applications*, 238:121861, 2024.

Nathan Vandemoortele, Bram Steenwinckel, Femke Ongenae, and Sofie Van Hoecke. From haystack to needle: Label space reduction for zero-shot classification. *arXiv preprint arXiv:2502.08436*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Gang Wang, Yajun Du, Yurui Jiang, Jia Liu, Xianyong Li, Xiaoliang Chen, Hongmei Gao, Chunzhi Xie, and Yan-li Lee. Label-text bi-attention capsule networks model for multi-label text classification. *Neurocomputing*, 588:127671, 2024a.

Jiyao Wang, Zijie Chen, Yang Qin, Dengbo He, and Fangzhen Lin. Multi-aspect co-attentional collaborative filtering for extreme multi-label text classification. *Knowledge-Based Systems*, 260: 110110, 2023a.

Xindi Wang, Robert Mercer, and Frank Rudzicz. Multi-stage retrieve and re-rank model for automatic medical coding recommendation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4881–4891, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.273. URL https://aclanthology.org/2024.naacl-long.273/.

Xindi Wang, Robert E. Mercer, and Frank Rudzicz. Auxiliary knowledge-induced learning for automatic multi-label medical document classification. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2006–2016, Torino, Italia, May 2024c. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.181/.

Zeqiang Wang, Yuqi Wang, Haiyang Zhang, Wei Wang, Jun Qi, Jianjun Chen, Nishanth Sastry, Jon Johnson, and Suparna De. Icdxml: enhancing icd coding with probabilistic label trees and dynamic semantic representations. *Scientific Reports*, 14(1):18319, 2024d.

Zihan Wang, Tianle Wang, Dheeraj Mekala, and Jingbo Shang. A benchmark on extremely weakly supervised text classification: Reconcile seed matching and prompting approaches. *arXiv preprint arXiv:2305.12749*, 2023b.

WHO. *International Statistical Classification of Diseases and Related Health Problems*. 11th edition, 2025. URL https://www.who.int/standards/classifications/classification-of-diseases.

John Wu, David Wu, and Jimeng Sun. Dila: Dictionary label attention for mechanistic interpretability in high-dimensional multi-label medical coding prediction. *arXiv preprint arXiv:2409.10504*, 2024.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. Ehr coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pp. 649–658, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357897. URL https://doi.org/10.1145/3357384.3357897.

Jie Xiong, Li Yu, Xi Niu, and Youfang Leng. Xrr: Extreme multi-label text classification with candidate retrieving and deep ranking. *Information Sciences*, 622:115–132, 2023.

Yan Yan, Junyuan Liu, and Bo-Wen Zhang. Labelcorank: Revolutionizing long tail multi-label classification with co-occurrence reranking. *arXiv preprint arXiv:2503.07968*, 2025.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022a.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. Knowledge injected prompt based fine-tuning for multi-label few-shot icd coding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, pp. 1767. NIH Public Access, 2022b.

Zhichao Yang, Sanjit Singh Batra, Joel Stremmel, and Eran Halperin. Surpassing gpt-4 medical coding with a two-stage approach. *arXiv preprint arXiv:2311.13735*, 2023a.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hong Yu. Multi-label few-shot icd coding as autoregressive generation with prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5366–5374, 2023b.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023c.

Hui Ye, Rajshekhar Sunderraman, and Shihao Ji. Matchxml: an efficient text-label matching framework for extreme multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4781–4793, 2024.

Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. Pecos: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98): 1–32, 2022.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

Kevin Yuan, Chang Ho Yoon, Qingze Gu, Henry Munby, A Sarah Walker, Tingting Zhu, and David W Eyre. Transformers and large language models are efficient feature extractors for electronic health record studies. *Communications Medicine*, 5(1):83, 2025.

Ling Yuan, Xinyi Xu, Ping Sun, Hai ping Yu, Yin Zhen Wei, and Jun jie Zhou. Research of multi-label text classification based on label attention and correlation networks. *PloS one*, 19 (9):e0311305, 2024.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 808–814, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.91. URL https://aclanthology.org/2022.acl-short.91.

Bin Zhang and Junli Wang. A novel icd coding framework based on associated and hierarchical code description distillation. *arXiv preprint arXiv:2404.11132*, 2024.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280, 2021.

Shurui Zhang, Bozheng Zhang, Fuxin Zhang, Bo Sang, and Wanchun Yang. Automatic ICD coding exploiting discourse structure and reconciled code embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2883–2891, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.254.

Xu Zhang, Kun Zhang, Wenxin Ma, Rongsheng Wang, Chenxu Wu, Yingtai Li, and S Kevin Zhou. A general knowledge injection framework for icd coding. *arXiv preprint arXiv:2505.18708*, 2025.

Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5948–5957, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.463. URL https://aclanthology.org/2021.acl-long.463.

Yaxin Zhu and Hamed Zamani. Icxml: An in-context learning framework for zero-shot extreme multi-label classification. *arXiv preprint arXiv:2311.09649*, 2023.

A Training Details

LLM Backbone We use the following pretrained instruction-tuned LLMs as base models $\mathcal{M}_{\text{base}}$ in our experiments, all publicly available on the Hugging Face Model Hub and compatible with the Transformers library: (1) **Mistral-7B-Instruct-v0.3** (7B, Mistral AI): `https://huggingface.co/mixtral-7b-instruct-v0.3`, (2) **Llama-3.1-8B** (8B, Meta AI): `https://huggingface.co/meta-llama/Llama-3.1-8B`, (3) **DeepSeek-R-336B** (336B, DeepSeek): `https://huggingface.co/deepseek/DeepSeek-R-336B`, and (4) **Phi-3-mini-3.8B** (3.8B, Microsoft): `https://huggingface.co/microsoft/Phi-3-mini-3.8B`. These serve as the LLM backbones for fine-tuning.

Quantization & LoRA Adaptation Starting with a pretrained model $\mathcal{M}_{\text{base}}$, such as `Mistral-7B`, `LLaMA3-8B`, or `Phi-3`, we apply Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) Hu et al. (2022); Dettmers et al. (2023); Liu et al. (2024b).

To enable memory-efficient fine-tuning on resource-constrained hardware, we first quantize $\mathcal{M}_{\text{base}}$ to 4-bit precision using the `NormalFloat4` format with double quantization, yielding $\mathcal{M}_{\text{quant}}$:

$$Q(\boldsymbol{W}) = \text{round}\left(\frac{\boldsymbol{W}}{s}\right) \cdot s,$$

where $\boldsymbol{W}$ is a model weight matrix and s is a learned scale. Inference is performed using `bfloat16` precision ($\mathbb{F}_{16}$b) (Refer to Frantar et al. (2022) for details).

We then apply LoRA to a subset of the attention projection layers (query, key, value, and output), introducing trainable low-rank matrices:

$$\Delta\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}, \quad \text{with } \boldsymbol{A} \in \mathbb{R}^{d\times r}, \boldsymbol{B} \in \mathbb{R}^{r\times d},$$

using rank $r = 16$, scaling factor $\alpha = 32$, and dropout $p = 0.05$. The adapted model becomes:

$$\mathcal{M}_{\text{adapt}} = \mathcal{M}_{\text{quant}} + \alpha \cdot \Delta\boldsymbol{W}.$$

Optimization and Training Regimen We train all configurations using the AdamW optimizer Loshchilov & Hutter (2017) with a learning rate of 2×10^{-5}, weight decay of 0.01, and a cosine annealing learning rate schedule without restarts. To balance memory constraints and gradient stability, we use a per-device batch size of 8 with 4-step gradient accumulation, effectively simulating a batch size of 32. All models are trained using mixed-precision (FP16) via PyTorch's `autocast` module to reduce memory usage and accelerate training, and we enable gradient checkpointing to further reduce memory overhead during backpropagation.

Each model—corresponding to Stage 1 (Attention Pretraining) and Stage 2 (End-to-End Fine-tuning) as described in Section 2—is trained for up to 10 epochs with early stopping. In Stage 1, we monitor validation ndcg@k and apply early stopping with a patience of 2 epochs. In Stage 2, early stopping is based on validation macro-F1 with the same patience setting. We apply gradient clipping with a maximum norm of 1.0 for stability. To ensure reproducibility, we fix random seeds across `random`, `numpy`, `torch`, and `torch.cuda`. All experiments are conducted on NVIDIA A100 GPUs using distributed data-parallelism (DDP) when applicable, and training metrics are logged via Weights & Biases for real-time monitoring and version tracking.

Token Selection Sensitivity To test sensitivity to token selection in the ranking loss (Equation 2), we vary the top-k token threshold with $k \in \{500, 1000, 2000\}$.

HYPERPARAMETERS IN ARCHITECTURE (SECTION 2)

The multi-head attention module MultiHead (Equation 1) uses $k = 8$ attention heads. The adaptive average pooling layer (Equation **??**) produces an output size of $p = 128$.

FOCAL LOSS WITH LABEL SMOOTHING

For completeness, we present the explicit formulation of the focal loss used in Stage 2 training (Section 2). The focal loss for an input $\mathbf{x}_i$ is defined as:

$$\mathcal{L}_{\text{focal}}^{(i)}(\tilde{y}, \hat{y}, \theta) = -\frac{1}{|\mathcal{L}|} \sum_{l=1}^{|\mathcal{L}|} \Big[\tilde{y}_{il} \left(1 - \sigma(\hat{y}_{il})\right)^{\gamma} \log\left(\sigma(\hat{y}_{il})\right) + \left(1 - \tilde{y}_{il}\right) \left(\sigma(\hat{y}_{il})\right)^{\gamma} \log\left(1 - \sigma(\hat{y}_{il})\right) \Big], \tag{3}$$

where θ denotes all trainable model parameters, $\sigma(\cdot)$ is the sigmoid function, $\hat{y}_{il} \in \mathbb{R}$ is the predicted logit for label l, $\gamma = 2$ is the focusing parameter that emphasizes harder examples, and $\tilde{y}_{il}$ is the smoothed label: $\tilde{y}_{il} = (1 - \epsilon)^{y_{il}} (\epsilon)^{1 - y_{il}}$ with $\epsilon = 0.1$ to prevent overconfidence in predictions.

B PRECOMPUTATIONS

B.1 MUTUAL INFORMATION GAIN IN XMTC

In extreme classification (e.g., ICD coding with MIMIC-III), MIG quantifies how informative a token t_j is for predicting the presence of a label l. We define:

$$r_{l,j} = \sum_{(x,y) \in \{0,1\}^2} P(x, y) \log\left(\frac{P(x, y)}{P(x)\, P(y)}\right),$$

where $x = 1[l \text{ present}]$, $y = 1[t_j \text{ present}]$, and probabilities are estimated from corpus co-occurrence statistics. The scores $r_{l,j}$ are normalized to $[0, 1]$ for numerical stability and used to rank tokens per label pairs.

C IMPLEMENTATION DETAILS

DATASETS

We compare PLANT to SOTA ICD coding models using the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets, which include rich textual and structured records from ICU settings, primarily discharge summaries annotated with ICD-9 (MIMIC-III) and ICD-10 (MIMIC-IV) codes. MIMIC-III contains 52,722 discharge summaries with 8,929 unique ICD-9 codes, and MIMIC-IV includes 122,279 summaries with 7,942 ICD-10 codes. We follow established methodologies for patient ID-based splits and frequent code subsets. For few-shot learning, we evaluate PLANT on the MIMIC-III-rare50 dataset (Yang et al., 2022b), which features 50 rare ICD codes, and the MIMIC-III-few dataset (Yang et al., 2023b), a subset with 685 unique ICD-9 codes occurring between 1 and 5 times in the training set. We denote these datasets as `MIMIC-III-full`, `MIMIC-III-top50`, `MIMIC-III-rare50`, `MIMIC-III-few`, and `MIMIC-IV-full` (refer to Table 7 for statistics). Following prior research (Mullenbach et al., 2018; Xie et al., 2019; Li & Yu, 2020), we tokenize and lowercase all text while eliminating non-alphabetic tokens containing numbers or punctuation.

To assess generalizability beyond the clinical domain, we also experiment with two large-scale extreme multilabel datasets. The `EURLEX-4K` dataset, comprising 15,449 training and 3,865 test European Union legal documents annotated with 3,956 EUROVOC labels, supports automated legal topic classification, compliance analysis, and cross-lingual information retrieval (`http://manikvarma.org/downloads/XC/XMLRepository.html`). The `WIKI10-31K` dataset, with 14,146 training and 6,616 test Wikipedia articles associated with 30,938 categories, facilitates automatic tagging, web-scale document organization, and content recommendation (`http://manikvarma.org/downloads/XC/XMLRepository.html`). Both datasets are used to study large-scale label spaces and imbalanced label distributions(refer to Table 8 for statistics).

	MIMIC-III-full	MIMIC-IV-full
Number of documents	52,723	122,279
Number of patients	41,126	65,659
Number of unique codes	8,929	7,942
Codes per instance: Median (IQR)	14(10–20)	14(9–20)
Words per document: Median (IQR)	1,375(965–1,900)	1,492(1,147–1,931)
Documents: Train/val/test [%]	90.5/3.1/6.4	72.9/10.9/16.2

Table 7: Descriptive statistics for MIMIC-III-full and MIMIC-IV-full discharge summary training sets.

	EURLEX-4K	WIKI10-31K
Number of train documents	15,449	14,146
Number of test documents	3,865	6,616
Number of unique labels	3,956	30,938
Average number of labels per instance	5.30	18.64
Average number of instances per label	20.79	8.52

Table 8: Descriptive statistics for publicly available XMTC datasets EURLEX-4K and WIKI10-31K.

Implementation and Hyperparameters

We ensure robustness across diverse XMTC datasets by fine-tuning hyperparameters on the MIMIC-III-full and MIMIC-IV-full validation sets. Experiments are conducted on an NVIDIA QUADRO RTX 8000 GPU with 48 GB VRAM. We utilize the AWD-LSTM LM with an embedding size of 400, 3 LSTM layers with 1152 hidden activations, and the Adam Optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay of 0.01. During fine-tuning, we apply dropout rates and weight dropout, with a batch size of 384, BPTT of 80, 20 epochs, and a learning rate of $1e-5$. Classifier training also includes dropout rates and weight dropout, with a batch size of 16, BPTT of 72, and discriminative fine-tuning with gradual unfreezing over 115 epochs (on MIMIC-III-full), alongside scheduled weight decay and learning rate ranges.

D Baselines for comparisons

ICD Baselines: We compare PLANT against a diverse set of ICD coding baselines spanning classical, recent, and few-shot paradigms.

Early deep learning models: CAML (Mullenbach et al., 2018), MSATT-KG (Xie et al., 2019), MUltiResCNN (Li & Yu, 2020), and HyperCore (Cao et al., 2020).

Attention- and hierarchy-based models: LAAT and JointLAAT (Vu et al., 2021), ISD (Zhou et al., 2021), Effective-CAN (liu et al., 2021), Hierarchical (Dai et al., 2022), and MSMN (Yuan et al., 2022).

Recent pretraining and architecture innovations: DiscNet (Zhang et al., 2022), KEPTLongformer (Yang et al., 2022b), PLM-ICD (Huang et al., 2022), AHDD (Zhang & Wang, 2024), CoRelation (Luo et al., 2024), Contrastive (Lu et al., 2023), MIMIC-IV-Benchmark (Nguyen et al., 2023), Tr-EHR (Yang et al., 2023c), and PLM-CA (Edin et al., 2024).

Few-shot ICD coding methods: AGMHT (Song et al., 2021), RareCodes (Chen et al., 2023a), GP (Yang et al., 2023b), and KEPT (Yang et al., 2022b).

Knowledge-injected models: KEMTL (Li et al., 2023), MRR (Wang et al., 2024b), AKIL (Wang et al., 2024c), and GKI-ICD (Zhang et al., 2025).

XMTC Baselines: We also compare PLANT against XMTC models like: PECOS (Yu et al., 2022), ICXML (Zhu & Zamani, 2023), XRR (Xiong et al., 2023), RDE (Shi et al., 2024), MatchXML (Ye et al., 2024), DE (Gupta et al.), InceptionXML (Kharbanda et al., 2023), GANDALF (Kharbanda et al., 2024), CG (Chai et al., 2024).

E Evaluation Metrics

We focus on micro-F1, macro-F1, micro-P, macro-P, micro-R, macro-R, micro-AUC, macro-AUC, P@k, and R@k to compare with prior ICD studies. Micro-averaging treats each (text, code) pair individually, aggregating true positives, false positives, and false negatives across all instances. Macro-averaging computes metrics per label, giving more weight to infrequent labels. micro-P is the ratio of aggregated true positives to the sum of true positives and false positives, while macro-P averages precision across all labels. micro-R is the ratio of aggregated true positives to the sum of true positives and false negatives, while macro-R averages recall across all labels. micro-AUC computes the area under the ROC curve for all instances aggregated together, while macro-AUC averages the AUC scores across all labels. P@k and R@k measure the proportion of the top k predicted labels that match the ground truth, focusing on precision and recall, respectively.

$$\text{micro-P} = \frac{\sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FP}_i)}$$

$$\text{micro-R} = \frac{\sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FN}_i)}$$

$$\text{micro-F1} = \frac{2 \cdot \sum_i \text{TP}_i}{\sum_i (\text{TP}_i + \text{FP}_i) + \sum_i (\text{TP}_i + \text{FN}_i)}$$

$$\text{micro-AUC} = \int_0^1 \text{TPR}_{\text{micro}}(\text{FPR}_{\text{micro}})\, d\text{FPR}_{\text{micro}}$$

$$\text{macro-P} = \frac{1}{L} \sum_{i=1}^{L} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$

$$\text{macro-R} = \frac{1}{L} \sum_{i=1}^{L} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

$$\text{macro-F1} = \frac{1}{L} \sum_{i=1}^{L} \frac{2 \cdot \text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{TP}_i + \text{FN}_i}$$

$$\text{macro-AUC} = \frac{1}{L} \sum_{i=1}^{L} \int_0^1 \text{TPR}_i(\text{FPR}_i)\, d\text{FPR}_i$$

$$\text{P@}k = \frac{1}{k} \sum_{i=1}^{k} 1\left[\text{pred}_i \in Y\right]$$

$$\text{R@}k = \frac{1}{\min(k, |Y|)} \sum_{i=1}^{k} 1\left[\text{pred}_i \in Y\right]$$

where TP_i, FP_i, and FN_i are the true positives, false positives, and false negatives for label i, respectively, L is the total number of labels, $\text{TPR}_{\text{micro}}$ and $\text{FPR}_{\text{micro}}$ are the true positive rate and false positive rate for the aggregated micro-averaged data, TPR_i and FPR_i are the true positive rate and false positive rate for label i, Y is the ground truth label set for an instance, and pred_i is the i-th top predicted label.

F Statistical Significance

Statistical Significance via Wilcoxon Signed-Rank Test. We assess statistical significance using the non-parametric Wilcoxon Signed-Rank Test (Demšar, 2006) for comparing paired model outputs. For metrics computed at the instance level (e.g., P@15), we apply the test directly to the paired per-instance scores between the base model and its PLANT-enhanced counterpart. For aggregate metrics such as F1 and AUC, which are reported as single values over the full test set, we first collect N paired scores—either from repeated evaluations (e.g., $N = 10$ in 10-fold cross-validation) or from N bootstrap resamples. Let $\{a_1, a_2, \ldots, a_N\}$ and $\{b_1, b_2, \ldots, b_N\}$ denote the scores of the base model and the PLANT-enhanced model, respectively. We compute the difference $d_i = b_i - a_i$ for each pair and rank the absolute values $|d_i|$ (excluding zeros), averaging ranks in the case of ties. Each rank is assigned the sign of d_i, and we compute the rank sums W^+ and W^- over positive and negative differences. The test statistic is $W = \min(W^+, W^-)$.

For small N, statistical significance is determined using exact Wilcoxon critical values; for larger N, we apply the normal approximation with

$$\mu = \frac{N(N+1)}{4}, \quad \sigma = \sqrt{\frac{N(N+1)(2N+1)}{24}},$$
$$z = \frac{W - \mu}{\sigma}.$$

We reject the null hypothesis of no difference if the resulting p-value is less than a threshold α (typically 0.05). In our tables, statistically significant improvements are marked using ▲. This test is readily implemented in standard libraries such as `scipy.stats.wilcoxon` in Python or `wilcox.test(paired=TRUE)` in R.

Reporting Gains with Confidence Intervals. We also report absolute gains along with 95% confidence intervals (CI) using paired bootstrap resampling. For each evaluation metric, we draw $B = 1000$ bootstrap samples from the test set and compute the difference $\Delta_b = \text{Metric}_b^{\text{PLANT}} - \text{Metric}_b^{\text{Base}}$ for each sample b. The reported gain is the mean $\hat{\mu}$ of $\{\Delta_b\}$, and the CI is computed using the percentile bootstrap method by taking the 2.5th and 97.5th percentiles of the empirical distribution of $\{\Delta_b\}$.

We mark results as statistically significant only if the Wilcoxon signed-rank test (α=0.05) is passed *and* the 95% CI excludes 0. In such cases, we annotate the score with a colored arrow: ▲ for statistically significant gains and ▼ for significant drops. If the CI includes 0, no arrow is shown. For example, 14.7▲ (+1.2, [0.6, 1.8]) indicates a statistically significant gain over the base model, while 70.1▼(-1.4, [-2.1, -0.7]) denotes a significant drop. In contrast, 73.8 (+0.3, [0.0, 0.6]) is not statistically significant and is shown without an arrow.

G Additional Results

Model	MIMIC-III-full					MIMIC-IV-full				
	AUC		F1		P@15	AUC		F1		P@15
	Macro	Micro	Macro	Micro		Macro	Micro	Macro	Micro	
Mistral-7B	90.8	98.9	13.5	62.0	63.5	90.2	98.7	20.0	57.0	53.8
Mistral-7B + PLANT	98.1▲ *(+7.3)*	**99.9**▲ *(+1.0)*	14.7▲ *(+1.2)*	64.1▲ *(+2.1)*	65.8▲ *(+2.3)*	97.4▲ *(+7.2)*	99.5▲ *(+0.8)*	23.0▲ *(+3.0)*	59.2▲ *(+2.2)*	56.9▲ *(+3.1)*
LLaMA3-8B	91.0	99.0	13.8	62.5	64.0	90.5	98.8	20.5	57.5	54.0
LLaMA3-8B + PLANT	**98.3**▲ *(+7.3)*	99.8▲ *(+0.8)*	**15.0**▲ *(+1.2)*	**64.5**▲ *(+2.0)*	**66.2**▲ *(+2.2)*	**97.6**▲ *(+7.1)*	**99.6**▲ *(+0.8)*	**23.5**▲ *(+3.0)*	**59.5**▲ *(+2.0)*	**57.0**▲ *(+3.0)*
DeepSeek-V3	90.6	98.8	13.2	61.8	63.2	90.0	98.6	19.8	56.8	53.5
DeepSeek-V3 + PLANT	97.9▲ *(+7.3)*	99.7▲ *(+0.9)*	14.5▲ *(+1.3)*	64.0▲ *(+2.2)*	65.5▲ *(+2.3)*	97.2▲ *(+7.2)*	99.4▲ *(+0.8)*	22.8▲ *(+3.0)*	59.0▲ *(+2.2)*	56.5▲ *(+3.0)*
Phi-3	90.4	98.7	13.0	61.5	63.0	89.8	98.5	19.5	56.5	53.2
Phi-3 + PLANT	97.7▲ *(+7.3)*	99.6▲ *(+0.9)*	14.3▲ *(+1.3)*	63.8▲ *(+2.3)*	65.3▲ *(+2.3)*	97.0▲ *(+7.2)*	99.3▲ *(+0.8)*	22.5▲ *(+3.0)*	58.8▲ *(+2.3)*	56.3▲ *(+3.1)*
Avg. gain with PLANT	Δ+7.3	Δ+0.9	Δ+1.3	Δ+2.2	Δ+2.3	Δ+7.2	Δ+0.8	Δ+3.0	Δ+2.2	Δ+3.1

Table 9: **Performance of LLMs with and without PLANT.** Each model is evaluated standalone and with PLANT on `MIMIC-III-full` and `MIMIC-IV-full`. Green rows denote results after integrating PLANT. Bold values indicate the best score for each metric. A compact version with only `MIMIC-IV-full` results is provided in Table 2 in the main paper.

Model	MIMIC-III-full						MIMIC-III-top50				P@5
	AUC		F1		Precision		AUC		F1		
	Macro	Micro	Macro	Micro	P@8	P@15	Macro	Micro	Macro	Micro	
Effective-CAN liu et al. (2021)	92.1	98.9	10.6	58.9	75.8	60.6	92.0	94.5	66.8	71.7	66.4
MSMN Yuan et al. (2022)	95.0	99.2	10.3	58.4	75.2	59.9	92.8	94.7	68.3	72.5	68.0
PLM-ICD (Huang et al., 2022)	92.6	98.9	10.4	59.8	77.1	61.3	91.0	93.4	66.3	71.9	66.0
Contrastive + JointLAAT (Lu et al., 2023)	94.1	98.8	11.5	58.3	73.9	59.4	91.3	93.7	67.2	72.0	67.9
KEMTL (Li et al., 2023)	95.3	99.6	12.7	58.3	75.6	59.3	94.8	95.5	69.5	72.9	70.8
AHDD (Zhang & Wang, 2024)	95.2	99.3	10.9	58.9	75.3	60.1	92.8	94.7	68.5	72.8	67.8
CoRelation (Luo et al., 2024)	95.2	99.2	10.2	59.1	76.2	60.7	93.3	95.1	69.3	73.1	68.3
PLM-CA (Edin et al., 2024)	91.6	98.9	10.3	59.9	77.2	61.6	91.6	93.6	67.1	71.0	66.4
MRR (Wang et al., 2024b)	94.9	99.5	11.4	60.3	77.5	62.3	92.7	94.7	68.7	73.2	68.5
AKIL (Wang et al., 2024c)	94.8	99.4	11.2	60.5	78.4	63.7	92.8	95.0	69.2	73.4	68.3
GKI-ICD (Zhang et al., 2025)	96.2	99.3	12.3	61.2	77.7	62.4	93.3	95.2	69.2	73.5	68.1
PLANT (Ours)	**98.1**	**99.9**	**14.7**	**64.1**	**80.3**	**65.8**	**95.1**	**96.1**	**69.9**	**73.8**	**70.9**
	▲ *+1.9*	▲ *+0.3*	▲ *+2.0*	▲ *+2.9*	▲ *+1.9*	▲ *+2.1*	*+0.3*	▲ *+0.9*	▲ *+0.6*	*+0.3*	*+0.1*
	[1.15, 2.72]	*[0.02, 0.61]*	*[1.26, 2.58]*	*[2.14, 3.41]*	*[1.02, 2.83]*	*[1.33, 2.74]*	*[-0.01, 0.58]*	*[0.47, 1.36]*	*[0.25, 0.84]*	*[-0.05, 0.63]*	*[-0.19, 0.39]*

Table 10: **PLANT vs. SOTA models on `MIMIC-III-full` and `MIMIC-III-top50`.** On `MIMIC-III-full`, PLANT achieves aggregate gains of **+1–3** across AUC, F1 (Macro), and Precision, including a **+2** (95% CI: **1.3–2.6**) gain in F1 (Macro). For `MIMIC-III-top50` (top 50 most frequent codes), gains are more modest, averaging around **+0.5** (e.g., **+0.6** in F1 (Macro), 95% CI: **0.3–0.8**).

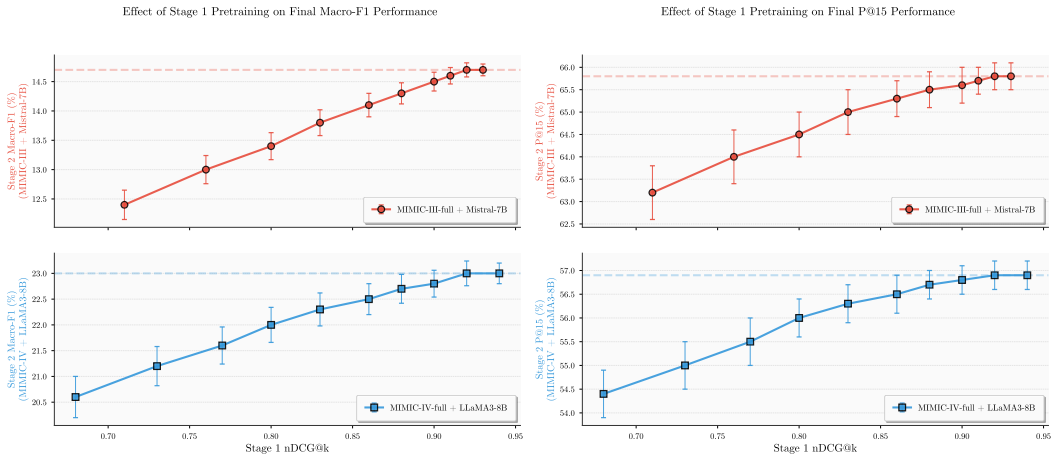


Figure 4: Effect of attention initialization quality on downstream performance across two dataset–LLM pairs: `MIMIC-III-full` with `Mistral-7B` and `MIMIC-IV-full` with `LLaMA3-8B`—as Stage 1 nDCG@k improves, final macro-F1 and P@15 after Stage 2 monotonically increase. The single-dataset view (`MIMIC-IV-full` with `LLaMA3-8B` on macro-F1) is shown in the main paper as Figure 3 (left).

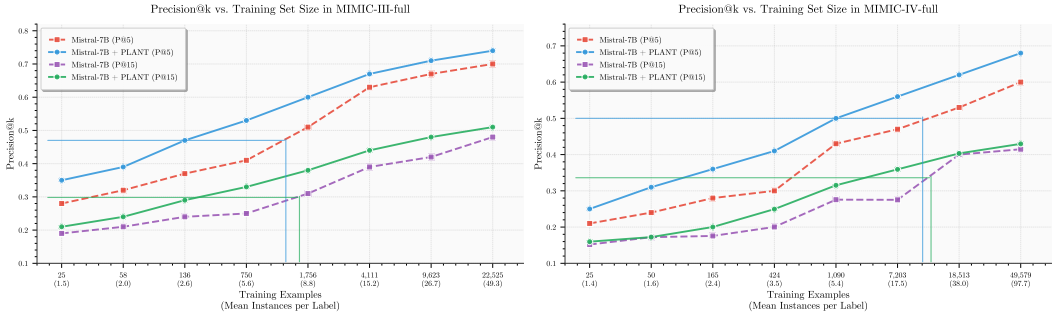


Figure 5: PLANT consistently boosts `Mistral-7B` on `MIMIC-III-full` (left) and `MIMIC-IV-full` (right) across training set sizes. Solid lines (`Mistral-7B` +PLANT) outperform dashed lines (`Mistral-7B` baseline) on both P@5 and P@15, with the largest gains appearing in low-data regimes. Reference lines highlight that PLANT reaches baseline performance using substantially fewer training examples. The single-dataset (`MIMIC-IV-full` only) view is shown in the main paper as Figure 3 (right).

Model	MIMIC-III-few						MIMIC-III-rare50			
	F1		Precision		Recall		AUC		F1	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
AGMHT (Song et al., 2021)	18.7	29.2	17.6	49.4	19.9	20.7	80.5	82.0	29.5	31.0
KEPTLongformer (Yang et al., 2022b)	20.5	31.0	19.2	51.0	22.0	22.5	82.7	83.3	30.4	32.6
MSMN + Contrastive (Lu et al., 2023)	4.3	8.5	4.5	**70.9**	4.2	4.5	–	–	31.2	30.6
GP (Yang et al., 2023b)	30.2	35.3	27.9	38.5	32.9	32.6	84.0	85.5	32.0	33.5
Tr-EHR (Yang et al., 2023c)	22.0	32.5	20.5	52.0	23.5	24.0	83.5	84.8	31.5	33.0
CoRelation (Luo et al., 2024)	25.0	34.0	23.5	50.5	26.5	27.0	85.0	86.0	33.0	34.5
PLM-CA (Edin et al., 2024)	26.5	35.0	24.5	51.5	28.0	28.5	86.0	87.0	34.0	35.5
GKI-ICD (Zhang et al., 2025)	24.0	33.5	22.5	49.0	25.5	26.0	84.5	85.8	32.5	34.0
PLANT (Ours)	**66.3**	**71.0**	**65.1**	68.6	**81.0**	**81.7**	**95.6**	**96.0**	**82.6**	**84.2**
	▲ *+36.1* *[30.5, 41.7]*	▲ *+35.7* *[29.8, 41.2]*	▲ *+37.2* *[31.0, 43.5]*	▼ *-2.3* *[-3.7, -1.0]*	▲ *+48.1* *[42.6, 54.0]*	▲ *+49.1* *[43.3, 54.8]*	▲ *+9.6* *[6.2, 12.4]*	▲ *+9.0* *[5.9, 11.7]*	▲ *+48.6* *[41.2, 56.4]*	▲ *+48.7* *[40.9, 55.5]*

Table 11: **Performance on rare labels.** PLANT achieves substantial improvements on most metric, with several gains exceeding +35 and percentile bootstrap CI well-separated from zero. A compact version with only the MIMIC-III-few results is provided in the main paper as Table 4.

Ablation Config	macro-AUC	macro-F1	P@15
Dataset: MIMIC-III-full, **LLM:** Mistral-7B			
[1]Single-Stage BCE	92.0▼ *(-6.1, [-6.9, -5.1])*	10.0▼ *(-4.7, [-5.6, -3.3])*	60.0▼ *(-5.8, [-6.6, -4.7])*
[1]Single-Stage Focal Loss	93.5▼ *(-4.6, [-5.5, -3.7])*	11.5▼ *(-3.2, [-4.1, -2.1])*	61.5▼ *(-4.3, [-5.2, -3.1])*
[1]PLANT w/ Vanilla BCE	95.5▼ *(-2.6, [-3.3, -1.9])*	12.7▼ *(-2.0, [-2.6, -1.3])*	63.0▼ *(-2.8, [-3.4, -2.1])*
[2]PLANT w/o Label Smoothing	97.8▼ *(-0.3, [-0.6, -0.1])*	13.8▼ *(-0.9, [-1.4, -0.4])*	64.5▼ *(-1.3, [-1.9, -0.7])*
[2]PLANT w/o Hard Neg Mining	97.5▼ *(-0.6, [-1.0, -0.2])*	13.0▼ *(-1.7, [-2.3, -1.1])*	62.5▼ *(-3.3, [-4.0, -2.4])*
[1]PLANT w/ Term Frequency	97.0▼ *(-1.1, [-1.8, -0.5])*	12.5▼ *(-2.2, [-2.9, -1.6])*	62.0▼ *(-3.8, [-4.5, -2.9])*
[1]PLANT w/ MSE	96.5▼ *(-1.6, [-2.4, -0.9])*	12.0▼ *(-2.7, [-3.4, -1.9])*	61.8▼ *(-4.0, [-4.7, -3.2])*
PLANT (full setup)	**98.1**	**14.7**	**65.8**
Dataset: MIMIC-IV-full, **LLM:** LLaMA3-8B			
[1]Single-Stage BCE	91.0▼ *(-6.4, [-7.2, -5.4])*	18.0▼ *(-5.0, [-5.9, -3.8])*	52.0▼ *(-4.9, [-5.7, -3.9])*
[1]Single-Stage Focal Loss	92.5▼ *(-4.9, [-5.8, -4.0])*	19.5▼ *(-3.5, [-4.2, -2.7])*	53.5▼ *(-3.4, [-4.0, -2.7])*
[1]PLANT w/ Vanilla BCE	95.0▼ *(-2.4, [-3.0, -1.7])*	21.0▼ *(-2.0, [-2.6, -1.4])*	54.8▼ *(-2.1, [-2.7, -1.5])*
[2]PLANT w/o Label Smoothing	97.0▼ *(-0.4, [-0.7, -0.2])*	21.8▼ *(-1.2, [-1.8, -0.7])*	55.8▼ *(-1.1, [-1.7, -0.6])*
[2]PLANT w/o Hard Neg Mining	96.8▼ *(-0.6, [-1.0, -0.3])*	21.0▼ *(-2.0, [-2.7, -1.3])*	54.5▼ *(-2.4, [-3.1, -1.7])*
[1]PLANT w/ Term Frequency	96.5▼ *(-0.9, [-1.4, -0.5])*	20.5▼ *(-2.5, [-3.2, -1.8])*	54.0▼ *(-2.9, [-3.6, -2.1])*
[1]PLANT w/ MSE	96.0▼ *(-1.4, [-2.1, -0.8])*	20.0▼ *(-3.0, [-3.9, -2.1])*	53.8▼ *(-3.1, [-3.9, -2.3])*
PLANT (full setup)	**97.4**	**23.0**	**56.9**

Table 12: Ablation results on MIMIC-III-full and MIMIC-IV-full with base LLMs (Mistral-7B, LLaMA3-8B). PLANT 's largest gains come from Stage 1 attention initialization via MIG+ranking, while Stage 2 refinements (label smoothing, HNM, focal loss) add complementary improvements. A compact version with only MIMIC-IV-full results using LLaMA3-8B is provided in Table 6 in the main paper.