NEURIPS 2026 MAIN TRACK • POSTER

Don’t Pay Attention, PLANT It:

Pretraining Attention via Learning-to-Rank



Debjyoti Saha Roy



Byron C. Wallace



Javed A. Aslam

Khoury College of Computer Sciences, Northeastern University

arXiv 2410.23066 deb-official/plant debjyoti.org



One long clinical note in, a handful of codes out, from thousands of possible codes

Input: the text

Discharge summary

72-year-old man admitted with fever and shortness of breath.
Chest X-ray shows pneumonia
History of type 2 diabetes
and congestive heart failure
… median 1,375 words

Output: the labels

`J18.9` Pneumonia

`E11.9` Type 2 diabetes

`I50.9` Heart failure

median 14 codes per note, from ≈154,000 in ICD-10

Most codes are rare: ≈5,411 of 8,929 appear fewer than 10 times.



LLMs can do this zero-shot, but a long label list in the prompt dilutes their attention

A zero-shot prompt

Note: 72-year-old man …

Which codes apply?

A00.0 A00.1 A00.9
A01.00 A01.01 A01.02
A01.03 A01.04 A01.05
A01.09 A01.1 A01.2
A01.3 A01.4 A02.0

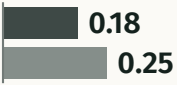
… hundreds more

≈1,000 codes in the prompt

Micro-F1 frequent codes dominate



Macro-F1 every code counts equally

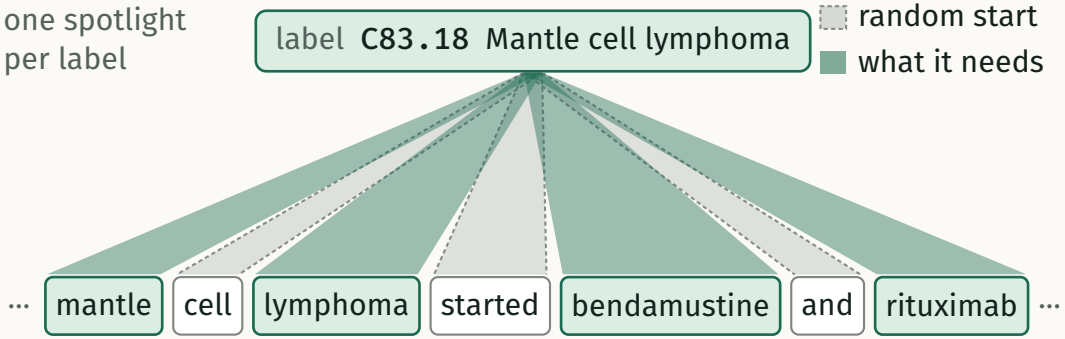


A 70B LLM trails a fine-tuned model that is 200× smaller.

LLM: DeepSeek-R1-Distill-Llama-70B
Motzfeldt et al., EMNLP Findings 2025



Fine-tuning gives every label its own attention spotlight, but each spotlight starts random



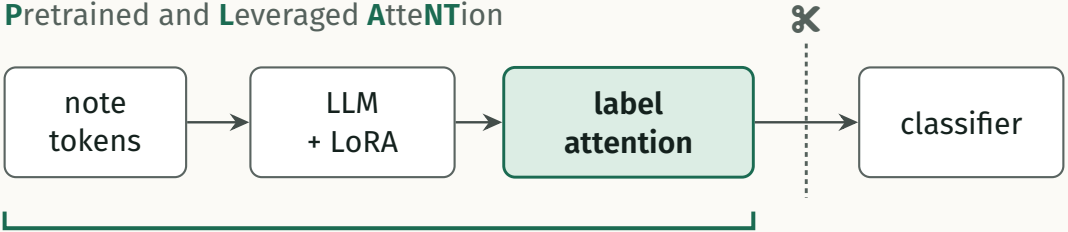
Taught only by the classifier: rare labels get little signal.

THE IDEA



PLANT plants the attention first: pretrain it as a ranker, then fine-tune as usual

Pretrained and **L**everaged **A**tte**NT**ion

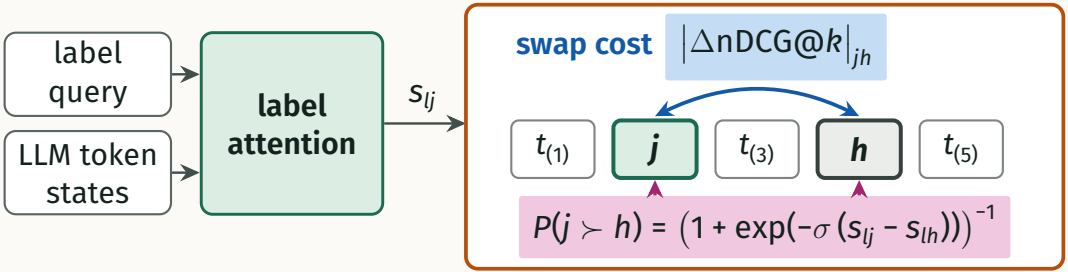


Stage 1: plant ranking loss on the attention; no classifier yet

Stage 2: fine-tune the usual classification loss, from the planted weights



Stage 1 trains attention to rank tokens the way mutual information ranks them



s_{lj}: score of token j

$\mathcal{T}_l$: top-k tokens by mutual information r; $r_{lj} > r_{lh}$

$$\mathcal{L}_{\text{rank}} = -\sum_l \sum_{j,h \in \mathcal{T}_l} |\Delta\text{nDCG@}k|_{jh} \times \log_2 P(j \succ h)$$



Ranking gives attention far denser supervision than a classification loss

Classification loss



$O(1)$

signal per label

Was the label right?

PLANT ranking loss



top-k tokens of label l

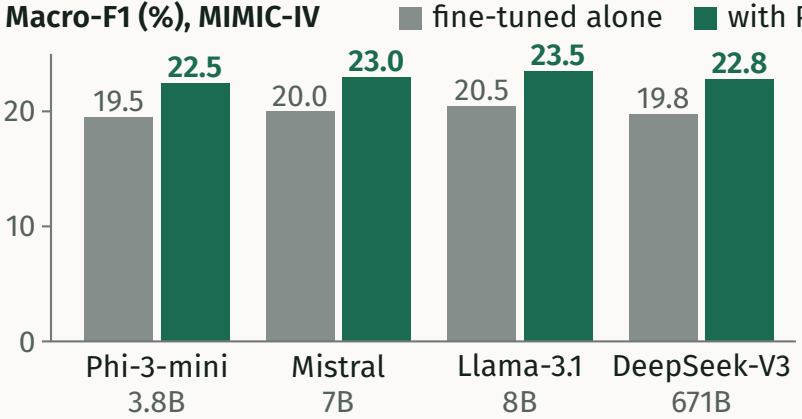
$O(k^2)$

signals per label

Which tokens outrank which?



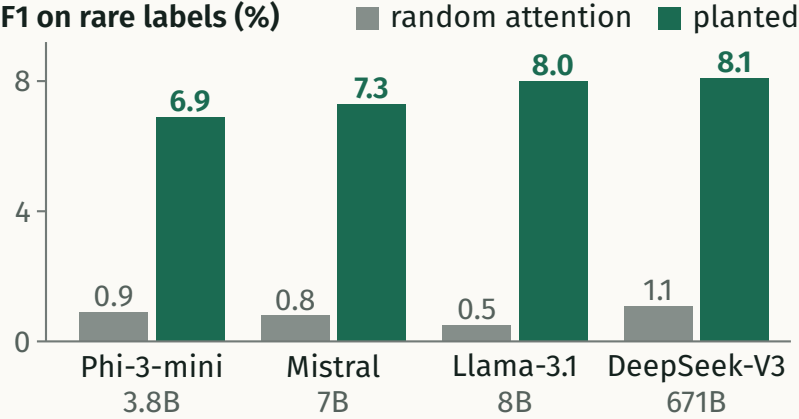
PLANT improves every LLM backbone we tested, by about 3 Macro-F1 points



+3.0
Macro-F1 on average



Planted attention reaches rare labels it never trained on; random attention does not



Setup

Train: common labels only

Test: rare labels, unseen

TAKEAWAYS



LLMs can label at extreme scale; planted attention makes them better at it

1. **Plug-and-play:** pretrain the label attention as a ranker, then fine-tune as usual
2. **Architecture-agnostic:** modern LLM backbones do better with PLANT; rare labels gain the most
3. **Across tasks:** also strong on legal topic classification and Wikipedia tagging

Limits: needs labeled data; adds 12–28% training time.



Don’t pay attention. PLANT it.

Pretrain label attention as a ranker, then fine-tune.

Poster Session 1, Sydney
Tuesday 8 December
10:00–13:00 AEDT
Hall 1-4

saharoy.d@northeastern.edu
debjyoti.org

Paper
arXiv 2410.23066

Code
deb-official/plant