





Should We Still Pretrain Encoder Models with MLM?

Hippolyte Gisserot-Boukhlef, Nicolas Boizard, et al.

ANLP - JANUARY 20, 2026

Encoders vs. Generative LLMs

In the modern era of NLP, two primary types of models dominate: encoder- and decoder-only generative models.

Generative LLM

How much do you like encoders?

↓

Causal Transformer

↓

I like encoders very much!

Encoder

How much do you like encoders?

↓

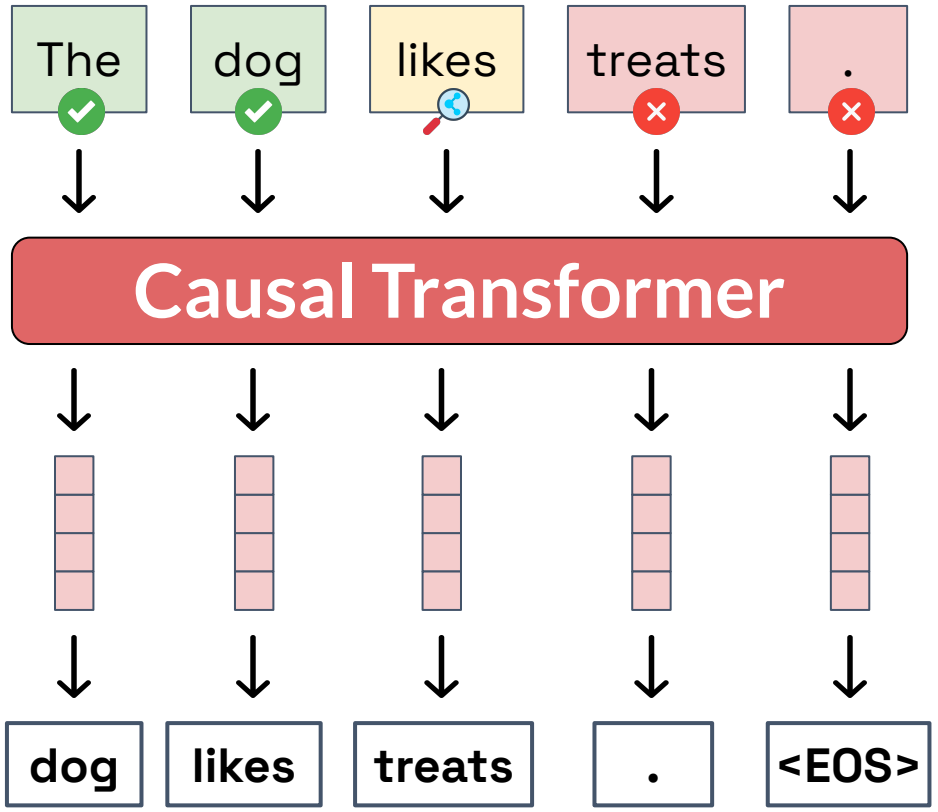
Bidirectional Transformer

↓



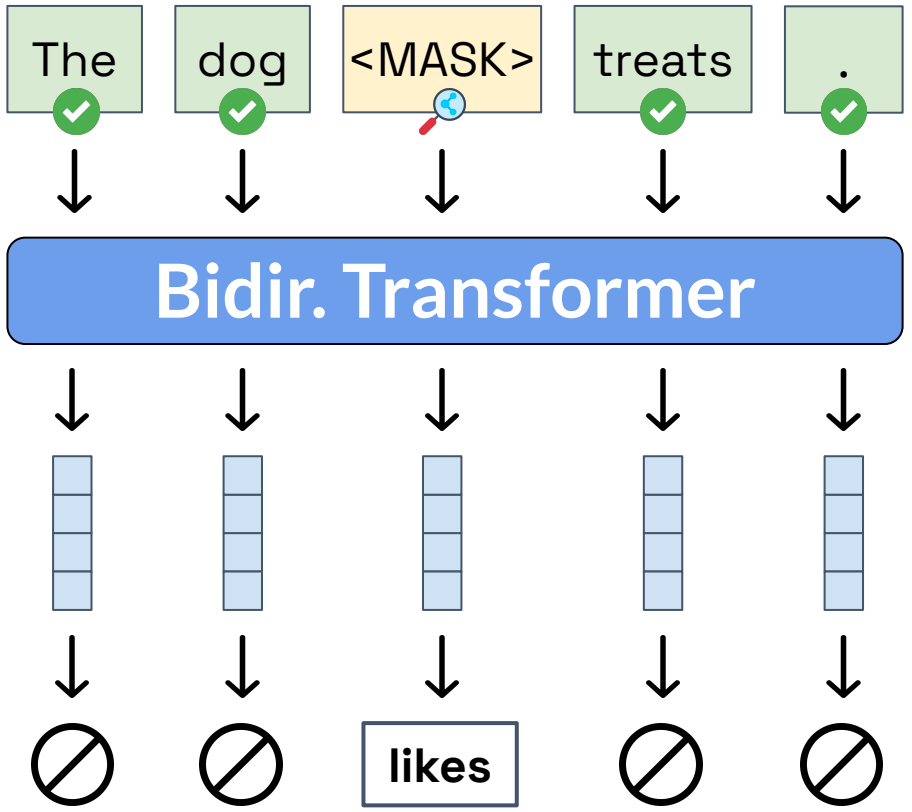
MLM vs. CLM Training Objectives

CLM (generation)



$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$

MLM (encoding)



$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}})$$

Why Encoder Models?



With the recent explosion of generative AI models, researchers tend to forget encoders are still by far the most widely used deep learning text models in the industry!



Versatility
Encoder representations can be used for tasks like classification, retrieval, NER, extractive Q&A...



Efficiency
Encoder models are inherently more compact than decoder models, making them significantly faster for both fine-tuning and inference.

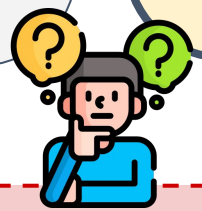
Motivations (1/2)



EuroBERT: Scaling Multilingual Encoders for European Languages
Accepted @ COLM (10/25)
N. Boizard, et al.

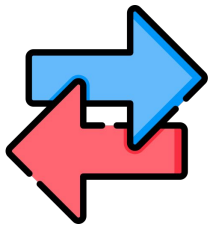


Should We Still Pretrain Encoders with Masked Language Modeling?
Preprint (07/25)
H. Gisserot-Boukhlef., et al.



> Minimal gains from 600M to 2B EuroBERT prompted questions about traditional encoder scaling
> EuroBERT is data-focused; next, we examine the role of the pretraining objective and bidirectionality

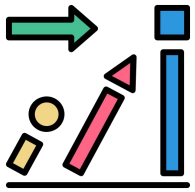
Motivations (2/2)



Bidirectional pretraining

> Bidirectional training has traditionally been the **standard approach** for learning strong text representations (e.g., BERT)

> Most bidirectional encoder models remain relatively small, typically capped at **600–700 million parameters**



Causal pretraining and representation learning

> More recently, generative models have been adapted for representation learning, achieving **SOTA results on benchmarks**

> However, such models generally operate at a **much larger scale** than classic bidirectional encoders

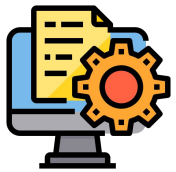


Research question

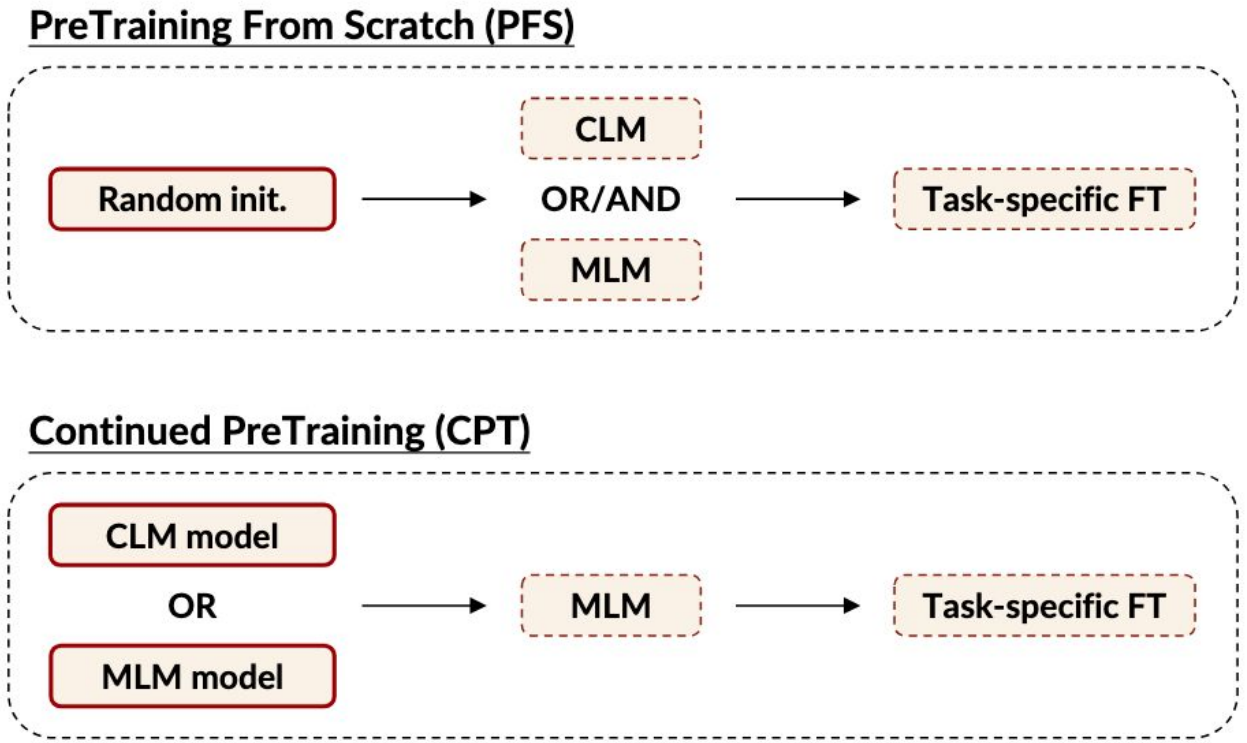
> To what extent does causal pretraining truly contribute to stronger representations, and to what extent is this effect driven by confounding factors such as model scale and data regime?

Experimental Setup (1/2)

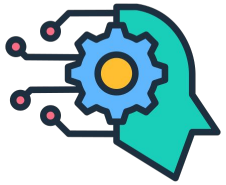
Goal: Assess the specific impact of the pretraining objective on downstream text representation quality



> **2 realistic setups**: PFS, CPT
> **3 model sizes**: 210M, 610M, 1B
> **Unified framework**: unified pretraining and fine-tuning / evaluation pipelines



Experimental Setup (2/2)



Pretraining

> **38 pretraining runs** with varied settings to study training objective impact

> **100B tokens** (42,000 steps) per run in default settings to achieve minimal loss convergence



Fine-tuning and evaluation

> **50 models** fine-tuned in total (including intermediate checkpoints)

> **4 tasks, 12 datasets** for comprehensive model evaluation

> **Fine-tuning protocol**: 1,000 steps, 6 learning rates, 5 seeds to capture variability



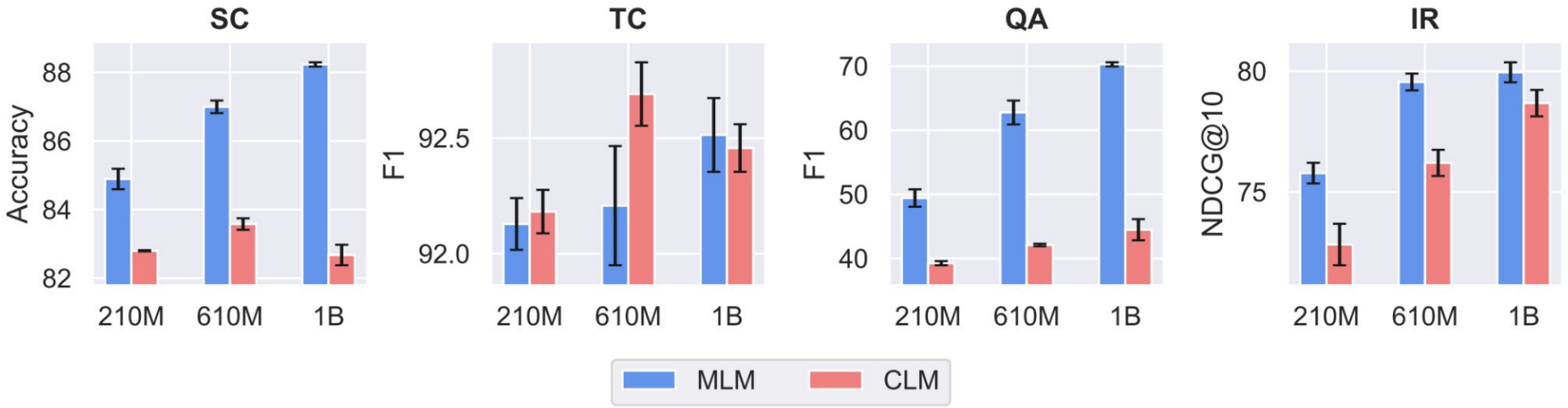
Experimental scale

> Ensure **statistically robust results** through sufficiently long pretraining and multiple fine-tuning runs

> Total compute budget: **110k MI250X GPU hours**

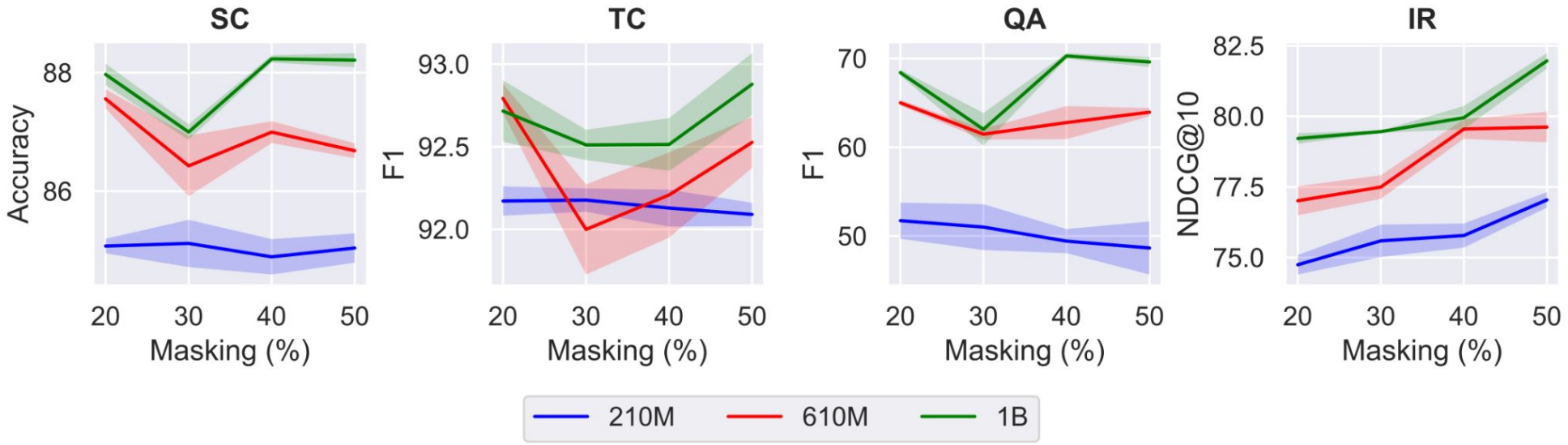
XP #1 – Pretraining with MLM or CLM (1/3)

Question: Between CLM and MLM, which objective is more effective for learning strong text representations?



> **MLM globally outperforms CLM** for learning text representations
> **Scaling behavior differs** by task (e.g., SC vs. IR)
> **CLM matches or outperforms MLM on TC**

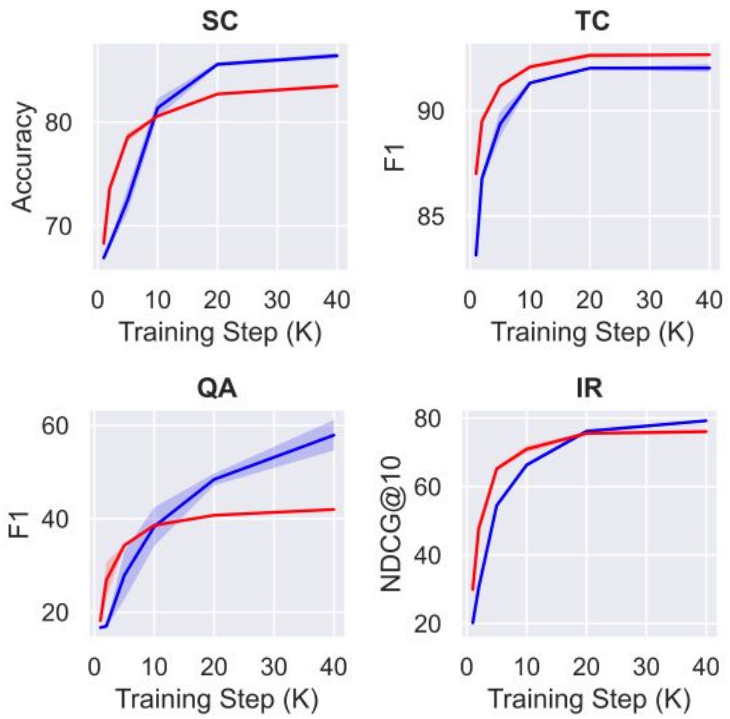
XP #1 – Pretraining with MLM or CLM (2/3)

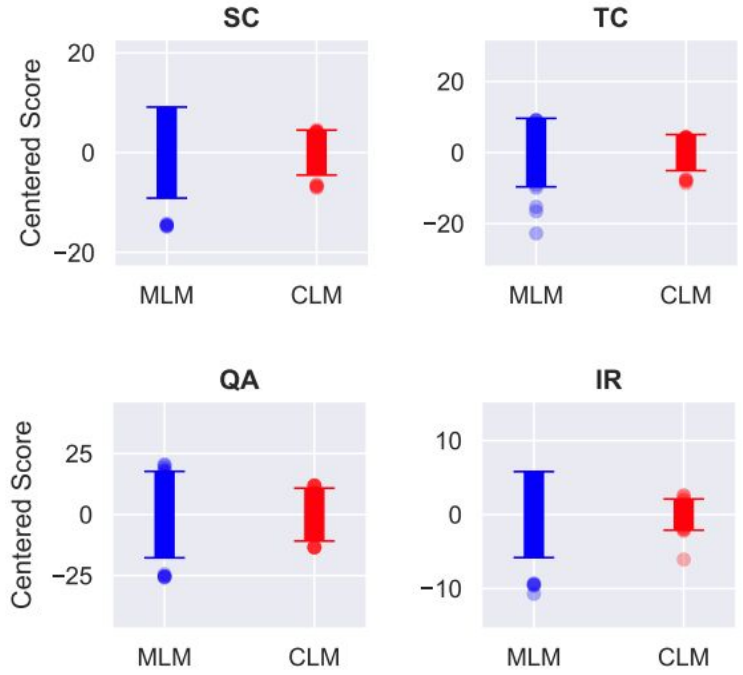




> There is **no universally optimal** masking ratio
> MLM demands **more careful pretraining design** than CLM

XP #1 – Pretraining with MLM or CLM (3/3)



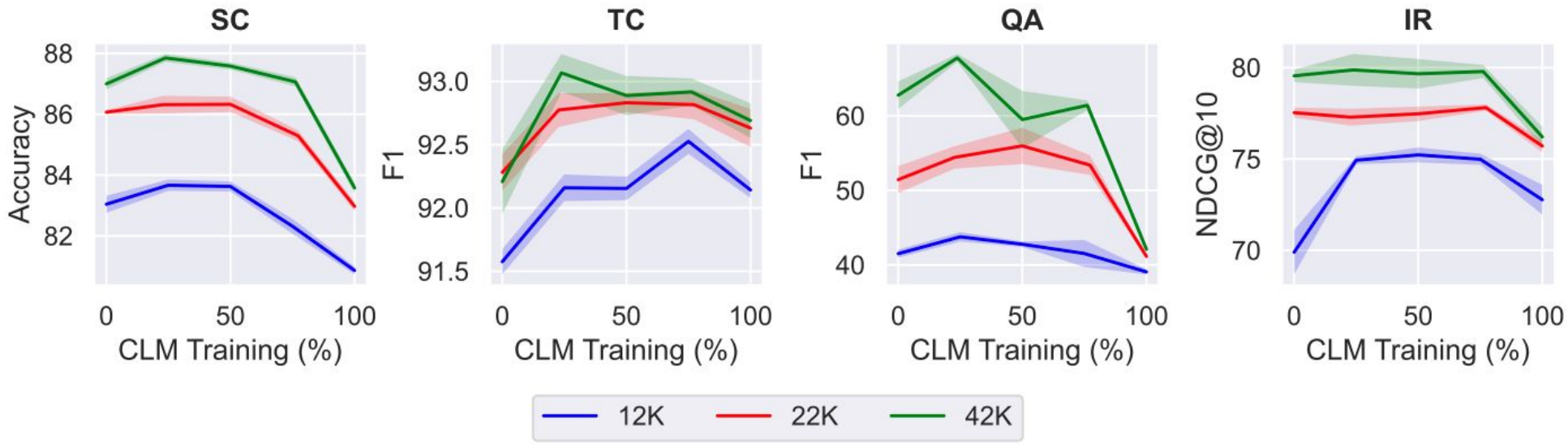




> CLM is more **data-efficient** than MLM (left)
> CLM yields models that are **easier to fine-tune** than MLM (right)

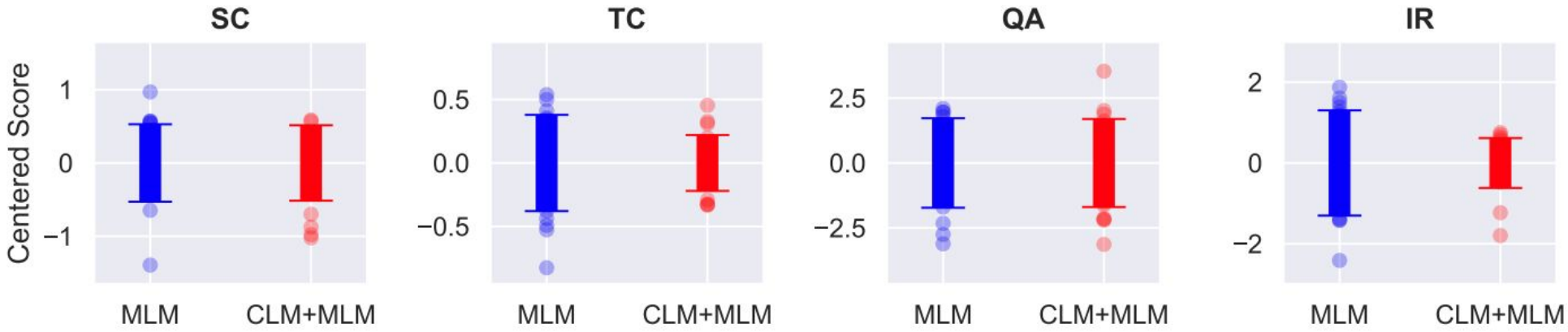
XP #2 – Two-Stage Pretraining (1/2)

Question: Under a fixed computational budget, how to optimally combine the MLM and CLM pretraining objectives?



> **MLM alone is never optimal**
> **Starting with 25% CLM steps** is generally effective, though task-specific differences still emerge

XP #2 - Two-Stage Pretraining (2/2)

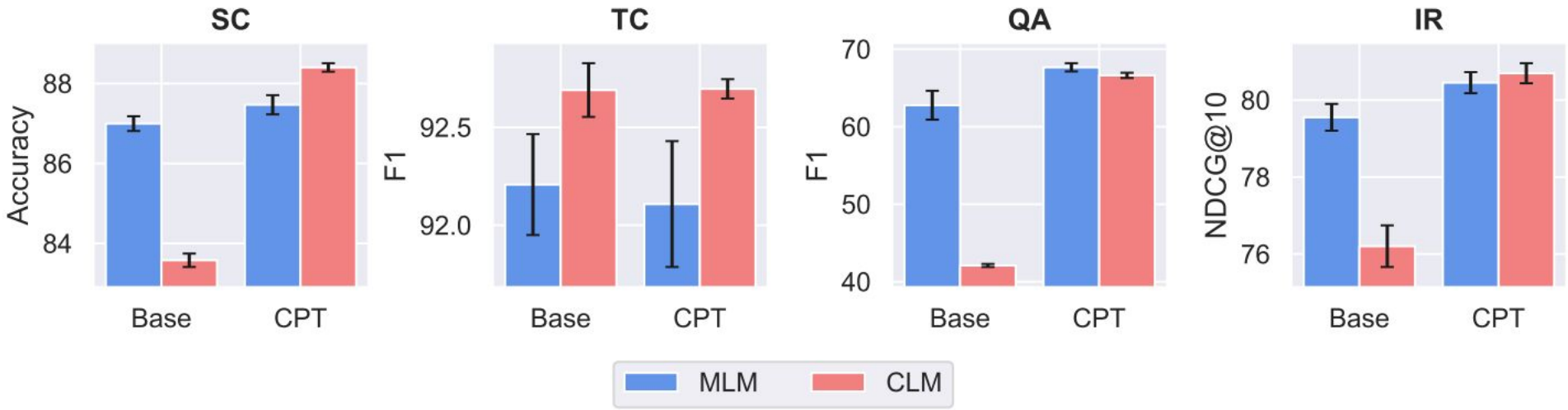




> Interestingly, starting with CLM **reduces the sensitivity** of MLM to the masking ratio

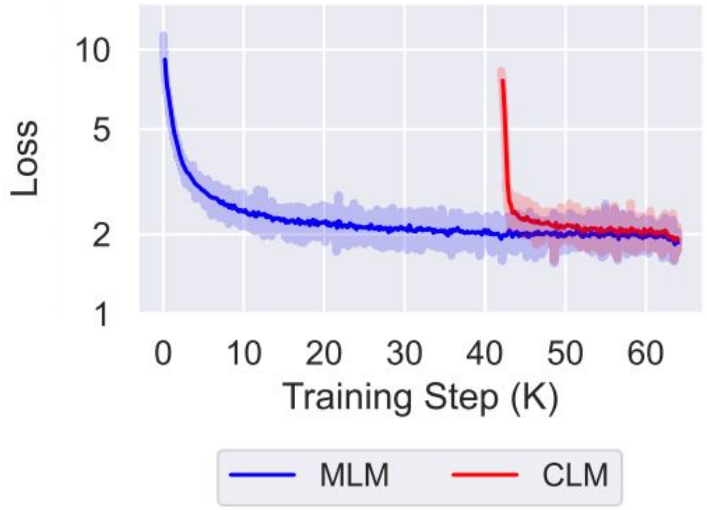
XP #3 - Continued Pretraining (1/2)

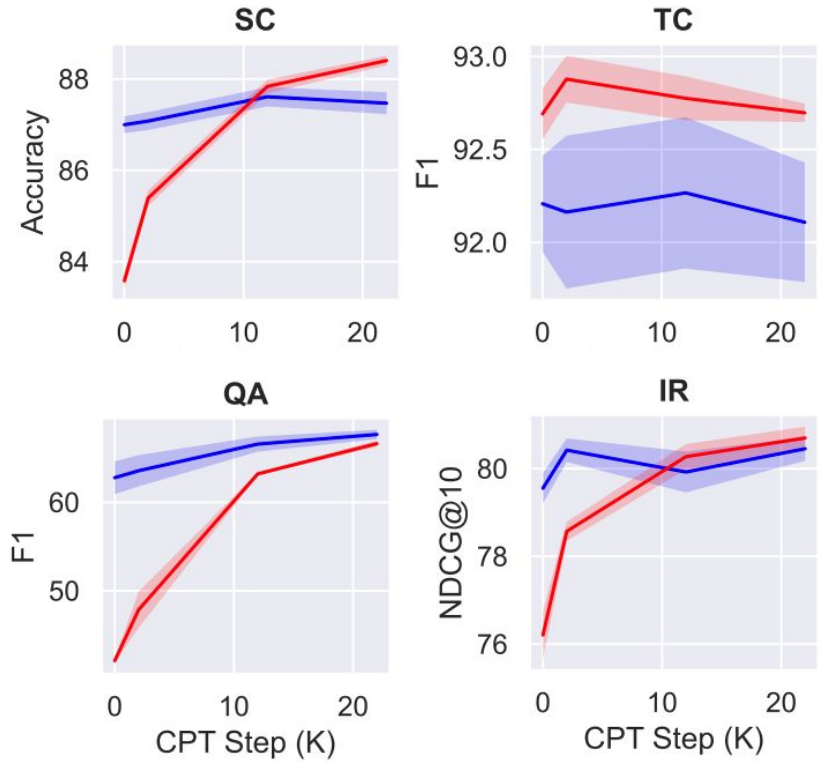
Question: If we want to continue pretraining an existing model, should we start from an MLM or a CLM checkpoint?



> Overall, **adapting a CLM model with MLM yields better results** than continuing MLM on an MLM-pretrained model.

XP #3 - Continued Pretraining (2/2)

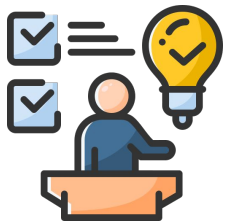






> **A few CPT steps** suffice to reach competitive performance
> Models pretrained with CLM show a **more promising trajectory**

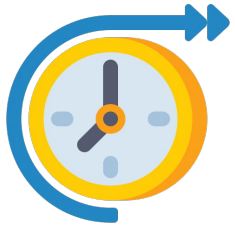
Conclusions & Future Work



Conclusions

> Causal pretraining leads to **more stable models**, making adaptation and fine-tuning easier

> When appropriately adapted and given the same data budget, CLM can yield **stronger representations** than MLM alone



Future work

> **Scaling up** both model size and training data to better understand objective dynamics at larger

> Extending the analysis to other domains, such as **VLMs** and multimodal pretraining







Thank you for your attention!