Novelty Assessment Report

Paper: Should We Still Pretrain Encoders with Masked Language Modeling?

PDF URL: https://openreview.net/pdf?id=jpz7e3jhRq

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Learning high-quality text representations is fundamental to a wide range of NLP tasks. While encoder pretraining has traditionally relied on Masked Language Modeling (MLM), recent evidence suggests that decoder models pretrained with Causal Language Modeling (CLM) can be effectively repurposed as encoders, often surpassing traditional encoders on text representation benchmarks. However, it remains unclear whether these gains reflect an inherent advantage of the CLM approach or arise from confounding factors such as model and data scale. In this paper, we address this question through a series of large-scale, carefully controlled pretraining ablations, training a total of 38 models ranging from 210 million to 1 billion parameters, and conducting over 15,000 fine-tuning and evaluation runs. We find that while training with MLM generally yields better performance across text representation tasks, CLM-trained models are more data-efficient and demonstrate improved fine-tuning stability. Building on these findings, we experimentally show that a biphasic training strategy that sequentially applies CLM and then MLM, achieves optimal performance under a fixed computational training budget. Moreover, we demonstrate that this strategy becomes more appealing when initializing from readily available pretrained CLM models, reducing the computational burden needed to train best-in-class encoder models. We release all project artifacts at \url{https://huggingface.co/XXX} to foster further research.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Comparing Masked Language Modeling and Causal Language Modeling for Encoder Pretraining**

A total of **50 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Pretraining Objectives and Architectures for Language Models**
- **Autoregressive and Bidirectional Modeling for Sequence Generation**
- **Application Domains of Pretrained Language Models**
- **Surveys, Comparative Studies, and Educational Resources**
- **Non-NLP Sequence Modeling Applications**

Complete Taxonomy Tree

- Comparing Masked Language Modeling and Causal Language Modeling for Encoder Pretraining Survey Taxonomy
- Pretraining Objectives and Architectures for Language Models
  - Direct Comparisons of MLM and CLM Pretraining ★ (5 papers)
  - [0] Should We Still Pretrain Encoders with Masked Language Modeling? (Anon et al., 2026) View paper
  - [8] What language model architecture and pretraining objective works best for zero-shot generalization? (T Wang, 2022) View paper
  - [12] What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization? (Wang, 2022) View paper
  - [16] Scaling Behavior of Encoder Language Models in Low-Resource Settings (R Visser, 2025) View paper
  - [20] On the role of bidirectionality in language model pre-training (Mikel Artetxe, 2022) View paper
  - Hybrid and Sequential Pretraining Strategies (5 papers)
  - [24] GPT or BERT: why not both? (Charpentier, 2024) View paper
  - [27] CMLM-CSE: Based on Conditional MLM Contrastive Learning for Sentence Embeddings (Wei Zhang, 2023) View paper
  - [31] Recipes for Sequential Pre-training of Multilingual Encoder and Seq2Seq Models (Saleh Soltan, 2023) View paper
  - [39] Low resource neural machine translation from english to khasi: A transformer-based approach (N. D. J. Thabah, 2021) View paper
  - [40] Causal and Masked Language Modeling of Javanese Language using Transformer-based Architectures (Wilson Wongso, 2021) View paper
  - Bidirectional Context Modeling in Autoregressive Frameworks (4 papers)
  - [5] Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer (Li, 2023) View paper
  - [19] Bi-apc: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's asr (Ruchao Fan, 2021) View paper
  - [33] The Bidirectional Awareness Induction in Autoregressive Sequence-To-Sequence Models (Jia Cheng Hu, 2025) View paper
  - [48] XLNet: Generalized Autoregressive Pretraining for Language Understanding (Yang Zhilin, 2019) View paper
  - Encoder-to-Decoder Adaptation and Transformation (2 papers)
  - [46] Entangled Bidirectional Encoder to Autoregressive Decoder for Sequential Recommendation (Taegwan Kang, 2021) View paper
  - [49] Extrapolating Multilingual Understanding Models as Multilingual Generators (Bohong Wu, 2023) View paper
- Autoregressive and Bidirectional Modeling for Sequence Generation
  - Autoregressive Generation with Unidirectional Attention (4 papers)

  - [7] SeqTrack: Sequence to Sequence Learning for Visual Object Tracking (Chen Xin, 2023) View paper
  - [37] Randomized Autoregressive Visual Generation (Yu, 2024) View paper
  - [41] GReF: A Unified Generative Framework for Efficient Reranking via Ordered Multi-token Prediction (Zhijie Lin, 2025) View paper
  - [43] Towards Sequence Modeling Alignment between Tokenizer and Autoregressive Model (Wu Pingyu, 2025) View paper
  - Bidirectional Autoregressive and Diffusion-Based Generation (6 papers)
  - [1] From slow bidirectional to fast autoregressive video diffusion models (Tianwei Yin, 2025) View paper
  - [2] Bamm: Bidirectional autoregressive motion model (Ekkasit Pinyoanuntapong, 2024) View paper
  - [3] Bad: Bidirectional auto-regressive diffusion for text-to-motion generation (Seyed Rohollah Hosseyni, 2025) View paper
  - [4] Diverse image inpainting with bidirectional and autoregressive transformers (Yingchen Yu, 2021) View paper
  - [15] Dense Policy: Bidirectional Autoregressive Learning of Actions (Su Yue, 2025) View paper
  - [21] Bidirectional Autoregressive Diffusion Model for Dance Generation (Canyu Zhang, 2024) View paper
  - Bidirectional Representations for Autoregressive Biological Sequences (2 papers)
  - [29] Bidirectional Representations Augmented Autoregressive Biological Sequence Generation:Application in De Novo Peptide Sequencing (Zhang Xiang, 2025) View paper
  - Hybrid Generative-Deterministic Rendering and Synthesis (2 papers)
  - [32] UMAMI: Unifying Masked Autoregressive Models and Deterministic Rendering for View Synthesis (Thanh-Tung Le, 2025) View paper
  - [38] EarthMapper: Visual Autoregressive Models for Controllable Bidirectional Satellite-Map Translation (Dong Zhe, 2025) View paper
- Application Domains of Pretrained Language Models
  - Text Classification and Sentiment Analysis (7 papers)
  - [11] ACOA: Archimedes conditional autoregressive optimization algorithm based RMDL for web data classification (Kante Ramesh, 2024) View paper
  - [17] Effective Analysis of Machine and Deep Learning Methods for Diagnosing Mental Health Using Social Media Conversations (Yashwanth Kasanneni, 2025) View paper
  - [23] InstructNet: A novel approach for multi-label instruction classification through advanced deep learning (Tanjim Taharat Aurpa, 2024) View paper
  - [26] Crypto-sentiment Detection in Malay Text Using Language Models with an Attention Mechanism (Nur Azmina Mohamad Zamani, 2023) View paper
  - [30] Comparison of BERT and XLNet accuracy with classical methods and algorithms in text classification (Neli Arabadzhieva - Kalcheva, 2022) View paper
  - [42] Offensive Language Detection on Social Media Using XLNet (Alothman Reem, 2025) View paper
  - [44] From BERT to Qwen: Hate Detection across architectures (Ariadna Mon, 2025) View paper
  - Sequence Labeling and Scoring (2 papers)
  - [10] Lightweight cross-lingual sentence representation learning (Zhuoyuan Mao, 2021) View paper
  - [28] Transcormer: Transformer for Sentence Scoring with Sliding Language Modeling (Song, 2022) View paper
  - Medical and Healthcare NLP (2 papers)
  - [9] Towards the Anonymization of the Language Modeling (Boutet, 2025) View paper
  - [14] Urinary Bladder Acute Inflammations and Nephritis of the Renal Pelvis: Diagnosis Using Fine-Tuned Large Language Models (M. Ma'aitah, 2025) View paper
  - Multimodal and Cross-Modal Applications (3 papers)
  - [6] Grid and Road Expressions Are Complementary for Trajectory Representation Learning (Silin Zhou, 2024) View paper
  - [18] A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network (Ying Lu, 2025) View paper
  - [47] BARTPredict: Empowering IoT Security with LLM-Driven Cyber Threat Prediction (Alaeddine Diaf, 2025) View paper
  - Diffusion Models for Text Embeddings (1 papers)
  - [35] Diffusion vs. Autoregressive Language Models: A Text Embedding Perspective (Zhang Siyue, 2025) View paper
- Surveys, Comparative Studies, and Educational Resources (4 papers)
  - [13] A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT (Ce Zhou, 2023) View paper
  - [22] Comparative Evaluation of GPT, BERT, and XLNet: Insights into Their Performance and Applicability in NLP Tasks (Zhou, 2024) View paper
  - [25] Impact of Transformer-Based Models in NLP: An In-Depth Study on BERT and GPT (Mustafa SalÄ±cÄ±, 2024) View paper
  - [50] ENHANCING EDUCATIONAL AI EDUCHAT WITH FINE-TUNED OPEN-SOURCE LANGUAGE MODELS (Sadiq, 2025) View paper
- Non-NLP Sequence Modeling Applications (2 papers)
  - [34] Ensemble Bidirectional Long Short-Term Memory Network Identification for Nonlinear Autoregressive Exogenous Model: Application to Dual Double-Acting Piston Pump (Cong Toai Truong, 2025) View paper
  - [45] Dynamic Positional Attention Fusion (DPAF): Adaptive Encoding and Weighted Attention for Ship Motion Attitude Prediction (Huachuan Zhao, 2024) View paper

Narrative

Core task: Comparing masked language modeling and causal language modeling for encoder pretraining. The field structure reflects a broad investigation into how different pretraining objectives shape language model capabilities. The taxonomy organizes work into several main branches: one focused on pretraining objectives and architectures, examining foundational design choices such as masked versus causal modeling; another on autoregressive and bidirectional modeling for sequence generation, exploring how directionality affects downstream performance; a third on application domains where pretrained models are deployed; and additional branches covering surveys and non-NLP sequence modeling. Within the pretraining objectives branch, many studies directly compare MLM and CLM strategies, assessing trade-offs in representation quality, computational efficiency, and task-specific performance. Representative works like Zero Shot Generalization[8] and Zero Shot Architecture[12] illustrate how architectural decisions influence generalization, while Encoder Low Resource[16] highlights the importance of pretraining choices in constrained settings.

A particularly active line of work examines the role of bidirectionality in pretraining, with studies such as Bidirectionality Pretraining Role[20] and Bidirectional Awareness Induction[33] investigating how bidirectional context improves understanding tasks. Pretrain Encoders MLM[0] sits squarely within this cluster, directly comparing MLM and CLM for encoder pretraining and emphasizing the empirical advantages of masked objectives for certain encoder architectures. This contrasts with works like GPT or BERT[24] and BERT GPT Impact[25], which take a broader view of the historical and practical distinctions between these paradigms. Meanwhile, papers such as CMLM Sentence Embeddings[27] explore hybrid or conditional masked approaches, suggesting that the boundary between pure MLM and CLM is not always sharp. The central question remains how to balance the rich bidirectional context of MLM with the simplicity and scalability of causal modeling, especially as models are adapted to diverse downstream tasks.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. What language model architecture and pretraining objective works best for zero-shot generalization?

Authors: T Wang, A Roberts, D Hesslow | **Year/Venue**: 2022 | **URL**: View paper

Abstract

â¦ language modeling and masked language modeling objectives applied to causal/non-causal decoder-only and encoder-â¦ , for encoder-decoder with prefix language modeling, we have â¦

Relationship Analysis

Both papers belong to the category of direct comparisons between masked language modeling (MLM) and causal language modeling (CLM) for encoder pretraining, conducting large-scale empirical evaluations of these objectives. They overlap in examining how MLM and CLM affect downstream task performance, data efficiency, and training stability, with both exploring sequential training strategies. The key difference is that the original paper focuses specifically on encoder models and investigates biphasic CLM-then-MLM pretraining and continued pretraining scenarios, while the candidate paper examines a broader range of architectures (causal decoder, non-causal decoder, encoder-decoder) and evaluates zero-shot generalization with and without multitask finetuning.

2. What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

Authors: Wang, Thomas, Thomas J. Wang, Roberts, Adam, et al. (25 authors total) | **Year/Venue**: 2022 • International Conference on Machine Learning | **URL**: View paper

Abstract

Large pretrained Transformer language models have been shown to exhibit zero-shot generalization, i.e. they can perform a wide variety of tasks that they were not explicitly trained on. However, the architectures and pretraining objectives used across state-of-the-art models differ significantly, and there has been limited systematic comparison of these factors. In this work, we present a large-scale evaluation of modeling choices and their impact on zero-shot generalization. In particular, we f...

Relationship Analysis

Both papers belong to the category of direct comparisons between MLM and CLM pretraining objectives for encoder models, conducting controlled empirical evaluations. They overlap in examining how causal versus masked language modeling affects downstream task performance through systematic ablations. However, the original paper focuses specifically on encoder pretraining with a biphasic CLM-then-MLM strategy and continued pretraining scenarios, while the candidate paper evaluates zero-shot generalization across different architectures (causal/non-causal decoders, encoder-decoders) with and without multitask prompted finetuning, representing a broader architectural comparison beyond pure encoder pretraining.

3. Scaling Behavior of Encoder Language Models in Low-Resource Settings

Authors: R Visser, T Grobler, M Dunaiski | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Similar to previous work that shows a smooth power law relationship between optimal model performance and CLM validation loss [11, 13, 18], we find that MLM accuracy scales â¦

Relationship Analysis

Both papers belong to the category of direct comparisons between MLM and CLM pretraining objectives for encoder models. They share an overlapping focus on empirically evaluating how these pretraining paradigms affect downstream task performance and model behavior. However, the original paper conducts large-scale controlled experiments (38 models, 100B tokens) comparing MLM-only, CLM-only, and sequential CLM+MLM strategies across multiple model sizes and tasks, while the candidate paper specifically investigates scaling behavior and power law relationships in low-resource settings, focusing on how MLM accuracy scales with limited data.

4. On the role of bidirectionality in language model pre-training

Authors: Mikel Artetxe, Jingfei Du, Naman Goyal, Luke Zettlemoyer, Veselin Stoyanov | **Year/Venue**: 2022 | **URL**: View paper

Abstract

Prior work on language model pre-training has explored different architectures and learning objectives, but differences in data, hyperparameters and evaluation make a principled comparison difficult. In this work, we focus on bidirectionality as a key factor that differentiates existing approaches, and present a comprehensive study of its role in next token prediction, text infilling, zero-shot priming and fine-tuning. We propose a new framework that generalizes prior approaches, including fully...

Relationship Analysis

Both papers belong to the same taxonomy category of direct comparisons between MLM and CLM pretraining objectives for encoder models. They share overlapping focus on empirically evaluating how bidirectional (MLM) versus unidirectional (CLM) pretraining affects downstream text representation tasks through controlled experiments. The key difference is that the original paper investigates sequential two-stage CLM+MLM pretraining strategies and continued pretraining scenarios with models up to 1B parameters on 100B tokens, while the candidate paper focuses on distinguishing bidirectional context from bidirectional attention as separate factors, training models up to 6.7B parameters and evaluating across language modeling, infilling, zero-shot priming, and fine-tuning with different attention configurations.

Contributions Analysis

Overall novelty summary. This paper contributes a large-scale controlled ablation study comparing masked language modeling (MLM) and causal language modeling (CLM) for encoder pretraining, training 38 models from 210M to 1B parameters with over 15,000 evaluation runs. It resides in the 'Direct Comparisons of MLM and CLM Pretraining' leaf, which contains five papers total including this one. This leaf sits within a moderately populated branch on 'Pretraining Objectives and Architectures,' suggesting the paper addresses a well-established but not overcrowded research direction focused on foundational design choices for language model pretraining.

The taxonomy reveals closely related work in adjacent leaves: 'Hybrid and Sequential Pretraining Strategies' (five papers) explores combined objectives, while 'Bidirectional Context Modeling in Autoregressive Frameworks' (four papers) examines bidirectionality within decoder architectures. The paper's biphasic CLM-then-MLM strategy bridges these areas, connecting direct objective comparisons with sequential training methods. Neighboring branches on autoregressive generation and application domains indicate the field's broader interest in how pretraining choices propagate to downstream tasks, though this work focuses specifically on encoder-level representation quality rather than generation or task-specific deployment.

Among 30 candidates examined, the large-scale ablation study (Contribution 1) shows no clear refutation across 10 candidates, suggesting this systematic scale and control may be distinctive. However, the biphasic CLM-then-MLM strategy (Contribution 2) and the demonstration of CLM-to-MLM superiority (Contribution 3) each found one refutable candidate among 10 examined, indicating prior work on sequential or hybrid training exists within this limited search scope. The statistics suggest the ablation methodology may be more novel than the biphasic training concept itself, though the search examined only top-30 semantic matches rather than exhaustive coverage.

Based on the limited 30-candidate search, the work appears to offer methodological rigor (large-scale ablations) in a moderately explored area, while the biphasic training strategy shows some overlap with existing hybrid approaches. The taxonomy structure confirms this sits in an active but not saturated research direction, with clear boundaries separating direct comparisons from hybrid methods and application studies. A broader literature search might reveal additional sequential training precedents not captured in this top-K semantic scope.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Large-scale controlled ablation study comparing MLM and CLM for encoder pretraining

Description: The authors conduct extensive controlled experiments training 38 models from 210M to 1B parameters with both MLM and CLM objectives, performing over 15,000 fine-tuning runs to isolate the effects of pretraining paradigm from confounding factors like model scale and data size.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. AntLM: Bridging Causal and Masked Language Models

URL: View paper

Brief Assessment

AntLM Bridging[54] focuses on the BabyLM challenge with small-scale models and limited training data, not large-scale controlled ablations with 38 models from 210M-1B parameters trained on 100B tokens.

---

2. Mask-Enhanced Autoregressive Prediction: Pay Less Attention to Learn More

URL: View paper

Brief Assessment

Mask Enhanced Autoregressive[56] focuses on integrating MLM into decoder-only models for in-context retrieval tasks, not on controlled ablation studies comparing MLM versus CLM for encoder pretraining across different model scales and downstream tasks.

---

3. Unilmv2: Pseudo-masked language models for unified language model pre-training

URL: View paper

Brief Assessment

Unilmv2 Pseudo Masked[53] focuses on joint pre-training of autoencoding and partially autoregressive objectives using pseudo-masked language models, not on controlled ablation studies comparing MLM vs CLM paradigms for encoder pretraining at scale.

---

4. Diverse image inpainting with bidirectional and autoregressive transformers

URL: View paper

Brief Assessment

Diverse Image Inpainting[4] focuses on image inpainting using transformers with bidirectional and autoregressive approaches for visual content generation, not on text encoder pretraining or controlled ablation studies comparing MLM and CLM objectives.

---

5. Fusing Sentence Embeddings Into LSTM-based Autoregressive Language Models

URL: View paper

Brief Assessment

Sentence Embeddings LSTM[57] focuses on fusing sentence embeddings from masked language models into LSTM-based autoregressive models for language modeling tasks, not on controlled ablation studies comparing MLM and CLM pretraining paradigms for encoder models.

---

6. What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

URL: View paper

Brief Assessment

Zero Shot Architecture[12] focuses on zero-shot generalization across different architectures (causal/non-causal decoder-only and encoder-decoder) rather than specifically on encoder pretraining with controlled ablations of MLM vs CLM at the scale described in the original paper.

---

7. What language model architecture and pretraining objective works best for zero-shot generalization?

URL: View paper

Brief Assessment

Zero Shot Generalization[8] focuses on zero-shot task generalization across different architectures (encoder-decoder vs decoder-only) rather than controlled ablation studies specifically for encoder pretraining with MLM vs CLM objectives.

---

8. Relative position prediction as pre-training for text encoders

URL: View paper

Brief Assessment

Relative Position Prediction[55] focuses on a position-centric perspective for self-supervised learning, adapting relative position encoding paradigms. It does not present controlled ablation studies comparing MLM and CLM objectives across different model scales and training configurations.

---

9. Look ahead or look around? a theoretical comparison between autoregressive and masked pretraining

URL: View paper

Brief Assessment

Look Ahead Around[51] provides theoretical comparisons between autoregressive and masked SSL paradigms, focusing on mathematical frameworks for classification and generation tasks. The original paper conducts large-scale empirical ablations with 38 models and 15,000+ fine-tuning runs on encoder pretraining, which is a fundamentally different methodological approach.

10. Enabling autoregressive models to fill in masked tokens

URL: View paper

Brief Assessment

Autoregressive Fill Masked[52] focuses on combining pre-trained MLM and AR models for masked infilling tasks, not on controlled ablation studies comparing pretraining paradigms for encoder models across text representation benchmarks.

Contribution 2: Biphasic CLM-then-MLM pretraining strategy

Description: The authors propose and validate a two-stage pretraining approach that first applies Causal Language Modeling followed by Masked Language Modeling, demonstrating that this sequential strategy outperforms MLM-only training under fixed compute budgets.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Generative Audio Language Modeling with Continuous-valued Tokens and Masked Next-Token Prediction

URL: View paper

Brief Assessment

Audio Language Modeling[62] focuses on audio generation using causal language models with continuous-valued tokens and masked next-token prediction for audio, not text encoder pretraining with sequential CLM-then-MLM strategies.

2. AntLM: Bridging Causal and Masked Language Models

URL: View paper

Prior Art Analysis

AntLM Bridging[54] demonstrates prior work on combining CLM and MLM in a sequential manner. The candidate explicitly proposes alternating between CLM and MLM training objectives during the pretraining process, which directly addresses the core concept of combining these two paradigms sequentially. While the original paper focuses on a strict CLM-then-MLM sequence, the candidate shows that the idea of combining these objectives in training was already explored, challenging the novelty of the biphasic strategy itself.

Evidence

Evidence 1 - **Rationale**: Both papers propose combining CLM and MLM paradigms to leverage their respective advantages, demonstrating that the concept of integrating these two approaches was already explored. - **Original**: we experimentally show that a biphasic training strategy that sequentially applies clm and then mlm, achieves optimal performance under a fixed computational training budget - **Candidate**: we propose a novel language modeling paradigm named $\textbf{antlm}$, which integrates both clm and mlm to leverage the advantages of these two classic paradigms

Evidence 2 - **Rationale**: The candidate describes alternating between CLM and MLM objectives during training, which demonstrates prior work on combining these paradigms in a sequential or alternating manner, directly related to the biphasic strategy. - **Original**: clm+mlm sequentially combines the former approaches in a two-stage setup, where clm pretraining is performed first, followed by mlm - **Candidate**: During the training process for specific foundation models, we alternate between applying clm or mlm training objectives and causal or bidirectional attention masks

Evidence 3 - **Rationale**: Both papers demonstrate that combining CLM and MLM objectives improves training performance, showing that the benefit of sequential combination was already established in prior work. - **Original**: building on these findings, we experimentally show that a biphasic training strategy that sequentially applies clm and then mlm, achieves optimal performance under a fixed computational training budget - **Candidate**: experimental results show that combining the two pretraining objectives leverages their strengths, enhancing overall training performance

3. Heptapod: Language Modeling on Visual Signals

URL: View paper

Brief Assessment

Heptapod Visual Signals[66] focuses on image generation using causal transformers for visual signals, not text representation learning. It does not address sequential CLM-then-MLM pretraining strategies for NLP encoders.

4. Unified Multimodal Pre-training and Prompt-based Tuning for Vision-Language Understanding and Generation

URL: View paper

Brief Assessment

Unified Multimodal Pretraining[65] focuses on vision-language multimodal tasks using causal masks for generation capabilities, not on sequential CLM-then-MLM pretraining strategies for text-only encoder models.

5. Mask more and mask later: Efficient pre-training of masked language models by disentangling the token

URL: View paper

Brief Assessment

Mask More Later[64] focuses on a two-stage learning method for efficient masked language modeling but does not appear to explore the specific CLM-then-MLM sequential strategy that the original paper investigates. The candidate's approach and objectives differ from the original's biphasic pretraining paradigm.

6. What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?

URL: View paper

Brief Assessment

Zero Shot Architecture[12] explores adaptation of pretrained models across architectures and objectives but does not propose or validate a sequential CLM-then-MLM pretraining strategy as a primary contribution for encoder training.

7. What language model architecture and pretraining objective works best for zero-shot generalization?

URL: View paper

Brief Assessment

Zero Shot Generalization[8] explores adaptation between objectives but does not propose or validate a biphasic CLM-then-MLM pretraining strategy as a primary contribution for encoder models.

8. Causal and Masked Language Modeling of Javanese Language using Transformer-based Architectures

URL: View paper

Brief Assessment

Javanese Language Modeling[40] focuses on transfer learning techniques (transferring English embeddings, gradual unfreezing) for low-resource Javanese language modeling, not on comparing or proposing sequential CLM-then-MLM pretraining strategies under controlled conditions.

9. NormFormer: Improved Transformer Pretraining with Extra Normalization

URL: View paper

Brief Assessment

NormFormer[63] focuses on architectural improvements through additional normalization operations in transformer layers, not on sequential pretraining strategies combining CLM and MLM objectives.

10. EmbedTurk: Leveraging Large Language Models as Text Encoders for Turkish Language

URL: View paper

Brief Assessment

EmbedTurk Turkish Encoders[67] focuses on Turkish language text encoders and does not provide sufficient detail about biphasic pretraining strategies combining CLM then MLM to challenge the original paper's novelty claim.

Contribution 3: Demonstration that CLM-to-MLM continued pretraining outperforms MLM-only training

Description: The authors show that applying MLM continued pretraining to decoder models initially trained with CLM yields better text representations than continuing to train models that were pretrained with MLM, suggesting that leveraging existing pretrained decoders is the most effective path to strong encoders.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Efficient Domain-adaptive Continual Pretraining for the Process Industry in the German Language

URL: View paper

Brief Assessment

Process Industry German[60] focuses on domain-adaptive continual pretraining using in-context learning augmentation for low-resource German industrial text, not on comparing CLM-to-MLM versus MLM-only pretraining strategies for general text encoders.

2. Unilmv2: Pseudo-masked language models for unified language model pre-training

URL: View paper

Brief Assessment

Unilmv2 Pseudo Masked[53] does not investigate continued pretraining scenarios where CLM models are adapted with MLM. Instead, it proposes a unified pre-training framework combining autoencoding and partially autoregressive modeling in a single forward pass.

3. BERTs are generative in-context learners

URL: View paper

Brief Assessment

BERTs Generative Learners[59] focuses on demonstrating that existing masked language models (specifically DeBERTa) can perform in-context learning without additional training. It does not address continued pretraining strategies or compare CLM-to-MLM adaptation versus MLM-only training approaches.

4. Bamm: Bidirectional autoregressive motion model

URL: View paper

Brief Assessment

Bamm Motion Model[2] focuses on motion modeling using encoders for 3D pose sequences and codebook quantization, not on language model pretraining strategies or comparing CLM-to-MLM versus MLM-only approaches.

5. Diverse image inpainting with bidirectional and autoregressive transformers

URL: View paper

Brief Assessment

Diverse Image Inpainting[4] addresses image completion tasks using transformers for visual data, not continued pretraining strategies for text encoders or adaptation of causal language models with masked language modeling.

6. Customization of Large Language Models for Causal Inference and Data Quality

URL: View paper

Brief Assessment

Causal Inference Customization[61] focuses on adapting LLMs for causal inference and data quality tasks, not on comparing CLM-to-MLM versus MLM-only pretraining strategies for text encoders. The candidate's limited context mentions MLM training but does not address the specific research question of whether continued pretraining from CLM models outperforms continued training from MLM models.

7. Bad: Bidirectional auto-regressive diffusion for text-to-motion generation

URL: View paper

Brief Assessment

Bad Text Motion[3] focuses on text-to-motion generation using a permutation-based diffusion approach that combines autoregressive and mask-based modeling. It does not address continued pretraining strategies for text encoders or compare CLM-to-MLM adaptation versus MLM-only training paradigms.

8. What language model architecture and pretraining objective works best for zero-shot generalization?

URL: View paper

Prior Art Analysis

Zero Shot Generalization[8] demonstrates that adapting causal decoder models pretrained with full language modeling (CLM/FLM) by continuing training with masked language modeling (MLM) as non-causal decoders significantly accelerates convergence and improves performance. The paper shows that 'non-causal masked language modeling adaptation: starting from a causal decoder model pretrained with flm, we then continue training the model as a non-causal decoder using a mlm objective' achieves '3.3x' speedup in convergence. This directly demonstrates that CLM-to-MLM adaptation outperforms training with MLM alone, establishing prior work on this approach.

Evidence

Evidence 1 - **Rationale**: This pair shows that Zero Shot Generalization[8] already demonstrated adapting CLM-pretrained models with MLM continuation training, achieving faster convergence than MLM-only training, which directly refutes the novelty of the original paper's claim about CLM-to-MLM continued pretraining. - **Original**: we demonstrate that this strategy becomes more appealing when initializing from readily available pretrained clm models, reducing the computational burden needed to train best-in-class encoder models. - **Candidate**: we now introduce non-causal masked language modeling adaptation: starting from a causal decoder model pretrained with flm, we then continue training the model as a non-causal decoder using a mlm objective. this is essentially the reverse of the language modeling adaptation setup, and the conversion ...

Evidence 2 - **Rationale**: This evidence pair demonstrates that Zero Shot Generalization[8] explicitly evaluated and showed that MLM-adapted models (starting from CLM) perform better than alternatives, establishing prior demonstration of this continued pretraining approach. - **Original**: in a continued pretraining setting, adapting a clm-pretrained model with mlm proves more effective than continuing mlm training from an mlm-pretrained model. - **Candidate**: we find that the mlm-adapted model performs best by a significant margin and outperforms every other model we considered on eai-eval. furthermore, the measured zero-shot generalization is in line with the mlmpretrained non-causal decoder reported in figure 4, though it still lags behind the mlm-pret...

Evidence 3 - **Rationale**: Both papers reach the same conclusion that starting from CLM-pretrained decoder models and adapting them with MLM is an effective strategy, showing Zero Shot Generalization[8] established this finding prior to the original paper. - **Original**: This result suggests that leveraging widely pretrained decoder models is currently the best approach to obtain a strong encoder model - **Candidate**: decoder-only models can be efficiently adapted from one architecture/objective prior to the other. specifically, we recommend starting with a causal decoder-only model, pretraining it with a full language modeling objective, and then using non-causal masked language modeling adaptation before taking...

9. Denoising token prediction in masked autoregressive models

URL: View paper

Brief Assessment

Denoising Token Prediction[58] focuses on masked autoregressive models for image generation using diffusion procedures, not on continued pretraining strategies for text encoders or comparing CLM-to-MLM versus MLM-only training paradigms.

10. GPT or BERT: why not both?

URL: View paper

Brief Assessment

GPT or BERT[24] focuses on hybrid training within a single model architecture for the BabyLM challenge, not on comparing continued pretraining strategies from different base models. The candidate does not provide evidence of prior work demonstrating that CLM-to-MLM continued pretraining outperforms continuing MLM training from MLM-pretrained models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Should We Still Pretrain Encoders with Masked Language Modeling? View paper
- [1] From slow bidirectional to fast autoregressive video diffusion models View paper
- [2] Bamm: Bidirectional autoregressive motion model View paper
- [3] Bad: Bidirectional auto-regressive diffusion for text-to-motion generation View paper
- [4] Diverse image inpainting with bidirectional and autoregressive transformers View paper
- [5] Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer View paper
- [6] Grid and Road Expressions Are Complementary for Trajectory Representation Learning View paper
- [7] SeqTrack: Sequence to Sequence Learning for Visual Object Tracking View paper
- [8] What language model architecture and pretraining objective works best for zero-shot generalization? View paper
- [9] Towards the Anonymization of the Language Modeling View paper
- [10] Lightweight cross-lingual sentence representation learning View paper
- [11] ACOA: Archimedes conditional autoregressive optimization algorithm based RMDL for web data classification View paper
- [12] What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization? View paper
- [13] A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT View paper
- [14] Urinary Bladder Acute Inflammations and Nephritis of the Renal Pelvis: Diagnosis Using Fine-Tuned Large Language Models View paper
- [15] Dense Policy: Bidirectional Autoregressive Learning of Actions View paper
- [16] Scaling Behavior of Encoder Language Models in Low-Resource Settings View paper
- [17] Effective Analysis of Machine and Deep Learning Methods for Diagnosing Mental Health Using Social Media Conversations View paper
- [18] A fake news detection model using the integration of multimodal attention mechanism and residual convolutional network View paper
- [19] Bi-apc: Bidirectional autoregressive predictive coding for unsupervised pre-training and its application to children's asr View paper
- [20] On the role of bidirectionality in language model pre-training View paper
- [21] Bidirectional Autoregressive Diffusion Model for Dance Generation View paper
- [22] Comparative Evaluation of GPT, BERT, and XLNet: Insights into Their Performance and Applicability in NLP Tasks View paper
- [23] InstructNet: A novel approach for multi-label instruction classification through advanced deep learning View paper
- [24] GPT or BERT: why not both? View paper

- [25] Impact of Transformer-Based Models in NLP: An In-Depth Study on BERT and GPT View paper
- [26] Crypto-sentiment Detection in Malay Text Using Language Models with an Attention Mechanism View paper
- [27] CMLM-CSE: Based on Conditional MLM Contrastive Learning for Sentence Embeddings View paper
- [28] Transcormer: Transformer for Sentence Scoring with Sliding Language Modeling View paper
- [29] Bidirectional Representations Augmented Autoregressive Biological Sequence Generation:Application in De Novo Peptide Sequencing View paper
- [30] Comparison of BERT and XLNet accuracy with classical methods and algorithms in text classification View paper
- [31] Recipes for Sequential Pre-training of Multilingual Encoder and Seq2Seq Models View paper
- [32] UMAMI: Unifying Masked Autoregressive Models and Deterministic Rendering for View Synthesis View paper
- [33] The Bidirectional Awareness Induction in Autoregressive Sequence-To-Sequence Models View paper
- [34] Ensemble Bidirectional Long Short-Term Memory Network Identification for Nonlinear Autoregressive Exogenous Model: Application to Dual Double-Acting Piston Pump View paper
- [35] Diffusion vs. Autoregressive Language Models: A Text Embedding Perspective View paper
- [36] Bidirectional Representations Augmented Autoregressive Biological Sequence Generation View paper
- [37] Randomized Autoregressive Visual Generation View paper
- [38] EarthMapper: Visual Autoregressive Models for Controllable Bidirectional Satellite-Map Translation View paper
- [39] Low resource neural machine translation from english to khasi: A transformer-based approach View paper
- [40] Causal and Masked Language Modeling of Javanese Language using Transformer-based Architectures View paper
- [41] GReF: A Unified Generative Framework for Efficient Reranking via Ordered Multi-token Prediction View paper
- [42] Offensive Language Detection on Social Media Using XLNet View paper
- [43] Towards Sequence Modeling Alignment between Tokenizer and Autoregressive Model View paper
- [44] From BERT to Qwen: Hate Detection across architectures View paper
- [45] Dynamic Positional Attention Fusion (DPAF): Adaptive Encoding and Weighted Attention for Ship Motion Attitude Prediction View paper
- [46] Entangled Bidirectional Encoder to Autoregressive Decoder for Sequential Recommendation View paper
- [47] BARTPredict: Empowering IoT Security with LLM-Driven Cyber Threat Prediction View paper
- [48] XLNet: Generalized Autoregressive Pretraining for Language Understanding View paper
- [49] Extrapolating Multilingual Understanding Models as Multilingual Generators View paper
- [50] ENHANCING EDUCATIONAL AI EDUCHAT WITH FINE-TUNED OPEN-SOURCE LANGUAGE MODELS View paper
- [51] Look ahead or look around? a theoretical comparison between autoregressive and masked pretraining View paper
- [52] Enabling autoregressive models to fill in masked tokens View paper
- [53] Unilmv2: Pseudo-masked language models for unified language model pre-training View paper
- [54] AntLM: Bridging Causal and Masked Language Models View paper
- [55] Relative position prediction as pre-training for text encoders View paper
- [56] Mask-Enhanced Autoregressive Prediction: Pay Less Attention to Learn More View paper
- [57] Fusing Sentence Embeddings Into LSTM-based Autoregressive Language Models View paper
- [58] Denoising token prediction in masked autoregressive models View paper
- [59] BERTs are generative in-context learners View paper
- [60] Efficient Domain-adaptive Continual Pretraining for the Process Industry in the German Language View paper
- [61] Customization of Large Language Models for Causal Inference and Data Quality View paper
- [62] Generative Audio Language Modeling with Continuous-valued Tokens and Masked Next-Token Prediction View paper
- [63] NormFormer: Improved Transformer Pretraining with Extra Normalization View paper
- [64] Mask more and mask later: Efficient pre-training of masked language models by disentangling the token View paper
- [65] Unified Multimodal Pre-training and Prompt-based Tuning for Vision-Language Understanding and Generation View paper
- [66] Heptapod: Language Modeling on Visual Signals View paper
- [67] EmbedTurk: Leveraging Large Language Models as Text Encoders for Turkish Language View paper