Beyond Multi-Token Prediction: Pretraining LLMs with Future Summaries

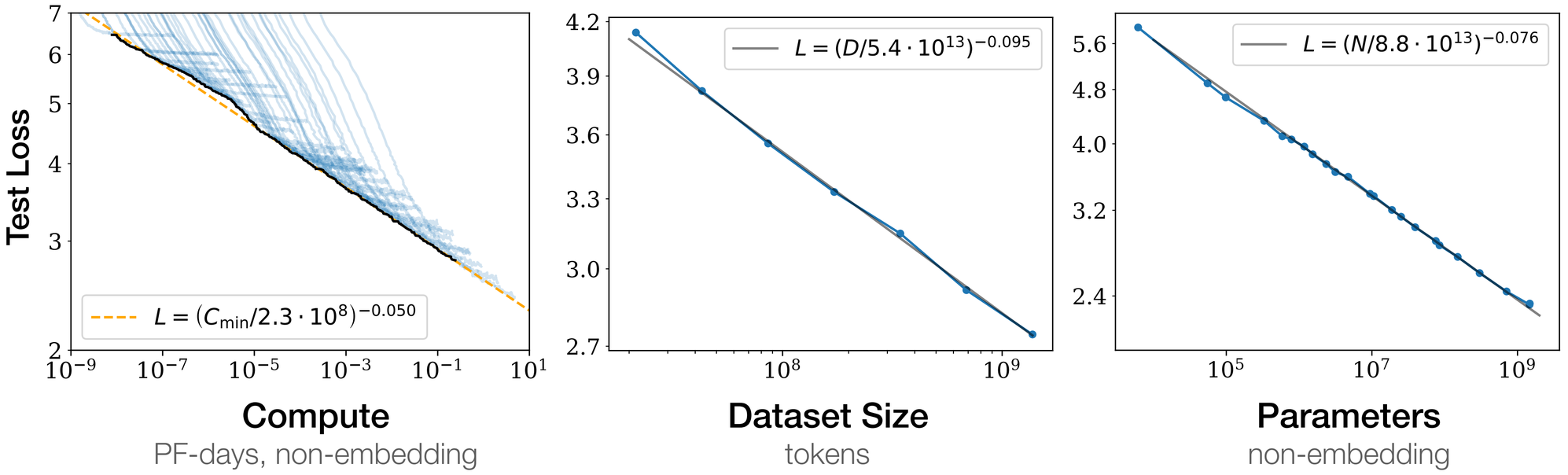
Divyat Mahajan[1], **Sachin Goyal**[2], **Badr Youbi Idrissi**[3], **Mohammad Pezeshki**[3], **Ioannis Mitliagkas**[1], **David Lopez-Paz**[3], **Kartik Ahuja**[3]

[1]Mila, Université de Montréal, [2]Carnegie Mellon University, [3]FAIR at Meta

International Conference on Learning Representations (ICLR) 2026

LLMs Scale Predictably

Scaling compute via parameters or data provides steady improvements



Scaling Laws for Neural Language Models— Kaplan et al., 2020

Challenges Ahead with Scaling

Need new recipes as scaling data is not sustainable

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

Design new objectives to extract more information from the same data

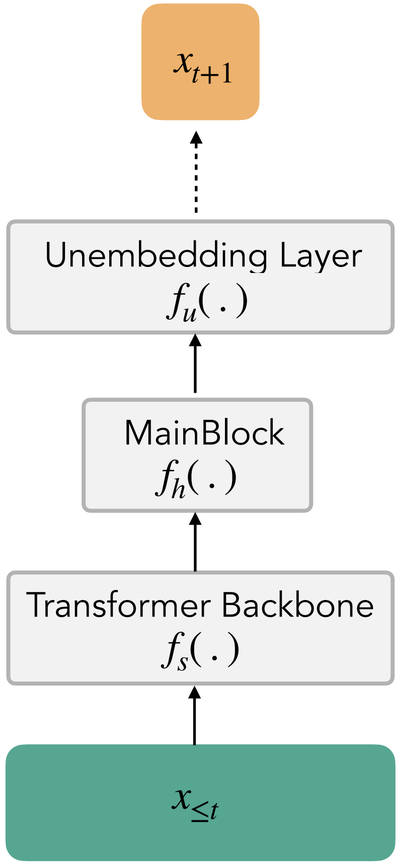
Ilya Sutskever— Test of time talk, NeurIPS 2024

Background

Going beyond Next-token Prediction

Next-Token Prediction (NTP)

NTP can in principle learn any distribution

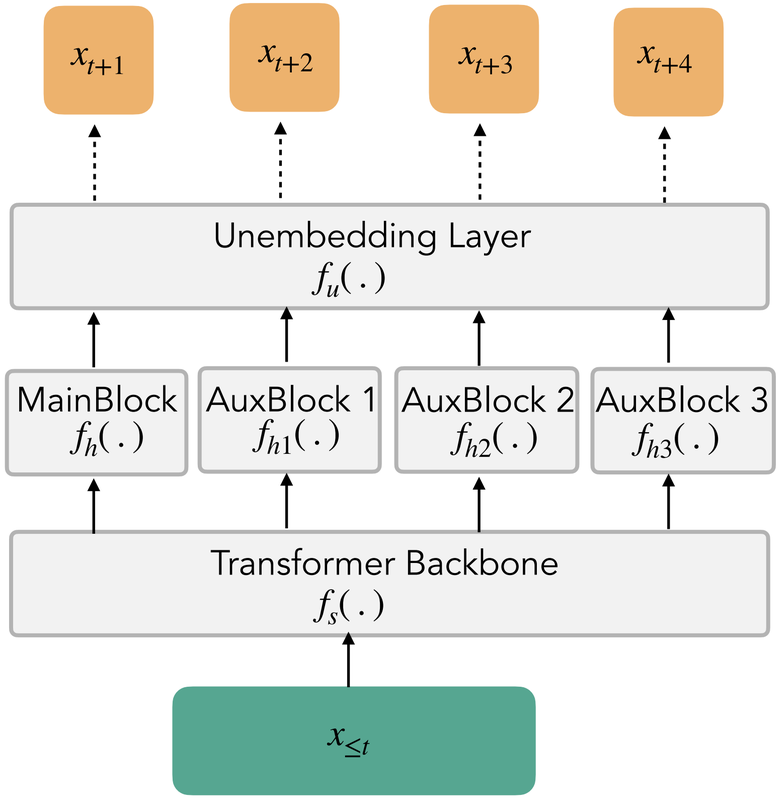


$$P_\theta(x_{t+1} \mid x_{\leq t}) = Softmax(f_u \circ f_h \circ f_s(x_{\leq t}))$$

$$L_{NTP}(x, P_\theta) = -\sum_{t=1}^{T-1} \log P_\theta(x_{t+1} \mid x_{\leq t})$$

Multi-token Prediction (MTP)

MTP provides a richer learning objective than NTP



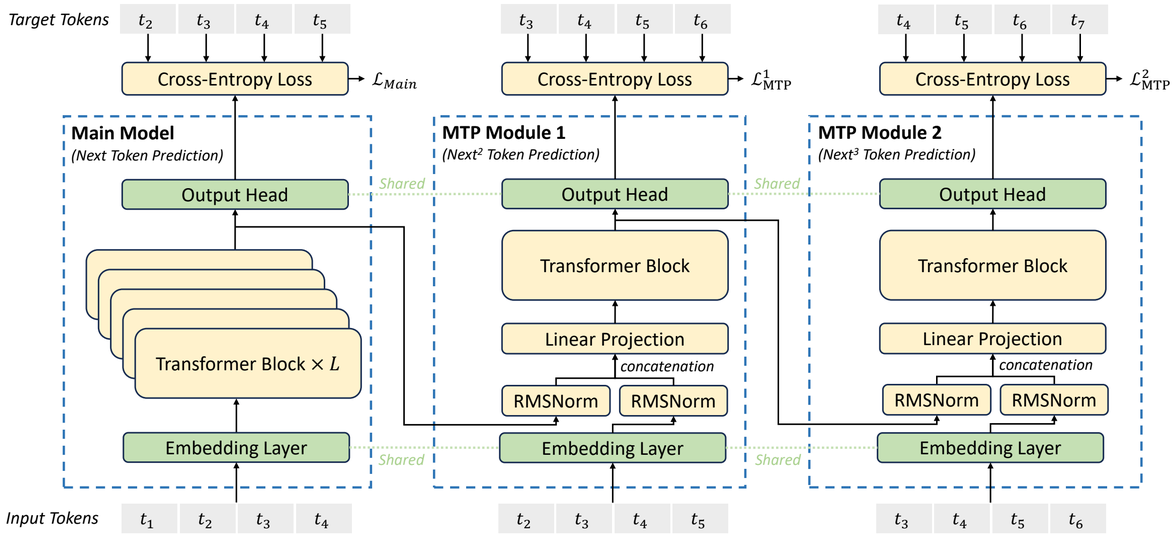
$$Softmax(f_u \circ f_h \circ f_s(x_{\leq t})) \quad Softmax(f_u \circ f_{h1} \circ f_s(x_{\leq t})) \quad Softmax(f_u \circ f_{h2} \circ f_s(x_{\leq t})) \quad Softmax(f_u \circ f_{h3} \circ f_s(x_{\leq t}))$$

$$L_{MTP}(x, P_\theta) = -\sum_{t=1}^{T-1} \log P_\theta(x_{t+1} \mid x_{\leq t}) \; -\sum_{t=1}^{T-2} \log P_\theta(x_{t+2} \mid x_{\leq t}) \; -\sum_{t=1}^{T-3} \log P_\theta(x_{t+3} \mid x_{\leq t}) \; -\sum_{t=1}^{T-4} \log P_\theta(x_{t+4} \mid x_{\leq t})$$

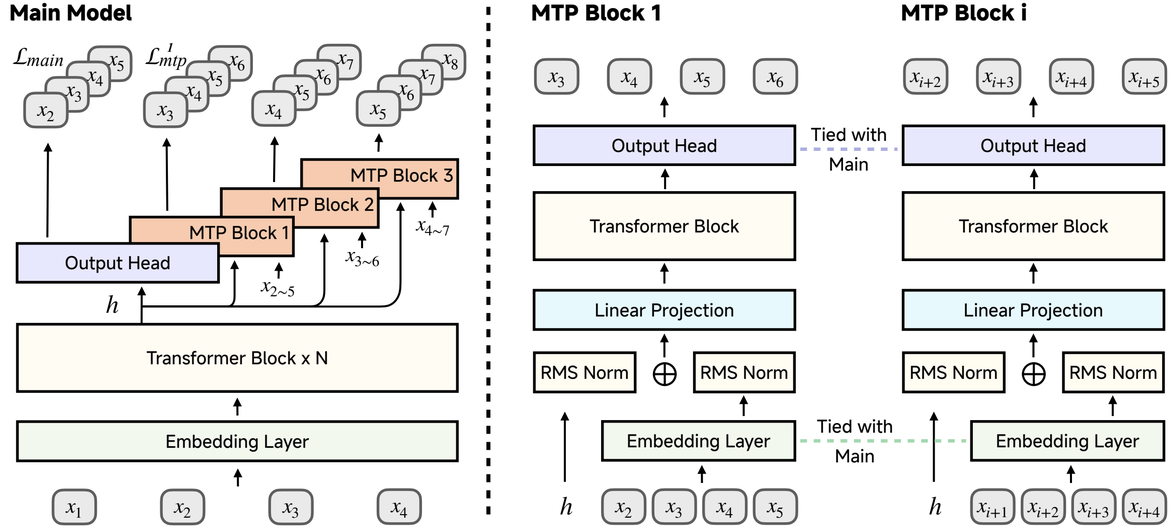
Better & Faster Language Models via Multi-token Prediction— Gloeckle et al., 2024

Multi-Token Prediction

Strong open source models are using MTP (variants)



DeepSeek-V3 Technical Report— DeepSeekAI, 2025



MiMo: Unlocking reasoning potential of Language Model— Xiaomi., 2025

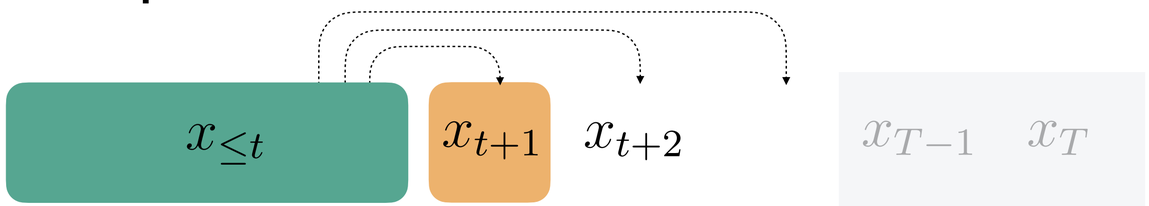
Future Summary Prediction

Going beyond Multi-Token Prediction

Issue with MTP: Scaling Prediction Horizon

Need auxiliary head for every additional future token

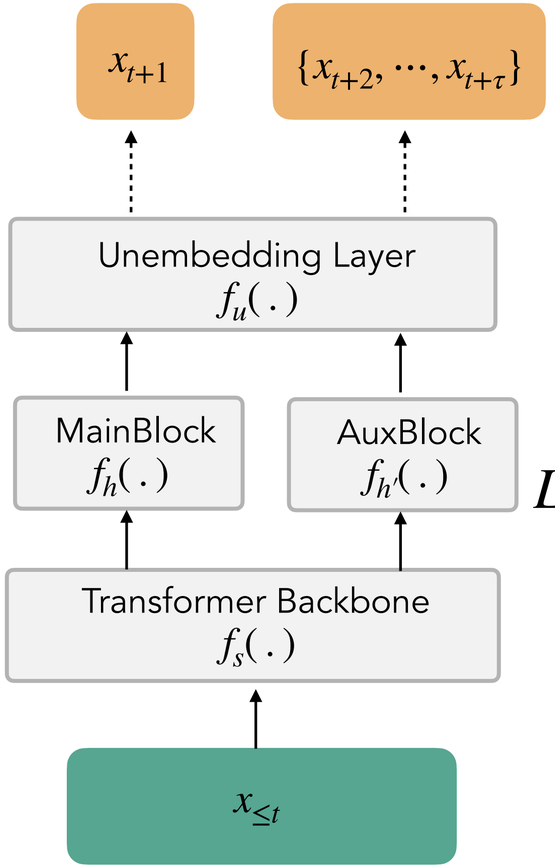
MTP: Uses multiple auxiliary heads, each predicting a specific future token



Instead of the entire future sequence, lets predict a future summary!

Future Summary Prediction: Bag-of-words (FSP-BoW)

Auxiliary target as bag-of-words summary of future (**Single auxiliary head!**)



$$Softmax(f_u \circ f_h \circ f_s(x_{\leq t})) \qquad Sigmoid(f_u \circ f_{h'} \circ f_s(x_{\leq t}))$$

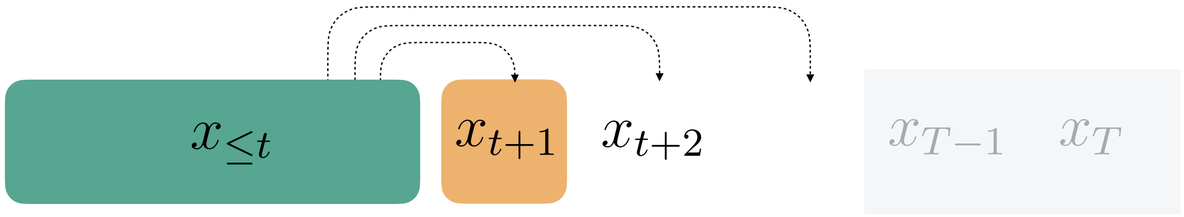
$$L_{FSP-BoW}(x, P_\theta) = -\sum_{t=1}^{T-1} \log P_\theta(x_{t+1} \mid x_{\leq t}) + BCE\big(A_\theta(x_{\leq t}), \{x_{t+2}, \cdots, x_{t+\tau}\} \big)$$

Binary Cross Entropy Loss

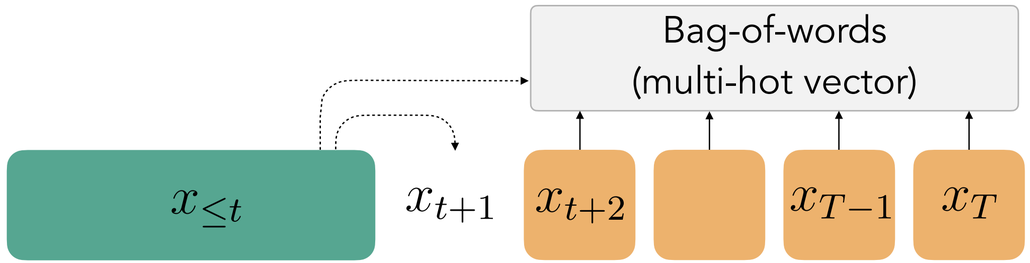
Issue with MTP: How much lookahead?

Future can have irrelevant or noise tokens

MTP: Uses multiple auxiliary heads, each predicting a specific future token



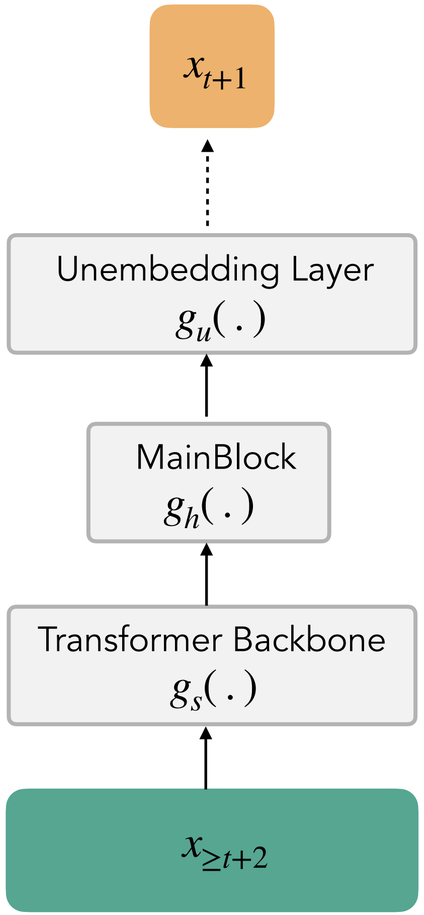
FSP-BoW: Predicts a "bag-of-tokens" summary

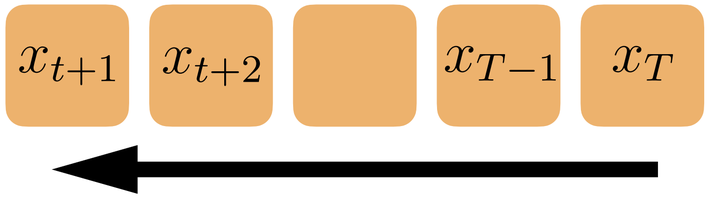


Need to learn a future summary instead of hand-crafted choices!

Future Summary Prediction: ReverseLM (FSP-RevLM)

Step 1. Train language model on reverse sequences (RevLM)





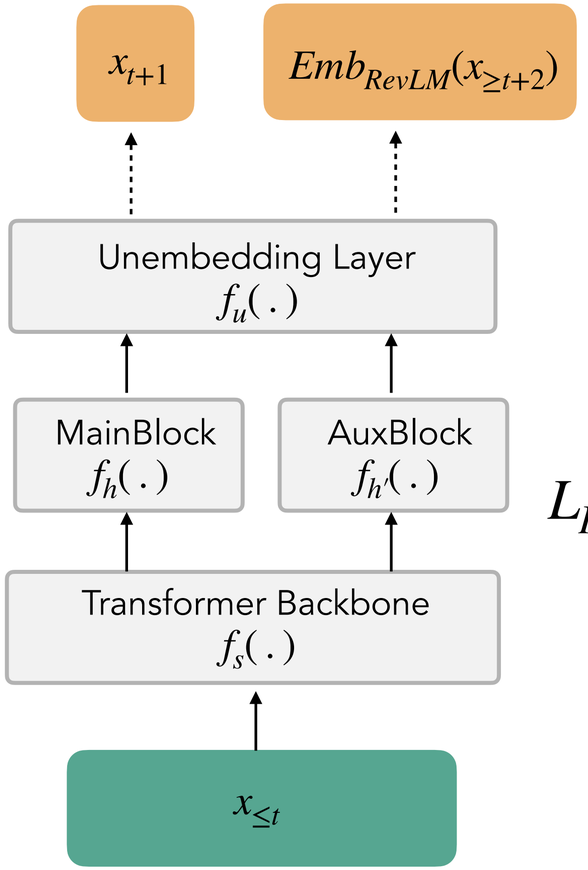
$$Q_\phi(x_{t+1} \mid x_{\geq t+2}) = Softmax(g_u \circ g_h \circ g_s(x_{\geq t+2}))$$

$$L_{RevLM}(x, Q_\phi) = -\sum_{t=0}^{T-2} \log Q_\phi(x_{t+1} \mid x_{\geq t+2})$$

Future Summary: $Emb_{RevLM}(x_{\geq t+2}) = g_h \circ g_s(x_{\geq t+2})$

Future Summary Prediction: ReverseLM (FSP-RevLM)

Step 2. Auxiliary target as a learned embedding of future (Single auxiliary head!)



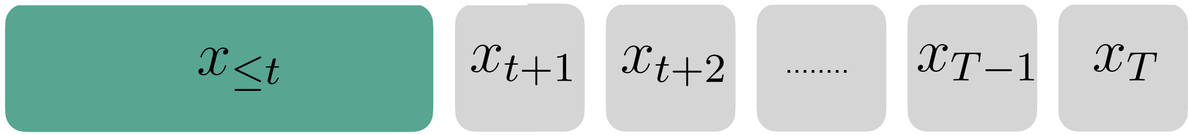
$$Softmax(f_u \circ f_h \circ f_s(x_{\leq t}))$$

$$f_{h'} \circ f_s(x_{\leq t})$$

$$L_{FSP-RevLM}(x, P_\theta) = -\sum_{t=1}^{T-1} \log P_\theta(x_{t+1} \mid x_{\leq t}) + || A_\theta(x_{\leq t}), Emb_{RevLM}(x_{\geq t+2}) ||_2^2$$

Embedding Matching Loss

Future Summary Prediction

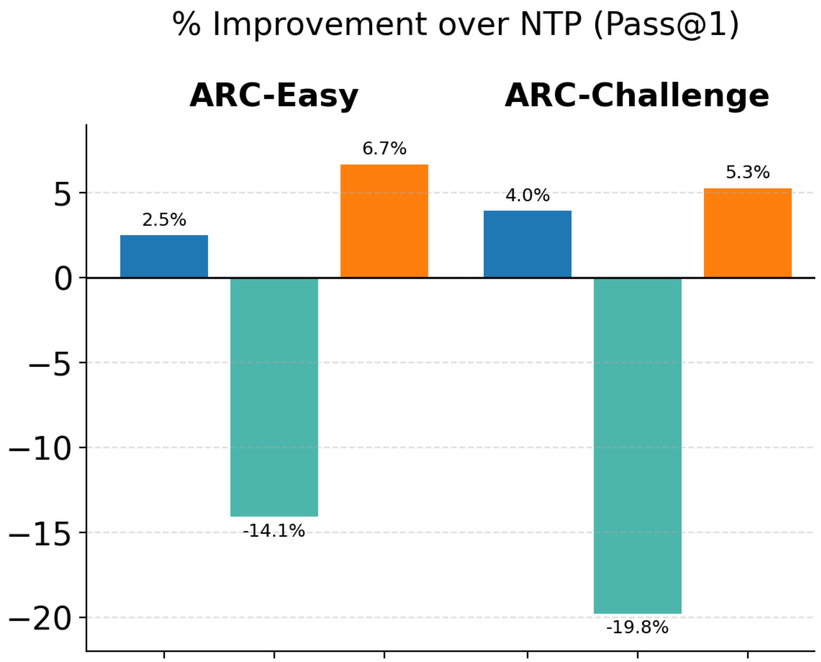


Future Summary Prediction

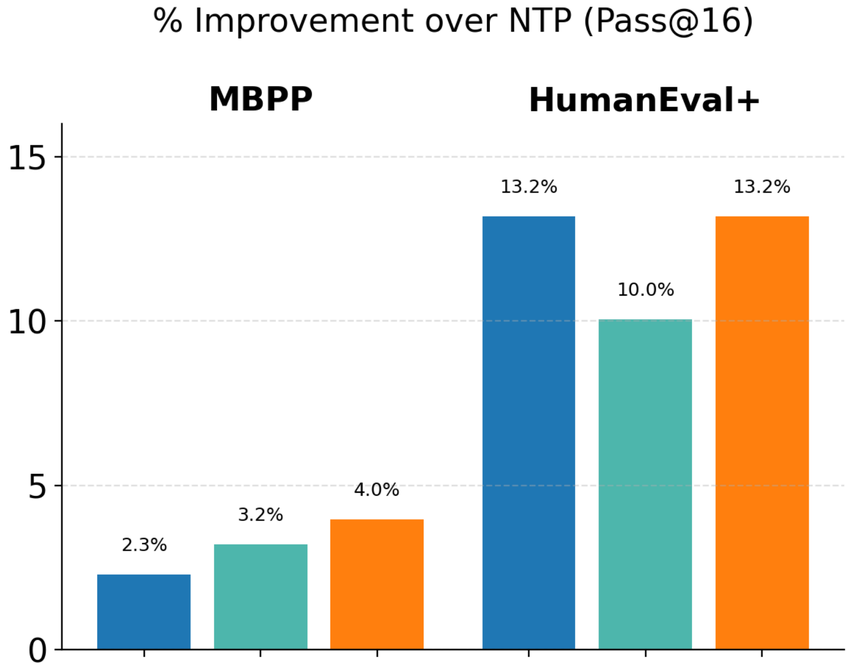
Real-world pretraining experiments

Pretraining Results: 8B

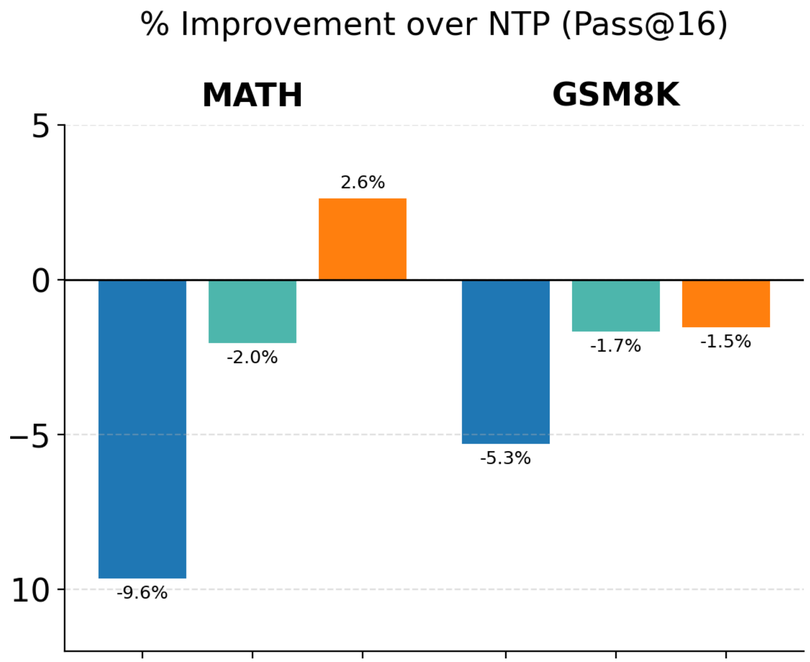
General Reasoning



Coding



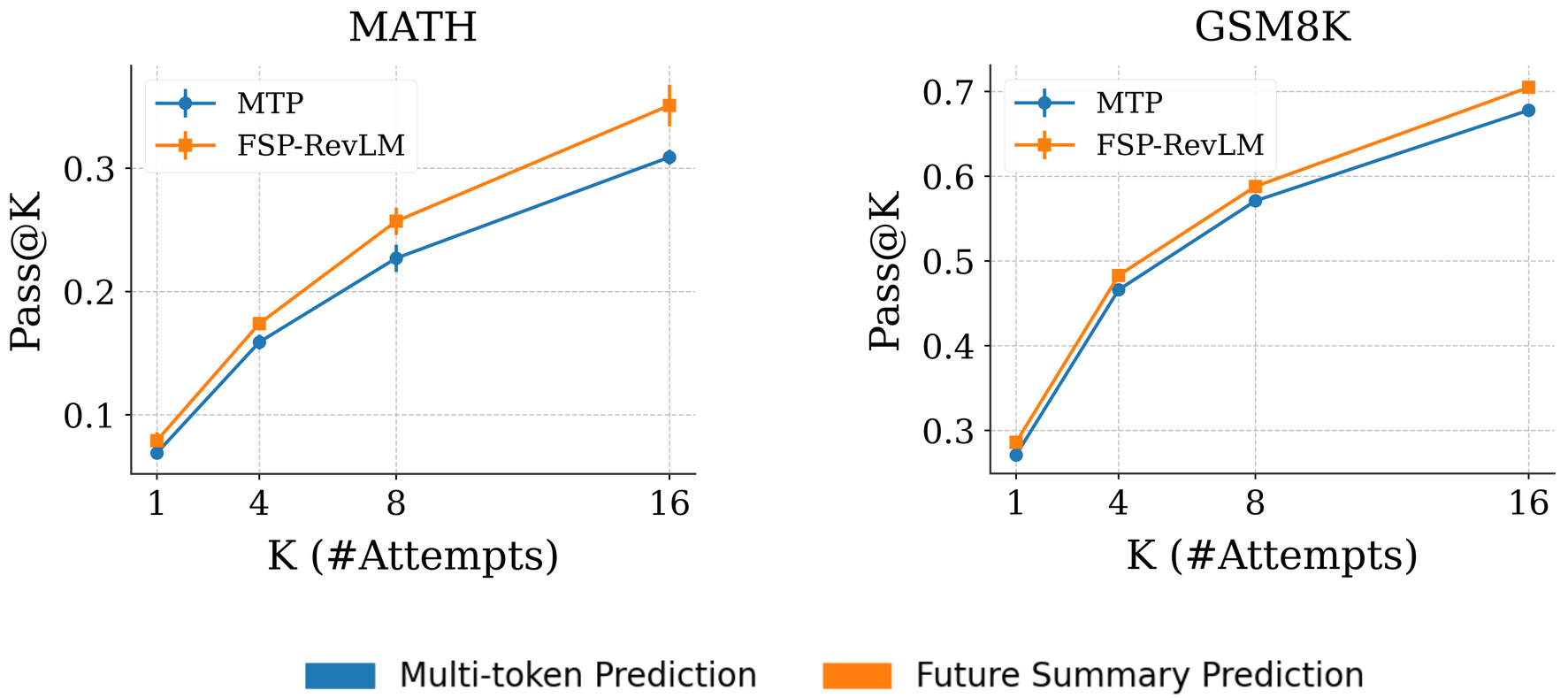
Math Reasoning



Multi-token Prediction DS Multi-token Prediction Future Summary Prediction

Pretraining Results: 8B

Future Summary Prediction leads to more diversity than MTP



Chat with us during the poster session!

Experiment Setup

- Data (DCLM, Github, Proof Pile, etc.) & Architecture (LLaMA 3)

- Scale
  - 3B Parameters, 250B Tokens
  - 8B Parameters, 1T Tokens

- Auxiliary Heads
  - Training: Single auxiliary head for MTP & DS-MTP for fair comparison with FSP
  - Inference: Discard the auxiliary head and only use the next-token (main) head

- ReverseLM (Teacher)
  - Same model size and trained on the same dataset as the baselines