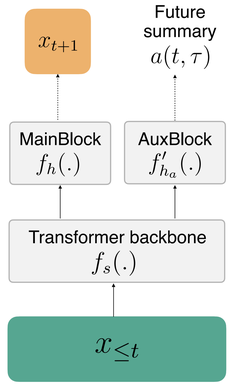
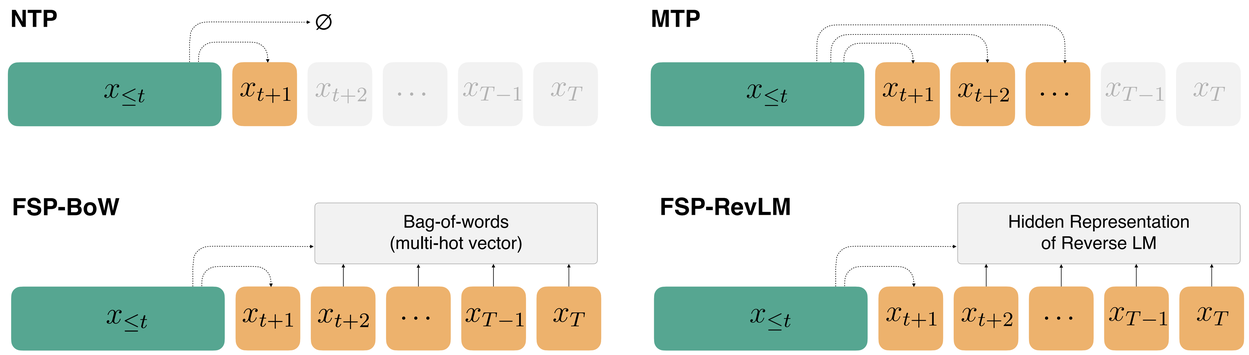
Beyond Multi-token Prediction: Pretraining LLMs with Future Summaries

Divyat Mahajan, Sachin Goyal, Badr Youbi Idrissi, Mohammad Pezeshki, Ioannis Mitliagkas, David Lopez-Paz, Kartik Ahuja

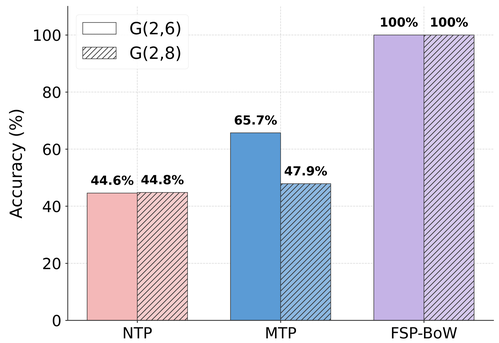
Contributions

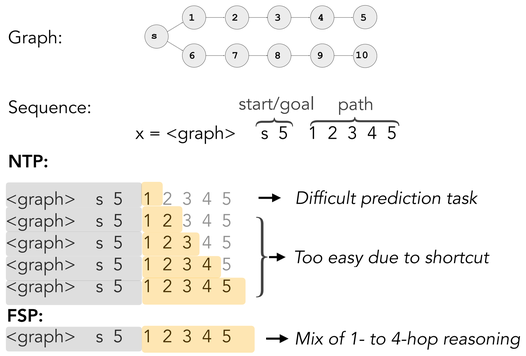
- Limitations of *Next-token prediction* (NTP) due to teacher forcing
  - Exposure bias & Shortcut learning
- *Multi-token prediction* (MTP): Predict short blocks of future tokens to reduce teacher forcing
  - One auxiliary head per future token is not scalable
- Proposal: *Future summary prediction* (FSP) to capture long-range dependencies
  - *Single auxiliary head* predicts a compact summary of future sequences
- Techniques for future summary prediction
  - *Hand-crafted summary:* Bag-of-words as auxiliary target
  - *Learned summary:* Embedding from reverse language model as auxiliary target



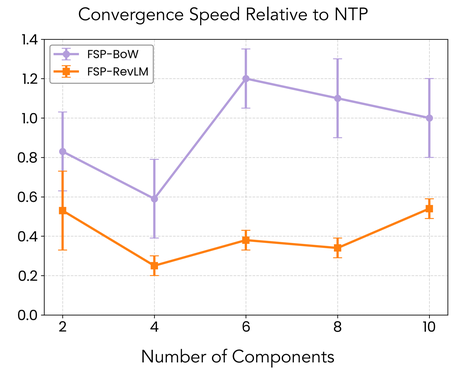


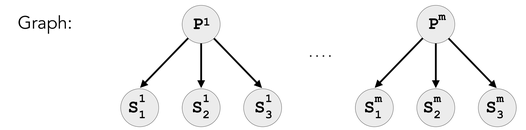
Long Future Summary is Important





Adaptive Future Summary is Important



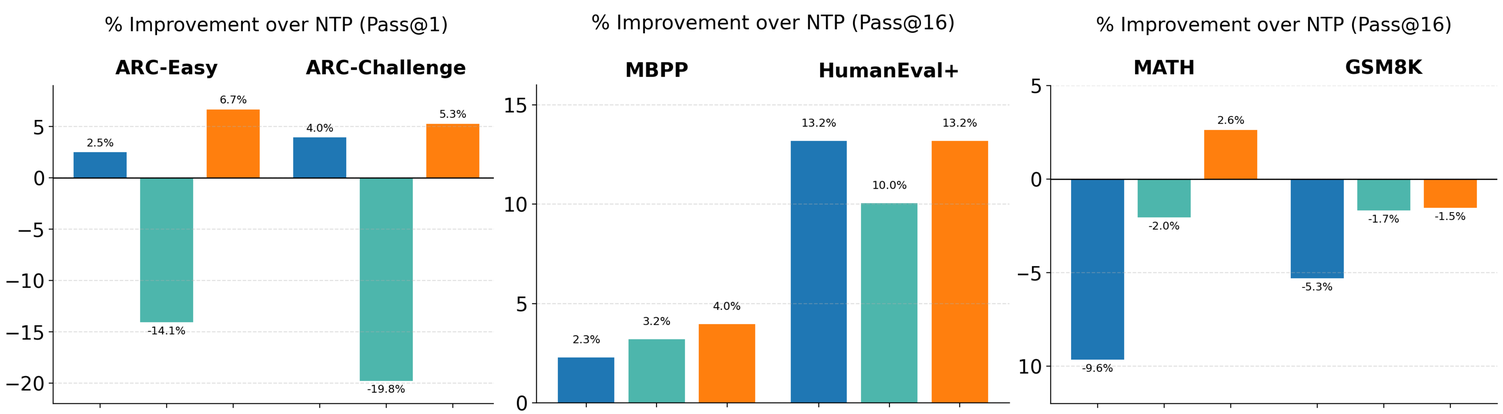


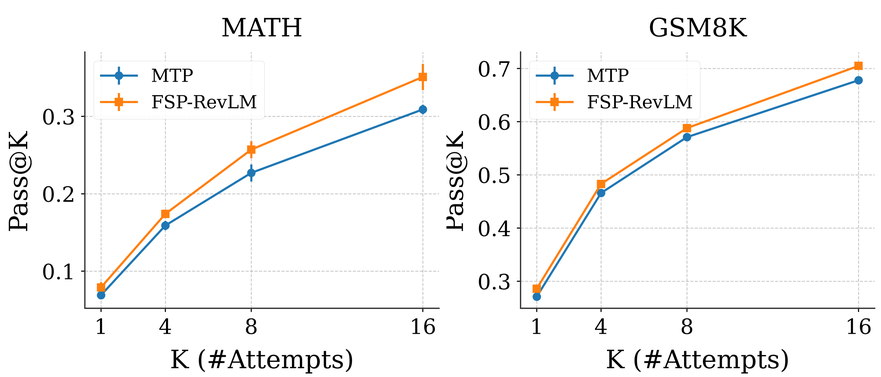
Example Sequence:

$x = S_1^1\ S_2^1\ S_3^1\ P^1\ S_1^2\ \ S_2^2\ S_3^2\ P^2\ \ S_1^3\ S_2^3\ S_3^3\ P^3$

If this is the prefix — *Relevant future* — *Not predictable from the prefix*

Pretraining Results: 8B Scale





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