ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

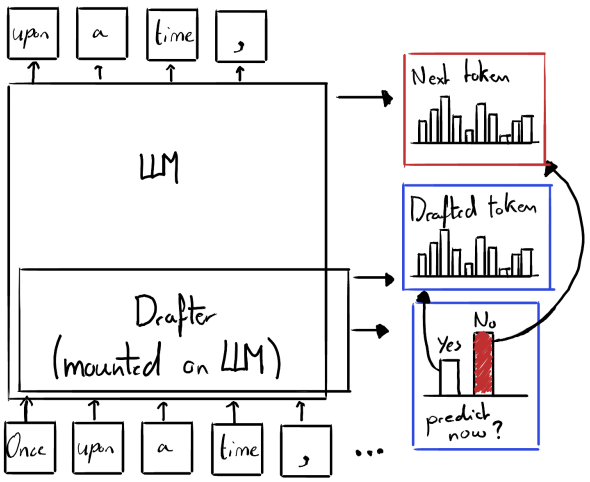
Prof. R. Wattenhofer

Unleashing Forward Thinking in Language Models

Motivation: Large Language Models (LLMs) are incredibly powerful, but their token-by-token generation limits their ability to *think ahead* or plan strategically during the forward pass. This leads to reactive generation, where each token is predicted based on immediate context, without explicit foresight. While existing methods like Mixture-of-Depths (MoD) [1] and early-exiting strategies [2] reduce compute, they don't re-purpose this saving for planning. The goal of this project is to modify the architecture and training of LLMs to allow forward thinking. More precisely, the idea is to introduce a gating mechanism that explicitly decides when to predict a token and when to dedicate the remaining compute to proactive *thinking ahead* within the same forward pass.

Method: We will develop a new LLM architecture with a forward-thinking mechanism:

1. We'll integrate a learnable gating mechanism into the LLM's forward pass. This gate will decide if the model predicts the next token and allocates the remaining computational budget to think ahead, or if the model allocates the full forward pass to preduct the next token.
2. Finally, we will train and evaluate our architecture on various benchmarks. We hypothesize that this forward-thinking capability will lead to improved coherence, logical consistency, and more strategically planned outputs.



Coarse architecture with early-exiting (drafter) and thinking-ahead module (remaining compute if the token is drafted early).

Requirements: Strong programming skills with deep learning frameworks (e.g., PyTorch) and familiarity with LLM architectures are ideal. We're looking for someone interested in novel research and contributing to a publishable paper. Weekly meetings will be held for discussion and brainstorming.

Contact:

- Frédéric Berdoz : fberdoz@ethz.ch, ETZ G60.1

References

[1] David Lepikhin et al. Mixture-of-Depths: Dynamically allocating compute in transformer-based language models. In: *arXiv preprint arXiv:2404.02258* (2024).
[2] Tal Schuster et al. Confident Adaptive Language Modeling. In: *arXiv preprint arXiv:2207.07061* (2022).