***Korol Vladyslav**, BA Student*
***Kanchura Yevheniia**, PhD (Philology), As.Prof.*
*Zhytomyr State Polytechnic University*

# ACCELERATING LARGE LANGUAGE MODEL PRETRAINING WITH META LEARNING

One of the most long-standing problems in large language models (LLMs) is the fact that it takes enormous data to pretrain them until they become anywhere close to being useful for human-assisted applications. We propose a novel approach to accelerate the training of language models by applying model-agnostic meta learning (MAML [1]) during the pretraining phase as a means of inducing inductive bias towards language proficiency.

We make an observation that language models (including, but not limited to GPT variants, Claude, Llama, etc.) take significantly more data to reach convergence compared to humans. Kaplan et. al [2] suggest that as models get larger and more capable, it takes proportionally longer to pretrain them, and Chinchilla scaling laws [3] even further estimate the specific figure of 20 tokens per parameter to converge. This makes model pretraining prohibitively expensive for small teams or universities and consolidates the market on artificial intelligence around a few key players (OpenAI, Google, Anthropic, and Meta).

We also note that despite the computational challenges, humans learn language in orders of magnitude faster than AI systems do. We attribute most of this ability to evolutionary inductive biases as humans evolved throughout millions of years to communicate with each other, instilling the ability to speak genetically. The observation that every human baby learns language after just a few years of exposure supports this claim. On the other hand, language models begin with tabula rasa (blank slate), an analogy of a clear state where they do not have any predetermined biases but learn language from scratch and first principles.

Our approach boils down to utilizing meta learning, more precisely model-agnostic meta learning, in order to induce linguistic inductive biases onto language models. Our method describes a learning pipeline that consists of two learning loops: in the inner loop, the model learns to predict the next token using cross-entropy loss, AdamW optimizer, and learning rate scheduler, as most LLMs are trained. In the outer loop, we optimize the model to learn language faster, with fewer examples, more robustly and reliably, using MAML. We frame the hypothesis that since humans learn language from fewer examples, and since meta learning makes models learn downstream tasks more efficiently, applying meta learning during pretraining would induce similar inductive biases on language models, just like humans do, which would make large language models cheaper and more accessible to train to the public. This would democratize and diversify the LLM scene and lead to further innovation in the domain.

Finally, we note that meta learning techniques have already been used with language models, albeit in fine-tuning stages to make models faster to adapt to specific domains than traditional fine-tuning. ProtoMAML, meta datasets [4], and few-shot prompt organization [5] demonstrate that meta learning is well-equipped to optimize language models, further strengthening the methodology outlined above.

Meta learning pretraining stands as the unexplored frontier in the domain of artificial intelligence that reveals profound implications for the future of the industry.

**REFERENCES**

1. Finn C., Abbeel P., Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. / C. Finn et al. ArXiv. 18.06.2017. URL: https://arxiv.org/pdf/1703.03400.
2. Scaling laws for neural language models / J. Kaplan et al. ArXiv. 23.01.2020. URL: https://arxiv.org/pdf/2001.08361.
3. Training compute-optimal large language models / J. Hoffmann et al. ArXiv. 29.03.2022. URL: https://arxiv.org/pdf/2203.15556.
4. Meta-dataset: a dataset of datasets for learning to learn from few examples / E. Triantafillou et al. ArXiv. 08.05.2020. URL: https://arxiv.org/pdf/1903.03096.
5. Comparing transfer and meta learning approaches on a unified few-shot classification benchmark / V. Dumoulin et al. ArXiv. URL: https://arxiv.org/pdf/2104.02638.