**Project Team: Oishi Banerjee, Shreya Johri**

**Title: Benchmarking Generation Length Prediction Accuracy from LLM Hidden Embeddings Across Models and Data Shift**

**Introduction:** Predicting inference times for LLM responses can significantly enhance job scheduling, overcoming the limitations of first-come-first-serve algorithms where individual long jobs can block many shorter ones. Some works have proposed training specialized models to predict the generation lengths. For example, Qiu et al. [1](https://paperpile.com/c/oPwtSK/NaYS) demonstrate a regression-based approach built on a fine-tuned BERT-base model coupled with a shallow two-layer network, achieving around 61.5% prediction accuracy on token count categories, which in turn leads to a 30–40% reduction in job completion times and a throughput boost of up to 3.6×. Similarly, Fu et al. [2](https://paperpile.com/c/oPwtSK/4Dgw) employ an OPT-based predictor with a linear output head, attaining Kendall’s Tau values of up to 0.62 on real-world conversation datasets, effectively ranking requests by predicted output lengths for improved scheduling. Training a separate model for prediction and subsequent scheduling of jobs leads to additional computational overhead. Contrary to this approach, recent studies have proposed using lightweight classifiers that leverage an LLM's intermediate representations to predict the number of tokens remaining in a response[3,4](https://paperpile.com/c/oPwtSK/bwri+WjOq). Using internal state probes leverages rich contextual information already computed by the LLM with minimal added latency. This approach not only allows predicting job times at the time of prompt, but also during the decoding phase, with the potential to offer insights into the model's processing dynamics. For instance, Shahout et al.[3](https://paperpile.com/c/oPwtSK/bwri) found that certain layers may have a disproportionate influence on response length, as suggested by the predictive power of intermediate responses from layers 10-15.

Building on this foundational work of utilizing LLM’s internal state for prediction of output lengths, we propose a benchmark to systematically evaluate response length predictors across three types of distribution shifts: model families, data distributions and LLM hyperparameters. Understanding factors that influence these prediction capabilities could enable dynamic resource allocation across data centers, optimizing energy usage during peak demand periods and prioritizing time-sensitive applications like financial trading or emergency response systems. Additionally, such predictions could improve user experience by providing more precise response time estimates and enable new techniques for early request cancellation when outputs are predicted to exceed desired lengths or computation budgets.

**Proposed Experiments:** We propose the following experiments to systematically evaluate the three kinds of distribution shifts – model families, data distributions and LLM hyperparameters. We will use a simple linear predictor (linear layer followed by ReLU activation) for all experiments.

*Model Families:* We hypothesize that variations in model architecture and pretraining will lead to differences in the capability of a linear predictor for generation length, despite similar model sizes. We propose to benchmarking the following 5 models

* meta-llama/Meta-Llama-3-8B-Instruct[5](https://paperpile.com/c/oPwtSK/aDXd) (as benchmarked in Shahout et al.[3](https://paperpile.com/c/oPwtSK/bwri))
* qwen/Qwen2.5-7B-Instruct[6](https://paperpile.com/c/oPwtSK/ZFgg),
* mistralai/Mistral-7B-Instruct-v0.3[7](https://paperpile.com/c/oPwtSK/qEQZ),
* deepseek-ai/DeepSeek-R1-Distill-Qwen-7B[8](https://paperpile.com/c/oPwtSK/at0W)
* deepseek-ai/DeepSeek-R1-Distill-Llama-8B[8](https://paperpile.com/c/oPwtSK/at0W).

The latter two are reasoning models, which also output a reasoning chain (sandwiched between <think> and </think> tokens) before its final response. We are interested in the following questions -

1. *Accuracy of predictions:*
   1. Is a linear predictor equally accurate between different non-reasoning models?
   2. For the reasoning models, how accurately can linear predictors estimate response length at the start vs. when in the “thinking” phase vs. when in the “response” phase?
2. *Best predictive layers:* Do LLMs show differences in which layers are the most useful predictors of response length (q1, q2, q3, q4 quantile layers).

*Data Distributions:* Shahout et al. focused their experiments on the Alpaca dataset. Despite diversity in topic (from science to language translation), they are all relatively concise prompts and answers (e.g. What are the three primary colors?; The three primary colors are red, blue, and yellow).

1. *Open-ended vs specific prompts:*
   1. We will investigate whether predictors maintain performance when answer lengths can vary widely (e.g. Translate the following news article …), similar to real world queries.
   2. Do predictors implicitly over rely on the length of the prompt to predict the length of the output? We will include data pairs where prompts and their desired answers have different lengths (e.g. “Write a five-paragraph essay on the Constitution”).
2. Does specifying an output length in the prompt improve prediction accuracy?

*We hypothesize that increasing variation in response length will lower predictor performance; conversely, we expect that predictors will be more accurate when prompts that explicitly order certain output lengths.*

*Hyperparameters:* Hyperparameters can substantially impact response length by changing how outputs are sampled. Predictors that rely solely on intermediate model representations may not account for these differences.

1. We will investigate how changes in temperature and top\_k effect the predictor performance. We hypothesize that changes in hyperparameters will cause predictor performance to deteriorate.
2. We will also investigate when predictors are provided with the hyper-parameters as an input, does it improve performance over a range of top\_k and temperature settings.

**Resources required:**

* *Models:* Selected models are open source and available freely through Hugging Face.
* *Compute:* To run LLM inference, we’ll utilize high-end GPUs (48-80G VRAM), which we have access to through our thesis labs (HMS compute cluster).
* *Datasets:* We will use a publicly available medical instruction tuning dataset collection on Hugging Face[9](https://paperpile.com/c/oPwtSK/YFm2), filtering with datapoints with diverse input and output lengths.

**Alternatives:** It is possible that the experiments take a longer time to run than anticipated. In that case, we will focus on the first two kinds of distribution shifts – model families and data distributions. We do not anticipate challenges in access to GPUs or models.

**References:**

1. [Qiu, H. *et al.* Efficient Interactive LLM Serving with Proxy Model-based Sequence Length Prediction. (2024).](http://paperpile.com/b/oPwtSK/NaYS)

2. [Fu, Y. *et al.* Efficient LLM Scheduling by Learning to Rank. (2024).](http://paperpile.com/b/oPwtSK/4Dgw)

3. [Shahout, R. *et al.* Don’t Stop Me Now: Embedding Based Scheduling for LLMs. (2024).](http://paperpile.com/b/oPwtSK/bwri)

4. [Ashok, D. & May, J. Language Models Can Predict Their Own Behavior. (2025).](http://paperpile.com/b/oPwtSK/WjOq)

5. [Grattafiori, A. *et al.* The Llama 3 Herd of Models. (2024).](http://paperpile.com/b/oPwtSK/aDXd)

6. [Bai, J. *et al.* Qwen Technical Report. (2023).](http://paperpile.com/b/oPwtSK/ZFgg)

7. [Jiang, A. Q. *et al.* Mistral 7B. (2023).](http://paperpile.com/b/oPwtSK/qEQZ)

8. [DeepSeek-AI *et al.* DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. (2025).](http://paperpile.com/b/oPwtSK/at0W)

9. [Medical Instruction Tuning Datasets - a lavita Collection.](http://paperpile.com/b/oPwtSK/YFm2) <https://huggingface.co/collections/lavita/medical-instruction-tuning-datasets-66398ea593fc99cf52a6e6e2>[.](http://paperpile.com/b/oPwtSK/YFm2)