PREDICTING LLM OUTPUT LENGTH VIA ENTROPY-GUIDED REPRESENTATIONS

Huanyi Xie, Yubin Chen, Liangyu Wang, Lijie Hu, and Di Wang
KAUST PRADA Lab MBZUAI
ICLR 2026

Motivation

The long-tailed distribution of sequence lengths in LLM serving and reinforcement learning (RL) sampling causes significant computational waste.

- Tensor shapes must align in batched inference, meaning shorter sequences are padded to match the longest one.
- This redundant padding leads to a marked "barrel effect", enlarging GPU time consumption and reducing effective computation.

Limitations of Existing Methods

Prior work relies on attaching a fine-tuned, lightweight auxiliary predictor (e.g., DistilBERT or OPT). This suffers from three key limitations:

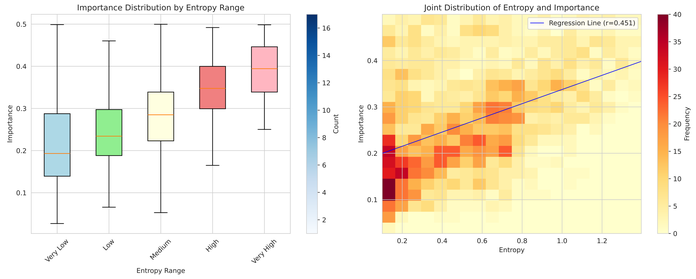
- **Instability in stochastic generation:** During RL sampling, a single prompt can yield multiple completions with widely differing lengths, rendering static prompt-based estimates unreliable.
- **Limited generalization:** They perform poorly on long sequences and complex reasoning.
- **High Overhead:** Each request must run a separate predictor instead of reusing the main LLM's rich hidden states.

Core Idea & Entropy Analysis

Insight: If an LLM can determine when to emit the <eos> token, then signals related to the eventual output length must be implicitly encoded in its internal states. We can reuse these activations with minimal additional cost.
We measured each token's importance I_t for prediction via gradient attribution:

$$I_t = ||\nabla_{h_t} \mathcal{L}_{MSE}||_2$$

Empirical validation demonstrates a significant positive correlation between a token's entropy and its importance for length prediction (Pearson's $r = 0.451$). High-entropy tokens are critical signals for forecasting output length.



Empirical analysis of token entropy and importance for length prediction.

Entropy-Guided Token Pooling (EGTP)

EGTP aggregates a sequence of input hidden states into a single feature vector h for static prediction.
Step 1: Compute Token Entropy. From the next-token probability distribution $P(v|x_{<i})$ over vocabulary V:

$$H_i = -\sum_{v \in V} P(v|x_{<i}) \log P(v|x_{<i})$$

Step 2: Generate Attention Weights. Transform entropies via a softmax function:

$$w_i = \frac{\exp(H_i)}{\sum_{j=1}^{n} \exp(H_j)}$$

Step 3: Aggregate Representation. Adaptively focus on the most informative parts of the prompt:

$$h = \sum_{i=1}^{n} w_i h_i$$

Length Regression via Soft Label Distribution

Standard MSE is sensitive to outliers, while pure classification ignores the distance between values. We discretize the target space into K bins and create a soft label distribution p:

$$p_j = \frac{\exp(-|j-i|)}{\sum_{k=1}^{K} \exp(-|k-i|)}$$

The model produces a classification prediction $\hat{p}$ and a continuous regression expected value $\hat{y}$. The joint loss balances cross-entropy and MSE:

$$\mathcal{L} = \lambda \mathcal{L}_{CE}(p, \hat{p}) + (1-\lambda)\mathcal{L}_{MSE}(y, \hat{y})$$

Progressive Length Prediction (PLP)

In RL sampling (a "one-to-many" scenario), static prediction fails. PLP leverages the autoregressive nature of LLMs by estimating the remaining length $y_{rem}^{(t)}$ at each decoding step.
It combines the prompt feature h with hidden states of generated tokens $\{h'_1, ..., h'_t\}$:

$$z_t = \text{Aggregate}(h, \{h'_1, ..., h'_t\})$$

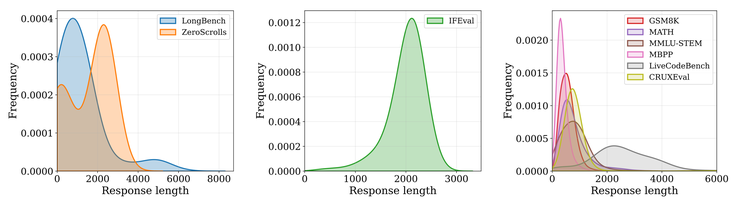
By iteratively refining its forecast, PLP enables dynamic adjustments to scheduling strategies.

ForeLen Benchmark & Experiments

We introduce **ForeLen**, the first comprehensive benchmark designed to test predictors on challenging Long-Sequence, Chain-of-Thought (CoT) reasoning, and RL sampling data.

- **Models tested:** Qwen2.5 (0.5B-7B) and Llama3.2 (1B, 3B).
- **Baselines:** SSJF-Reg, SSJF-MC, S3, PiA, TPV, TRAIL, LTR-C.

Response Length Distribution in ForeLen



(a) Long-Sequence Scenario **(b) Reasoning Scenario** **(c) RL Scenario**

ForeLen covers three distinct settings with markedly different response-length distributions, including heavy-tailed long-sequence data, concentrated reasoning data, and highly diverse RL sampling data.

Main Results: MAE

Model / Method	Long Seq	Reasoning	RL
Qwen2.5 7B (EGTP)	**81.60**	133.57	**95.24**
Best Baseline	129.37	**124.19**	155.51
Llama3.2 3B (EGTP)	**78.83**	**111.23**	**114.53**
Best Baseline	143.62	160.36	131.75

✓ EGTP substantially improves Long Seq and RL prediction, while remaining competitive on Reasoning across representative 7B/3B settings.