AIRI Skoltech

fall into ML 2025
4th conference on machine learning & AI

30th ANNIVERSARY EMNLP 2025 Suzhou, China November 4-9 中国苏州 11月4日-9日

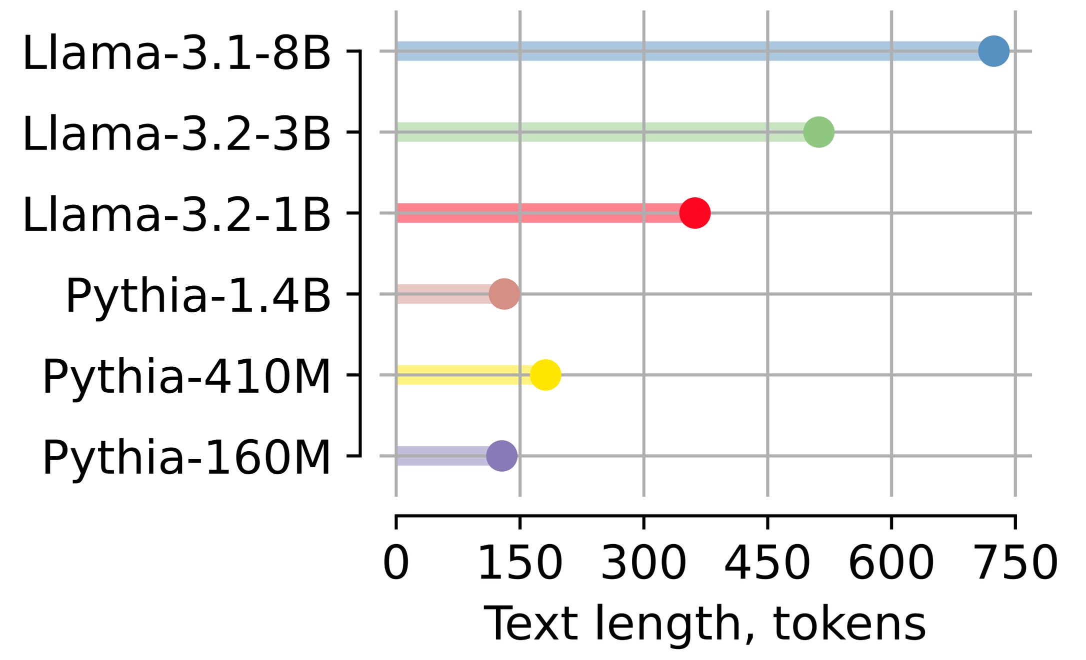
Exploring the Hidden Capacity of LLMs for One-Step Text Generation

Gleb Mezentsev, Ivan Oseledets

arXiv GitHub

Main result

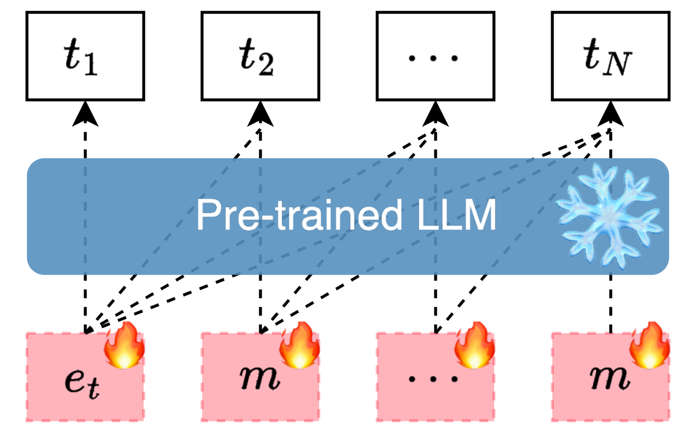
Frozen LLMs can generate **hundreds** of accurate **tokens** with **non-autoregressive** generation in a **single forward pass** if conditioned on a "proto-token" (special embedding).



Each dot shows the maximum exact reconstruction length in a single non-autoregressive forward pass with frozen weights, conditioned only on a single learned embedding — evidence of hidden multi-token capabilities.

Method

Two proto-tokens (trainable embeddings) are fed into frozen LLM and optimized in such a way that LLM predicts an arbitrary token-sequence in a **single forward pass**. e_t is trained for each text separately, while m is universal.



$$L_{CE} = -\sum_{i=1}^{N} log\mathbb{P}_{LM}(t_i \mid e_t, \underbrace{m, \dots, m}_{i-1})$$

The loss function we optimise to find e_t and m

Arrangement	$N=1$	$N=2$	$N=4$	$N=256$
$[e]_{\times N}$	$1.00_{\pm 0.00}$	$0.45_{\pm 0.31}$	$0.17_{\pm 0.18}$	$0.01_{\pm 0.01}$
$[e]_{\times (N/2)}[m]_{\times (N/2)}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.12_{\pm 0.13}$	$0.01_{\pm 0.01}$
$[e, m]_{\times (N/2)}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.17_{\pm 0.34}$
$[e][m]_{\times N}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.97_{\pm 0.15}$
$[e][m]_{\times (N-1)}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.99_{\pm 0.10}$

Reconstruction accuracies for different input token arrangements across sequence lengths. Subscripts indicate the number of copies for each proto-token.

Shared	Agg	$S_g=1$	$S_g=16$	$S_g=256$
e	max	$1.00_{\pm 0.00}$	$0.99_{\pm 0.01}$	$0.99_{\pm 0.02}$
	avg	$0.98_{\pm 0.08}$	$0.90_{\pm 0.17}$	$0.86_{\pm 0.20}$
m	max	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.01}$
	avg	$0.98_{\pm 0.07}$	$0.86_{\pm 0.19}$	$0.83_{\pm 0.18}$

Reconstruction accuracy with one of proto-tokens shared within groups for different group sizes. "max" is maximum accuracy across ten random seeds, "avg" is the average accuracy.

Quantitative results

Main metrics: $C_{tokens} = \sum_{i=1}^{N} \mathbb{1}(\arg\max \mathbb{P}_{LM}(\cdot \mid e_t, \underbrace{m, \dots, m}_{i-1}) = t_i)$ $\quad H_{LM} = -\sum_{i=1}^{N} log\mathbb{P}_{LM}(t_i \mid t_{<i})$

Maximum generation capacity for **random/unseen/seen/generated** texts across models of different sizes:

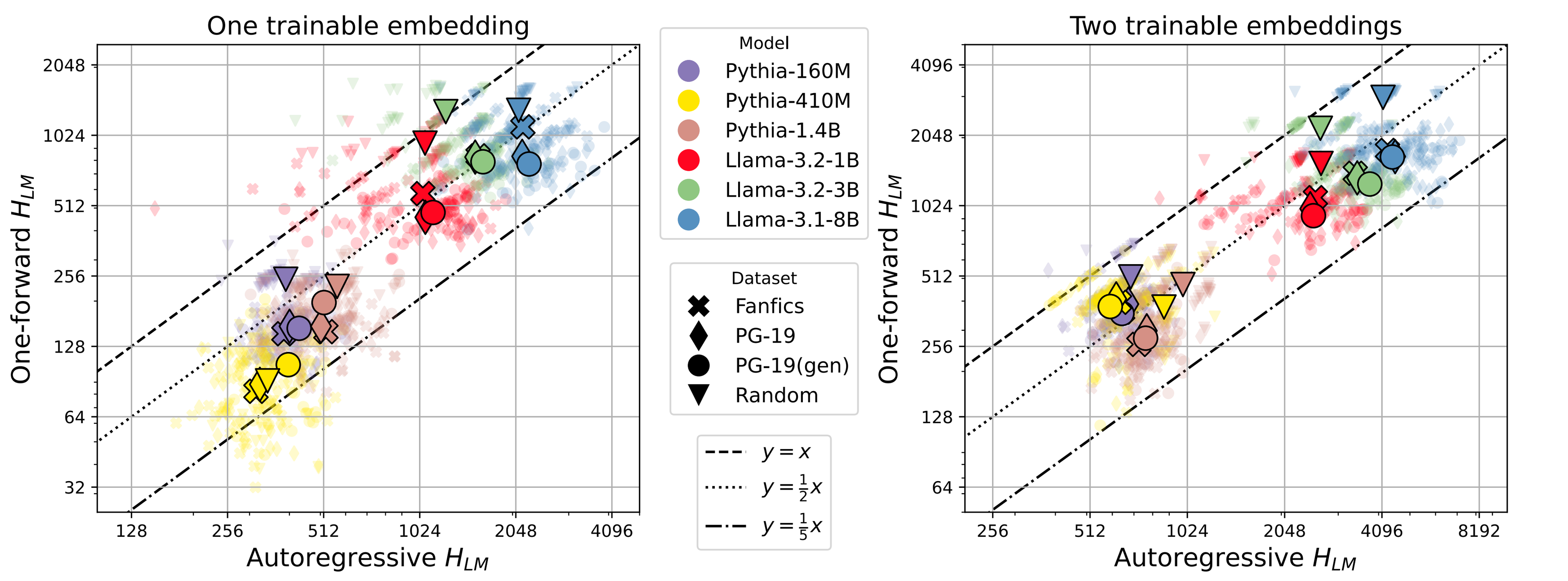
All that matters is whether it is real text or random tokens.

		Share m	Pythia 160M	Pythia 410M	Pythia 1.4B	Llama 3.2-1B	Llama 3.2-3B	Llama 3.1-8B
Random	C_{tokens}	False	90	92	90	256	362	512
		True	45	22	45	181	256	256
	H_{LM}	False	$507.5_{\pm 105.9}$	$377.1_{\pm 133.1}$	$470.7_{\pm 103.1}$	$1551.3_{\pm 159.5}$	$2193.4_{\pm 190.2}$	$2974.4_{\pm 298.3}$
		True	$247.9_{\pm 32.0}$	$91.1_{\pm 30.8}$	$231.0_{\pm 37.9}$	$947.7_{\pm 155.0}$	$1292.2_{\pm 217.4}$	$1309.4_{\pm 234.6}$
Fanfics	C_{tokens}	False	128	128	131	362	512	724
		True	45	45	45	181	288	362
	H_{LM}	False	$358.9_{\pm 73.3}$	$395.4_{\pm 97.8}$	$261.0_{\pm 56.4}$	$1107.6_{\pm 129.1}$	$1408.4_{\pm 179.5}$	$1763.3_{\pm 280.2}$
		True	$145.0_{\pm 26.2}$	$82.3_{\pm 28.1}$	$147.9_{\pm 29.7}$	$576.4_{\pm 90.4}$	$835.9_{\pm 121.7}$	$1112.8_{\pm 168.6}$
PG-19	C_{tokens}	False	128	167	128	362	512	724
		True	45	32	64	181	256	362
	H_{LM}	False	$388.4_{\pm 66.4}$	$408.8_{\pm 96.3}$	$298.4_{\pm 77.4}$	$993.8_{\pm 183.4}$	$1346.0_{\pm 218.4}$	$1659.8_{\pm 344.5}$
		True	$156.0_{\pm 33.9}$	$88.1_{\pm 30.3}$	$156.0_{\pm 30.2}$	$456.5_{\pm 56.5}$	$826.1_{\pm 117.6}$	$832.3_{\pm 171.0}$
PG-19 (gen)	C_{tokens}	False	128	181	128	362	512	724
		True	45	32	64	181	362	362
	H_{LM}	False	$354.1_{\pm 72.0}$	$379.2_{\pm 82.6}$	$277.6_{\pm 71.3}$	$927.3_{\pm 103.4}$	$1266.6_{\pm 125.9}$	$1653.1_{\pm 211.4}$
		True	$153.0_{\pm 17.8}$	$106.9_{\pm 38.5}$	$197.1_{\pm 39.3}$	$478.7_{\pm 85.7}$	$788.6_{\pm 130.8}$	$771.7_{\pm 143.0}$

Maximum reconstruction capacities for different models on different datasets.

Maximum generation capacity **compared to autoregressive** setup:

You can fit half of the information compared to autoregressive generation.

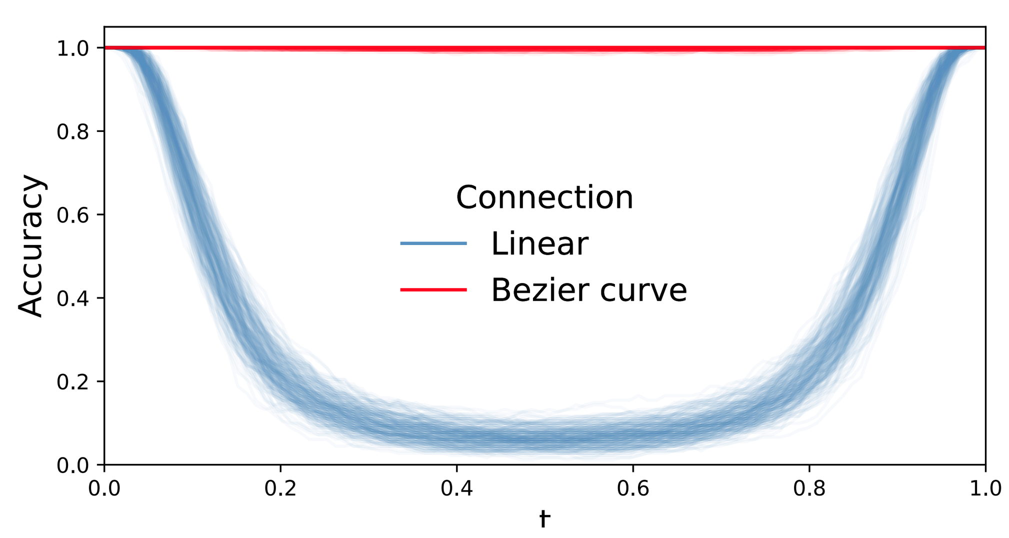


Maximum language information (H_{LM} for a maximum text prefix that is accurately reconstructed) for different models and datasets.
On the left plot, a single [mem] token is used in the autoregressive setting, and in the non-autoregressive one, m proto-token is shared between all texts within each model.
On the right plot, two [mem] tokens are used and m proto-tokens are not shared.
Each small point on the plots represents a single text, larger points indicate the average within each (model, dataset) pair.

Solution-space structure

For a given text, solution is not unique and the **solution set** is non-convex, but **connected and localized:**

A solid potential for training practical encoder.



Pairwise interpolation accuracies between 10 solutions for 5 texts ($5 \times 10 \times 9/2$ pairs in total).

Each pair of solutions could be connected via degree-two Bezier curve with perfect accuracy along the curve.

$$\phi_\pi(\tau) = (1-\tau)^2 p_1 + 2\tau(1-\tau)\pi + \tau^2 p_2$$

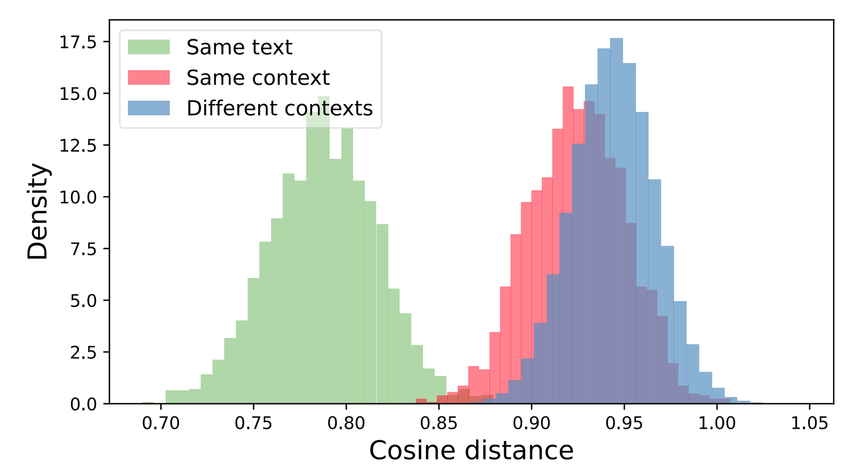
$$l_\pi = \mathop{\mathbb{E}}_{\tau \sim \mathcal{U}[0,1]} \sum_{i=1}^{N} -log\mathbb{P}_{LM}(t_i \mid \phi_\pi(\tau))$$

Bezier curve parameterisation and optimisation problem

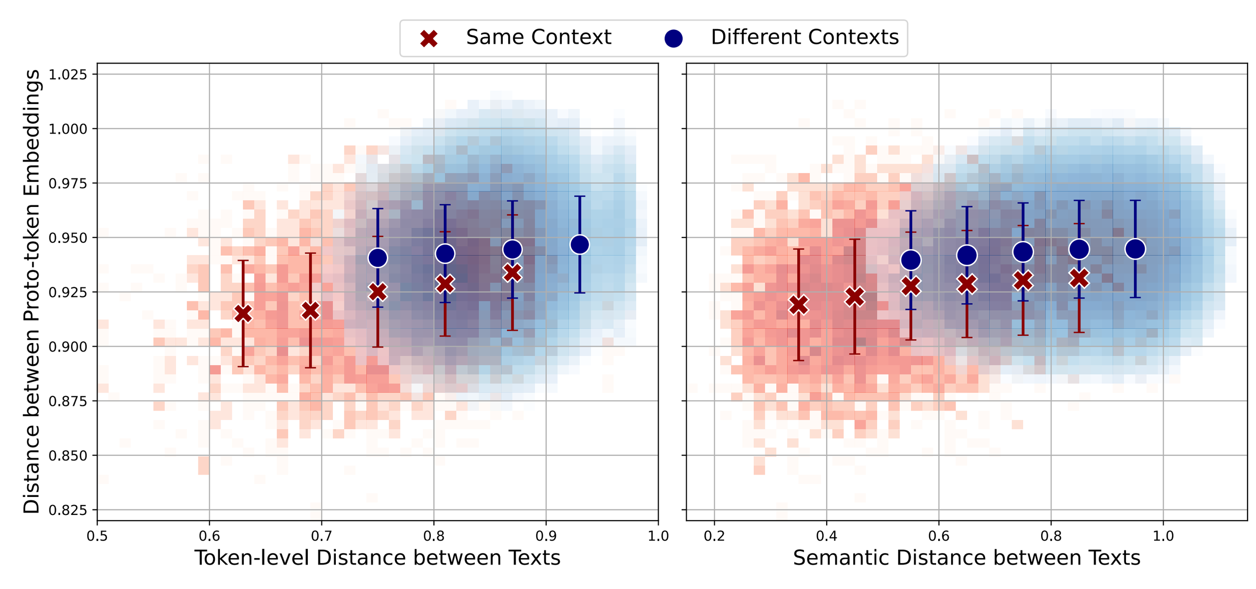
Solution interpretation

Embeddings seem to **contain** information beyond the target text itself, with some **traces of the potential context:**

The representation is useful — not just token ids.



Cosine embedding distances for different pairings of proto-tokens. We select 50 contexts from PG19 and for each context, generate 10 continuation texts. We find one solution for each of the first 9 generations and 10 different-seed solutions for the last generation.



We compare proto-token embedding distances for same context text pairs and different-context text pairs. Token-level distance is measured as cosine distance between TF-IDF embeddings. Semantic distance is measured as cosine distance between semantic text embeddings