Fine-Tuned Language Models Exhibit Improved Information Conveyance

Yuwei Cheng, Weiyi Tian*, Haifeng Xu*

The University of Chicago

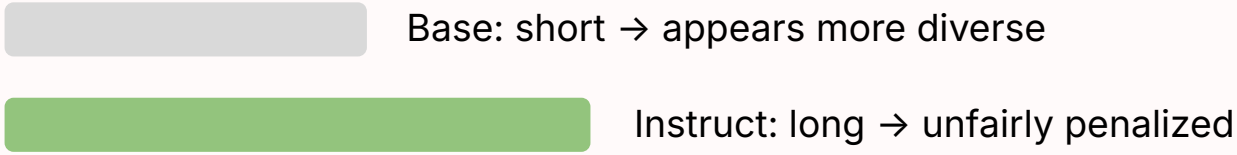
SIGMA LAB
Strategic IntelliGence for Machine Agents

COMMON BELIEF

Fine-Tuning Reduces Diversity ?

Existing diversity metrics do not account for generation **length** and often favor shorter sequences.

RLHF models tend to produce longer outputs - length isn't just noise; it's a feature of alignment.



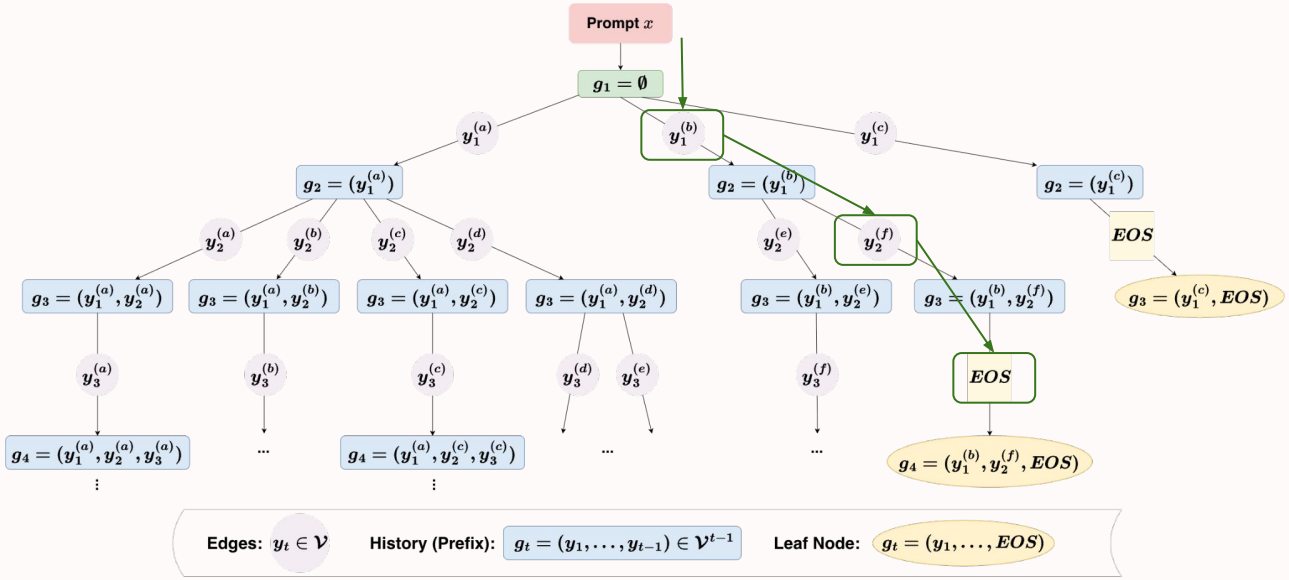
Base: short → appears more diverse

Instruct: long → unfairly penalized

What happens when we control for length?

Autoregressive Generation Tree

Every generation is a path from the prompt *x* down to an *EOS* leaf.



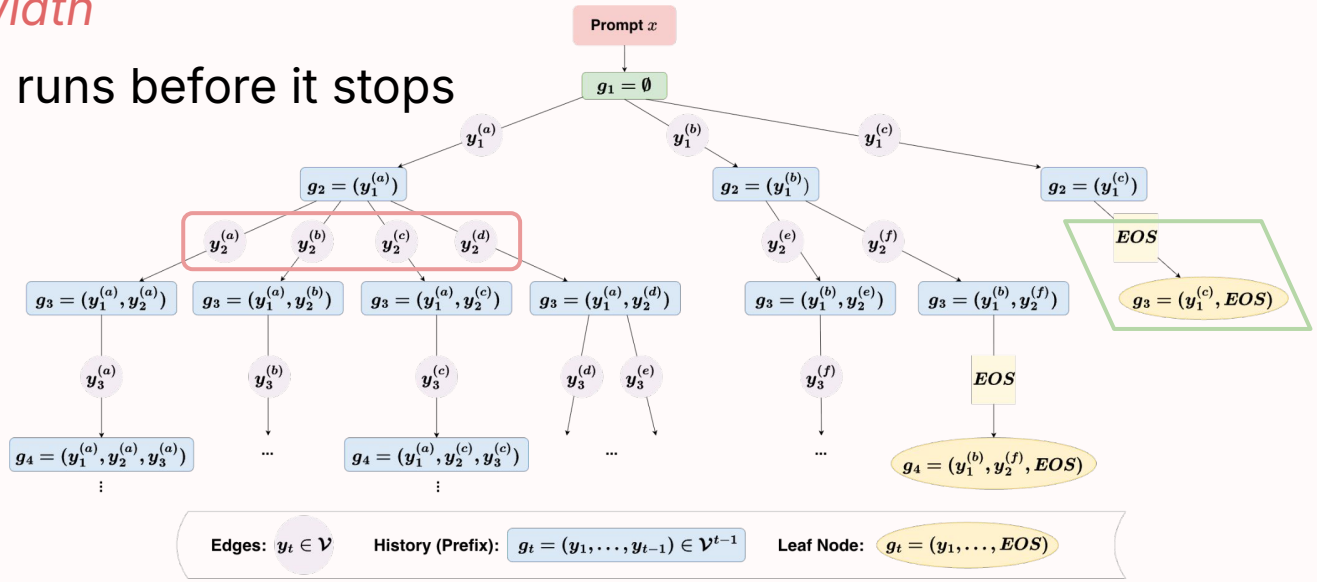
SIGMA LAB
Strategic IntelliGence for Machine Agents

Autoregressive Generation Tree

Every generation is a path from the prompt *x* down to an *EOS* leaf.

Uncertainty (size of the tree) comes from two aspects:

- How many options at each step
  Local branching width
- How long the path runs before it stops
  Trajectory depth



SIGMA LAB
Strategic IntelliGence for Machine Agents

Canopy Entropy $\mathrm{CE}^\star$

First information-theoretic measure of the full generation space (length + content)

length

$$\mathrm{CE}^\star := \mathbb{E}\Big[\sum_{t=1}^{N(x)} \widetilde{H}(g_t \mid x)\Big]$$

sum along path

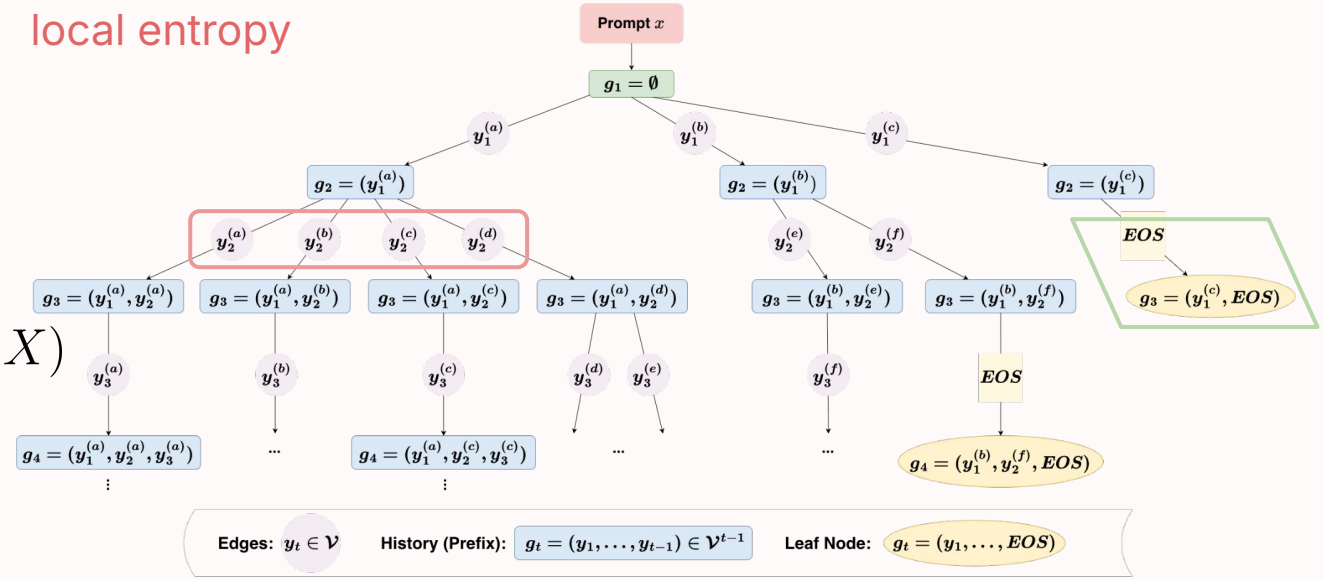
local entropy

$$\widetilde{H}(g_t \mid x) := H(Z_t \mid G_t = g_t, X = x) = -\sum_{z \in \mathcal{Z}} q(z \mid g_t, x) \log q(z \mid g_t, x)$$

Theorem 1

Total Shannon Entropy

$$\mathrm{CE}^\star = H(\underbrace{N}_{\text{length}}, \underbrace{Y_{1:N}}_{\text{content}} \mid X)$$



SIGMA LAB
Strategic IntelliGence for Machine Agents

Fine-Tuning Reduces Uncertainty Non-Uniformly

Per-token normalization $\mathrm{GenPPL} := \exp\left(\frac{\mathrm{CE}^\star}{\mathbb{E}[N]}\right)$

Per-trajectory normalization $\mathrm{BF} := \exp\left(\mathbb{E}[r_N]\right)$ where $r_N := \frac{H(Y_{1:N} \mid N = n, X)}{n}$ is entropy rate

- Fine-tuning reduces local branching diversity.
- Total generation space shifts by task - even increasing in creative generation.

Task	Type	GenPPL			BF			CE*		
		Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B
Math	Base	1.23 ± 0.0052	3.57 ± 0.041	4.56 ± 0.052	1.23 ± 0.0064	5.31 ± 0.12	4.98 ± 0.055	55.7 ± 2.0	458 ± 18	487 ± 11
	Instruct	1.17 ± 0.0067	1.76 ± 0.076	1.16 ± 0.0091	1.16 ± 0.0057	1.51 ± 0.025	1.15 ± 0.0052	40.5 ± 2.5	108 ± 12	56.7 ± 4.7
	% Change	$-4.44 \pm 0.47\%$	$-50.8 \pm 2.3\%$	$-74.5 \pm 0.33\%$	$-5.76 \pm 0.38\%$	$-71.5 \pm 0.71\%$	$-76.9 \pm 0.24\%$	$-27.3 \pm 3.3\%$	$-76.3 \pm 2.8\%$	$-88.4 \pm 1.0\%$
Coding	Base	1.68 ± 0.017	3.16 ± 0.055	2.60 ± 0.041	2.08 ± 0.034	4.96 ± 0.057	3.24 ± 0.038	207 ± 4.8	326 ± 13	492 ± 25
	Instruct	1.23 ± 0.0053	1.27 ± 0.0075	1.04 ± 0.0023	1.21 ± 0.0061	1.27 ± 0.0070	1.04 ± 0.0018	81.2 ± 4.0	126 ± 3.9	16.2 ± 1.0
	% Change	$-27.1 \pm 0.79\%$	$-59.6 \pm 0.71\%$	$-59.9 \pm 0.64\%$	$-41.9 \pm 0.99\%$	$-74.4 \pm 0.33\%$	$-67.8 \pm 0.38\%$	$-60.7 \pm 1.6\%$	$-61.4 \pm 1.7\%$	$-96.7 \pm 0.28\%$
Sentence Completion	Base	4.77 ± 0.16	6.73 ± 0.12	8.81 ± 0.11	4.20 ± 0.11	7.80 ± 0.15	9.49 ± 0.11	826 ± 55	969 ± 40	1084 ± 36
	Instruct	1.68 ± 0.025	2.32 ± 0.060	1.78 ± 0.019	1.55 ± 0.018	2.53 ± 0.042	1.71 ± 0.024	112 ± 11	188 ± 9.4	513 ± 29
	% Change	$-64.9 \pm 1.1\%$	$-65.6 \pm 1.1\%$	$-79.8 \pm 0.30\%$	$-63.0 \pm 0.88\%$	$-67.6 \pm 0.81\%$	$-82.0 \pm 0.30\%$	$-86.5 \pm 1.0\%$	$-80.6 \pm 1.2\%$	$-52.6 \pm 2.9\%$
Story Generation	Base	3.57 ± 0.11	7.40 ± 0.071	7.98 ± 0.062	4.03 ± 0.092	8.24 ± 0.073	8.68 ± 0.070	445 ± 20	1320 ± 41	692 ± 23
	Instruct	2.28 ± 0.037	2.93 ± 0.083	2.22 ± 0.023	2.11 ± 0.039	3.07 ± 0.077	2.16 ± 0.027	454 ± 28	471 ± 19	892 ± 39
	% Change	$-36.0 \pm 1.3\%$	$-60.4 \pm 1.2\%$	$-72.2 \pm 0.28\%$	$-47.6 \pm 1.3\%$	$-62.7 \pm 1.1\%$	$-75.1 \pm 0.30\%$	$1.92 \pm 3.5\%$	$-64.3 \pm 1.3\%$	$29.0 \pm 8.0\%$

- decrease
- increase

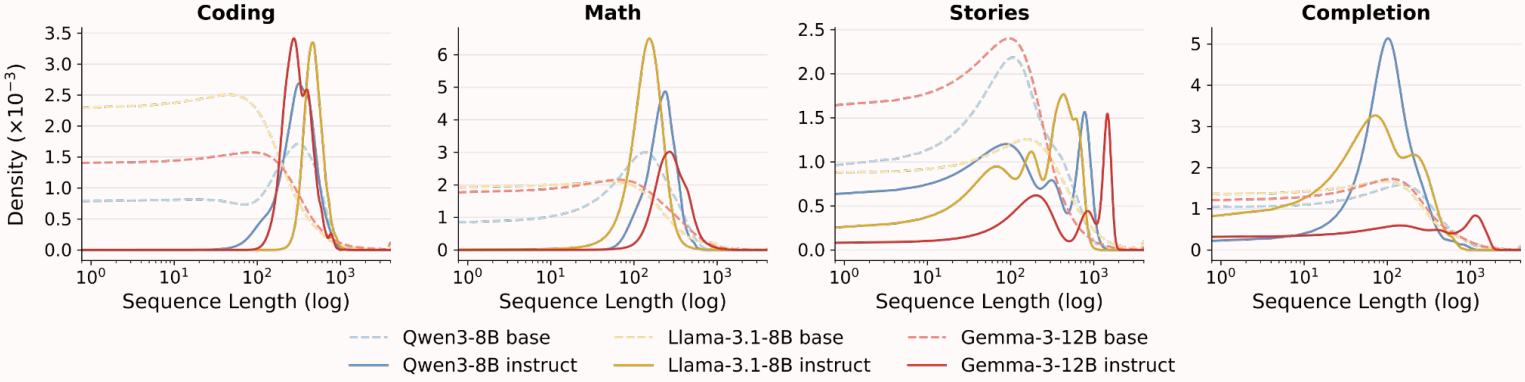
Length-normalized metrics drop, but the total generation space can grow.

SIGMA LAB
Strategic IntelliGence for Machine Agents

Fine-Tuning Sustains Information Along Longer Trajectories

Entropy rate $r_N := \dfrac{H(Y_{1:N} \mid N = n, X)}{n}$ (information density)

Length-entropy correlation $\rho(N, r_N) := \mathbb{E}_X\left[\mathrm{Corr}(N, r_N \mid X = x)\right]$



- Fine-tuned models generate longer sequences.
- Instruct models' longer outputs stay informative rather than redundant.

Metric	Model	MATH			CODING			SENTENCE COMPLETION			STORY GENERATION		
		Base	Instruct	Abs. Change	Base	Instruct	Abs. Change	Base	Instruct	Abs. Change	Base	Instruct	Abs. Change
$\rho(N, r_N)$	Qwen	-0.018 ± 0.016	0.087 ± 0.037	0.11 ± 0.042	-0.41 ± 0.014	0.61 ± 0.034	1.0 ± 0.037	0.062 ± 0.016	0.52 ± 0.023	0.46 ± 0.027	-0.27 ± 0.017	0.16 ± 0.021	0.43 ± 0.027
	Llama	-0.27 ± 0.016	0.38 ± 0.041	0.65 ± 0.045	-0.38 ± 0.011	0.26 ± 0.018	0.64 ± 0.019	-0.26 ± 0.015	-0.29 ± 0.021	-0.028 ± 0.027	-0.24 ± 0.012	-0.13 ± 0.016	0.11 ± 0.022
	Gemma	-0.13 ± 0.014	0.23 ± 0.039	0.36 ± 0.041	-0.27 ± 0.012	0.40 ± 0.036	0.68 ± 0.038	-0.23 ± 0.015	0.15 ± 0.038	0.39 ± 0.040	-0.17 ± 0.012	-0.015 ± 0.026	0.16 ± 0.031

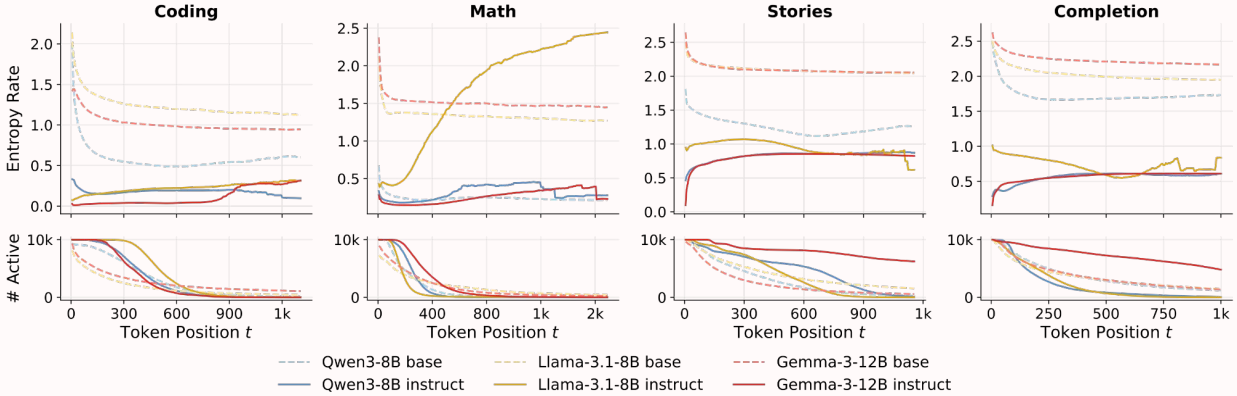
- ρ < 0: longer => redundant
- ρ > 0: longer ⇒ informative

SIGMA LAB
Strategic IntelliGence for Machine Agents

Fine-Tuning Sustains Information Along Longer Trajectories

Entropy rate $r_N := \dfrac{H(Y_{1:N} \mid N = n, X)}{n}$ (information density)

Length-entropy correlation $\rho(N, r_N) := \mathbb{E}_X\left[\mathrm{Corr}(N, r_N \mid X = x)\right]$



In base models, entropy tends to decay as generation progresses.

In fine-tuned models, entropy stays much more stable.

Metric	Model	MATH			CODING			SENTENCE COMPLETION			STORY GENERATION		
		Base	Instruct	Abs. Change	Base	Instruct	Abs. Change	Base	Instruct	Abs. Change	Base	Instruct	Abs. Change
$\rho(N, r_N)$	Qwen	-0.018 ± 0.016	0.087 ± 0.037	0.11 ± 0.042	-0.41 ± 0.014	0.61 ± 0.034	1.0 ± 0.037	0.062 ± 0.016	0.52 ± 0.023	0.46 ± 0.027	-0.27 ± 0.017	0.16 ± 0.021	0.43 ± 0.027
	Llama	-0.27 ± 0.016	0.38 ± 0.041	0.65 ± 0.045	-0.38 ± 0.011	0.26 ± 0.018	0.64 ± 0.019	-0.26 ± 0.015	-0.29 ± 0.021	-0.028 ± 0.027	-0.24 ± 0.012	-0.13 ± 0.016	0.11 ± 0.022
	Gemma	-0.13 ± 0.014	0.23 ± 0.039	0.36 ± 0.041	-0.27 ± 0.012	0.40 ± 0.036	0.68 ± 0.038	-0.23 ± 0.015	0.15 ± 0.038	0.39 ± 0.040	-0.17 ± 0.012	-0.015 ± 0.026	0.16 ± 0.031

- increase: longer ⇒ informative

SIGMA LAB
Strategic IntelliGence for Machine Agents

Fine-Tuning Converts Uncertainty Into Semantic Diversity More Efficiently

Semantic diversity: average pairwise cosine distance between generated outputs (ModernBERT embeddings)

Beta regression: models *semantic diversity* vs. *entropy rate*, controlling for length, task, and model.

- Fine-tuned models start from lower baseline semantic diversity.
- Fine-tuning nearly triples the entropy rate v.s. semantic diversity slope.

