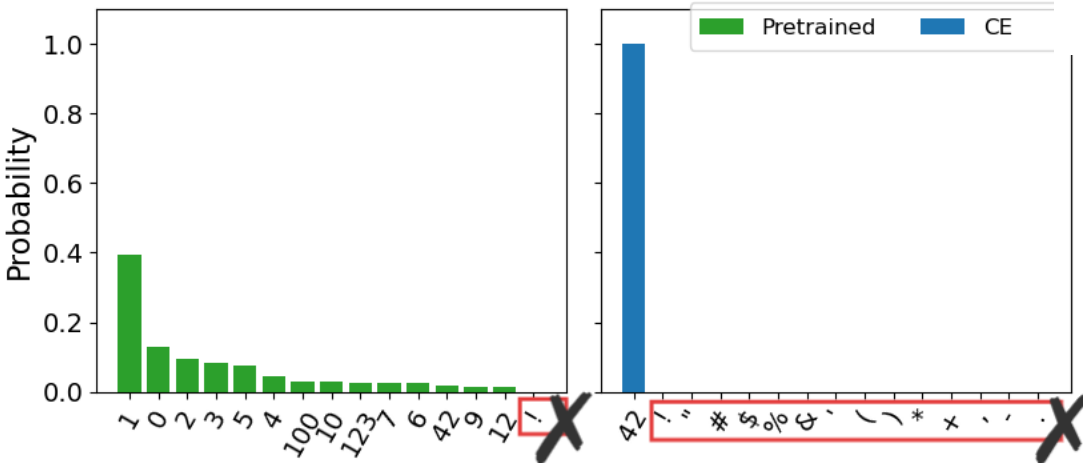
TS^2: Training with Sparsemax+, Testing with Softmax for Accurate and Diverse LLM Fine-Tuning

Ziyang Xu[1*], Ananthu Rajendran Pilla[2*], Yinghua Yao[3†], Yuangang Pan[3]

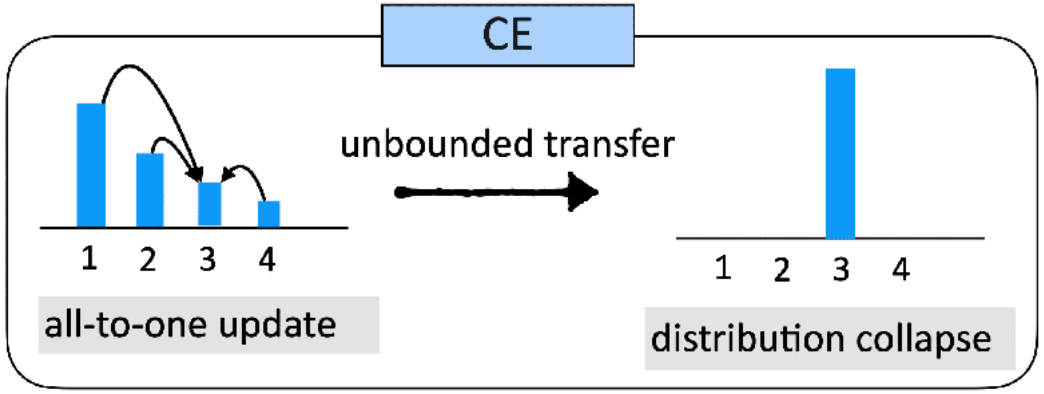
[1]National University of Singapore, Singapore
[2]Nanyang Technological University, Singapore
[3]Agency for Science, Technology and Research, Singapore

Output Diversity is Lost in LLM SFT

- Supervised Fine-Tuning (SFT) with Cross-Entropy (CE) improves instruction-following but often causes token distribution collapse



Token distribution for the question
"… Give me a single-digit number"



Softmax push convergence towards one-hot

Right figure from "Preserving Diversity in Supervised Fine-Tuning of Large Language Models," ICLR 2025

Tail-Suppressed Plausible Diversity (TSPD)

Define $\boldsymbol{p}$ as a token predicted probability distribution and $Top_m(\boldsymbol{p})$ as the indices of the m largest coordinates of $\boldsymbol{p}$.

- If $y \in Top_m(p)$, let $S \coloneqq Top_m(\boldsymbol{p})$
- Otherwise, let $S \coloneqq Top_{m-1}(\boldsymbol{p}) \cup y$

$$\textbf{\textit{(Head Preservation)}} \quad \min_{j \in \mathcal{S}} p_j \geq \varepsilon_{\text{head}},$$

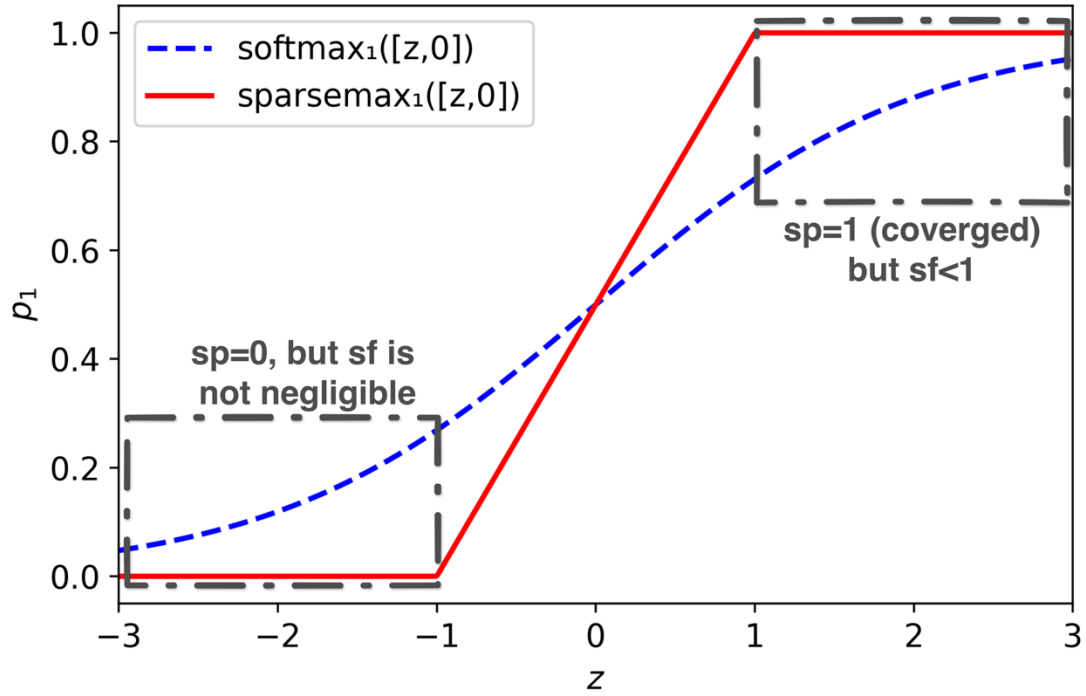
→ Preserve diversity for non-trivial $\varepsilon_{\text{head}}$

$$\textbf{\textit{(Tail Suppression)}} \quad \sum_{j \notin \mathcal{S}} p_j \leq \varepsilon_{\text{tail}}.$$

→ Preserve accuracy for small $\varepsilon_{\text{tail}}$

Training with Sparsemax+, Testing with Softmax

- Decouple the mapping of logit to prob $\boldsymbol{z} \rightarrow \boldsymbol{p}$
  - Training with sparsemax for converged supervision
  - Testing with softmax for output diversity
- Design a sparsemax+ loss with tail suppression



softmax vs. sparsemax

sparsemax mapping:

$$p_i^{sp}(\boldsymbol{z}) = \text{sparsemax}(\boldsymbol{z})_i \coloneqq \max\{z_i - \tau(\boldsymbol{z}), 0\}$$

Sparsemax+ Training Loss

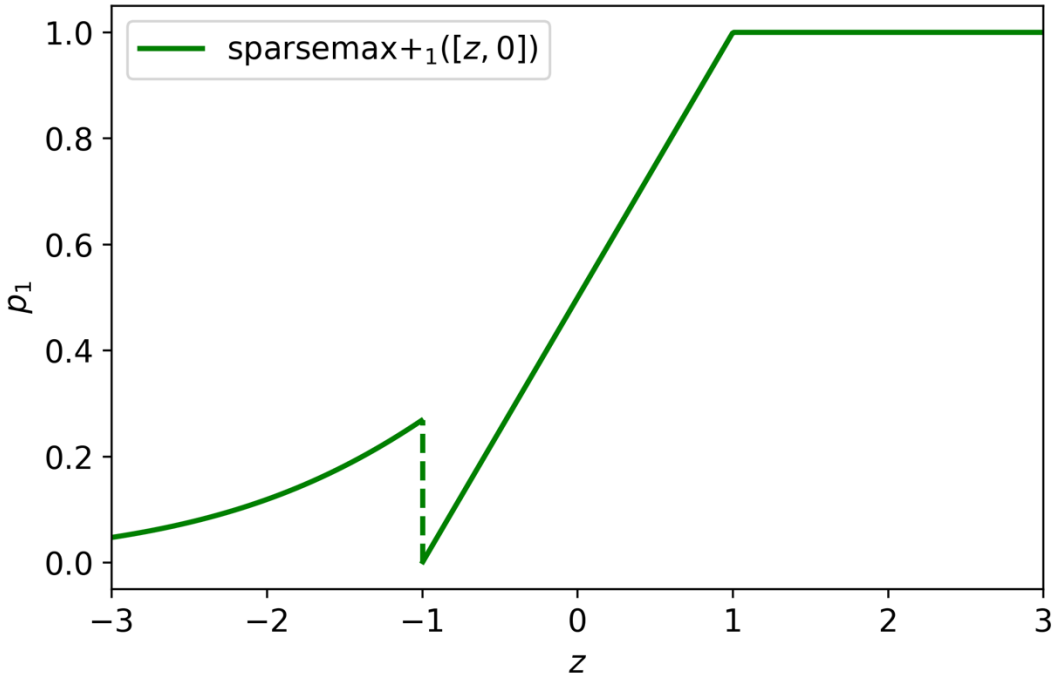
Sparsemax+ loss

$$L_{\mathrm{spm}+}(z;y)$$

$$= L_{\mathrm{spm}}(z;y) + \alpha\left(-\log\left(1 - \sum_{i\notin \mathrm{S}_{\mathrm{sp}}(z), i\neq y} p_i^{\mathrm{sf}}\right)\right)$$

$$L_{\mathrm{spm}}(\boldsymbol{z};y) = -z_y + \frac{1}{2}\sum_{j\in S^{sp}(\boldsymbol{z})}\left(z_j^2 - \tau^2(\boldsymbol{z})\right) + \frac{1}{2}$$

where $p^{\mathrm{sf}} = \mathrm{softmax}(z)$ and $S_{\mathrm{sp}}(z)$ is the sparsemax support.



Experiments

- Base models: Llama-3.1-8B and Qwen-2-7B
- SFT on UltraFeedback dataset
- Baselines: GEM (reverse KL+entropy regularization), CE, CE with Weight Decay (CE+WD), NEFT (adding noise to word embeddings), CE+label smoothing, sparsemax, 1.5-Entmax
- Evaluation
  - win rate & diversity on the **chat** task AlpacaEval dataset
  - pass@k & diversity on the **code** generation HumanEval benchmark
  - diversity on the **creative writing** task
  - generalization on the **OpenLLM** Leaderboard
  - capability on **multi-turn dialogue** benchmark MT-Bench-101

Results

Win rate (Best of N@32) and diversity metrics for Llama-3.1-8B and Qwen-2-7B on `AlpacaEval`.

Model	Method	Win Rate (%) ↑	N-gram ↑	100 - Self-BLEU ↑	Sent-BERT ↑
LLaMA-3.1-8B	CE	29.77	17.78	47.04	9.97
	CE+WD	29.72	17.78	47.14	10.03
	NEFT	29.77	17.74	47.41	10.07
	GEM	31.53	20.32	49.82	11.16
	TS^2 (Ours)	**33.12**	**23.78**	**53.87**	**12.80**
Qwen-2-7B	CE	31.41	17.23	16.77	7.95
	CE+WD	31.05	17.43	17.08	8.06
	NEFT	30.36	16.59	24.59	8.06
	GEM	33.89	24.35	31.19	9.25
	TS^2 (Ours)	**37.48**	**30.15**	**39.04**	**9.81**

Comparison with GEM, CE, CE+label smoothing, sparsemax and Ent-max

	TS^2 (ours)	GEM	CE	CE + label smoothing	sparsemax	1.5-Entmax
Winrate (%)	**33.12**	31.53	29.77	28.25	12.87	13.51

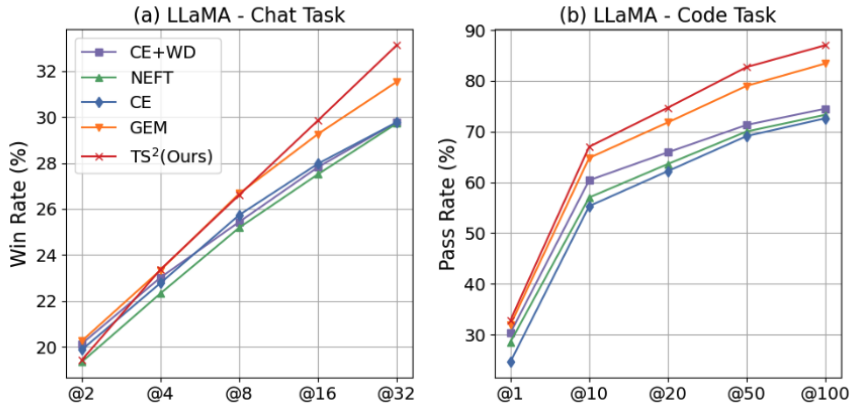
Diversity evaluation on creative writing tasks for Llama-3.1-8B

Method	Poem			Story		
	N-gram ↑	100 - Self-BLEU ↑	Sent-BERT ↑	N-gram ↑	100 - Self-BLEU ↑	Sent-BERT ↑
CE	38.87	55.38	14.83	44.47	67.20	22.15
CE+WD	38.92	55.69	14.17	44.43	67.26	22.22
NEFT	38.80	55.68	14.13	44.31	67.21	22.04
GEM	46.59	57.50	14.70	50.05	69.15	24.02
TS^2 (Ours)	**49.70**	**59.41**	**16.52**	**52.10**	**70.36**	**24.98**

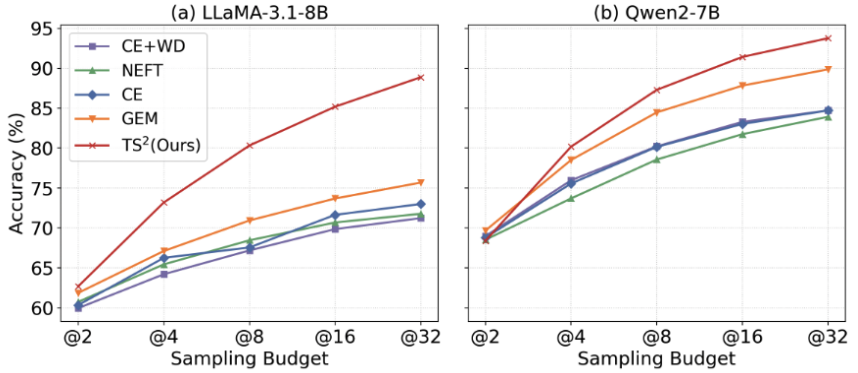
Performance on the multi-turn dialogue dataset `MT-Bench-101`

Model	Avg.	Perceptivity					Adaptability						Interactivity	
		Memory	Understanding		Interference		Rephrasing		Reflection		Reasoning		Questioning	
		CM	SI	AR	TS	CC	CR	FR	SC	SA	MR	GR	IC	PI
CE	5.99	5.01	4.59	6.03	5.10	5.03	7.33	7.02	6.34	7.38	6.37	4.67	7.49	5.46
GEM	6.24	4.63	4.43	6.88	5.95	5.36	7.19	7.76	7.2	8.05	6.09	5.05	6.54	5.98
TS^2(Ours)	**6.65**	**5.96**	**4.76**	6.58	5.94	**6.70**	**8.27**	6.60	**7.42**	**8.53**	6.03	**6.14**	6.80	**6.76**

Performance vs. sampling budget



Generalization on `OpenLLM`



Please find more results in our paper.

Code Link:

Take-away Messages

- Tail-Suppressed Plausible Diversity as a target property for useful generation diversity.

- TS^2, a decoupled recipe: train with Sparsemax+ and test with Softmax for achieving TSPD.
  - This decoupled recipe can extend to more logit->prob mapping functions

Thank you!