Beyond Temperature

Hyperfitting as a Late-Stage Geometric Expansion

M. Li Y. Ding E. Garces Arias C. Heumann

ICML 2026, Seoul

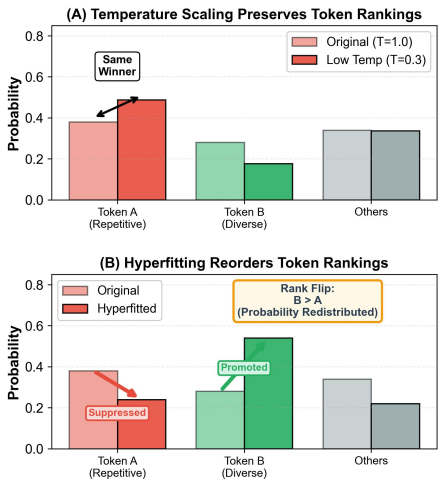
LMU Munich · Henan University · MCML, Munich

The Problem: Greedy Decoding Degenerates

Deterministic decoding (greedy, beam) tends to fall into repetitive loops.

Hyperfitting (Carlsson et al., 2024): fine-tune to near-zero loss on a small dataset. Greedy generation becomes diverse, yet the output stays low-entropy ($H \approx 0.86$ nats).

Is this simply learned temperature scaling?



Research Questions

1. Is Hyperfitting equivalent to temperature scaling?
2. Is the reordering a static vocabulary bias or dynamic and context-dependent?
3. Where in the network does the mechanism live?
4. Can localization enable parameter-efficient fine-tuning?

Experimental Setup

Hyper-saturation

- 2,000 samples, 260 epochs
- train until $\mathcal{L} \to 0$

Models

- TinyLlama-1.1B *(main)*
- Qwen2.5-1.5B, LLaMA-3.2-3B, Gemma-2-2B
- LLaMA-3.1-8B / Qwen2.5-7B Instruct

Metrics

- TTR, bi-/tri-gram repetition, MAUVE
- Top-1 agreement, Spearman ρ, entropy
- Effective dim. (Participation Ratio)

Control

- entropy-matched temperature $T^\star$ per context
- Fiction-Stories, WritingPrompts, AG News

Evidence I: The Entropy–Quality Paradox

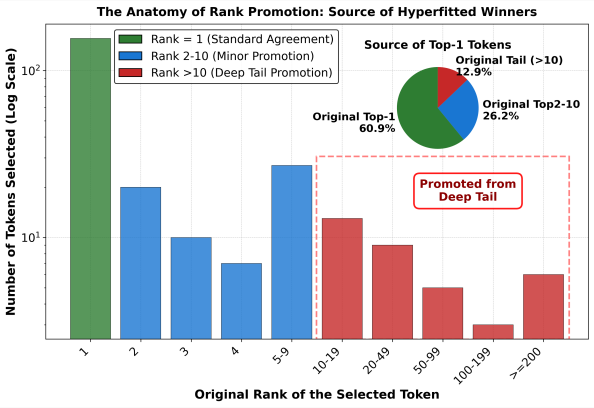
At matched entropy, temperature scaling does not recover diversity; however, Hyperfitting does.

Metric	Orig. T=1.0	Orig. T=0.59	Hyperfit
TTR ↑	0.400	0.397	**0.684**
Bigram Rep. ↓	0.592	0.604	**0.140**
Top-1 Agree. ↓	1.000	0.997	**0.570**
Spearman ρ ↓	1.000	0.998	**0.430**
Entropy (nats)	2.083	0.875	0.862

Same entropy, very different output.

Spearman ρ drops from 1.0 to 0.43: Hyperfitting changes the token ranking, which temperature scaling never does.

Evidence II: Where the Winners Come From



Origin (in the base model) of tokens the hyperfitted model selects:

- 60.9% Rank-1 (linguistic anchor)
- 26.2% Rank 2–10
- 12.9% deep tail (Rank >10)

Temperature scaling is rank-preserving, so it cannot reach the deep tail.

Not a Static Bias

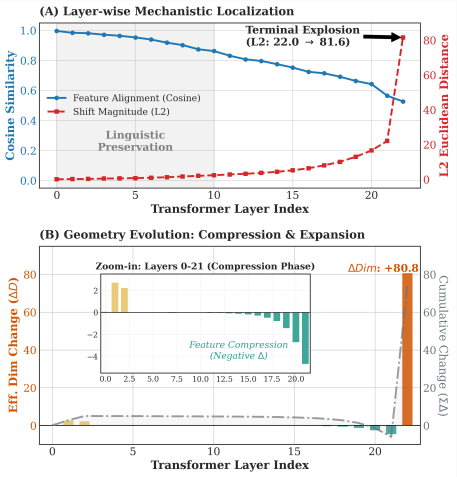
Replaying the learned logit shift *without* training only makes things worse.

Configuration	TTR ↑	Bigram Rep. ↓
Original baseline	0.449	0.588
Synthetic (α=0.01)	0.409	0.609
Synthetic (α=0.10)	0.408	0.616
Synthetic (α=0.50)	0.215	0.706
Hyperfitted	0.679	0.136

A static, global logit bias monotonically degrades quality ($\rho_{\text{TTR vs. } \alpha} = -0.94$).

The reordering is context-dependent.

Mechanistic Localization: Terminal Expansion



- Early layers preserve linguistic features (cosine > 0.86)
- Intermediate layers compress
- Final block expands: L_2 $22.0 \rightarrow 81.6$, $\Delta\mathrm{Dim} \approx +\mathbf{80.8}$

Positive final-layer $\Delta\mathrm{Dim}$ for all models (LLaMA-3.2 $+36.4$, Qwen2.5 $+28.4$, Gemma-2 $+18.3$).

Deterministic, Yet Diverse

Hyperfitted greedy decoding is the only deterministic method that also reaches high MAUVE.

Method	TTR ↑	BiRep ↓	MAUVE ↑	Det.
Orig. greedy	.28	.71	.01–.08	Yes
Min-p (p=.1)	.49–.51	.30–.35	.67–.74	No
GUARD (w=7)	∼.80	∼.07	.46–.53	No
Hyper greedy	.67–.69	.12–.17	**.82–.91**	**Yes**

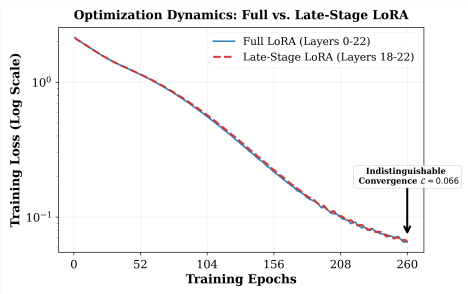
LLaMA-3.2-3B, Gemma-2-2B

Mechanism-Inspired Intervention: Late-Stage LoRA

Model	Full TTR	Late-St. TTR	Param. cut
TinyLlama-1.1B	0.508	0.469	−78%
LLaMA-3.2-3B	0.634	0.600	∼80%
Qwen2.5-1.5B	0.575	**0.591**	−83%
Gemma-2-2B	0.559	**0.608**	∼80%

Tune only the final 5 layers: same loss as Full LoRA, comparable or better diversity.

LLM-judge: preferred in 57.3% of pairs (p=0.02), mainly on coherence (+16.1 pp).



Late-Stage LoRA matches the Full LoRA loss trajectory.

Takeaways

- Hyperfitting is context-dependent rank reordering, not temperature scaling.
- It localizes to an effective-dimensional expansion in the final block.
- Late-Stage LoRA reaches comparable results by tuning only the last 5 layers.

Diverse, deterministic generation by adapting only the final 5 layers.

Thank you!

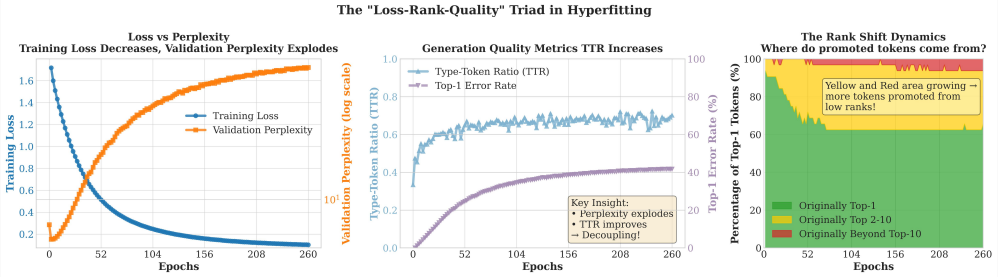
Paper

Code & Data

github.com/YecanLee/Beyond-Temperature

Backup Slides

Backup: Loss–Rank–Quality Dynamics



Training loss → 0 and validation perplexity rises, yet TTR improves; the hyperfitted Top-1 tokens increasingly originate from lower base-model ranks.

Backup: The Paradox Generalizes Across Architectures

Model	TTR (orig)	TTR (hyper)	Top-1 Agree	Final ΔDim
Qwen2.5-1.5B	0.315	**0.434**	0.545	+28.4
LLaMA-3.2-3B	0.311	**0.626**	0.601	+36.4
Gemma-2-2B	0.276	**0.610**	0.629	+18.3

The entropy-matched temperature control fails on every architecture; the ranking change and final-layer expansion are consistent throughout.

Backup: Robustness and Training Efficiency

Cross-domain (TTR / MAUVE)

Model	FS	WP	AG
TinyLlama	.66/.94	.60/.91	.71/.92
Gemma-2-2B	.67/.82	.68/.91	.78/.96
LLaMA-3.2-3B	.65/.27	.61/.91	.70/.97

Training cost (LLaMA-3.2-3B)

Method	Time	Speedup
Full SFT	20.7h	1.00×
Full LoRA	14.4h	1.44×
Late-Stage LoRA	10.4h	1.99×

The effect holds across domains, including the low-entropy AG News.