✦ ICML 2026 SPOTLIGHT

The Tell-Tale Norm: ℓ_2 as Reasoning Signal in LLMs

Jinyang Zhang*[123] Hongxin Ding*[12] Yue Fang*[12] Weibin Liao*[12] Muyang Ye[4] Junfeng Zhao[135] Yasha Wang[12]

[1]School of CS, Peking University [2]National Key Lab of Software Engineering, PKU [3]Key Lab of HCI, PKU Zhejiang University [5]PKU IT Institute (Tianjin Binhai)

Interpretability | Theory-Backed | Training-Free | 📍 COEX, Seoul R

LLMs can reason now — but how do we know WHEN they're actually thinking?

People use **output entropy** to detect reasoning at the decoding level... but that only sees the surface! What's happening **inside** the hidden layers?

Let's start from SAE!

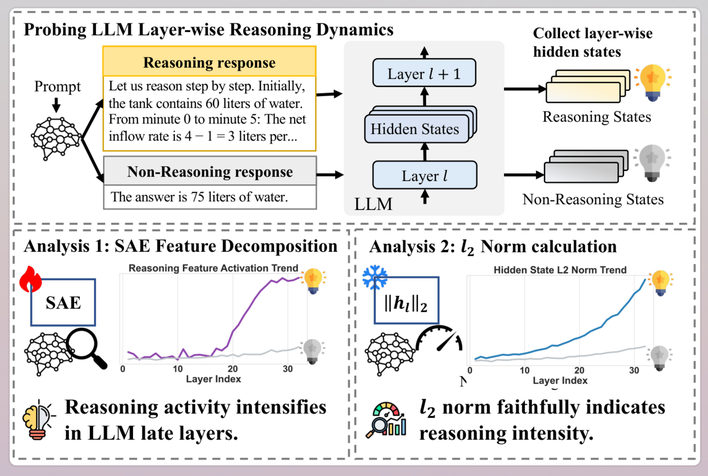
We use **Sparse Autoencoders** to probe hidden layers, comparing **thinking** vs **non-thinking** on same prompts.

Our exploration:

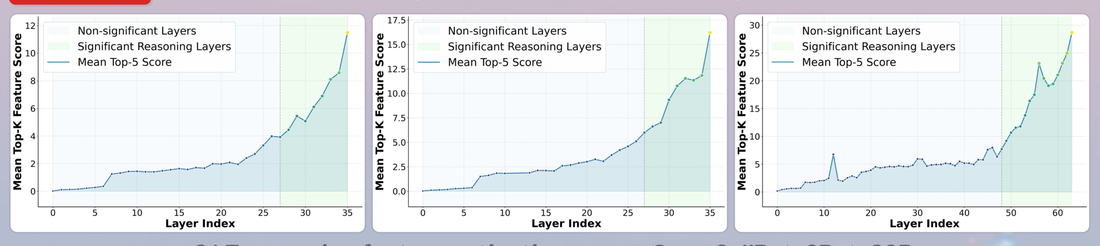
① SAE reveals late-layer dynamics

② ℓ_2 norm = same signal, **free!**

③ Three test-time techniques



Key finding! Reasoning features **spike in final ~25% of layers!**



SAE reasoning feature activation across Qwen3: 4B → 8B → 32B

Fascinating! But SAE needs **extra training for each model** — that's expensive and model-specific 😩

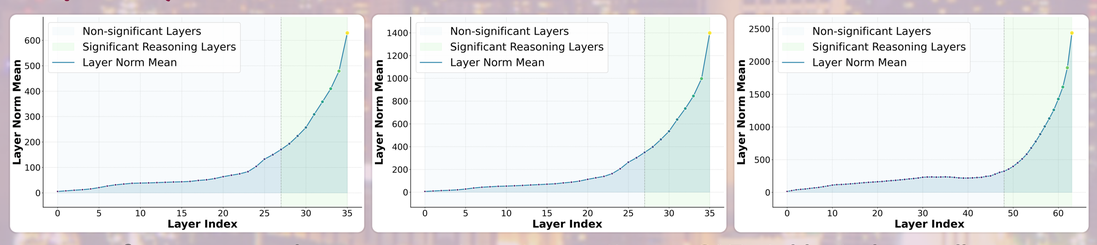
Is there an **endogenous metric** that correlates with SAE reasoning activation? Something we can compute for free?

Good question! Let me compare several layer-wise metrics against my SAE activations:

- **Layer Hidden Entropy** — captures state uncertainty
- **Layer Output Entropy** — measures prediction uncertainty
- **Attention Entropy** — reflects attention dispersion
- **ℓ_2 Norm** — just the magnitude of hidden states

But hidden/output/attention entropy all show **negative or weak** correlation with reasoning... 🤔

Pick me! I'm just $\|h\|_2$ — the simplest possible metric. But look at my trends — **I mirror SAE perfectly!**



ℓ_2 norm across layers — same two-stage pattern as SAE: stable → sharp spike

Looks promising visually... but can you **prove it formally?** And show quantitative evidence?

Theorem proved! ✅

$$\sum_{i \in \mathcal{I}_{reason}} \Delta z_i = \Theta\left(\sqrt{\|h_{think}\|_2^2 - \|h_{non}\|_2^2}\right)$$

Two-sided bound: ℓ_2 norm shift ⟺ reasoning feature activation (tight!)

Intuition: When reasoning activates, extra features fire on top of semantic representations. Due to near-orthogonality, they add constructive energy → $\|h\|_2$ grows!

What **assumptions** does your proof rely on? Are they actually valid? 🧐

All verified! ✅ **4 assumptions — all empirically validated:**

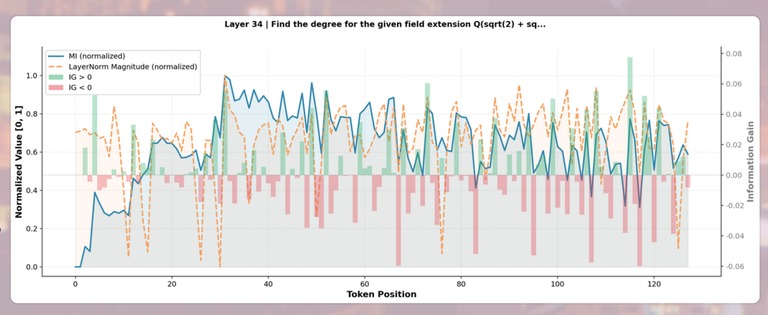
Assumption	Verification
A1: Decoder Orthogonality	Cosine sim < 0.02 across all layers
A2: Reconstruction Fidelity	$R^2 > 0.95$, residual < 5%
A3: Semantic ⊥ Reasoning	Cross-component cosine sim < 0.1
A4: Disentanglement	< 3% overlap in activation patterns

Validated across Qwen3 (1.7B–32B) & LLaMA3 — Appendix B

Numbers don't lie!

Model	SAE ρ	Ent. ρ
Qwen-1.7B	**86.5**	63.5
Qwen-8B	**85.2**	63.3
Qwen-14B	**87.6**	66.7
LLaMA-8B	**84.1**	52.1

ℓ_2 norm: ρ = 84–88% with SAE.
Entropy metrics all negative!



Is there prior theoretical support? Why would norm grow with reasoning? 🤔

ICCV insight!

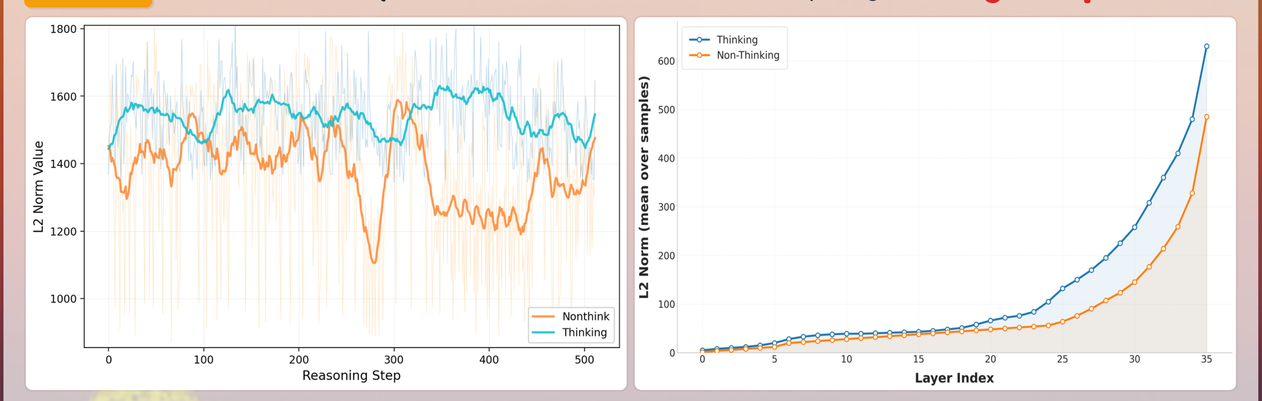
Xu et al. proved that cross-entropy loss **inherently drives feature norms up** — minimizing loss = larger norm + confident alignment.

In LLMs, reasoning tokens are the **hardest to predict** (low p_k, large gradient). After training on reasoning data, these tokens accumulate the most optimization pressure → **highest ℓ_2 norms!**

$$\frac{\partial \mathcal{L}}{\partial h} = \sum_j (p_j - \mathbb{1}_{[j=k]}) \cdot e_j \rightarrow \text{ large gradient when } p_k \ll 1$$

The norm records cumulative optimization force — reasoning tokens get the most!

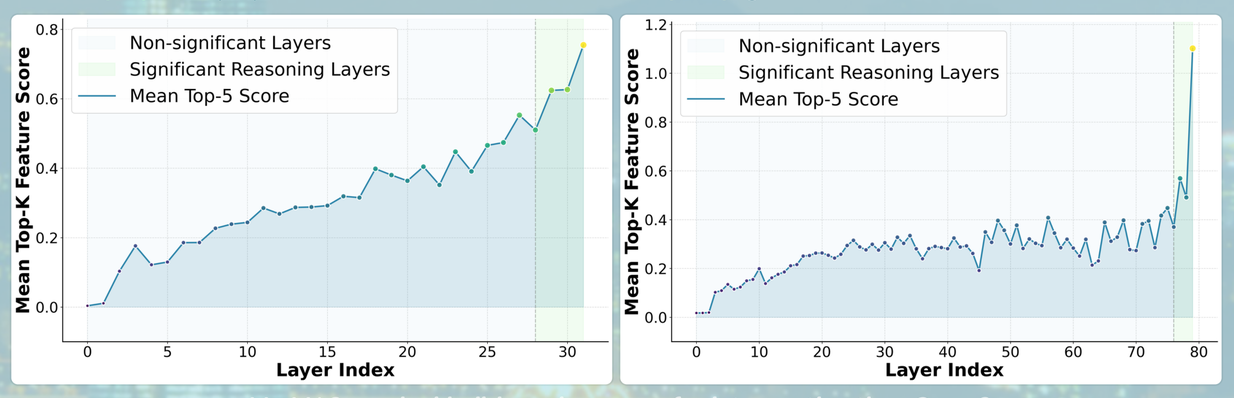
See it live! Watch me **token by token** — when the model thinks, my magnitude **lights up!** 🔥



Left: Token-wise norms at Layer 34 · Right: Layer-wise norms — thinking always higher

Training paradigm shapes reasoning geometry!

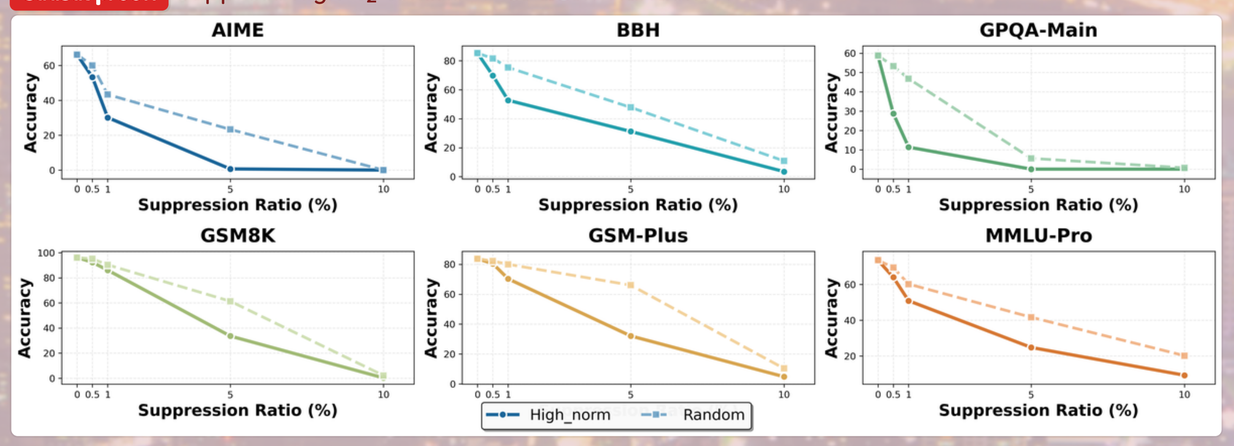
Qwen3 (RL): sharp spike → **+10% AIME!** **LLaMA3 (SFT)**: gradual → **+2.8% uniform**



LLaMA3: gradual buildup — less room for intervention than Qwen3

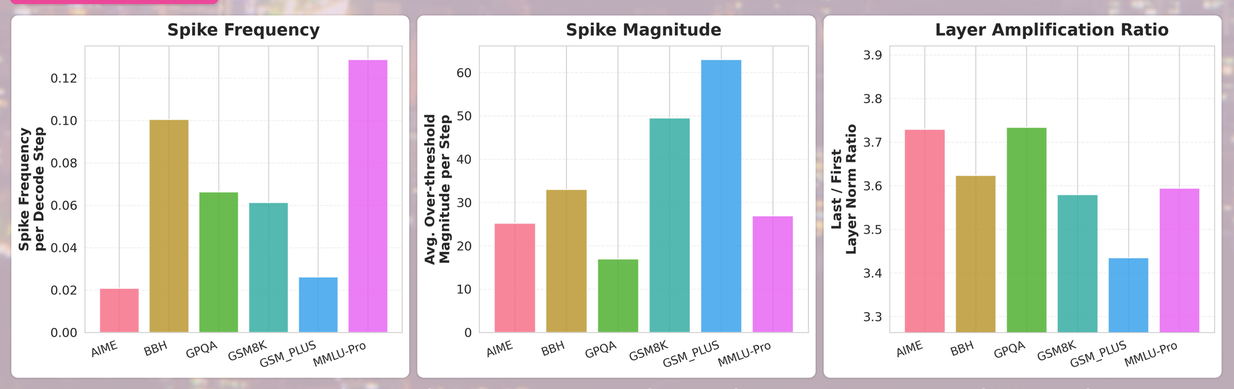
Convincing! But do we have **causal evidence?**

Causal proof! Suppress high-ℓ_2 tokens → **crash!** Random → **fine.**



Qwen3-14B: high-norm suppression (red) vs random (blue)

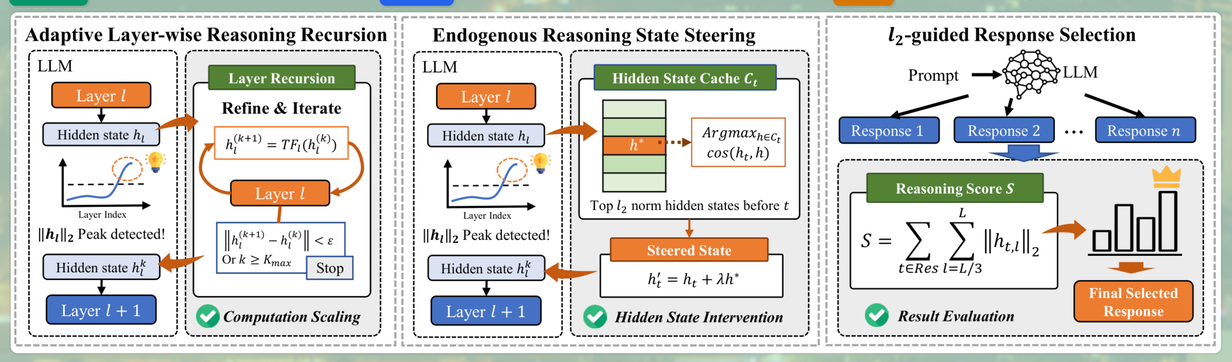
Cross-Benchmark! **ℓ_2 reveals task domain AND difficulty:**

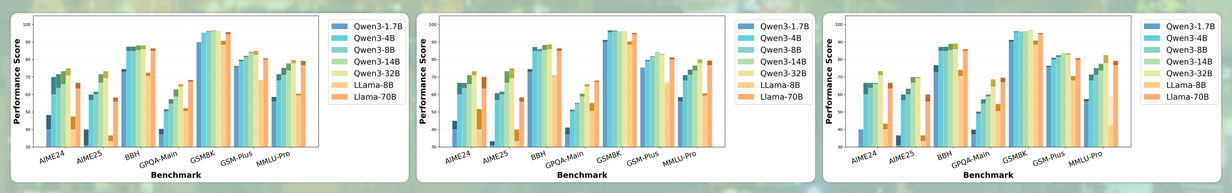


Per-layer norms (domain) · Spike magnitude (domain) · Amplification ratio (difficulty)

Applications! **3 training-free methods:**

ALRR Re-process at ℓ_2 peaks **ERSS** Steer toward reasoning states **LRS** Rank by aggregated ℓ_2





ALRR · ERSS · LRS — consistent gains with zero training data