

ZeroTuning (ICLR'26)

Question: How can we steer attention to improve LLM performance and reduce hallucinations?

Prior approaches:

- Manually (or via LLMs) identify important tokens, then **increase** their attention (*PASTA*, *AutoPASTA*)
- Use threshold-based rules to detect sink tokens (e.g., punctuation) and **decrease** their attention (*ACT*)

Problems / limitations:

- Brittle token-localization logic (easy to misidentify which tokens matter)
- Extra overhead to locate tokens / compute steering signals
- Requires attention maps, which limits compatibility with optimized kernels like **SDPA** / **FlashAttention**

ZeroTuning (ICLR'26)

Research Question: *Is it possible to enhance model performance by steering a **universal, task-agnostic** token, thereby bypassing the need for fragile, task-specific token identification?*

The Answer is YES!

Key finding: Steering the initial token (BOS) outperforms prior methods.

ZeroTuning (ICLR'26)

2023

Observation: BOS (initial token) identified as an “attention sink” [ICLR]

2024 - 2025

Understanding

- Why sinks emerge (softmax normalization, causal masking, and architectural biases) [ACL, ICLR, COLM]
- When sinks matter (keeping sink-like anchors can stabilize performance for long-context settings) [Neurips]
- More sink tokens discovered beyond BOS (e.g., punctuation / special markers) [ICML]

Exploitation

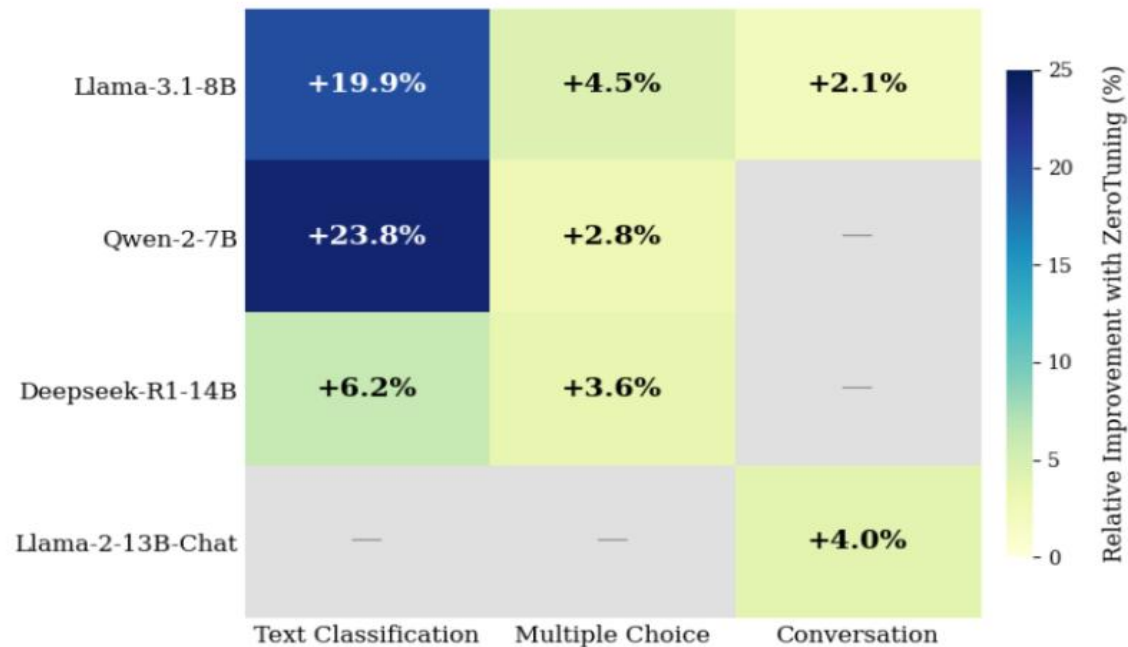
- Sink-aware techniques applied to KV-cache efficiency, quantization stability [ICLR, Neurips, ACL]

2026

Sink-aware techniques applied to reduce hallucinations and improve LLMs directly (**Our paper**) [ICLR]

ZeroTuning (ICLR'26)

```
Class LlamaAttention(nn.Module):
  def forward(self, target_layers, target_heads, scaling_factor, ...):
    # ... omitting unmodified LLamaAttention code
    # 1. Standard attention weight calculation
    attn_weights = F.softmax(torch.matmul(query_states,
                                          key_states.transpose(2, 3)), dim=-1)
    # 2. Our [ZeroTuning] Method
    if self.layer_idx in target_layers:
      # Shape: (bsz, num_heads, q_len, kv_len)
      attn_weights[:, target_heads, :, 0] *= scaling_factor
      # Re-normalize the Attention
      attn_weights[:, target_heads] =
        F.normalize(attn_weights[:, target_heads], p=1, dim=-1)
    # 3. Compute attention output
    attn_output = torch.matmul(attn_weights, value_states)
    # omitting unmodified LLamaAttention code ...
```



Core pipeline (3 steps):

- Head profiling: Split heads into up-effective vs down-effective (which direction of initial-token scaling helps)
- Selective rescaling: Search a scale γ , then apply it only to the dominant head group
- Renormalization: Re-normalize attention (softmax) to keep a valid distribution

Calibration options:

- Supervised: Choose heads + γ by maximizing accuracy on a labeled calibration set
- Unsupervised: Choose heads + γ by minimizing next-token entropy on unlabeled inputs

Implementation note: If attention scores aren't accessible (e.g., SDPA/FlashAttention), apply scaling to Q/K states instead

ZeroTuning (ICLR'26)

$$\mathbf{a} = [a_0, a_1, \dots, a_{T-1}], \quad \text{where } a_i \geq 0 \quad \text{and} \quad \sum_{i=0}^{T-1} a_i = 1.$$

$$a'_0 = \frac{\gamma a_0}{D}, \quad a'_i = \frac{a_i}{D} \quad \text{for } i = 1, \dots, T-1,$$

$$\frac{a'_i}{\sum_{j=1}^{T-1} a'_j} = \frac{\frac{a_i}{D}}{\sum_{j=1}^{T-1} \frac{a_j}{D}} = \frac{a_i}{\sum_{j=1}^{T-1} a_j}, \quad \text{for } i \geq 1,$$

This rescaling preserves the relative importance among all non-initial tokens

$$a'_i - a'_j = \frac{a_i - a_j}{D} = \frac{a_i - a_j}{(\gamma - 1)a_0 + 1}$$

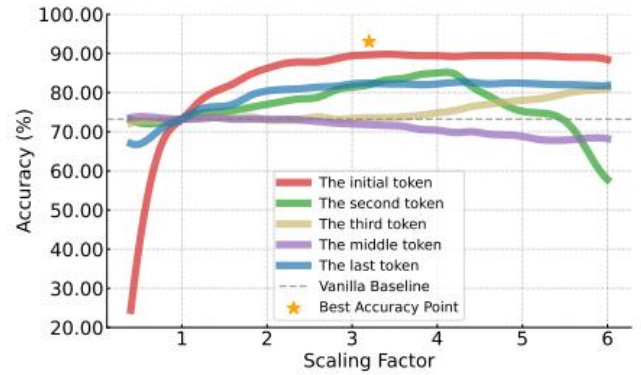
but compresses or expands their differences

$$E_{\text{diff},i,j} = |(a'_i - a'_j) - (a_i - a_j)| = |a_i - a_j| \left| \frac{1}{(\gamma - 1)a_0 + 1} - 1 \right| = |a_i - a_j| \frac{|\gamma - 1|a_0}{(\gamma - 1)a_0 + 1}$$

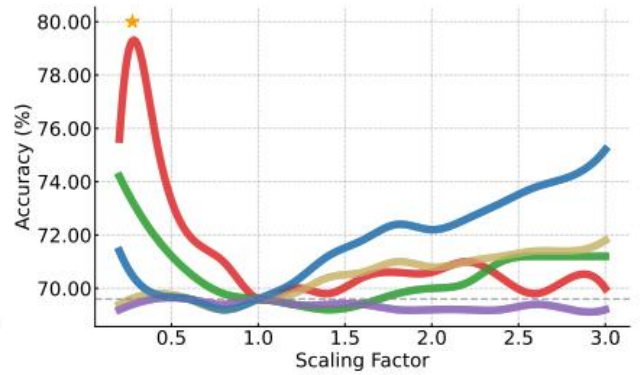
$$\frac{\partial E_{\text{diff},i,j}}{\partial a_0} = |a_i - a_j| |\gamma - 1| \cdot \frac{1}{((\gamma - 1)a_0 + 1)^2}$$

The larger the initial token's attention, the more effectively it can act as a lever to control the attention distribution

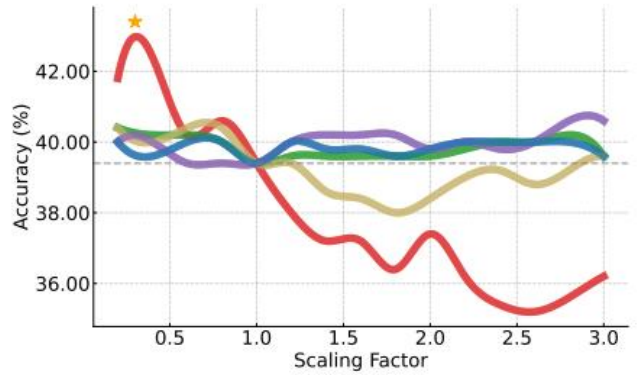
ZeroTuning (ICLR'26)



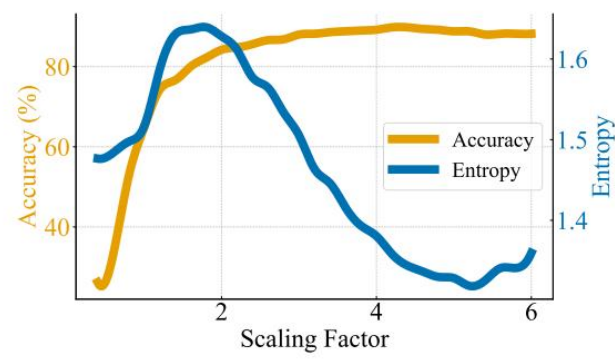
(a) SST-2



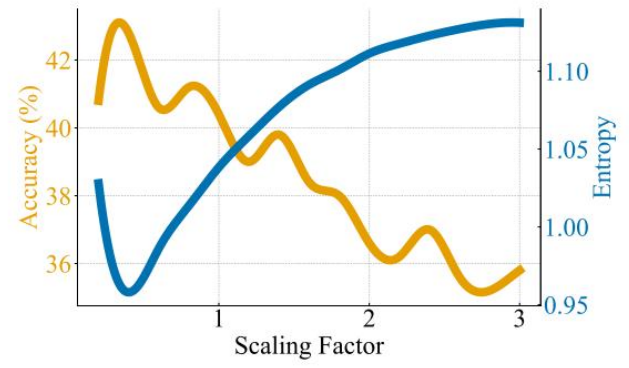
(b) BoolQ



(c) LogiQA



(a) SST-2



(b) LogiQA

ZeroTuning (ICLR'26)

Up-scaling ($\gamma > 1$) benefits tasks requiring holistic context integration.

SST2 Q: Classify the sentence into one of the following sentiments: positive or negative.

Sentence: “**hardly** a masterpiece, **but** it introduces viewers to a **good** charitable enterprise and some **interesting** real people.” **Sentiment:**

Vanilla Output: negative. Explanation: The sentence describes the movie as "hardly a masterpiece", which is a negative description. The positive aspects mentioned in the sentence are secondary to the negative description. Therefore, the overall sentiment of the sentence is negative.

ZeroTuning Output ($\gamma = 4$): positive. Explanation: The word "good" in the sentence has a positive connotation, and the word "interesting" also has a positive connotation. This indicates that the overall sentiment of the sentence is positive. The word "hardly" is a negative word, but in this context, it is used to express a mild criticism, so it does not dominate the overall sentiment of the sentence. Therefore, the sentence is classified as having a positive sentiment.

Ground Truth: positive

ZeroTuning (ICLR'26)

Up-scaling ($\gamma > 1$) benefits tasks requiring holistic context integration.

SST2 Q: Classify the sentence into one of the following sentiments: positive or negative.

Sentence: “generally, clockstoppers will **fulfill** your **wildest fantasies** about being a different kind of time traveler, **while** happily **killing** 94 minutes.” **Sentiment:**

Vanilla Output: negative. Explanation: Although the sentence mentions "wildest fantasies," which can be interpreted as positive, the overall tone is negative due to the phrase "happily killing 94 minutes." This phrase implies that the movie is a waste of time and that the viewer will lose 94 minutes of their life watching it. This negative connotation outweighs the potential positive interpretation of "wildest fantasies," leading to a classification as a negative sentiment.

ZeroTuning Output ($\gamma=4$): positive. Reasoning skill: Emotion Recognition And Sentiment Analysis. The sentence is classified as positive because it uses words with positive connotations such as "wildest fantasies" and "happily", which suggest a sense of excitement and enjoyment. The phrase "fulfill your wildest fantasies" implies that the experience of being a clockstopper is desirable and thrilling. Additionally, the word "happily" reinforces this positive sentiment.

Ground Truth: positive

ZeroTuning (ICLR'26)

Down-scaling ($\gamma < 1$) benefits tasks requiring sharp focus on critical evidence

BoolQ Read the text and answer the question by True or False.

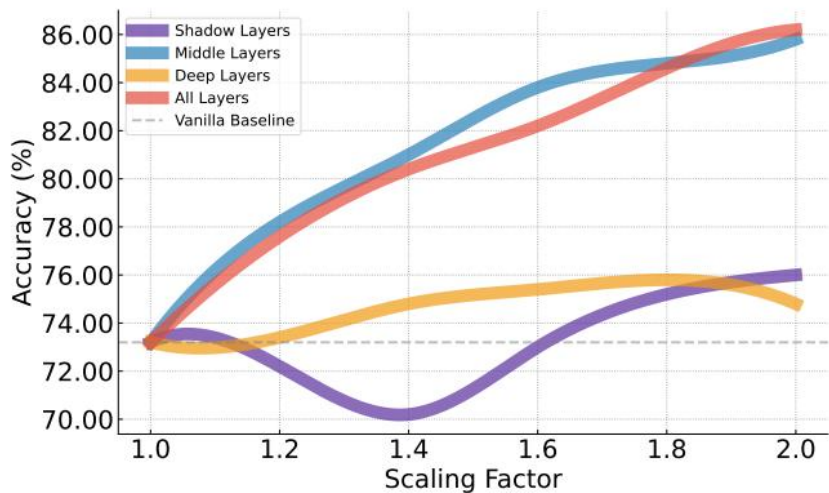
Text: Hawaii Five-0 (2010 TV series, **season 8**) -- The **eighth season** of the CBS police procedural drama series Hawaii Five-0 premiered on September 29, 2017 for the 2017--18 television season. CBS renewed the series for a 23 episode eighth season on March 23, 2017. However, on November 6, 2017 CBS ordered an additional episode for the season and did the same again on February 8, 2018 bringing the count to 25 episodes. The season concluded on May 18, 2018. The eighth season ranked \#18 for the 2017-18 television season and had an average of 11 million viewers. The series was also renewed for a ninth season. **Question:** will hawaii five o have a season 8? **Answer:**

Vanilla Output: False. The text says the series has already concluded its 8th season. (Note: this question is not about whether the series will be renewed for a 9th season, but whether it will have an 8th season, which it has already done). (Note: the series was renewed for a 9th season, but that is not relevant to the question).

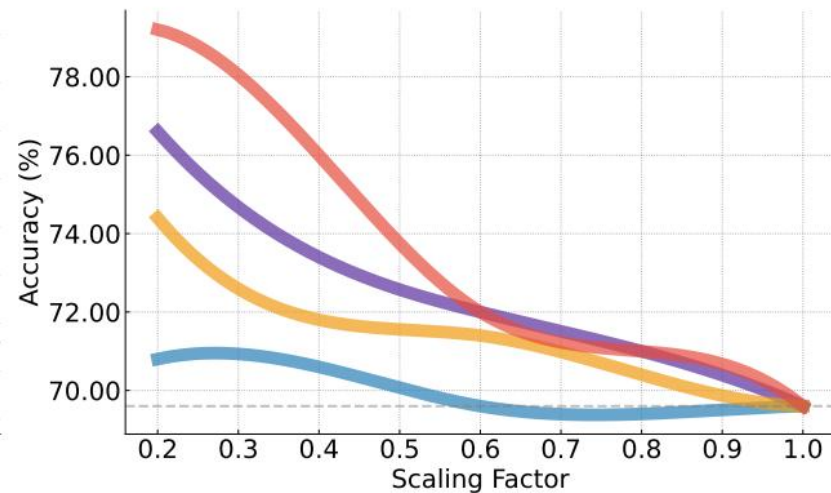
ZeroTuning Output ($\gamma=0.3$): True. (It is already confirmed by CBS). The season premiered on September 29, 2017 and concluded on May 18, 2018. It was renewed for a 23 episode eighth season on March 23, 2017 and had 25 episodes in total. (The extra two episodes were ordered on November 6, 2017 and February 8, 2018). The series also ranked 18 for the 2017-18 television season and had an average of 11 million viewers.

Ground Truth: True

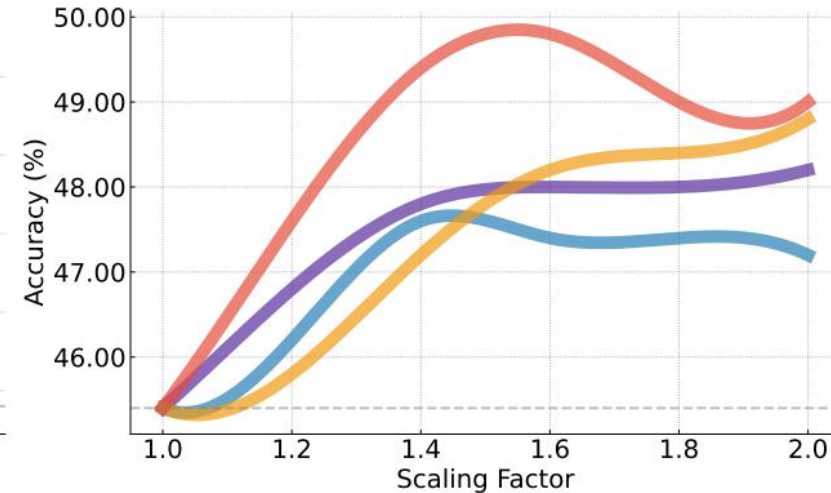
ZeroTuning (ICLR'26)



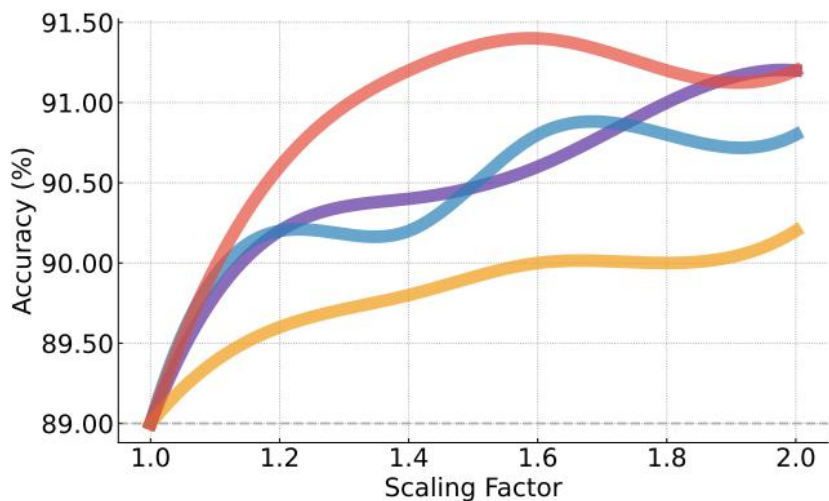
(a) SST-2



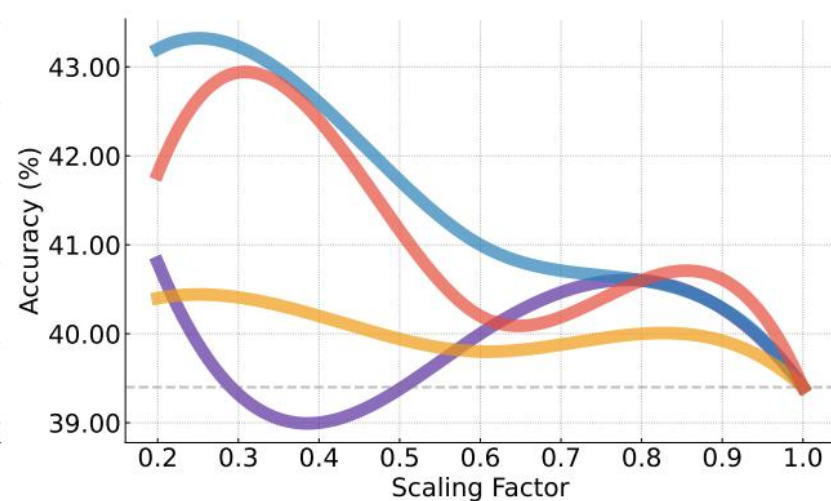
(b) BoolQ



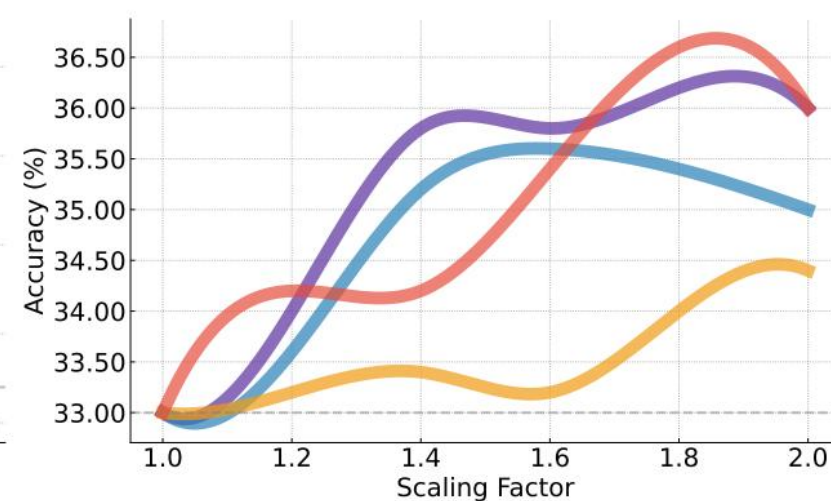
(c) SST-5



(d) MR

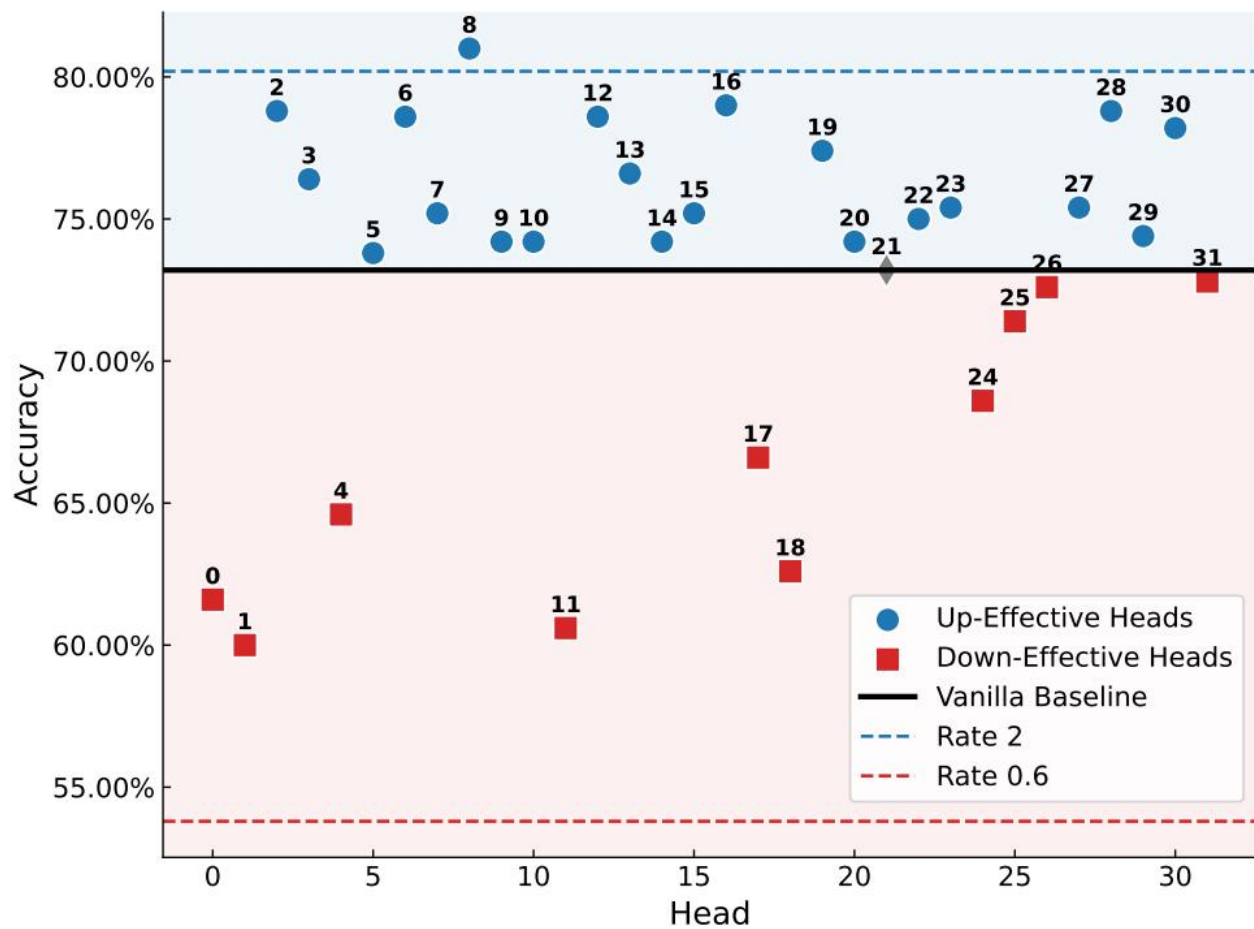


(e) LogiQA

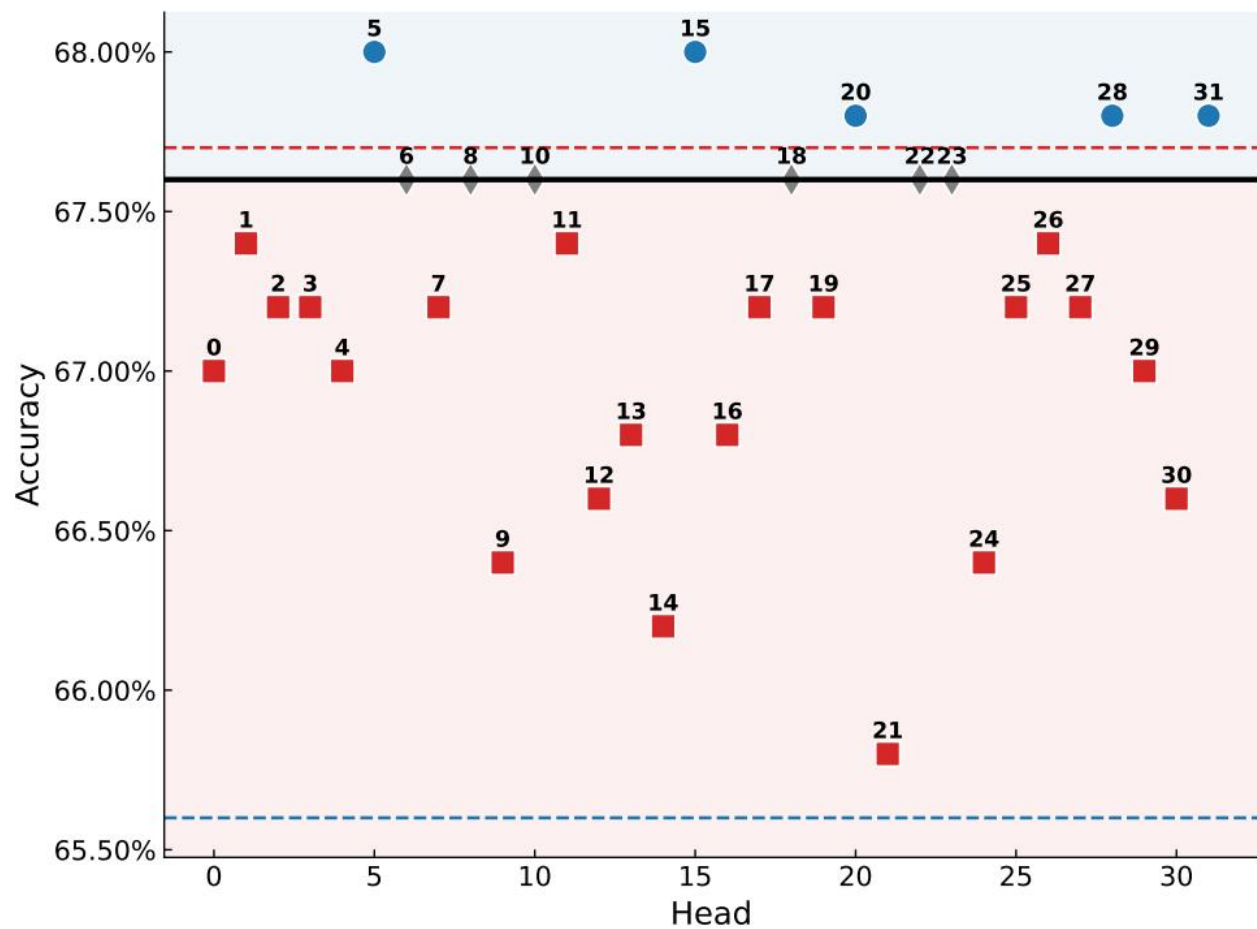


(f) MathQA

ZeroTuning (ICLR'26)

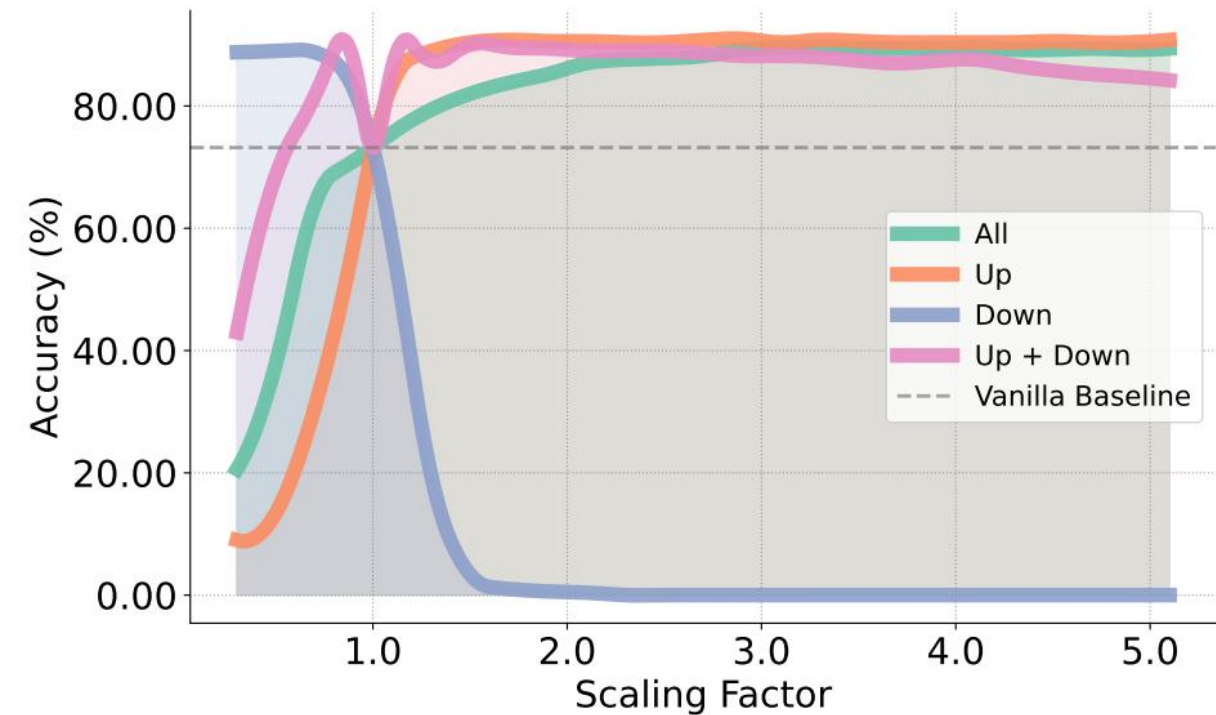


(a) SST-2

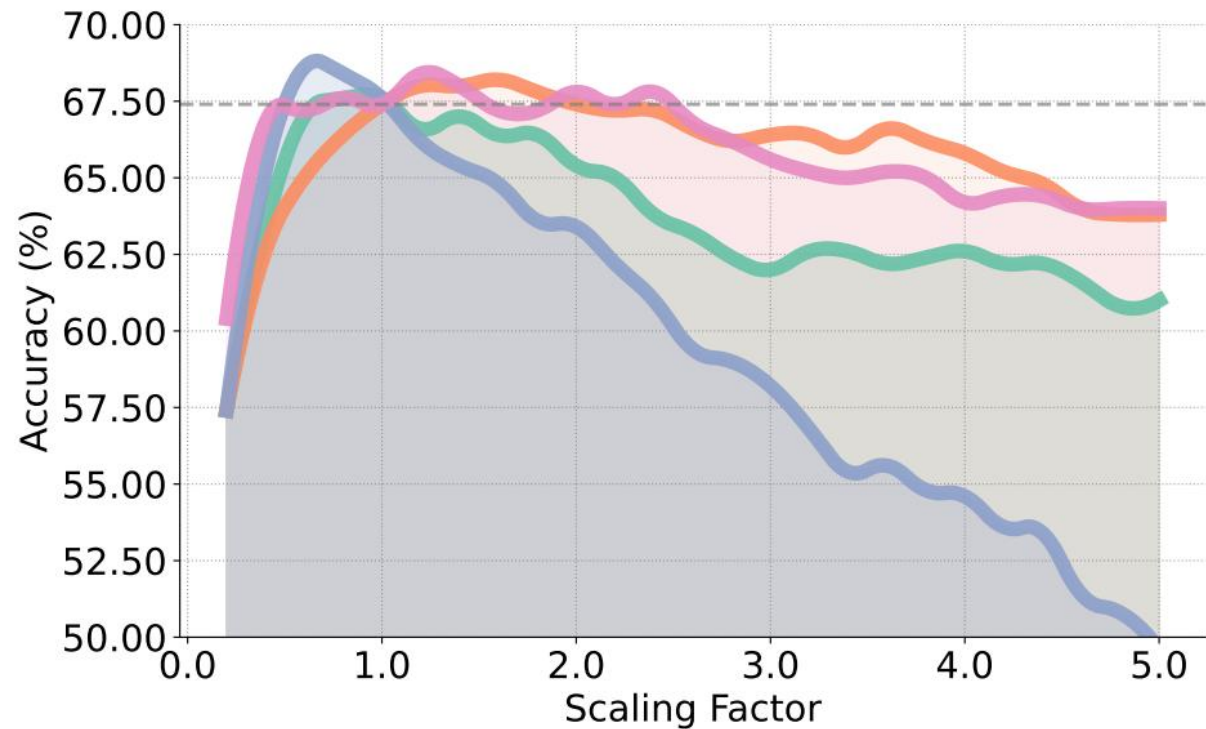


(b) MMLU

Representative Research Project: ZeroTuning (ICLR'26)



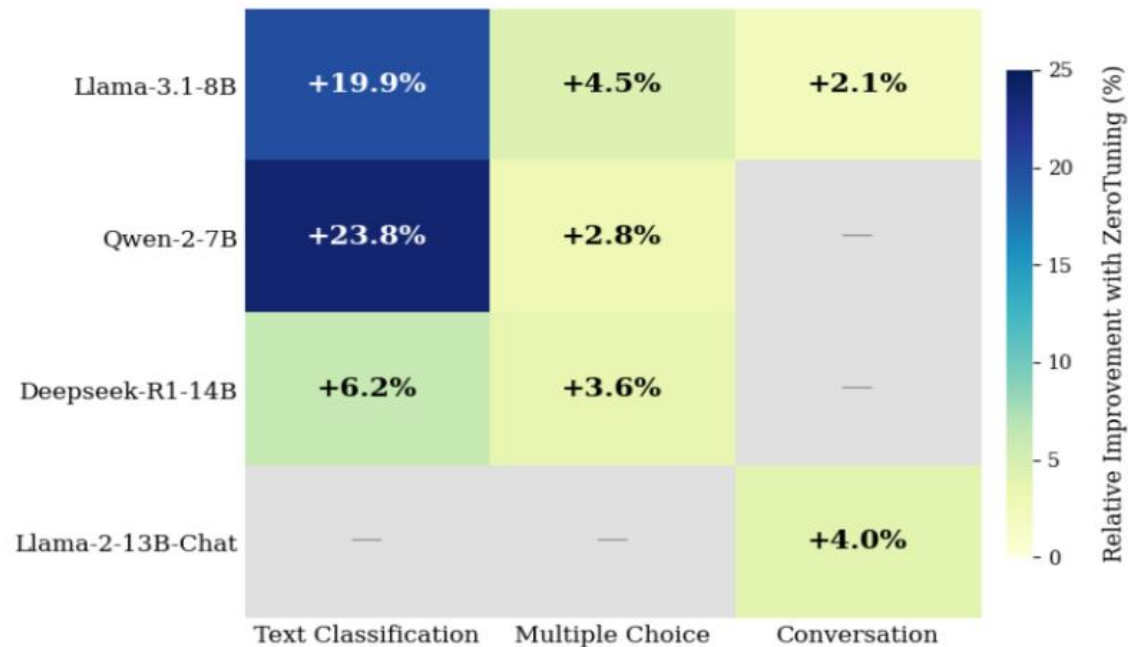
(a) SST-2



(b) MMLU

ZeroTuning (ICLR'26)

```
Class LlamaAttention(nn.Module):
  def forward(self, target_layers, target_heads, scaling_factor, ...):
    # ... omitting unmodified LLamaAttention code
    # 1. Standard attention weight calculation
    attn_weights = F.softmax(torch.matmul(query_states,
                                          key_states.transpose(2, 3)), dim=-1)
    # 2. Our [ZeroTuning] Method
    if self.layer_idx in target_layers:
      # Shape: (bsz, num_heads, q_len, kv_len)
      attn_weights[:, target_heads, :, 0] *= scaling_factor
      # Re-normalize the Attention
      attn_weights[:, target_heads] =
        F.normalize(attn_weights[:, target_heads], p=1, dim=-1)
    # 3. Compute attention output
    attn_output = torch.matmul(attn_weights, value_states)
    # omitting unmodified LLamaAttention code ...
```



Core pipeline (3 steps):

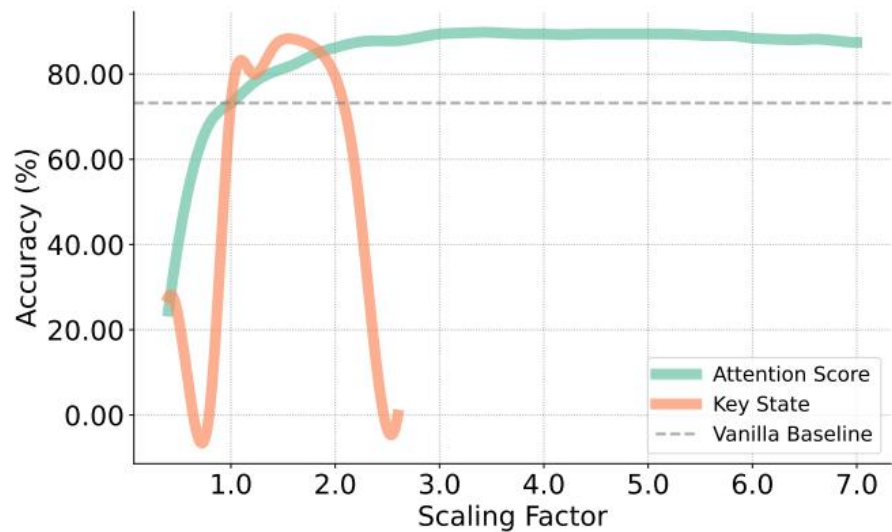
- Head profiling: Split heads into up-effective vs down-effective (which direction of initial-token scaling helps)
- Selective rescaling: Search a scale γ , then apply it only to the dominant head group
- Renormalization: Re-normalize attention (softmax) to keep a valid distribution

Calibration options:

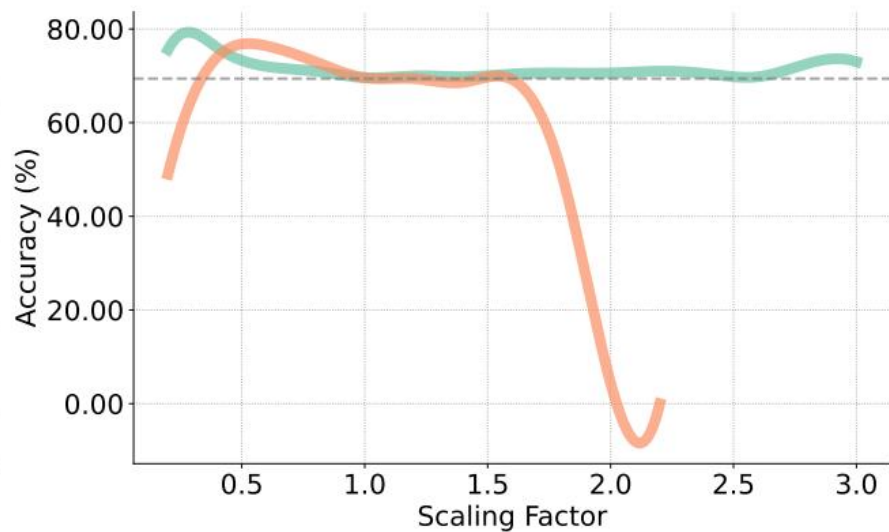
- Supervised: Choose heads + γ by maximizing accuracy on a labeled calibration set
- Unsupervised: Choose heads + γ by minimizing next-token entropy on unlabeled inputs

Implementation note: If attention scores aren't accessible (e.g., SDPA/FlashAttention), apply scaling to Q/K states instead

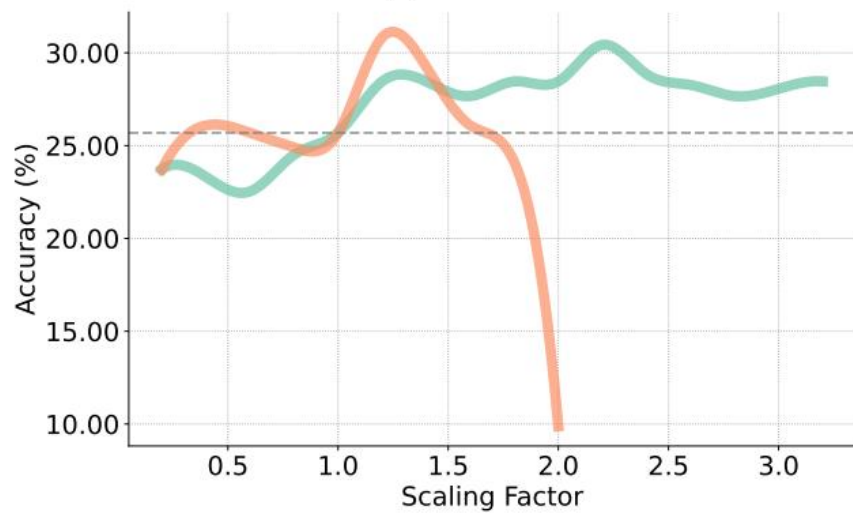
ZeroTuning (ICLR'26)



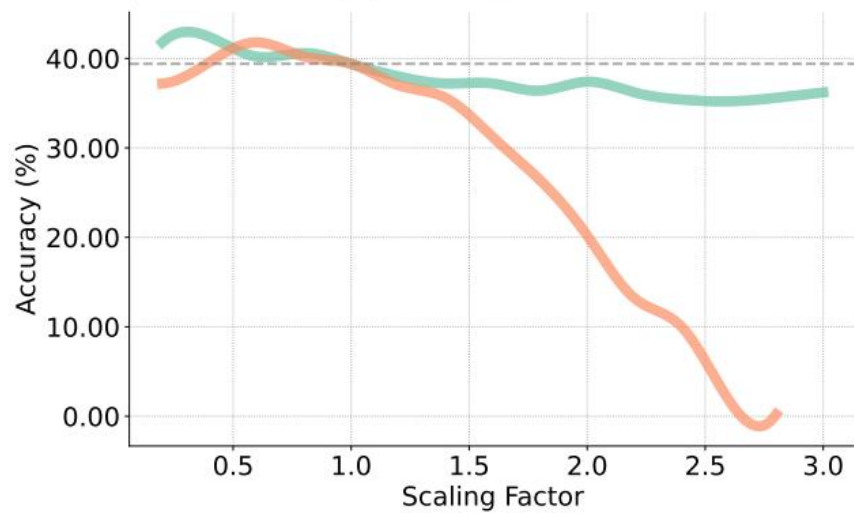
(a) SST-2



(b) BoolQ

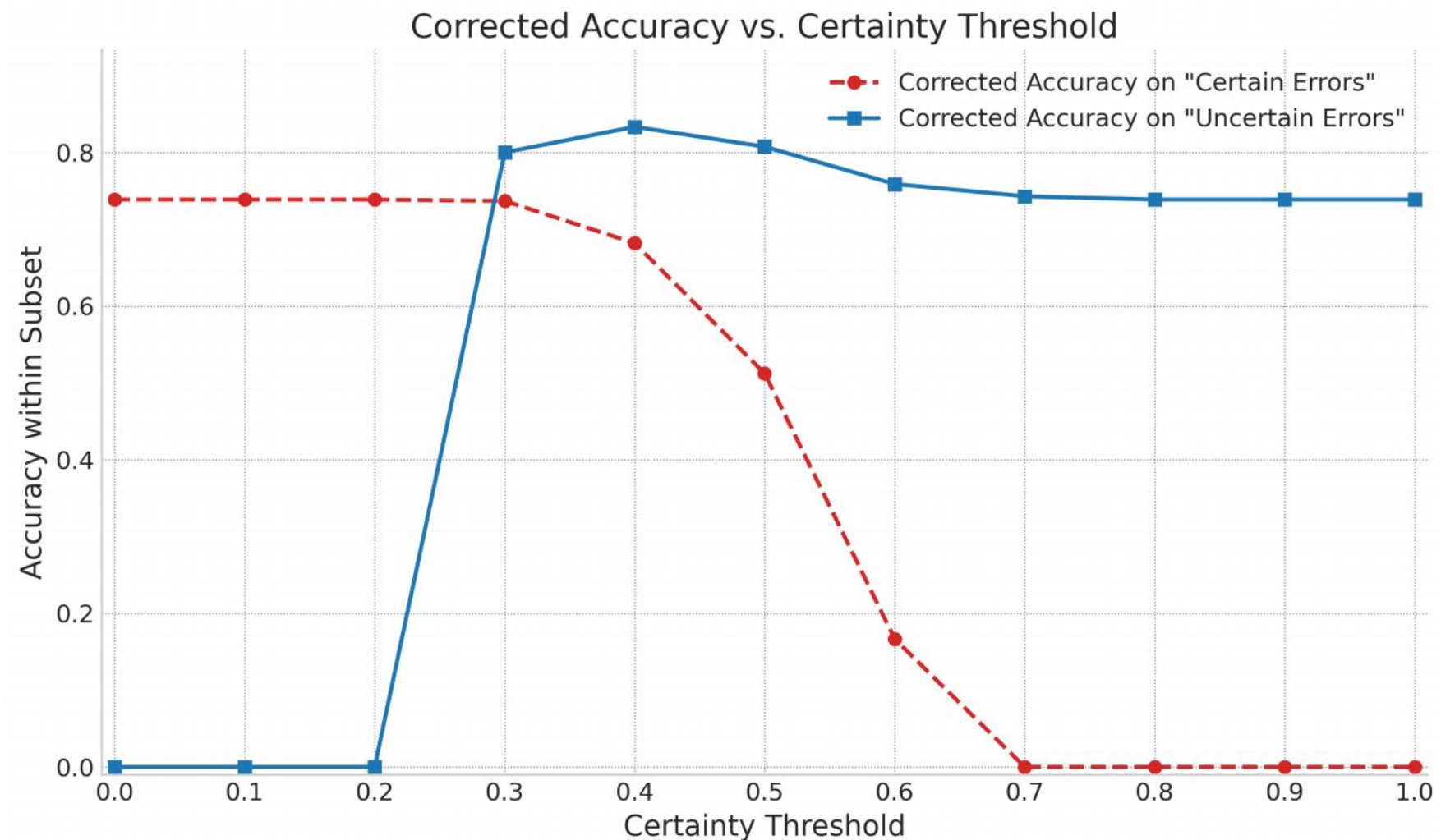


(c) AQUA



(d) LogiQA

ZeroTuning (ICLR'26)



ZeroTuning (ICLR'26)

We run a controlled experiment on BoolQ where both SFT and supervised ZeroTuning use the same 500 labeled examples. We fine-tune the model with a lightweight LoRA setup (rank $r=4$, 3 epochs)

Method	BoolQ Accuracy
Vanilla	69.60
Vanilla + SFT (LoRA)	81.20
Vanilla + ZeroTuning (Ours)	82.40
Vanilla + SFT + ZeroTuning (Ours)	83.60

5 FURTHER ANALYSIS AND KEY FINDINGS

Robustness Across Diverse Conditions. ZeroTuning demonstrates remarkable stability. It maintains strong performance gains even when faced with: (1) long contexts padded with irrelevant distractors, where it stabilizes information flow better than the vanilla model (Appendix I); (2) few-shot scenarios, where it consistently improves instruction-following and reduces invalid outputs (Appendix J); (3) significant prompt variations, including missing or altered instructions (Appendix M); and (4) low-precision 4-bit and 8-bit quantization, where it partially mitigates the associated accuracy degradation (Appendix N).

Practicality and Methodological Choices. Our method is not only robust but also highly practical. We theoretically and empirically confirm that tuning key states provides a viable, kernel-agnostic alternative to direct attention score manipulation, proving effective in optimized environments like FlashAttention (Appendix G). Furthermore, ZeroTuning is adaptable to resource-constrained settings, delivering gains even with a minimal, search-free scaling approach (Appendix H). We also analyze key methodological choices, showing that tuning a moderate subset of heads (40%–70%) is optimal, providing a clear and efficient configuration (Appendix L).

Boundaries of Efficacy. Finally, we analyze the method’s boundaries and potential side effects. We quantitatively demonstrate that ZeroTuning excels at correcting a model’s uncertain errors but cannot override high-confidence mistakes rooted in flawed pretrained knowledge. We also characterize the negative effects of extreme tuning, which provides a clearer picture of the method’s operational limits (Appendix D). This positions our method as a powerful tool for unlocking a model’s latent knowledge, rather than a substitute for fine-tuning. Intriguingly, we also find that within a safe operational range, the scaling factor can modulate output diversity in a manner analogous to temperature, but with the unique ability to alter the rank-ordering of logits and thereby correct errors that temperature scaling cannot fix. This aligns with our experiments on temperature tuning, where pure temperature adjustment fails to improve benchmark performance, while ZeroTuning yields consistent gains substantially.

Representative Research Project: ZeroTuning (ICLR'26)

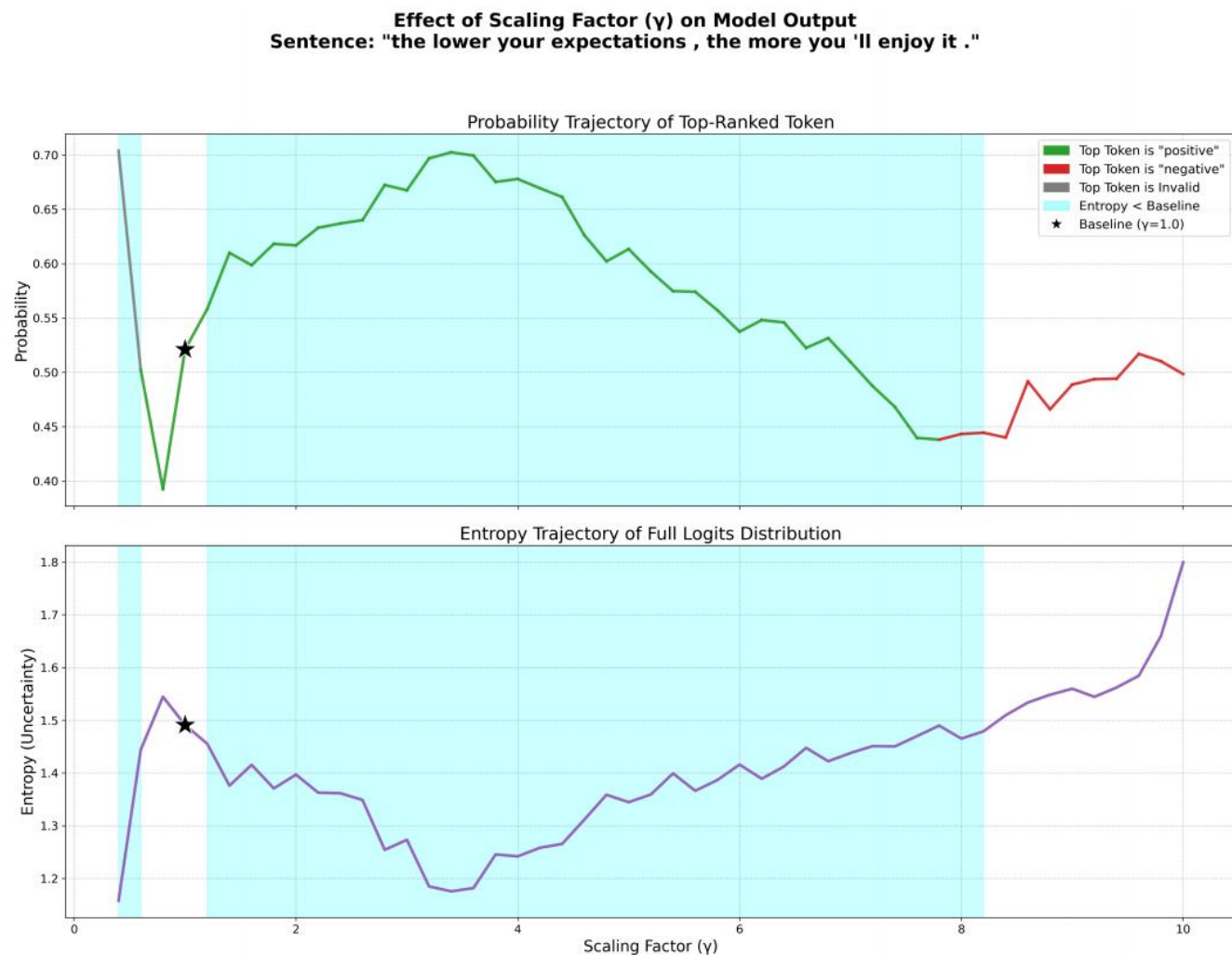


Figure 11: An uncertain sample where the correct answer is "negative". The vanilla model incorrectly outputs "positive". For $\gamma < 1$, the model outputs an invalid format. For $\gamma > 1$, it first amplifies the probability of the initial incorrect answer before flipping to the correct one. The entropy minima are misleadingly located in the regions of the invalid and incorrect answers.

Representative Research Project: ZeroTuning (ICLR'26)

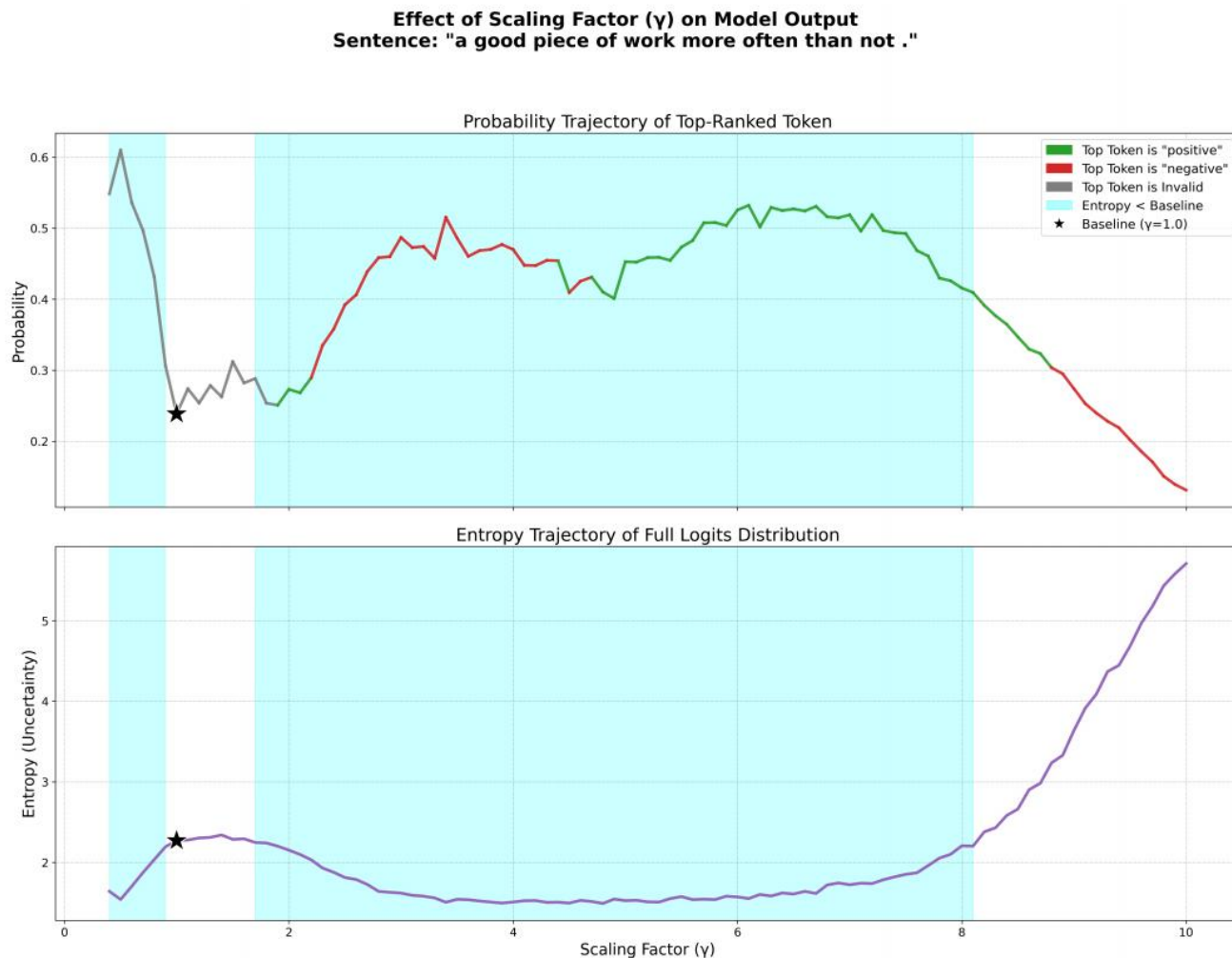


Figure 12: A more uncertain sample where the correct answer is "positive". The vanilla model incorrectly outputs "neutral" (an invalid format). As γ increases, the model first transitions to a valid but incorrect answer ("negative") before finally flipping to the correct answer ("positive"). The entropy landscape exhibits multiple local minima corresponding to the invalid, incorrect, and correct answers, making direct judgment based on the global minimum unreliable.

Representative Research Project: ZeroTuning (ICLR'26)

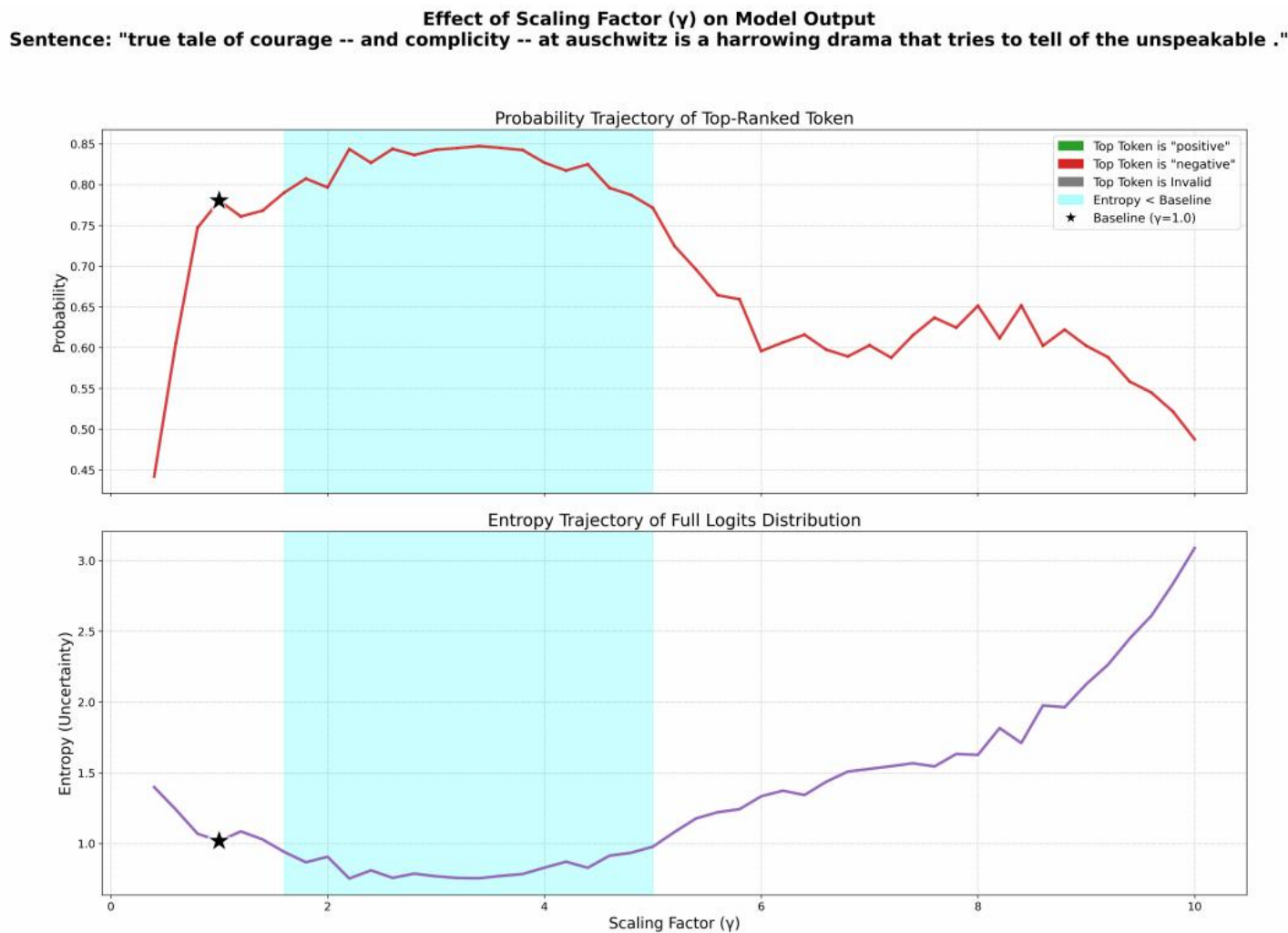


Figure 13: A **recalcitrant** certain sample. The vanilla model (\star) predicts an incorrect but in-format answer ("negative") with high confidence. Modulating γ only reinforces this conviction without an answer flip.

Representative Research Project: ZeroTuning (ICLR'26)

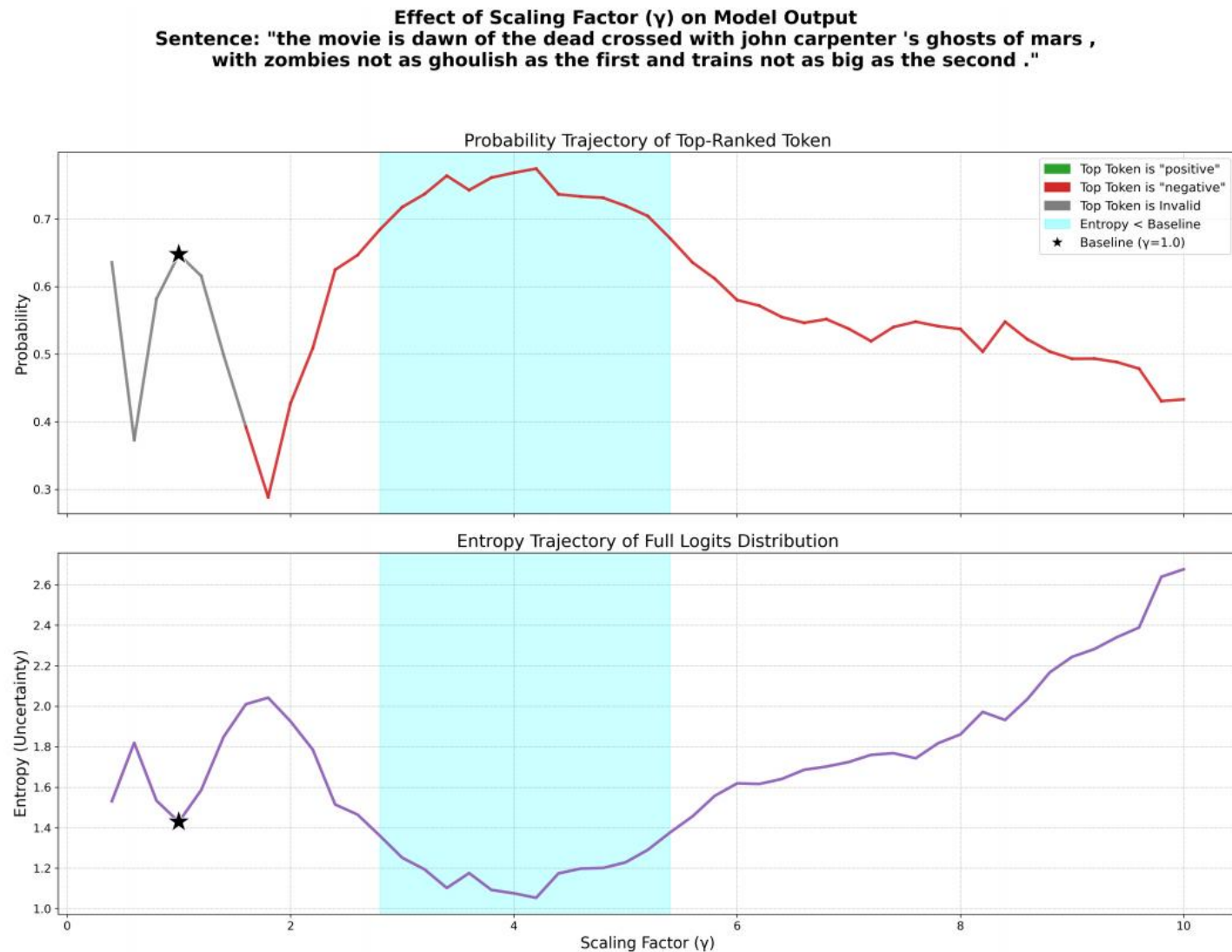
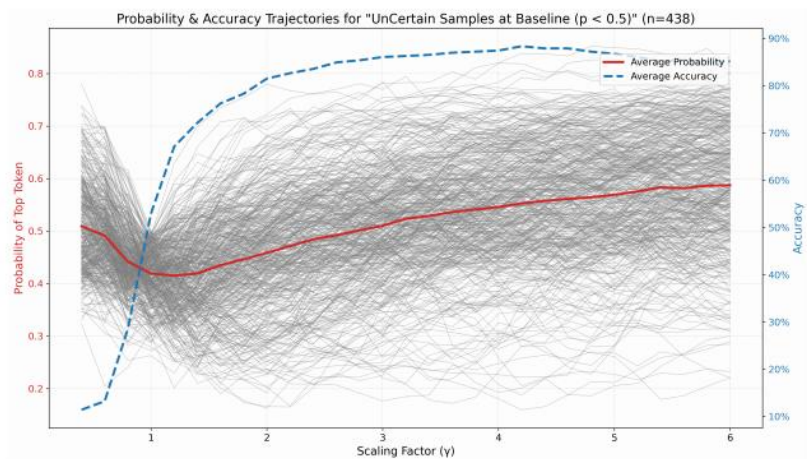
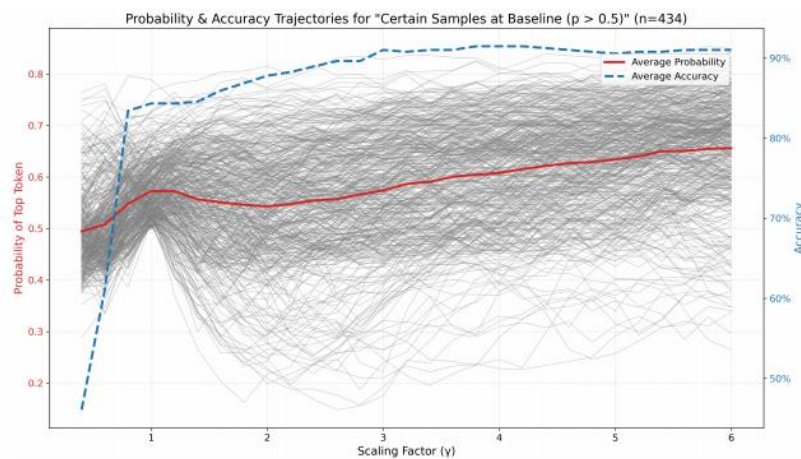


Figure 14: A **correctable** certain sample. The vanilla model (\star) predicts an out-of-format answer ("neutral") with high confidence. Increasing α guides the model to overcome the format error and output the correct answer ("negative").

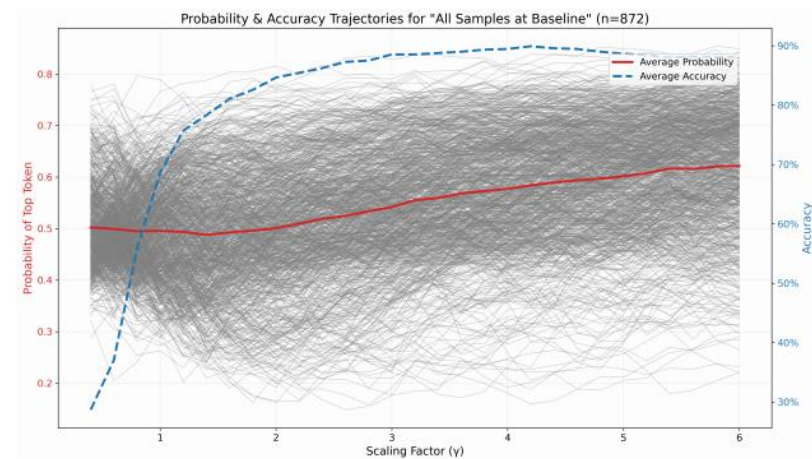
Representative Research Project: ZeroTuning (ICLR'26)



(a) Uncertain Population ($p < 0.5$)



(b) Certain Population ($p > 0.5$)



(c) All Samples (Aggregated)

Figure 15: Population-level probability and accuracy trajectories. Gray lines represent individual sample trajectories. Red: average probability of the top token. Blue: average accuracy.

Representative Research Project: ZeroTuning (ICLR'26)

Key empirical finding: Tuning only BOS attention gives the large stable gains; optimal direction is task-dependent ($\gamma > 1$ or $\gamma < 1$).

Mechanism (bias \leftrightarrow focus): BOS scaling reshapes downstream attention:

- $\gamma > 1$ \rightarrow flatter attention \rightarrow less fixation on misleading local cues (e.g., negative bias)
- $\gamma < 1$ \rightarrow sharper attention \rightarrow better evidence localization in diffuse / long context

Entropy signal: Accuracy is anti-correlated with next-token entropy; entropy is U-shaped vs. γ , and the entropy minimum \sim peak accuracy.

Layer-wise: The γ -performance trend is consistent across depth, with stronger effects in shallow/mid layers; multi-layer tuning is often best.

Head-wise: Heads split into up-effective vs. down-effective; scaling only one head group outperforms uniform scaling.

Practical recipe: Profile heads \rightarrow selectively rescale \rightarrow renormalize; supports supervised tuning and unsupervised calibration via entropy minimization.

Thanks!