Adaptive GoGI-Skip: Coupling Goal-Gradient Importance with Dynamic Uncertainty for Efficient Reasoning

Ren Zhuang
School of Information Science and Technology, Hangzhou Normal University
venomrko97@gmail.com, 2022112011050@stu.hznu.edu.cn

Abstract

Chain-of-Thought (CoT) prompting trades inference speed for reasoning accuracy. Existing compressors force a compromise as static gradient techniques treat tokens independently, severing sequential logic, while uncertainty-based pruning ignores the final answer. We introduce Adaptive GoGI-Skip, a framework that resolves this tension by non-linearly coupling Goal-Gradient Importance (GoGI) with Adaptive Dynamic Skipping (ADS). GoGI quantifies each token's functional contribution to answer correctness via gradient sensitivity. ADS leverages runtime entropy to dynamically modulate the GoGI threshold, preserving low-gradient tokens essential for structural coherence at high-uncertainty junctions. Trained on 7,472 MATH traces, our policy transfers zero-shot to AIME, GPQA, and GSM8K, reducing token volume by >45% and accelerating inference up to 2.0× without accuracy loss. These results suggest that thinking-optimal compression demands synergy between teleological goals and epistemic uncertainty.

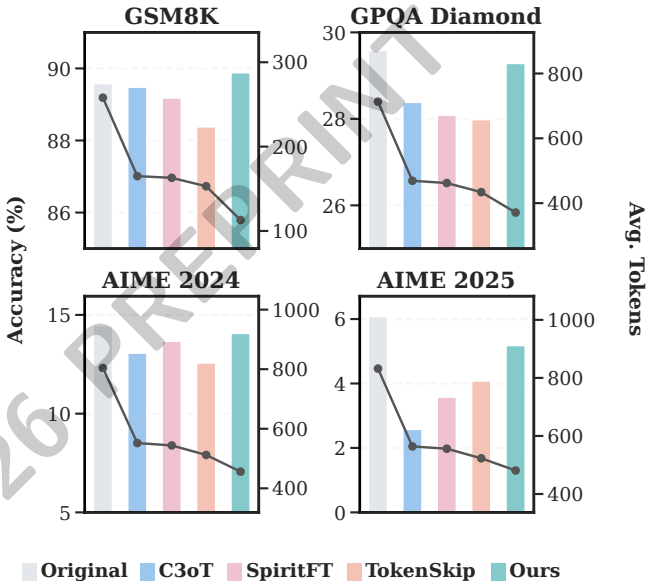

Figure 1: **Accuracy–Token Trade-off on Gemma3-4B-Instruct.** Adaptive GoGI-Skip shifts the trade-off curve, achieving superior accuracy retention compared to static baselines. The method dominates on rigorous benchmarks, preserving reasoning capability despite 45% token reduction.

1 Introduction

Chain-of-Thought (CoT) prompting unlocks advanced reasoning in Large Language Models [Wei *et al.*, 2022] but incurs prohibitive inference latency and memory costs [Qu *et al.*, 2025; Sui *et al.*, 2025]. Standard efficiency paradigms [Shazeer *et al.*, 2017; Katharopoulos *et al.*, 2020; Su *et al.*, 2024; Wang *et al.*, 2024a] overlook CoT's unique structural redundancy.

CoT traces often contain repetition, circular validation, and excessive elaboration [Fatemi *et al.*, 2025; Sui *et al.*, 2025], sometimes degrading accuracy on simpler tasks [Yang *et al.*, 2025]. A thinking-optimal equilibrium, where models reason only as much as strictly necessary, requires a compression paradigm that is both goal-aware and dynamically adaptive.

Existing Supervised Fine-Tuning (SFT) compression approaches [Xia *et al.*, 2025] rely on generic importance metrics like semantic similarity [Pan *et al.*, 2024] or perplexity [Cui *et al.*, 2025] that often mask structurally critical yet semantically simple tokens. Static compression rates compound the problem by ignoring heterogeneous step difficulty [Ge *et al.*, 2024].

A straightforward combination of gradient-based attribution and adaptive computation can be brittle. In long-context generation, gradient signals are often sparse and noisy [Wang *et al.*, 2024b], and pruning solely by gradient magnitude may disrupt the sequential dependencies required for multi-step deduction. Conversely, uncertainty-based criteria capture local ambiguity but do not necessarily reflect a token's contribution to final answer correctness. These observations motivate coupling the two signals: using epistemic uncertainty to dynamically regulate the teleological goal-alignment threshold.

To bridge this gap, we propose **Adaptive GoGI-Skip**, a framework for learning efficient, adaptive CoT compression. As Figure 2 illustrates, our approach integrates two synergistic innovations. Goal-Gradient Importance (GoGI) is a metric ($\mathcal{G}(x_i, l) = ||\nabla_{\mathbf{h}_i^l} \mathcal{L}_{\text{ans}}||_1$) that quantifies each token's func-

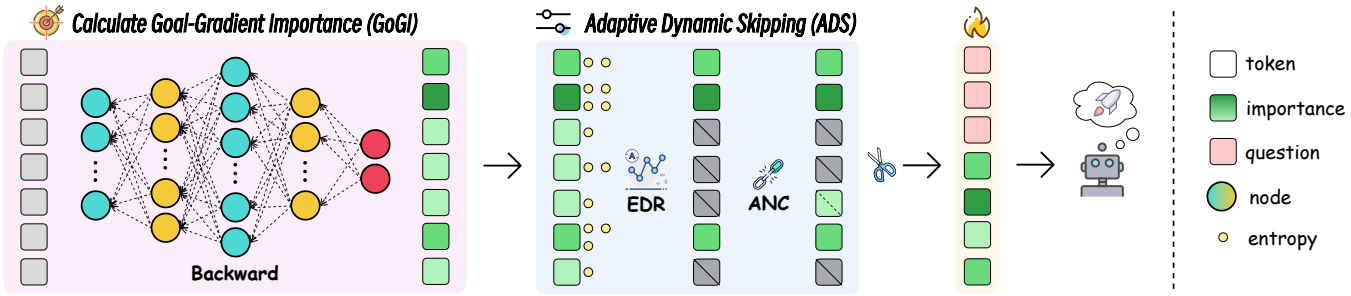

Figure 2: **The Adaptive GoGI-Skip Framework.** (a) **GoGI** measures token utility via $\nabla\mathcal{L}_{\text{ans}}$. (b) **ADS** combines GoGI with runtime uncertainty and local stability to decide K_t. (c) **SFT** distills the policy into model weights for overhead-free speedups.

tional utility via its gradient influence on the final answer loss. Adaptive Dynamic Skipping (ADS) employs Entropy-Driven Rate (EDR) Regulation, using predictive entropy H_t to adjust retention rates, and an Adaptive N-Constraint (ANC) that modulates consecutive deletions based on local complexity, preserving coherence.

Our key contributions are summarized below:

- We introduce Adaptive GoGI-Skip, a goal-aligned and uncertainty-regulated CoT compression framework that couples Goal-Gradient Importance (GoGI) with Adaptive Dynamic Skipping (ADS) to remove functionally redundant tokens while preserving reasoning-critical structure.
- We establish zero-shot universality by training a single compression policy on 7,472 MATH traces and transferring it across diverse benchmarks (AIME, GPQA, GSM8K) and model families (Gemma, Qwen) without task-specific tuning.
- We push the accuracy–efficiency frontier, achieving $>45\%$ token reduction and up to $2.0\times$ speedup while matching or improving original accuracy on rigorous evaluations (Figure 1).

2 Related Work

CoT reasoning's efficacy incurs substantial computational cost [Qu *et al.*, 2025; Sui *et al.*, 2025]. Optimization efforts bifurcate into SFT-based compression and alternative reasoning paradigms.

Token Importance Estimation. Existing compression methods rely on proxy metrics for token utility. External distillers such as C3OT [Kang *et al.*, 2025] leverage GPT-4. Internal methods instead target semantic redundancy, and TokenSkip [Xia *et al.*, 2025] builds on LLMLingua [Pan *et al.*, 2024]. Other approaches exploit activation sparsity, including LazyLLM [Fu *et al.*, 2025] and SlimInfer [Long *et al.*, 2025]. SPIRIT [Cui *et al.*, 2025] uses perplexity-based signals. These heuristics decouple importance from functional relevance and can remove structurally vital yet semantically simple logical connectors [Razzhigaev *et al.*, 2025]. Gradient-based saliency is well-established in interpretability [Simonyan *et al.*, 2013; Sundararajan *et al.*, 2017], but naive application to pruning fails since attribution does not necessarily reflect generation stability. High-gradient tokens are often high-entropy surprises that, if pruned, cause catastrophic drift; discarding low-gradient tokens can sever reasoning's connective tissue.

GoGI departs from generic attribution by non-linearly coupling gradient magnitude with predictive entropy, creating a dynamic threshold that adapts to the chain's structural fragility.

Adaptive Compression Strategies. Current SFT-based approaches employ static compression rates [Xia *et al.*, 2025] or coarse step-level skipping [Wang *et al.*, 2025c]. These rigid policies struggle to accommodate heterogeneous reasoning complexity. Recent work explores adaptive reasoning steps [Wang *et al.*, 2025b] or compute allocation [Manvi *et al.*, 2024], but often requires complex routers.

Adaptive GoGI-Skip addresses this via ADS. By modulating retention rates based on real-time predictive entropy and regulating consecutive deletions via local context, ADS achieves fine-grained balance between sparsity and coherence unattainable by static baselines.

Alternative Efficiency Paradigms. Parallel tracks explore different trade-offs. Latent-space reasoning [Saunshi *et al.*, 2025; Hao *et al.*, 2024] bypasses linguistic decoding entirely but sacrifices interpretability [Ma *et al.*, 2025]. Reinforcement Learning (RL) approaches [Luo *et al.*, 2025; Shen *et al.*, 2025] integrate length penalties into reward functions, yet RL often fails to incentivize genuine reasoning beyond the base model [Yue *et al.*, 2025], induces training instability [Melo *et al.*, 2026], and is prone to reward hacking [Gao *et al.*, 2024] or increased hallucination [Yao *et al.*, 2025].

Our SFT-based approach offers a distinct advantage through stable, direct optimization that learns precise compression policies without RL's fragility. It also serves as a complementary post-processor that trims verbose traces produced by RL-optimized reasoners.

3 Adaptive GoGI-Skip for Reasoning-Intensive CoT Optimization

Adaptive GoGI-Skip is an SFT framework for dynamic pruning toward Reasoning-Intensive CoT Optimization (RICO). The method combines GoGI, which measures each token's contribution to the answer loss, with ADS, which adapts the retention policy from predictive uncertainty and local stability. This coupling yields finer-grained compression than prior

approaches.

Notation. We denote sequences by $(x_t)_{t=1}^m$ and write vectors in bold lowercase. The hidden state of token t at layer l is $\mathbf{h}_t^l$. Key metrics include the GoGI score $\mathcal{G}_t$, predictive entropy H_t, dynamic retention rate γ_t, dynamic GoGI threshold τ_t, and the adaptive N-constraint N_t. The keep decision is $K_t \in \{0, 1\}$. We denote the answer loss by $\mathcal{L}_{\text{ans}}$ and model parameters by θ. The operator $||\cdot||_1$ denotes the L1 norm, ∇ the gradient, and Q_p the p-th percentile operator. We use $[\cdot]_a^b$ for clipping to $[a, b]$, $\lfloor\cdot\rfloor$ for the floor, and $\mathbb{I}$ for the indicator function. We write $\vee$ for logical OR and $\wedge$ for logical AND.

3.1 Goal-Gradient Importance (GoGI)

Existing importance metrics [Xia *et al.*, 2025; Pan *et al.*, 2024; Cui *et al.*, 2025] ignore the reasoning objective. GoGI directly measures each token's contribution to the target objective via gradient-based sensitivity analysis, grounded in the principle that gradients inherently quantify functional importance.

Given a CoT sequence $c = (x_t)_{t=1}^m$ and answer $A = (a_t)_{t=1}^k$, the answer prediction loss $\mathcal{L}_{\text{ans}}$ for model M_θ is:

$$\mathcal{L}_{\text{ans}}(A|c;\theta) = -\sum_{j=1}^{k} \log P_\theta(a_j|c, a_{<j}) \tag{1}$$

The GoGI score for token x_t at target layer l^* is the L1 norm of the gradient:

$$\mathcal{G}_t^{(l^*)} \triangleq \left\| \frac{\partial \mathcal{L}_{\text{ans}}(A|c;\theta)}{\partial \mathbf{h}_t^{l^*}} \right\|_1 \tag{2}$$

This score measures the local sensitivity of the final answer loss to the intermediate representation $\mathbf{h}_t^{l^*}$.

We select the target layer l^* by analyzing layer-wise gradient contributions rather than relying on a single isolated maximum. Peak sensitivity occurs in mid-to-late layers, and downstream performance remains empirically stable within this high-contribution region. On Gemma3-4B-Instruct, Layers 23 and 28 exhibit comparable peaks; we select Layer 23 because it balances local sensitivity with globally integrated representations for GoGI computation.

3.2 Adaptive Dynamic Skipping (ADS)

ADS dynamically adjusts the pruning strategy based on runtime context via EDR and ANC components.

Entropy-Driven Rate (EDR) Regulation

EDR regulation leverages predictive entropy H_t (Eq. 3) as an information density signal. Since only a fraction of CoT tokens exhibit high entropy—signaling critical branching points [Wang *et al.*, 2025a]—high H_t warrants conservative pruning. Crucially, H_t and $\mathcal{G}_t$ provide complementary signals.

$$H_t = -\sum_{v \in \mathcal{V}} p_{t,v} \log p_{t,v}, \tag{3}$$

where $p_{t,v} = P_\theta(X_{t+1} = v|x_1, \ldots, x_t)$. We map H_t to $\hat{H}_t \in [0, 1]$ via $\mathcal{M}(\cdot;\theta_{\mathcal{M}})$:

$$\hat{H}_t = \mathcal{M}(H_t;\theta_{\mathcal{M}}) \tag{4}$$

Algorithm 1 Adaptive GoGI-Skip Pruning

Input: CoT sequence $c = (x_t)_{t=1}^m$, answer A, model M_θ
Output: Compressed sequence c'_{comp}, keep mask $(K_t)_{t=1}^m$

Compute $\mathcal{L}_{\text{ans}}(A|c;\theta)$ via forward pass
$\mathcal{G}_t \leftarrow \|\nabla_{\mathbf{h}_t^{l^*}} \mathcal{L}_{\text{ans}}\|_1$ for all t {GoGI scores}
$H_t \leftarrow -\sum_v p_{t,v} \log p_{t,v}$ for all t {Entropy}
Determine $\mathcal{I}_{\text{valid}}$ by excluding non-substantive tokens
$C_0 \leftarrow 0$ {Consecutive prune counter}
for $t = 1$ to m **do**
  if $t \notin \mathcal{I}_{\text{valid}}$ **then**
    $K_t \leftarrow 1$
    $C_t \leftarrow 0$
    continue
  end if
  $\hat{H}_t \leftarrow \mathcal{M}(H_t;\theta_{\mathcal{M}})$
  $\gamma_t \leftarrow [\gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \cdot \hat{H}_t]_{\gamma_{\min}^{\text{abs}}}^{\gamma_{\max}^{\text{abs}}}$ {EDR}
  $\tau_t \leftarrow Q_{(1-\gamma_t)\times 100}(\{\mathcal{G}_j \mid j \in \mathcal{I}_{\text{valid}}\})$
  $\bar{H}_{t,W} \leftarrow \frac{1}{W} \sum_{j=\max(1,t-\lfloor W/2 \rfloor)}^{\min(m,t+\lfloor W/2 \rfloor)} H_j$
  $\hat{\bar{H}}_{t,W} \leftarrow \mathcal{M}'(\bar{H}_{t,W};\theta_{\mathcal{M}'})$
  $\tilde{N}_t \leftarrow N_{\min} + (N_{\max} - N_{\min})(1 - \hat{\bar{H}}_{t,W})$
  $N_t \leftarrow \lfloor [\tilde{N}_t]_{N_{\min}}^{N_{\max}} + 0.5 \rfloor$ {ANC}
  $\mathcal{G}'_t \leftarrow f_{\text{weight}}(\mathcal{G}_t, x_t)$ {Optional type-weighting}
  if $\mathcal{G}'_t \geq \tau_t$ **or** $C_{t-1} + 1 \geq N_t$ **then**
    $K_t \leftarrow 1$
    $C_t \leftarrow 0$
  else
    $K_t \leftarrow 0$
    $C_t \leftarrow C_{t-1} + 1$
  end if
end for
$c'_{\text{comp}} \leftarrow (x_t)_{K_t=1}$
return $c'_{\text{comp}}, (K_t)_{t=1}^m$

The local retention rate γ_t is then:

$$\gamma_t = [\gamma_{\min} + (\gamma_{\max} - \gamma_{\min}) \cdot \hat{H}_t]_{\gamma_{\min}^{\text{abs}}}^{\gamma_{\max}^{\text{abs}}} \tag{5}$$

where $\gamma_{\min}, \gamma_{\max}$ are soft bounds, and $[\cdot]_{\min}^{\max}$ denotes clipping. γ_t determines the dynamic GoGI threshold τ_t:

$$\tau_t = Q_{(1-\gamma_t)\times 100}(\{\mathcal{G}_j | j \in \mathcal{I}_{\text{valid}}\}) \tag{6}$$

where $\mathcal{I}_{\text{valid}}$ excludes non-substantive tokens.

Adaptive N-Constraint (ANC) for Coherence

To preserve coherence, ANC limits the number of consecutive pruned tokens (N_t) based on local complexity, estimated via windowed entropy $\bar{H}_{t,W}$:

$$\bar{H}_{t,W} = \frac{1}{W} \sum_{j=\max(1,t-\lfloor W/2 \rfloor)}^{\min(m,t+\lfloor W/2 \rfloor)} H_j \tag{7}$$

$$\hat{\bar{H}}_{t,W} = \mathcal{M}'(\bar{H}_{t,W};\theta_{\mathcal{M}'}) \tag{8}$$

The adaptive constraint N_t is:

$$\tilde{N}_t = N_{\min} + (N_{\max} - N_{\min}) \cdot (1 - \hat{\bar{H}}_{t,W}) \tag{9}$$

Method	AIME 2025 [MAA, 2025]				AIME 2024 [MAA, 2024]			
	Acc.↑	Ratio ↓	Tokens↓	Speedup↑	Acc.↑	Ratio ↓	Tokens↓	Speedup↑
Original	$6.1_{(0.0\downarrow)}$	1.0	832	1.0×	$14.5_{(0.0\downarrow)}$	1.0	804	1.0×
Prompting	$1.8_{(4.3\downarrow)}$	0.9	764	1.0×	$10.2_{(4.3\downarrow)}$	0.9	728	1.1×
C3ot	$2.6_{(3.5\downarrow)}$	0.7	564	1.4×	$13.1_{(1.4\downarrow)}$	0.7	552	1.4×
Spiritft	$3.6_{(2.5\downarrow)}$	0.6	556	1.5×	$13.7_{(0.8\downarrow)}$	0.7	544	1.4×
TokenSkip ($\gamma = 0.6$)	$4.1_{(2.0\downarrow)}$	0.6	523	1.6×	$12.6_{(1.9\downarrow)}$	0.6	512	1.5×
TokenSkip ($\gamma = 0.5$)	$2.7_{(3.4\downarrow)}$	0.6	457	1.7×	$11.3_{(3.2\downarrow)}$	0.5	448	1.6×
Adaptive GoGI-Skip (Ours)	**$5.2_{(0.9\downarrow)}$**	**0.6**	**481**	**1.6×**	**$14.1_{(0.4\downarrow)}$**	**0.6**	**456**	**1.7×**

Method	GPQA Diamond [Rein *et al.*, 2024]				GSM8K [Cobbe *et al.*, 2021]			
	Acc.↑	Ratio↓	Tokens↓	Speedup↑	Acc.↑	Ratio↓	Tokens↓	Speedup↑
Original	$29.6_{(0.0\downarrow)}$	1.0	713	1.0×	$89.6_{(0.0\downarrow)}$	1.0	258	1.0×
Prompting	$21.7_{(7.9\downarrow)}$	0.9	651	1.0×	$80.2_{(9.4\downarrow)}$	0.9	225	1.1×
C3ot	$28.4_{(1.2\downarrow)}$	0.7	469	1.5×	$89.5_{(0.1\downarrow)}$	0.6	165	1.5×
Spiritft	$28.1_{(1.5\downarrow)}$	0.6	462	1.5×	$89.2_{(0.4\downarrow)}$	0.6	163	1.5×
TokenSkip ($\gamma = 0.6$)	$28.0_{(1.6\downarrow)}$	0.6	434	1.6×	$88.4_{(1.2\downarrow)}$	0.6	153	1.6×
TokenSkip ($\gamma = 0.5$)	$26.5_{(3.1\downarrow)}$	0.5	378	1.8×	$84.3_{(5.0\downarrow)}$	0.5	133	1.9×
Adaptive GoGI-Skip (Ours)	**$29.3_{(0.3\downarrow)}$**	**0.5**	**371**	**1.8×**	**$89.9_{(0.3\uparrow)}$**	**0.4**	**113**	**1.9×**

Table 1: **Main Efficiency Results on Gemma3-4B-Instruct.** Adaptive GoGI-Skip achieves consistent speedups (1.6–1.9×) with negligible accuracy loss (<0.9%), outperforming static baselines that degrade significantly on reasoning-intensive tasks.

$$N_t = \left\lfloor \left[\tilde{N}_t\right]_{N_{\min}}^{N_{\max}} + 0.5 \right\rfloor \tag{10}$$

High local entropy thus yields a smaller N_t, enforcing conservative pruning.

3.3 Adaptive GoGI-Skip Pruning Algorithm

Algorithm 1 integrates GoGI with EDR and ANC into a unified pruning procedure. Tokens are retained if intrinsically important ($\mathcal{G}'_t \geq \tau_t$) or if pruning would violate the ANC constraint ($C_{t-1} + 1 \geq N_t$). This dual-safety mechanism prevents fragmented chains common in pure importance sampling. The resulting compressed sequences c'_{comp} form the SFT training set $\mathcal{D}_{SFT}$.

4 Experiments

We evaluate Adaptive GoGI-Skip to demonstrate its efficiency and accuracy across benchmarks and model architectures. We further analyze the contribution of core components through ablation studies.

4.1 Experimental Setup

Models Our experiments employ two open-source instruction-tuned model families: Gemma3-Instruct [Team *et al.*, 2025] and Qwen2.5-Instruct [Yang *et al.*, 2024]. This selection enables a thorough evaluation of scalability and generalizability.

Training Data We sourced samples from the MATH training dataset [Hendrycks *et al.*, 2021] only, isolating transfer from the effects of broad multi-domain supervision. After standardizing formats and validating parsing, 7,472 valid samples remained. We compress each original CoT trace with Adaptive GoGI-Skip as an SFT set. We then report zero-shot results on AIME, GPQA, and GSM8K across architectures without additional task-specific training.

Baselines We compare against CoT compression baselines compatible with the SFT pipeline, including the standard CoT, denoted Original; zero-shot prompting, referred to as Prompting; C3oT, an external model–based compression method [Kang *et al.*, 2025]; Spiritft, which performs perplexity-guided token pruning [Cui *et al.*, 2025]; and TokenSkip, which applies static token skipping at ratios of 0.5 and 0.6 [Xia *et al.*, 2025].

Implementation Details We computed GoGI scores at model-specific optimal layers. ADS hyperparameters were fixed across experiments. Token type-weighting was disabled. All models were fine-tuned using Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] on 3 NVIDIA 4090 GPUs.

4.2 Main Results

We evaluated Adaptive GoGI-Skip's efficiency and accuracy balance. All metrics are averaged over 3 independent fine-tuning runs.

Table 1 shows that Adaptive GoGI-Skip consistently outperforms baselines, achieving ∼1.7× speedups on AIME and 1.9× on GSM8K. This efficiency stems from dynamic retention: the model retains ∼60% of tokens for complex AIME tasks versus ∼45% for GSM8K. Accuracy remains robust, with negligible drops around 0.4% on GPQA/AIME'24 and a 0.3% gain on GSM8K. Even on AIME'25, the 0.9% drop yields a competitive score. In contrast, static baselines like TokenSkip face a rigid trade-off that high speedup costs significant accuracy, while conservative settings limit speedup. Adaptive GoGI-Skip breaks this compromise, matching original accuracy with aggressive efficiency.

Method	Gemma3-1B-It								Gemma3-12B-It							
	AIME'25		AIME'24		GPQA		GSM8K		AIME'25		AIME'24		GPQA		GSM8K	
	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.
Original	0.7	1.0×	3.1	1.0×	18.7	1.0×	70.2	1.0×	8.1	1.0×	22.7	1.0×	38.4	1.0×	95.2	1.0×
Prompting	0.1	1.1×	1.2	1.1×	12.5	1.1×	65.7	1.1×	4.2	1.1×	15.3	1.1×	28.7	1.1×	85.6	1.1×
C3ot	0.5	1.4×	2.3	1.4×	16.1	1.4×	68.7	1.5×	7.4	1.4×	21.6	1.5×	37.4	1.5×	94.1	1.5×
Spiritft	0.4	1.4×	2.1	1.4×	17.3	1.5×	69.6	1.5×	7.2	1.5×	21.2	1.5×	36.6	1.5×	93.8	1.6×
TokenSkip (γ=0.6)	0.5	1.5×	2.2	1.5×	16.5	1.5×	69.9	1.6×	7.5	1.5×	21.9	1.6×	37.1	1.6×	94.3	1.6×
TokenSkip (γ=0.5)	0.3	1.6×	1.9	1.6×	15.8	1.6×	64.6	1.7×	6.8	1.6×	21.2	1.8×	36.1	1.7×	92.7	1.8×
Adaptive GoGI-Skip	**0.6**	**1.5×**	**2.9**	**1.5×**	**18.4**	**1.6×**	**70.2**	**1.8×**	**8.0**	**1.6×**	**22.9**	**1.7×**	**38.4**	**1.7×**	**95.5**	**2.0×**

Method	Qwen2.5-3B-It								Qwen2.5-7B-It							
	AIME'25		AIME'24		GPQA		GSM8K		AIME'25		AIME'24		GPQA		GSM8K	
	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.	Acc.	SU.
Original	2.1	1.0×	7.5	1.0×	25.3	1.0×	78.5	1.0×	6.2	1.0×	18.1	1.0×	35.7	1.0×	84.5	1.0×
Prompting	0.8	1.1×	4.1	1.1×	17.2	1.1×	65.3	1.1×	2.8	1.1×	11.2	1.1×	25.8	1.1×	77.1	1.1×
C3ot	1.7	1.4×	6.8	1.4×	23.5	1.5×	76.1	1.5×	5.6	1.5×	17.0	1.5×	34.3	1.5×	82.7	1.5×
Spiritft	1.6	1.5×	6.5	1.5×	23.0	1.5×	75.2	1.5×	5.3	1.5×	16.6	1.5×	33.6	1.5×	88.2	1.6×
TokenSkip (γ=0.6)	1.8	1.5×	6.9	1.6×	23.4	1.6×	76.5	1.6×	5.7	1.6×	17.2	1.6×	34.1	1.6×	83.1	1.6×
TokenSkip (γ=0.5)	1.4	1.6×	5.8	1.6×	22.5	1.7×	73.8	1.9×	5.4	1.7×	15.8	1.7×	32.9	1.7×	81.5	1.9×
Adaptive GoGI-Skip	**2.0**	**1.5×**	**7.3**	**1.5×**	**24.6**	**1.7×**	**78.9**	**1.8×**	**6.1**	**1.5×**	**18.1**	**1.5×**	**35.7**	**1.6×**	**85.1**	**1.9×**

Table 2: **Zero-shot generalization across models and scales.** Adaptive GoGI-Skip generalizes across Gemma and Qwen models from 1B to 12B. Efficiency gains increase with model size, consistent with stronger separation between reasoning signal and noise at larger scales.

Method Variant	AIME 2025		AIME 2024		GPQA Diamond		GSM8K	
	Acc.	Speedup	Acc.	Speedup	Acc.	Speedup	Acc.	Speedup
Full Method (GoGI+EDR+ANC)	5.2	1.6×	14.1	1.7×	29.3	1.8×	89.9	1.9×
w/o ANC (GoGI+EDR only)	4.5$_{(0.7\downarrow)}$	1.7×	13.6$_{(0.5\downarrow)}$	1.8×	28.5$_{(0.8\downarrow)}$	1.9×	89.0$_{(0.9\downarrow)}$	2.0×
w/o EDR (GoGI-Static, $\gamma \approx 0.5$)	4.3$_{(0.9\downarrow)}$	1.6×	13.5$_{(0.6\downarrow)}$	1.5×	28.6$_{(0.7\downarrow)}$	1.6×	89.5$_{(0.4\downarrow)}$	1.6×
w/o ADS (GoGI-Static, $\gamma \approx 0.5$)	4.0$_{(1.2\downarrow)}$	1.9×	13.0$_{(1.1\downarrow)}$	1.9×	28.0$_{(1.3\downarrow)}$	1.9×	88.5$_{(1.4\downarrow)}$	1.9×
w/o GoGI (LLMLingua+ADS)	4.6$_{(0.6\downarrow)}$	1.6×	13.8$_{(0.3\downarrow)}$	1.7×	28.8$_{(0.5\downarrow)}$	1.8×	89.1$_{(0.8\downarrow)}$	1.9×

Table 3: **Component Ablation Analysis on Gemma3-4B.** Identifying the contribution of each module. Removing dynamic regulation (w/o ADS/ANC) causes the maximal performance drop, confirming that static compression cannot accommodate the variable complexity of reasoning chains.

Table 2 confirms strong generalization across model scaling and architectures. Efficiency gains scale positively with model size from 1B to 12B. For Gemma3-12B, we achieve a 2.0× speedup on GSM8K while improving accuracy by 0.3% relative to Original as the pruned tokens function mainly as reasoning noise. The method works equally well on Qwen2.5, maintaining exact parity on AIME'24 and GPQA while accelerating GSM8K by 1.9×. This consistency indicates that the learned compression policy captures universal features of logical redundancy.

Table 3 quantifies each module's contribution. Removing ADS causes the sharpest accuracy decline despite achieving highest speedup, confirming that static pruning cannot accommodate variable information density. Without ANC, performance drops on harder tasks by 0.7% on AIME'25, where longer-range dependencies require periodic anchor tokens to maintain logical coherence. The w/o EDR variant underperforms despite matching compression rates since entropy-driven modulation prevents cascading failures at high-entropy branching points. Replacing GoGI with LLMLingua degrades accuracy at comparable speedups, indicating that perplexity-based salience does not reliably recover the steps that drive answer correctness. GoGI selects tokens by their sensitivity to the answer loss, and ADS turns this objective-aligned signal into stable, uncertainty-aware pruning. Together, they yield the best accuracy retention at matched speedups in our ablations.

4.3 Analysis

Robustness on Efficiency–Accuracy Frontier. Absolute low scores on AIME reflect base model limitations. Even so, Adaptive GoGI-Skip delivers the strongest accuracy retention at comparable speedups across baselines. While baselines such as C3oT and TokenSkip achieve speedups by sacrificing significant accuracy, Adaptive GoGI-Skip preserves performance even at high compression rates. Results on rigorous benchmarks like AIME are particularly telling that they represent robust evidence that our method identifies functionally critical tokens with greater precision than prior SFT approaches. Superior accuracy preservation under these adversarial conditions demonstrates structural alignment with the model's reasoning topology that static heuristic pruning fails

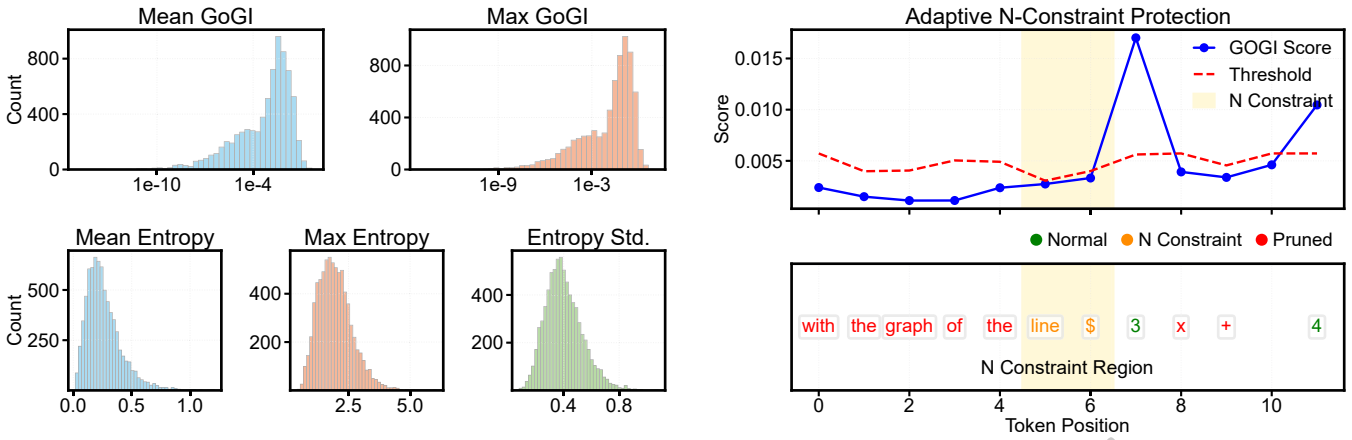

Figure 3: **Distributional Dynamics of Information. Left**: GoGI scores (top) follow a long-tail distribution, while entropy (bottom) identifies mode switches. **Right**: ANC dynamically relaxes pruning strictness during entropy spikes, preserving structural coherence at critical junctions.

to capture.

Effectiveness of Adaptive GoGI-Skip. Figure 3 (left) summarizes the signals that make compression reliable. GoGI scores follow a long-tailed distribution, which indicates that most tokens contribute little to the answer loss. The entropy distribution separates stable segments from uncertain transitions, where pruning errors are more likely to cascade. This separation explains why Adaptive GoGI-Skip prunes aggressively in stable regions while remaining conservative around high-uncertainty steps.

Dynamic Behavior of ADS. Figure 3 (right) visualizes the dynamic adaptability of our strategy. EDR dynamically expands the retention window (γ_t) during high-entropy segments, allocating compute to ambiguous transitions. Simultaneously, ANC acts as a structural stabilizer. ANC intervenes at critical transitions, such as shifting from calculation to conclusion, preventing the model from skipping logically vital but semantically stable connectors like "therefore" and implies. This dual-regulation inversely correlates pruning intensity with local reasoning difficulty, maintaining thinking-optimal density.

Joint Compression Logic. Figure 4 illustrates the visualization which elucidates the joint regulation of compression. We observe a hierarchical interaction that in high-entropy regimes, retention rates remain high regardless of GoGI scores, as EDR enforces safety during uncertainty. Conversely, in low-entropy confident zones, the surface dips significantly, allowing GoGI to aggressively prune tokens with low utility. This pattern confirms that our mechanism effectively prioritizes structural stability over local efficiency when necessary, preventing the blind skipping common in static methods.

Complexity Analysis. Adaptive GoGI-Skip reduces the theoretical attention complexity from $O(L^2)$ to $O((1-\bar{\gamma})^2 L^2)$, yielding an approximate $4\times$ reduction in FLOPs for $\bar{\gamma} = 0.5$. In practice, since LLM inference is memory-bandwidth bound, the observed $2.0\times$ speedup stems primarily from reduced memory access frequency for weights and KV-caches. Crucially, because all decision parameters are baked into the SFT weights, this acceleration incurs zero runtime overhead: standard decoding simply outputs a shorter, denser sequence.

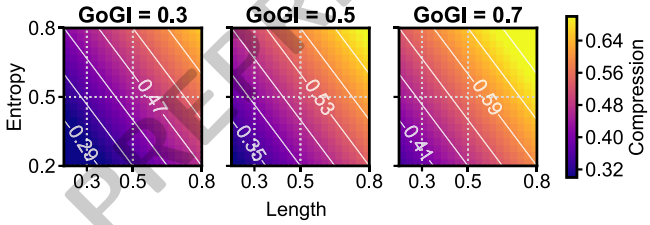

Figure 4: **Decision Surface of Adaptive Compression.** The joint policy $\gamma(H_t, \mathcal{G}_t)$ illustrates the safety mechanism: high uncertainty forces token retention regardless of GoGI score, while low uncertainty allows aggressive pruning of low-utility tokens.

5 Discussion

Our experiments validate that Adaptive GoGI-Skip addresses the efficiency–accuracy trade-off in CoT reasoning. Here we synthesize the implications.

5.1 Implications and Insights

Beyond uncertainty-based proxies. The near-zero correlation between GoGI scores and predictive entropy (Figure 5, left) challenges reliance on uncertainty as an importance proxy. Statistical surprise (H_t) signals branching points or ambiguity but cannot distinguish necessary functional steps from linguistic idiosyncrasies. As Jain and Wallace [Jain and Wallace, 2019] note, surface-level attention patterns often diverge from true model reasoning; uncertainty similarly fails to predict valid pruning candidates. By anchoring importance to the answer-loss gradient, GoGI enables aggressive pruning of semantically distinct but functionally redundant tokens, a capability missed by perplexity-based methods that conflate surprising with important.

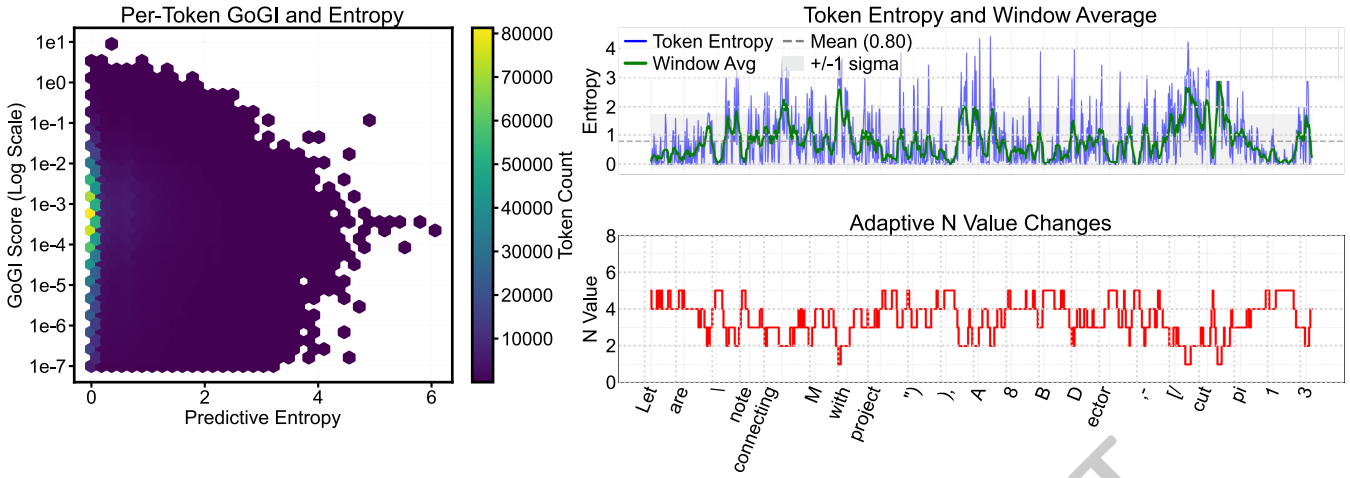

Figure 5: **Orthogonality and Regulation. Left:** The lack of correlation between GoGI and Entropy implies they capture distinct feature dimensions (Teleological vs. Epistemic). **Right:** The ANC windowing acts as a low-pass filter on skipping decisions, smoothing the compression rate to prevent fragmented chains.

The case for adaptivity. Rigid trade-offs in static baselines underscore the profound heterogeneity of reasoning steps: some require deep computational verification; others are routine syntactic bridges. ADS decouples token count from task difficulty, echoing Adaptive Computation Time principles [Graves, 2016]. As Figure 5 (right) shows, ANC preserves local structural integrity in high-entropy regions even when GoGI signals are low. This synergy enables heavy compression of routine logic while retaining full fidelity for crux moments, approximating thinking-optimal density essential for sustainable Green AI [Schwartz *et al.*, 2020].

Compatibility with RL-based reasoners. RL-based reasoning models such as DeepSeek-R1 [Guo *et al.*, 2025] achieve strong performance but produce increasingly verbose traces [Yeo *et al.*, 2025]. Adaptive GoGI-Skip offers a natural post-processing choice: it compresses RL-generated chains without modifying the base model or reward function. This decoupled design preserves RL's performance benefits while recovering inference efficiency, making our method applicable to any verbose reasoner regardless of its training paradigm.

5.2 Broader Impact

Our work advances practical AI deployment along two axes. Detailed CoT generation multiplies the energy cost per query, and reducing token count by $\sim$45% lowers kilowatt-hour consumption proportionally, supporting Green AI initiatives [Luccioni *et al.*, 2025; Patterson *et al.*, 2025]. Speedups reduce end-to-end latency, increase system throughput, and improve resource utilization. This supports higher concurrency and makes it easier to deploy more capable models under fixed infrastructure constraints.

5.3 Limitations

Since GoGI computation requires direct access to model parameters and gradients during data preparation, the method applies to all open models with full parameter access, but it does not extend to models that are only available through closed-source APIs. While inference accelerates, the offline GoGI score calculation incurs a one-time preprocessing cost that scales with the training corpus size, yet this cost amortizes over repeated deployments of the accelerated model.

6 Conclusion

Adaptive GoGI-Skip preserves CoT reasoning while enabling efficient deployment. By unifying goal-gradient importance with dynamic uncertainty regulation, it achieves a principled compression that minimizes redundancy without sacrificing accuracy. Trained on only 7,472 MATH traces, the policy transfers zero-shot across benchmarks and architectures, delivering >45% token reduction and up to 2.0× speedup with negligible accuracy loss. These results show that structural redundancy in reasoning is a learnable, task-agnostic property distinct from semantics. As RL-driven scaling yields ever-longer traces, GoGI-Skip isolates core deductive signals from superfluous content and embeds compact reasoning into model weights via standard SFT, setting a new baseline for resource-efficient inference.

Acknowledgments

I am grateful to Professor Ben Wang and Professor Shuifa Sun from Hangzhou Normal University for their guidance, encouragement, and continued support throughout the course of this work.

References

[Cobbe *et al.*, 2021] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[Cui *et al.*, 2025] Yingqian Cui, Pengfei He, Jingying Zeng, Hui Liu, Xianfeng Tang, Zhenwei Dai, Yan Han, Chen Luo, Jing Huang, Zhen Li, et al. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2502.13260*, 2025.

[Fatemi *et al.*, 2025] Mehdi Fatemi, Banafsheh Rafiee, Mingjie Tang, and Kartik Talamadupula. Concise reasoning via reinforcement learning. *arXiv preprint arXiv:2504.05185*, 2025.

[Fu *et al.*, 2025] Qichen Fu, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, and Mahyar Najibi. Lazyllm: Dynamic token pruning for efficient long context llm inference. In *ICLR*, 2025.

[Gao *et al.*, 2024] Jiaxuan Gao, Shusheng Xu, Wenjie Ye, Weilin Liu, Chuyi He, Wei Fu, Zhiyu Mei, Guangju Wang, and Yi Wu. On designing effective rl reward at training time for llm reasoning. *arXiv preprint arXiv:2410.15115*, 2024.

[Ge *et al.*, 2024] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

[Graves, 2016] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Hao *et al.*, 2024] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[Hu *et al.*, 2022] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Jain and Wallace, 2019] Sarthak Jain and Byron C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[Kang *et al.*, 2025] Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24312–24320, 2025.

[Katharopoulos *et al.*, 2020] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[Long *et al.*, 2025] Lingkun Long, Rubing Yang, Yushi Huang, Desheng Hui, Ao Zhou, and Jianlei Yang. Sliminfer: Accelerating long-context llm inference via dynamic token pruning. *arXiv preprint arXiv:2508.06447*, 2025.

[Luccioni *et al.*, 2025] Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference. *arXiv preprint arXiv:2505.00001*, 2025.

[Luo *et al.*, 2025] Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025.

[Ma *et al.*, 2025] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*, 2025.

[MAA, 2024] MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME*, 2024.

[MAA, 2025] MAA. American invitational mathematics examination - aime. In *American Invitational Mathematics Examination - AIME*, 2025.

[Manvi *et al.*, 2024] Rohin Manvi, Anikait Singh, and Stefano Ermon. Adaptive inference-time compute: Llms can predict if they can do better, even mid-generation. *arXiv preprint arXiv:2410.02725*, 2024.

[Melo *et al.*, 2026] Luckeciano C. Melo, Alessandro Abate, and Yarin Gal. Stabilizing policy gradients for sample-efficient reinforcement learning in llm reasoning. In *International Conference on Learning Representations*, 2026.

[Pan *et al.*, 2024] Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *arXiv preprint arXiv:2403.12968*, 2024.

[Patterson *et al.*, 2025] David Patterson, Joseph Gonzalez, Quoc Le, and Chen Liang. The energy impact of llms — where are we in 2025? *Journal of Green AI*, 3(2), 2025.

[Qu *et al.*, 2025] Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, et al. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*, 2025.

[Razzhigaev *et al.*, 2025] Anton Razzhigaev, Matvey Mikhalchuk, Temurbek Rahmatullaev, Elizaveta Goncharova, Polina Druzhinina, Ivan Oseledets, and Andrey Kuznetsov. Llm-microscope: Uncovering the hidden role of punctuation in context memory of transformers. *arXiv preprint arXiv:2502.15007*, 2025.

[Rein *et al.*, 2024] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[Saunshi *et al.*, 2025] Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. *arXiv preprint arXiv:2502.17416*, 2025.

[Schwartz *et al.*, 2020] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12), 2020.

[Shazeer *et al.*, 2017] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[Shen *et al.*, 2025] Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, and Shiguo Lian. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*, 2025.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[Su *et al.*, 2024] DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuandong Tian, and Qinqing Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. In *The Thirteenth International Conference on Learning Representations*, 2024.

[Sui *et al.*, 2025] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Hanjie Chen, Xia Hu, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Team *et al.*, 2025] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[Wang *et al.*, 2024a] Xinglin Wang, Shaoxiong Feng, Yiwei Li, Peiwen Yuan, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Make every penny count: Difficulty-adaptive self-consistency for cost-efficient reasoning. *arXiv preprint arXiv:2408.13457*, 2024.

[Wang *et al.*, 2024b] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*, 2024.

[Wang *et al.*, 2025a] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

[Wang *et al.*, 2025b] Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, Yujun Zhou, and Xiangliang Zhang. Adareasoner: Adaptive reasoning enables more flexible thinking. *arXiv preprint arXiv:2505.17312*, 2025.

[Wang *et al.*, 2025c] Yibo Wang, Haotian Luo, Huanjin Yao, Tiansheng Huang, Haiying He, Rui Liu, Naiqiang Tan, Jiaxing Huang, Xiaochun Cao, Dacheng Tao, and Li Shen. R1-compress: Long chain-of-thought compression via chunk compression and search. *arXiv preprint arXiv:2505.16838*, 2025.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Xia *et al.*, 2025] Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

[Yang *et al.*, 2024] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[Yang *et al.*, 2025] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.

[Yao *et al.*, 2025] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

[Yeo *et al.*, 2025] Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *arXiv preprint arXiv:2502.03373*, 2025.

[Yue *et al.*, 2025] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *NeurIPS*, 2025.