



Paper 2008



Code

Main Conference



Beyond Truncation: Rethinking LLM Decoding as Ensemble Pruning

Dunyao Xue¹, Chengshuo Du¹, Zhengbo Wang¹, Wenlin Dai^{12†}, Cheng Meng^{13†}¹ Institute of Statistics and Big Data, Renmin University of China (RUC) | ² Big Data and Responsible AI for National Governance, RUC | ³ Center for Applied Statistics, RUC

Motivation

Why does LLM decoding need diversity?

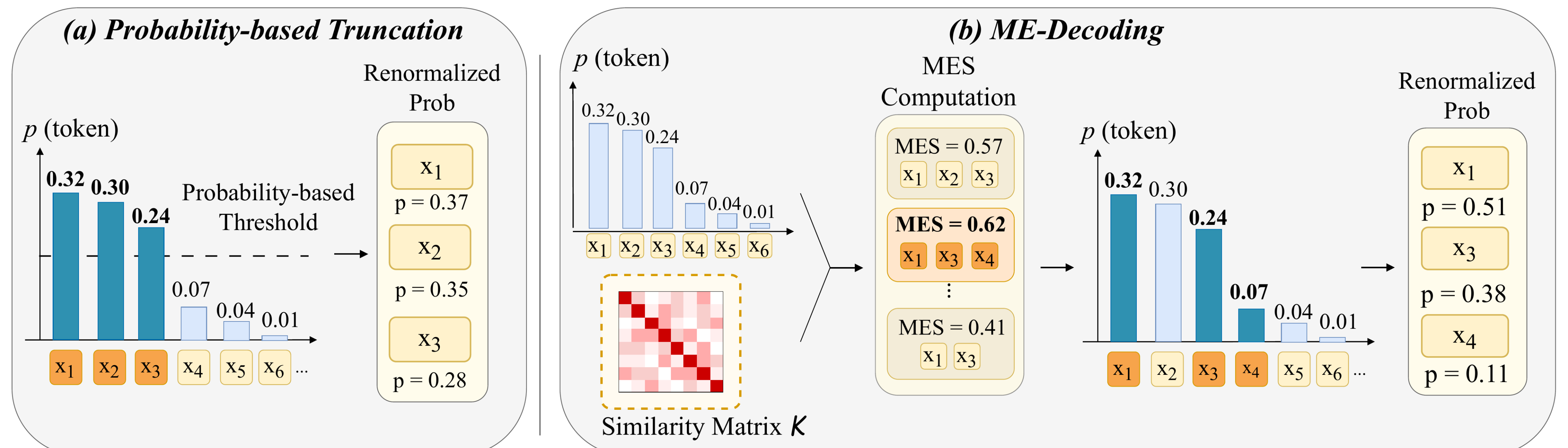
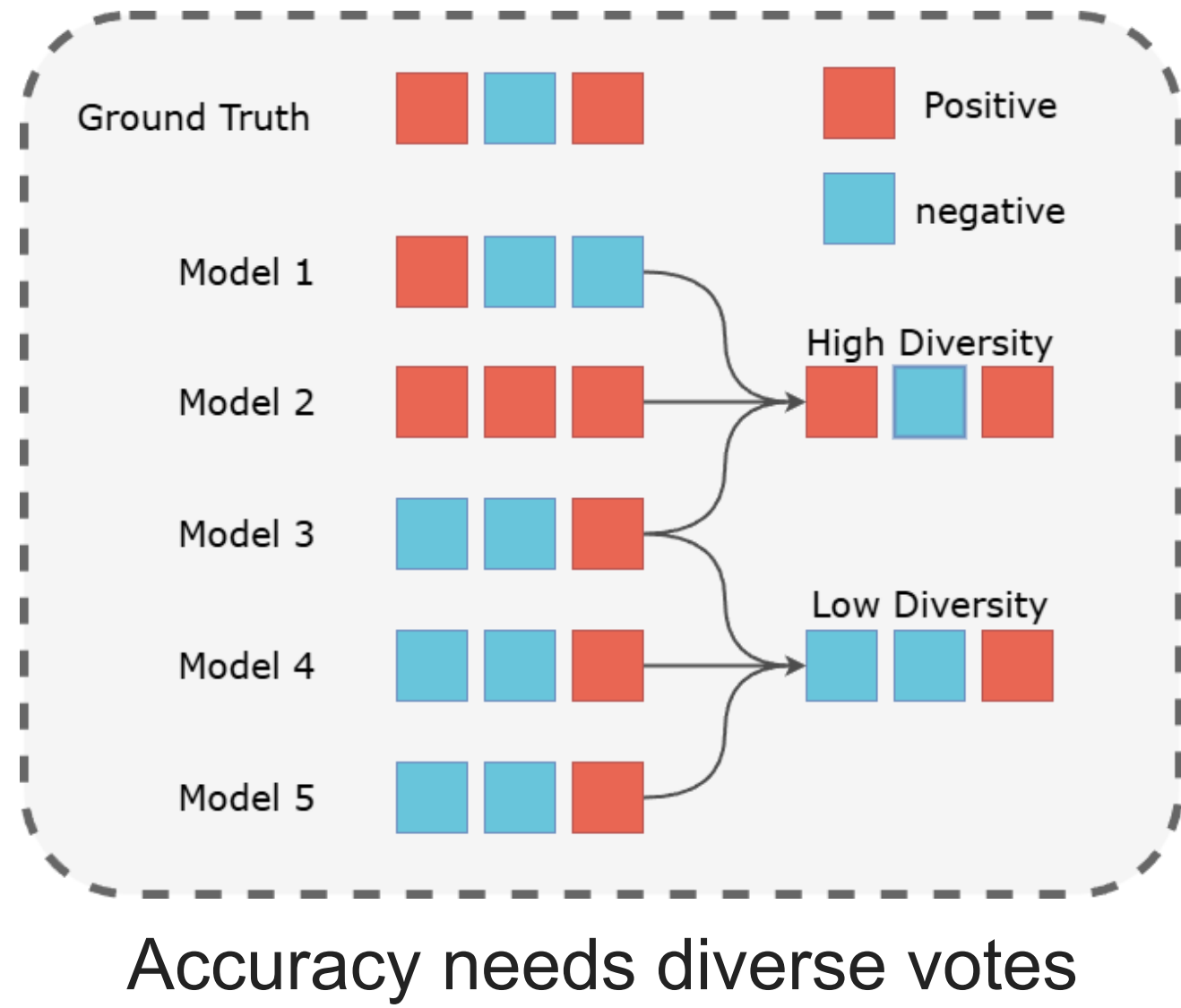
1. Diversity means exploring different generation paths for the same prompt.
2. At equal per-sample accuracy, diversity may raise the chance of finding a correct answer.

Limitations of existing methods

1. Probability alone overlooks semantic similarity.
2. Truncation struggles to adapt to distribution changes.
3. Redundant candidates crowd the retained support.

Method	Decoding Strategy	Geometry Aware ¹	Adaptive Selection	Redundancy Control ²
Greedy	Determin.	✗	✗	✗
Top- <i>k</i>	Prob. Trunc.	✗	✗	✗
Min- <i>p</i>	Prob. Trunc.	✗	✗	✗
<i>p</i> -less	Prob. Trunc.	✗	✓	✗
Top- <i>W</i>	Dist. Match.	✓	✓	✗
CraEG	Reweighting	✓	✗	✓
Ours	Ensemble	✓	✓	✓

Ensemble-learning inspiration:



Main Contributions

- Reformulate LLM decoding as ensemble pruning.
- Develop efficient greedy selection with conditional guarantees.
- Select compact token ensembles with a Mahalanobis objective.
- Improve performance on reasoning and generation tasks.

Method: ME-Decoding

1. Probability and redundancy

$$\text{MEE}(\mathcal{S}) = \mathbf{p}_{\mathcal{S}}^{\top} \mathbf{K}_{\mathcal{S}}^{-1} \mathbf{p}_{\mathcal{S}}, \quad \text{MES}_{\lambda}(\mathcal{S}) = \frac{\text{MEE}(\mathcal{S})}{[1 + \lambda H_2^T(\mathbf{p})]^{|\mathcal{S}|}},$$

$$H_2^T(\mathbf{p}) = 1 - \sum_i p_i^2.$$

MEE combines token confidence with a redundancy discount. MES adds an uncertainty-dependent size penalty.

2. Adaptive semantic kernel

$$C_{ij} = 1 - \mathbf{e}_i^{\top} \mathbf{e}_j, \quad K_{ij} = \exp(-C_{ij}/\epsilon), \quad \epsilon = \frac{1}{2} \sum_{i,j} p_i p_j C_{ij}.$$

Use normalized input embeddings and probabilities renormalized over the candidate pool. The bandwidth adapts to local semantic dispersion.

3. Greedy selection and stopping

$$\Delta_j(\mathcal{S}) = \text{MEE}(\mathcal{S} \cup \{j\}) - \text{MEE}(\mathcal{S}), \quad j^* = \arg \max_{j \notin \mathcal{S}} \Delta_j(\mathcal{S}),$$

$$\text{accept } j^* \iff \text{MES}_{\lambda}(\mathcal{S} \cup \{j^*\}) > \text{MES}_{\lambda}(\mathcal{S}).$$

Start with the most probable token. Accept the largest MEE gain only while MES improves.

$$\tilde{p}_i = \frac{p_i \mathbf{1}[i \in \mathcal{S}]}{\sum_{j \in \mathcal{S}} p_j}.$$

Theoretical Guarantees

Greedy-path unimodality

$$\kappa_N(\mathbf{K}) \leq 1 + \lambda H_2^T(\mathbf{p}) \implies \text{MES}_{\tau}^g = \max_{0 \leq t \leq N} \text{MES}_t^g.$$

Under the kernel condition, the first non-increasing step τ gives the highest MES along the greedy path.

Approximation to the global optimum

$$\text{MES}_{\tau}^g \geq \left(1 - e^{-\lambda_{\min}(\mathbf{K}, N)}\right) \text{MES}_{\lambda}^*.$$

The bound compares stopped greedy MES with the global optimum and holds under the same kernel condition.

Experiments

GSM8K accuracy (%)

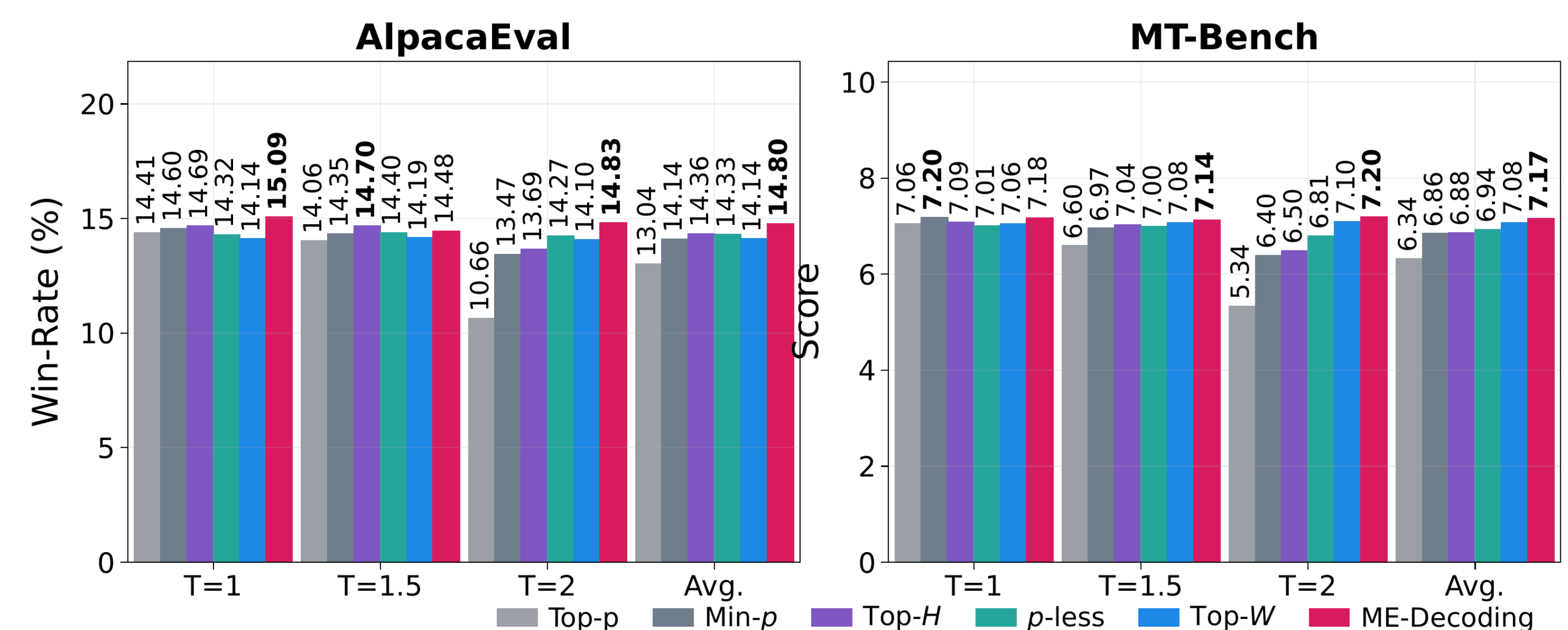
Method	Qwen3-4B-Inst.			Phi-4-mini-Inst.			Mistral-7B-Inst.			Avg.
	<i>T</i> = 1.0	<i>T</i> = 1.5	<i>T</i> = 2.0	<i>T</i> = 1.0	<i>T</i> = 1.5	<i>T</i> = 2.0	<i>T</i> = 1.0	<i>T</i> = 1.5	<i>T</i> = 2.0	
Min- <i>p</i>	70.74	62.02	54.06	79.68	70.81	30.86	48.67	36.32	17.97	52.35
Top- <i>p</i>	68.01	55.19	21.00	80.06	11.75	0.30	48.90	23.05	0.45	34.30
<i>p</i> -less	78.92	72.71	63.15	84.05	83.09	69.52	54.06	50.80	46.47	66.97
Top- <i>H</i>	76.80	67.17	58.53	81.65	73.77	34.65	52.01	46.78	22.52	57.10
Top- <i>W</i>	77.71	76.65	74.98	82.64	83.24	82.18	53.98	53.15	51.02	70.62
Ours	80.67	79.76	80.06	84.08	84.23	82.41	55.34	53.45	53.90	72.66

GPQA accuracy (%)

Method	Qwen3-4B-Inst.			Phi-4-mini-Inst.			Mistral-7B-Inst.			Avg.
	<i>T</i> = 1.0	<i>T</i> = 1.5	<i>T</i> = 2.0	<i>T</i> = 1.0	<i>T</i> = 1.5	<i>T</i> = 2.0	<i>T</i> = 1.0	<i>T</i> = 1.5	<i>T</i> = 2.0	
Min- <i>p</i>	34.38	32.59	35.04	27.01	29.02	26.56	26.12	28.12	25.22	29.34
Top- <i>p</i>	34.60	35.49	30.58	31.25	27.01	15.40	25.89	26.41	14.06	26.74
<i>p</i> -less	35.27	34.82	35.49	33.26	33.26	27.90	27.68	27.01	24.78	31.08
Top- <i>H</i>	34.38	33.93	34.15	34.38	33.93	29.91	<u>28.12</u>	27.23	23.88	31.10
Top- <i>W</i>	34.82	36.83	35.27	31.70	29.46	31.47	27.23	29.02	28.23	31.56
Ours	35.49	<u>35.49</u>	35.71	34.38	36.16	33.93	28.35	<u>28.79</u>	28.35	32.96

ME-Decoding achieves the highest average accuracy on GSM8K and GPQA. Accuracy is more stable as sampling temperature increases.

AlpacaEval and MT-Bench



DeepSeek-V4-Pro evaluation: highest average AlpacaEval win rate and MT-Bench score.

Matched controls

Method	GSM8K Acc. (%)	GPQA Acc. (%)
Size-matched	61.79 ± 1.26	28.47 ± 1.17
Entropy-matched	61.79 ± 0.99	28.47 ± 1.17
Cosine-matched	62.70 ± 0.35	29.25 ± 0.91
ME-Decoding	63.20 ± 0.57	32.66 ± 1.05

Qwen2.5-1.5B, *T* = 1.0. Mean ± SD over 3 seeds.

Gains remain after approximately matching size, entropy or cosine.

Sampling efficiency

2.7–7.4× faster CPU sampling

Versus Top-*W*, for sampler time only. GPU total-latency overhead: 1.5%–5.8% versus the fastest baseline.