Test-Time Augmentation for LLMs: Where to Inject Diversity at Inference Time

Nikita Kozodoi, Zainab Afolabi, Jack Butler
AWS Forward Deployed Engineering

amazon | science

1 Problem

More LLM answers raise accuracy, but **each answer costs a call**.

- **Self-consistency** varies only the reasoning path, from a fixed input.
- **Test-Time Augmentation (TTA)** varies the input instead.

Question: at a matched number of calls, where should diversity go?

Key Findings

- Paraphrases gain **+1.8pp**, beating self-consistency on **5 of 6** tasks.
- Character noise disagrees most (**37.8%**) but gains less.
- Only meaning-preserving diversity helps voting.

2 Test-Time Augmentation

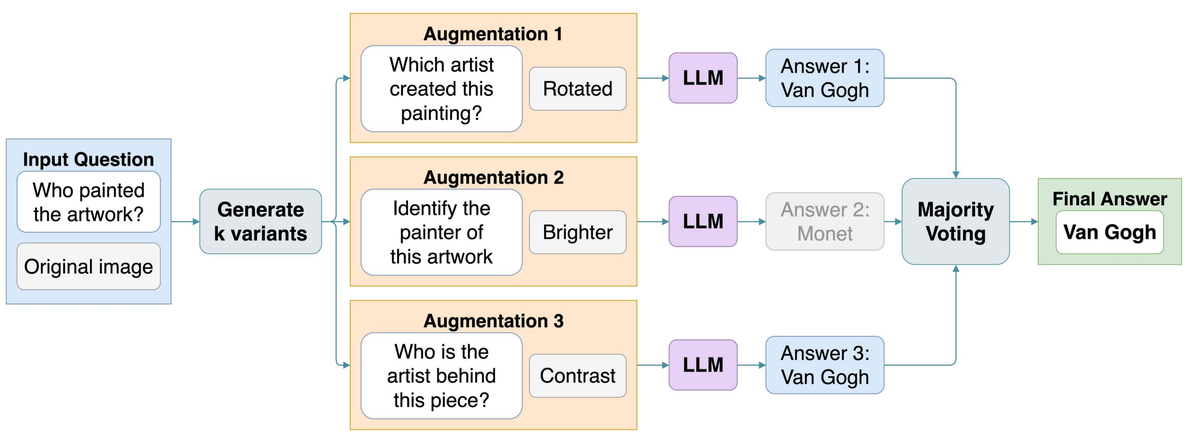

Figure 1. Each of the k variants gets one LLM call, and majority voting picks the answer.

Method	Varies	Example input
Self-consistency	output	What is the capital of France?
Semantic TTA	meaning	Which city is France's capital?
Lexical TTA	spelling	What is the captial of Frnace?
Visual TTA	image	rotation, brightness, contrast

Table 1. Methods compared. Typos are intentional.

3 Setup

Benchmarks: MMLU, MMMLU, MMMU, HLE, Math500, IMDB (400 each).

Model: Claude 4.5 Haiku at $T = 0.75$, with k of 2, 4 or 6 tuned on held-out data.

Baselines: single-call CoT prompting and self-consistency at the same k.

4 Model Size

MMMLU	Haiku	Sonnet	Opus
CoT prompting	79.75	85.50	88.25
Self-consistency	81.00	86.75	88.25
Lexical TTA	81.25	86.75	87.00
Semantic TTA	**82.50**	**87.50**	**88.50**

Table 2. Accuracy (%) across Claude 4.5 model sizes.

Gains shrink near the ceiling: TTA is a **mid-tier tool**.

Rephraser: Sonnet or Opus rephrasers land within 0.4pp of Haiku rephrasing itself, at 3 to 5× the cost.

5 Accuracy Gains

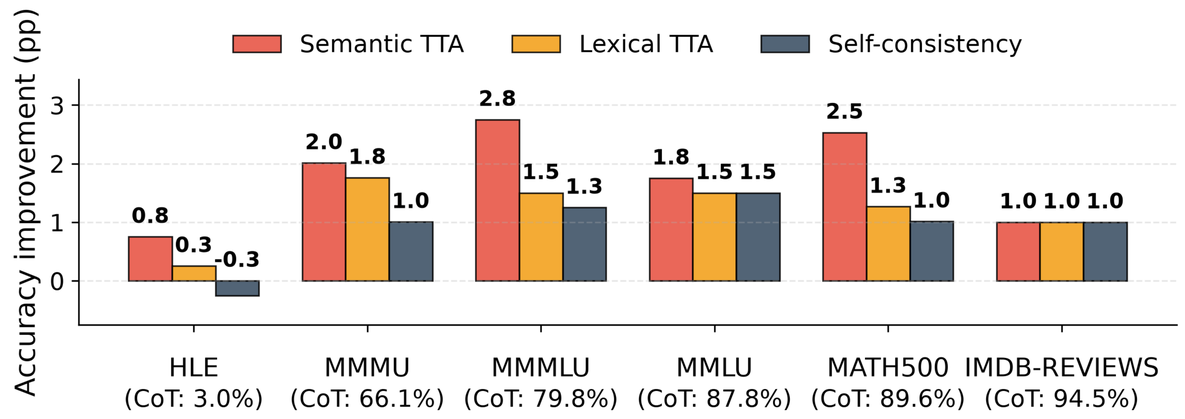

Figure 2. Gain over one CoT call (pp). Semantic TTA beats self-consistency at $p < 0.05$.

6 More Disagreement Is Not Better

Method	Disagree	Gain
Lexical TTA	**37.8%**	+1.25pp
Semantic TTA	27.4%	**+1.79pp**
Self-consistency	22.8%	+0.92pp

Table 3. Share of questions where the $k = 4$ answers split.

Paraphrases keep the answer intact, so voting can choose between them. Character noise yields answers that are often **wrong**.

7 Cost and Modality

Method	pp per $	pp per call	Best k
Semantic TTA	**133.1**	**0.40**	**4.3**
Lexical TTA	104.3	0.23	5.3
Self-consistency	91.8	0.21	4.7

Table 4. Gain per extra dollar and call, and mean selected k.

On MMMU, paraphrasing the text (68.1%) beats image transforms (67.6%). Perturbing both falls **below one call** (65.1% vs 66.1%).

8 Take-aways

Where to inject diversity

- Paraphrase the question, starting at $k = 4$.
- Augment text alone on multi-modal inputs.
- The answering model can rephrase itself.

Paper
arxiv.org/abs/2608.09351

Python library
github.com/aws-samples

kozodoi@amazon.com