

# AT-TTA: Dual-View Anchor-Tethered Test-Time Adaptation for Large Language Models

Yaxin Wang<sup>1</sup>, Junfeng He<sup>1</sup>, Zhihui Lu<sup>1</sup>, Weishen Lu<sup>1</sup>, Shijing Hu<sup>1</sup>,  
Peng Chen<sup>2</sup>, Ye Wang<sup>3</sup>, Chao Huang<sup>4</sup>, Yao Cheng<sup>5</sup>, Yudong Cheng<sup>1,†</sup>

<sup>1</sup>Fudan University    <sup>2</sup>Nanjing University of Information Science and Technology

<sup>3</sup>Renmin University of China    <sup>4</sup>Sun Yat-sen University

<sup>5</sup>Nanyang Technological University

{yaxinwang25, jfhe25, wslu24, sjhu24, ydcheng25}@m.fudan.edu.cn, lzh@fudan.edu.cn,  
pengchen@nuist.edu.cn, yewang@ruc.edu.cn, huangch253@mail.sysu.edu.cn, yao.cheng@ntu.edu.sg

## Abstract

Large Language Models (LLMs) frequently exhibit performance degradation in specialized domains due to distribution shifts. While unsupervised Test-Time Adaptation (TTA) offers a lightweight solution, existing methods suffer from critical flaws: input-centric adaptation lacks reasoning constraints, while output-centric optimization falls into a “Confidence Trap” of reinforced hallucinations. We instead study Few-Shot Anchored TTA, where a small labeled support set is available, and propose Anchor-Tethered Test-Time Adaptation (AT-TTA), a framework that synergizes linguistic alignment with reliable reasoning. AT-TTA employs Dual-View Uncertainty Calibration, minimizing input perplexity for domain alignment and sequence-level output energy for generation consistency. Furthermore, to mitigate catastrophic drift, we construct a “Semantic Anchor” by fine-tuning on the labeled support set and its filtered augmentations, which serves as a stable domain prior and constrains the adaptation trajectory via an Energy-Guided Adaptive Replay strategy. Extensive experiments on the AdaptEval benchmark demonstrate strong performance across diverse domains, effectively preventing model collapse while maintaining high inference efficiency.

## 1 Introduction

Large Language Models (LLMs) have achieved strong performance on a wide range of knowledge-intensive tasks, particularly in open-domain question answering (OpenAI, 2023; Llama Team, 2024; Yang et al., 2025). However, their performance drops significantly in specialized domains such as clinical medicine, financial forecasting, and agricultural science due to distribution shifts. Unlike general pre-training data, these fields require strict ter-

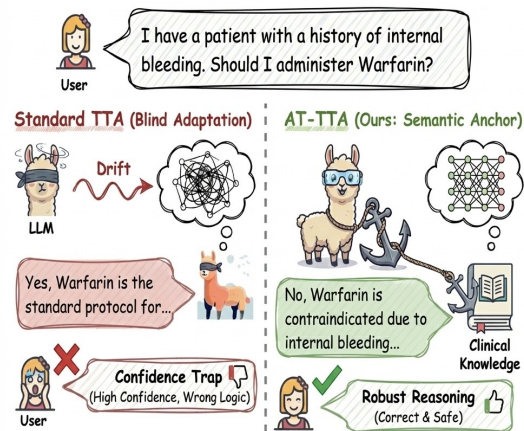


Figure 1: Comparing standard unsupervised TTA with AT-TTA in the few-shot anchored setting. While unsupervised methods risk drifting into a “Confidence Trap” (Left), AT-TTA utilizes a Semantic Anchor as a safety tether (Right) to constrain the model, facilitating robust performance in specialized domains.

minologies and rigorous reasoning (Singhal et al., 2023; Wu et al., 2023; Kuska et al., 2024). Consequently, the general capabilities of LLMs fail to transfer effectively, leading to poor performance in these target domains. Such adaptation capability is particularly valuable for specialized applications such as financial technology, where models may encounter evolving domain distributions and task requirements.

A common solution to this problem is Supervised Fine-Tuning (SFT) (Wei et al., 2022; Ouyang et al., 2022) on domain-specific data. While effective, this approach relies on large amounts of high-quality annotations and assumes a relatively static target distribution, making it costly and inflexible in dynamic deployment scenarios. Test-Time Adaptation (TTA) (Wang et al., 2021; Hu et al., 2025) has therefore emerged as an appealing alternative. By updating model parameters online

<sup>†</sup>Corresponding author.

using unlabeled test data, TTA enables models to adjust to distribution shifts without requiring massive explicit supervision or full retraining. Despite its promise, test-time adaptation for generative language models introduces unique challenges.

Current TTA methods (Wang et al., 2021; Niu et al., 2022; Hu et al., 2025) for LLMs largely follow a single-objective, unsupervised paradigm. They update the model by optimizing a loss computed on the current test inputs and the model’s own predictions, without any external supervision. In generative settings, however, this purely unsupervised approach often induces a fragile feedback loop: the model must treat its own outputs as implicit supervision under distribution shift. Input-centric approaches, such as minimizing input perplexity (Hu et al., 2025), improve alignment with domain-specific linguistic patterns but do not explicitly regulate the reasoning process during generation. As a result, improved fluency does not necessarily translate into reliable or correct outputs. Conversely, output-based objectives that encourage confident predictions, such as entropy minimization (Wang et al., 2021; Niu et al., 2022), lack semantic grounding and thus cannot distinguish correct generations from hallucinations. As illustrated in Figure 1, such blind optimization drives the model into a “Confidence Trap”, where disconnected optimization drives the model toward increasingly confident but erroneous outputs (e.g., ignoring critical medical contraindications), causing degradation in reasoning quality.

To address these limitations, recognizing that purely unsupervised TTA is prone to catastrophic drift for generative tasks, we clarify our problem setting as Few-Shot Anchored Test-Time Adaptation and present the AT-TTA (Anchor-Tethered Test-Time Adaptation) framework, implementing a dual-view anchored approach. Specifically, AT-TTA constructs a “Semantic Anchor” from a small labeled support set and its filtered augmentations, which serves as a stable prior to guide adaptation. During inference, the framework performs Dual-View Uncertainty Calibration, jointly optimizing an input-view objective for linguistic alignment and an output-view objective to ensure reasoning reliability. To prevent catastrophic drift, we introduce an Energy-Guided Adaptive Replay mechanism, which dynamically modulates the anchor’s influence based on the model’s uncertainty, acting as a “safety tether”.

Our contributions are summarized as follows:

- We propose AT-TTA, a few-shot anchored TTA framework that effectively bridges linguistic alignment and reliable reasoning. By coupling input- and output-view objectives with a Semantic Anchor, our approach explicitly mitigates the risk of model distribution shift.
- We introduce Dual-View Uncertainty Calibration alongside Energy-Guided Adaptive Replay. The former leverages complementary signals from inputs (perplexity) and outputs (energy), while the latter dynamically activates anchor-based regularization only when high uncertainty is detected, ensuring a balance between plasticity and stability.
- We validate AT-TTA on benchmarks covering diverse specialized domains, including medicine, finance, and agriculture. Experiments demonstrate strong performance across models of varying scales and consistent gains over both input-centric and output-centric baselines, including comparisons under matched few-shot initialization.

## 2 Related Work

### 2.1 Adaptation Strategies for Large Language Models

Existing strategies for adapting LLMs to specialized domains fall into two categories. Gradient-based approaches, such as SFT and RLHF (Wei et al., 2022; Ouyang et al., 2022; Chung et al., 2024), update model parameters using annotated corpora, but are computationally expensive and assume static target distributions. Inference-time methods, such as RAG (Lewis et al., 2020; Fan et al., 2024) and ICL (Brown et al., 2020), inject domain knowledge without parameter updates, but incur high latency and depend on external retrieval quality (Gao et al., 2023). These strategies are effective but either require costly offline supervision or depend on external retrieval/context quality. In contrast, TTA offers a lightweight parameter-updating alternative for adapting to shifting test distributions (Liang et al., 2025).

### 2.2 Test-Time Adaptation

Current TTA methods (Wang et al., 2021; Niu et al., 2022; Hu et al., 2025) for LLMs primarily distinguish themselves via their self-supervised optimization objectives, which can be divided into output-

centric and input-centric approaches. Output-centric methods, originally developed for computer vision, focus on sharpening the model’s prediction distribution. For instance, Tent (Wang et al., 2021) minimizes the Shannon entropy of model predictions, while EATA (Niu et al., 2022) incorporates Fisher regularization to selectively update reliable samples. COME (Zhang et al., 2025) further mitigates overconfidence and model collapse by conservatively minimizing entropy with uncertainty-aware regularization. Input-centric methods, specifically designed for generative tasks, optimize the input representation. The SOTA method, TLM (Hu et al., 2025), minimizes the perplexity of the input prompt to align the model with the target domain’s linguistic patterns. More recently, test-time fine-tuning has been framed as active data selection (Hübotter et al., 2025), while MLP projection matrices have been treated as fast weights adaptable at inference (Feng et al., 2026), reflecting a growing interest in lightweight parameter-efficient TTA. In summary, existing TTA methods rely on a single unsupervised learning signal. Despite recent efforts to improve their stability, output-centric methods lack explicit semantic grounding, while input-centric methods do not explicitly regulate the generation trajectory. In contrast, we study a few-shot anchored setting that combines input and output views and uses a semantic anchor derived from limited labeled support to regularize online adaptation.

### 3 Problem Formulation

We consider a pre-trained LLM  $P_\theta$  that generates responses autoregressively, encountering a stream of unlabeled test data  $\mathcal{D}_{\text{test}}$  from a target domain with distribution shift  $\mathcal{D}_{\text{test}} \neq \mathcal{D}_{\text{train}}$ . Standard TTA dynamically updates  $\theta \rightarrow \theta'$  by minimizing a self-supervised objective:

$$\min_{\theta'} \mathcal{L}_{\text{self}}(x; \theta'), \quad x \in \mathcal{D}_{\text{test}}. \quad (1)$$

However, under severe distribution shift, blindly optimizing a single unsupervised signal risks catastrophic forgetting or reinforcement of hallucinations.

**Few-Shot Anchored TTA Setting.** We assume access to a minimal labeled support set  $\mathcal{S} = \{(x_k, y_k)\}_{k=1}^K$  from the target domain, where  $K \ll M$  (e.g.,  $K = 16$ ). Unlike standard TTA,  $\mathcal{S}$  is not merely used for initialization: it serves as a persistent regularization constraint that anchors the

adaptation trajectory, preventing drift into hallucination during online updates.

## 4 AT-TTA: Dual-View Anchored Test-Time Adaptation

In this section, we propose AT-TTA, a few-shot anchored test-time adaptation framework for generative models under distribution shift. As illustrated in Figure 2, our method operates in two distinct stages: (1) **Anchor Initialization**, where a “Semantic Anchor” is constructed from the labeled support set and its filtered augmentations to provide a stable domain prior; and (2) **Synergistic Adaptation**, where the model is adapted using a “Dual-View Uncertainty Calibration” mechanism coupled with “Energy-Guided Adaptive Replay”. This design ensures the model adapts to new linguistic patterns while remaining tethered to reliable reasoning.

### 4.1 Anchor Initialization

Few-shot fine-tuning directly on a tiny support set  $\mathcal{S}$  ( $K \ll M$ ) is prone to overfitting and instability. To obtain a more robust prior, we first expand  $\mathcal{S}$  into an augmented set  $\mathcal{S}_{\text{aug}}$ . Given the original support pairs  $\mathcal{S} = \{(x_k, y_k)\}_{k=1}^K$ , we apply a suite of instruction-preserving augmentation strategies driven by a strong LLM (e.g., GPT-4 (OpenAI, 2023)). Concretely, we generate variants of each query by: Rephrasing, Complexity, Decomposition and Constraint. Detailed configurations and examples of the augmentation strategies are provided in Appendix D.3. These strategies diversify linguistic patterns and reasoning structures while keeping the underlying semantics anchored to the ground-truth  $y_k$ .

However, generative augmentation inevitably introduces noise. To ensure that the augmented variations maintain semantic consistency with the original user intent, we implement a rigorous quality control mechanism based on Natural Language Inference (NLI). Specifically, we employ a pre-trained NLI model to evaluate the logical relationship between the original and augmented queries, strictly filtering out samples that contradict the source instruction (see Appendix D.4 for detailed formulations). We denote the filtered, high-quality set as  $\mathcal{S}_{\text{valid}}$ . Consequently, the final anchor set is:

$$\mathcal{S}_{\text{anchor}} = \mathcal{S} \cup \mathcal{S}_{\text{valid}}. \quad (2)$$

We then perform parameter-efficient supervised fine-tuning with LoRA (Hu et al., 2022) on  $\mathcal{S}_{\text{anchor}}$ ,

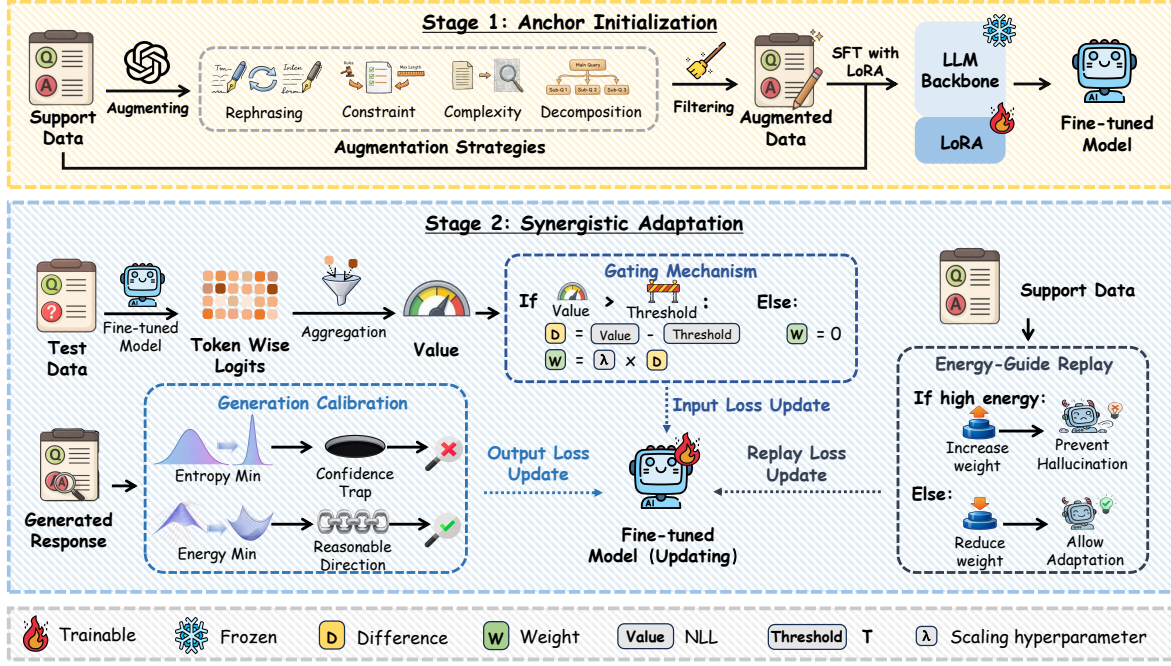


Figure 2: The overall framework of AT-TTA. It consists of two stages: (1) Anchor Initialization and (2) Synergistic Adaptation, which incorporates both Dual-View Uncertainty Calibration and Energy-Guided Adaptive Replay.

starting from the pre-trained parameters  $\theta_{\text{pre}}$ :

$$\theta_0 \leftarrow \text{LoRA-SFT}(\theta_{\text{pre}}, \mathcal{S}_{\text{anchor}}). \quad (3)$$

The resulting model provides a weak but stable domain prior, which we refer to as the *Semantic Anchor*. Importantly, this anchor is not the final adapted model; instead, it acts as a stable reference that constrains the subsequent test-time updates.

## 4.2 Synergistic Adaptation

### 4.2.1 Dual-View Uncertainty Calibration

During the test-time phase, the model adapts to the unlabeled stream  $\mathcal{D}_{\text{test}}$  via dynamic parameter optimization. We calibrate uncertainty from two complementary views: **input view**, which aligns the model with the target-domain query distribution; **output view**, which evaluates the sequence-level compatibility of generated responses. This dual-view design addresses the limitations of purely input-centric or purely output-centric TTA.

**Input View: Linguistic Alignment.** Following prior work (Hu et al., 2025), we use token-averaged negative log-likelihood (NLL), equivalently log perplexity, to measure how well an input aligns with the target domain. For an input sequence  $x = (x_1, \dots, x_m)$ , we compute

$$\mathcal{L}_{\text{input}}(\theta) = -\frac{1}{m} \sum_{i=1}^m \log P_{\theta}(x_i | x_{<i}), \quad (4)$$

where  $m$  is the number of tokens. To focus adaptation on challenging inputs, let  $\text{NLL}(x)$  denote this quantity and define the gating weight

$$W(x) = \lambda_{\text{input}} \cdot \mathbf{1}[\text{NLL}(x) > \tau] \cdot \exp(\text{NLL}(x) - \tau), \quad (5)$$

where  $\tau$  is the gating threshold and  $\lambda_{\text{input}}$  controls the update scale.

This gating mechanism has two effects. For below-threshold inputs, where  $\text{NLL}(x) \leq \tau$ , the gate sets  $W(x) = 0$ , avoiding unnecessary updates to inputs that are already well aligned. By contrast, for above-threshold inputs, where  $\text{NLL}(x) > \tau$ , the exponential term assigns larger weights as NLL increases, prioritizing adaptation to poorly aligned, high-shift queries. Together, these two behaviors improve the efficiency and stability of input-side adaptation.

**Output View: Generation Calibration.** While prompt perplexity captures how well the input matches the target domain, it does not provide information about the correctness or reliability of the generated output. We therefore introduce an output-view signal that evaluates the model’s own responses based on energy (Liu et al., 2020b).

Given a generated output  $\hat{y}$ , we define the

sequence-level energy as:

$$\begin{aligned}\mathcal{L}_{\text{output}}(\theta) &= E(x, \hat{y}) \\ &\approx -\frac{1}{|\hat{y}|} \sum_{j=1}^{|\hat{y}|} \log P_{\theta}(\hat{y}_j \mid \hat{y}_{<j}, x).\end{aligned}\quad (6)$$

This length-normalized quantity measures how strongly the model supports its own generation: low energy corresponds to high sequence likelihood, whereas high energy indicates weaker support along the generation trajectory. Importantly, energy is not treated as a ground-truth correctness signal. Instead, it serves as an empirical reliability indicator that complements input perplexity, as examined in Section 5.6. Together, the two signals characterize uncertainty from the input and output views, respectively.

#### 4.2.2 Energy-Guided Adaptive Replay

While dual-view calibration aligns the model with the test stream, relying solely on self-supervised signals carries the risk of ‘‘Self-Reinforced Hallucination,’’ where the model minimizes  $\mathcal{L}_{\text{output}}$  for factually incorrect generations. To address this issue, we introduce ‘‘Energy-Guided Adaptive Replay,’’ which tethers the adaptation trajectory to the ‘‘Semantic Anchor’’. The key idea is to modulate the strength of anchor supervision based on the reliability of the current generation.

We use the generation energy  $E(x, \hat{y})$  as an uncertainty indicator. High energy increases the influence of the anchor, while low energy allows more self-supervised adaptation. Specifically, we define a ‘‘Dynamic Replay Weight’’  $\gamma(x)$  that scales with the output energy of the current test sample  $x$ . The weight is calculated as:

$$\gamma(x) = \gamma_{\text{base}} \cdot (1 + \alpha \cdot \Psi(E(x, \hat{y}))), \quad (7)$$

where  $\gamma_{\text{base}}$  is the base regularization coefficient,  $\alpha$  controls the sensitivity, and  $\Psi(\cdot)$  represents a normalization function mapping energy scores to a relative scale  $\Psi(E(x, \hat{y})) = \text{Sigmoid}\left(\frac{E(x, \hat{y}) - \mu_E}{\sigma_E + \epsilon}\right)$ . Here,  $\mu_E$  and  $\sigma_E$  denote the mean and standard deviation of energy scores, and  $\epsilon$  is a small constant for numerical stability.

The final objective at each test step combines the synergistic adaptation loss with the anchored replay loss:

$$\mathcal{L}_{\text{total}} = W(x)\mathcal{L}_{\text{input}} + \beta\mathcal{L}_{\text{output}} + \gamma(x)\mathcal{L}_{\text{CE}}(\mathcal{S}_{\text{anchor}}), \quad (8)$$

where  $W(x)$  is the input weighting coefficient (defined in Eq. 5), and  $\beta$  balances the output confidence constraint.  $\mathcal{L}_{\text{CE}}(\mathcal{S}_{\text{anchor}})$  denotes the standard cross-entropy loss computed on a sampled batch from the Semantic Anchor set. This mechanism dynamically balances plasticity by adapting to new patterns in  $x$  and stability by preserving reasoning logic from  $\mathcal{S}_{\text{anchor}}$ , acting as a ‘‘safety tether’’ against model collapse. Algorithm 1 summarizes the complete AT-TTA procedure, connecting Anchor Initialization with the per-sample Synergistic Adaptation updates.

---

#### Algorithm 1 AT-TTA: Dual-View Anchor-Tethered Test-Time Adaptation

---

**Require:** Unlabeled test data  $\mathcal{D}_{\text{test}}$ , labeled support set  $\mathcal{S}$ , and pre-trained model parameters  $\theta_{\text{pre}}$ .

**Ensure:** Adapted model parameters  $\theta$ .

- 1: **Stage 1: Anchor Initialization**
  - 2: Generate  $\mathcal{S}_{\text{valid}}$  and construct  $\mathcal{S}_{\text{anchor}} \leftarrow \mathcal{S} \cup \mathcal{S}_{\text{valid}}$ .
  - 3: Initialize  $\theta_0 \leftarrow \text{LoRA-SFT}(\theta_{\text{pre}}, \mathcal{S}_{\text{anchor}})$ .
  - 4: **Stage 2: Synergistic Adaptation**
  - 5: Initialize  $\theta \leftarrow \theta_0$ .
  - 6: **for** each test sample  $x \in \mathcal{D}_{\text{test}}$  **do**
  - 7:   Generate a pseudo-response  $\hat{y} \sim P_{\theta}(\cdot \mid x)$ .
  - 8:   Compute  $\mathcal{L}_{\text{input}}$  and  $\mathcal{L}_{\text{output}} = E(x, \hat{y})$ .
  - 9:   Compute replay weight  $\gamma(x)$  via Eq. 7.
  - 10:   Compute input weight  $W(x)$  via Eq. 5.
  - 11:   Sample  $(x_s, y_s) \sim \mathcal{S}_{\text{anchor}}$  and compute  $\mathcal{L}_{\text{CE}}(x_s, y_s)$ .
  - 12:   Compute  $\mathcal{L}_{\text{total}}$  via Eq. 8.
  - 13:   Update  $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{\text{total}}$ .
  - 14: **end for**
- 

## 5 Experiments

### 5.1 Experimental Settings

**Datasets and Evaluation Metrics.** For generation evaluation, we utilize the AdaptEval benchmark (Hu et al., 2025), comprising 7 datasets across two categories: (1) **DomainBench** (Geography, Agriculture, Medicine, Finance) assesses adaptation to specialized vertical domains; (2) **InstructionBench** (Alpaca-GPT4, Dolly, InstructionWild) evaluates instruction-following under distribution shifts. The statistics and details of these datasets can be found in Appendix A. We primarily report ROUGE-Lsum (Lin, 2004) as it measures generation quality through summary-level lexical overlap. To ensure comprehensiveness, auxiliary metrics

| Model       | Method                   | DomainBench  |              |              |              |              | InstructionBench |              |                 |              |
|-------------|--------------------------|--------------|--------------|--------------|--------------|--------------|------------------|--------------|-----------------|--------------|
|             |                          | Geography    | Agriculture  | Medicine     | Finance      | Avg.         | Alpaca-GPT4      | Dolly        | InstructionWild | Avg.         |
| Qwen2.5-7B  | Base Model               | 25.26        | 9.42         | 12.52        | 26.08        | 18.32        | 42.06            | 28.55        | 28.10           | 32.90        |
|             | Base Model + SFT         | 26.27        | 10.25        | 23.00        | 29.45        | 22.24        | 45.76            | 33.33        | 30.60           | 36.56        |
|             | Tent (Wang et al., 2021) | 22.15        | 11.05        | 5.34         | 15.42        | 13.49        | 20.15            | 18.67        | 16.54           | 18.45        |
|             | EATA (Niu et al., 2022)  | 24.01        | 10.89        | 11.56        | 25.43        | 17.97        | 22.11            | 24.67        | 15.11           | 20.63        |
|             | TLM (Hu et al., 2025)    | <u>29.45</u> | <u>12.49</u> | <u>25.28</u> | <u>33.21</u> | <u>25.11</u> | <u>46.22</u>     | <u>34.56</u> | <u>32.42</u>    | <u>37.73</u> |
|             | <b>AT-TTA (Ours)</b>     | <b>34.73</b> | <b>15.51</b> | <b>32.72</b> | <b>34.92</b> | <b>29.47</b> | <b>47.16</b>     | <b>36.21</b> | <b>35.84</b>    | <b>39.74</b> |
| Qwen2.5-14B | Base Model               | 26.42        | 10.63        | 14.43        | 26.32        | 19.45        | 42.60            | 29.07        | 28.48           | 33.38        |
|             | Base Model + SFT         | 30.65        | 13.38        | 29.28        | 31.80        | 26.28        | 47.43            | 38.19        | 33.27           | 39.63        |
|             | Tent (Wang et al., 2021) | 24.11        | 8.44         | 12.12        | 22.11        | 16.70        | 25.66            | 24.11        | 22.34           | 24.04        |
|             | EATA (Niu et al., 2022)  | 25.99        | 9.11         | 10.01        | 25.11        | 17.56        | 28.45            | 29.33        | 27.56           | 28.45        |
|             | TLM (Hu et al., 2025)    | <u>33.15</u> | <u>13.40</u> | <u>31.43</u> | <u>34.21</u> | <u>28.05</u> | 47.12            | 35.73        | 31.04           | 37.96        |
|             | <b>AT-TTA (Ours)</b>     | <b>33.35</b> | <b>21.74</b> | <b>32.16</b> | <b>34.57</b> | <b>30.46</b> | <b>48.63</b>     | <b>38.63</b> | <b>36.59</b>    | <b>41.28</b> |
| Llama-3-8B  | Base Model               | 24.32        | 8.55         | 12.65        | 23.31        | 17.21        | 37.54            | 36.37        | 26.06           | 33.32        |
|             | Base Model + SFT         | 26.10        | 9.73         | 22.72        | 24.21        | 20.69        | 40.53            | 36.63        | 25.82           | 34.33        |
|             | Tent (Wang et al., 2021) | 9.44         | 0.85         | 1.12         | 4.33         | 3.94         | 18.55            | 0.44         | 8.11            | 9.03         |
|             | EATA (Niu et al., 2022)  | 20.11        | 0.25         | 1.33         | 12.89        | 8.65         | 14.22            | 16.11        | 10.11           | 13.48        |
|             | TLM (Hu et al., 2025)    | <b>32.50</b> | <u>13.09</u> | <u>22.89</u> | <u>32.33</u> | <u>25.20</u> | <u>42.82</u>     | 35.74        | <u>29.37</u>    | <u>35.98</u> |
|             | <b>AT-TTA (Ours)</b>     | <u>31.64</u> | <b>19.83</b> | <b>24.19</b> | <b>34.51</b> | <b>27.54</b> | <b>45.27</b>     | <b>39.90</b> | <b>32.45</b>    | <b>39.21</b> |

Table 1: Main results on DomainBench and InstructionBench. ROUGE-Lsum scores ( $\times 100$  for readability). **Bold** indicates the best performance and underlined denotes the second-best results.

such as BLEU and BERTScore are provided in Appendix B. We separately evaluate general reasoning retention on ReasoningBench, which includes GSM8K (Cobbe et al., 2021), MetaMath (Yu et al., 2024), and LogiQA (Liu et al., 2020a), following the exact-match (EM) protocol used by TLM (Hu et al., 2025).

**Models and Baselines.** We conduct experiments on Qwen2.5-7B/14B (Yang et al., 2024) and Llama-3-8B (Llama Team, 2024). We compare AT-TTA against the unadapted Base Model, its few-shot SFT variant, and three representative TTA baselines: **Tent** (Wang et al., 2021), **EATA** (Niu et al., 2022), and **TLM** (Hu et al., 2025). We report both their standard unsupervised configurations and variants initialized from the same  $K = 16$  SFT checkpoint as AT-TTA. Detailed implementation settings can be found in Appendix C.

**Implementation Details.** We conduct all experiments under the Offline Setting (Hu et al., 2025). Our framework is implemented using LLaMA-Factory (Zheng et al., 2024) with vLLM (Kwon et al., 2023) acceleration. We apply LoRA (rank 8) to  $W_q$  and  $W_v$  projections, with learning rates of  $2 \times 10^{-5}$  for Anchor Initialization and  $5 \times 10^{-5}$  for Synergistic Adaptation. We fix the labeled support-set size at  $K = 16$ . Detailed dataset statistics and hyperparameters are provided in Table 6 (see Appendix A) and Appendix D. Following prior LLM-TTA protocols, we report large-scale results as single runs and include a targeted multi-seed stability analysis in Appendix E.7.

## 5.2 Main Results

To demonstrate the effectiveness of our method, Table 1 compares AT-TTA with the original LLM, its few-shot SFT variant, and standard unsupervised TTA baselines. Table 2 separately provides matched-initialization comparisons. Results for other evaluation metrics are provided in Appendix E. Across all three backbones, AT-TTA achieves the highest average score on both DomainBench and InstructionBench. From the tables, we make the following observations:

(1) AT-TTA obtains stronger end-to-end performance than the standard unsupervised TLM configuration. On Qwen2.5-7B, the average absolute differences are 4.36 points on DomainBench and 2.01 points on InstructionBench; on Medicine, the relative difference is 29.43%. Because AT-TTA uses labeled support, we treat the matched-initialization results in Table 2, rather than this comparison alone, as the controlled evidence for the benefit of its test-time mechanism.

(2) AT-TTA remains stable where entropy-based baselines collapse. Methods such as Tent and EATA, which rely heavily on confidence minimization, often fall into the ‘‘Confidence Trap,’’ leading to model collapse or significant degradation—for instance, Tent deteriorates from 12.52 to 5.34 on the Qwen2.5-7B Medicine benchmark. In contrast, our energy-guided anchoring acts as a critical regularization term, adaptively replaying Semantic Anchors to suppress the drift typical of unsupervised adaptation, thereby achieving a consistent balance between plasticity and stability.

(3) Under matched few-shot initialization, AT-TTA outperforms both the frozen SFT model and the SFT-initialized TTA baselines. To separate the effect of test-time adaptation from labeled support exposure, we initialize all TTA methods from the same  $K = 16$  SFT checkpoint. As shown in Table 2, Tent+SFT and EATA+SFT continue to suffer from catastrophic drift on Medicine and Agriculture. On Llama-3-8B Medicine, for example, performance decreases from 22.72 for the frozen SFT model to 13.85 and 14.90, respectively. TLM+SFT is more stable but remains below AT-TTA, showing that the gains cannot be attributed solely to few-shot initialization.

| Method               | Qwen2.5-7B   |              | Llama-3-8B   |              |
|----------------------|--------------|--------------|--------------|--------------|
|                      | Med.         | Agr.         | Med.         | Agr.         |
| Base Model           | 12.52        | 9.42         | 12.65        | 8.55         |
| Tent (Unsupervised)  | 5.34         | 11.05        | 1.12         | 0.85         |
| EATA (Unsupervised)  | 11.56        | 10.89        | 1.33         | 0.25         |
| TLM (Unsupervised)   | 25.28        | 12.49        | 22.89        | 13.09        |
| Base + SFT           | 23.00        | 10.25        | 22.72        | 9.73         |
| Tent + SFT           | 11.45        | 10.50        | 13.85        | 3.55         |
| EATA + SFT           | 15.20        | 11.85        | 14.90        | 4.20         |
| TLM + SFT            | 25.40        | 13.20        | 23.17        | 13.55        |
| <b>AT-TTA (Ours)</b> | <b>32.72</b> | <b>15.51</b> | <b>24.19</b> | <b>19.83</b> |

Table 2: Matched few-shot initialization comparison across models and domains (ROUGE-Lsum; Med.: Medicine, Agr.: Agriculture). The SFT-initialized TTA methods and AT-TTA use the same  $K = 16$  labeled support and SFT checkpoint.

### 5.3 Ablation Study

To evaluate the contribution of each component, we conduct ablation studies by removing one module at a time. Table 3 reports results on three representative DomainBench datasets.

(1) Omitting anchor initialization (W/O STAGE 1) causes substantial degradation, notably in the Medicine domain. Without the Semantic Anchor as a stable reference, self-supervised adaptation becomes significantly more susceptible to drift.

(2) Eliminating augmentation (W/O AUGMENT) leads to a consistent decline in performance, especially on semantically diverse domains such as Geography. This confirms that augmentation effectively expands anchor coverage to mitigate the sparsity of tiny support sets.

(3) Removing either output energy (W/O ENERGY) or replay (W/O REPLAY) consistently harms performance. Energy calibration improves gen-

| Variant              | Geography    | Agriculture  | Medicine     |
|----------------------|--------------|--------------|--------------|
| W/O STAGE 1          | 30.66        | 12.54        | 26.53        |
| W/O AUGMENT          | 30.58        | 12.03        | 29.24        |
| W/O ENERGY           | 32.73        | 14.39        | 31.39        |
| W/O REPLAY           | 32.73        | 14.18        | 31.72        |
| <b>AT-TTA (Full)</b> | <b>34.73</b> | <b>15.51</b> | <b>32.72</b> |

Table 3: Ablation study of AT-TTA on three representative domains from DomainBench using Qwen2.5-7B.

eration reliability, while replay enforces anchor constraints to prevent instability, validating their synergistic roles in test-time adaptation.

### 5.4 Robustness under True Online Constraints

To further assess AT-TTA’s robustness, we conduct experiments in the True Online setting, where the model processes each test input sequentially while updating its parameters, without revisiting previous samples. As shown in Table 4, standard unsupervised methods like Tent and EATA completely collapse, dropping to an average ROUGE-Lsum of 4.00 and 5.69, respectively. In the few-shot anchored setting, AT-TTA achieves 26.73, compared with 20.41 for the standard unsupervised TLM configuration. Once initialized from the labeled support, the Semantic Anchor provides a persistent domain reference without requiring access to past test samples, making AT-TTA compatible with non-repeatable streams.

| Method        | Geog.        | Agric.       | Med.         | Fin.         | Avg.         |
|---------------|--------------|--------------|--------------|--------------|--------------|
| Base Model    | 24.32        | 8.55         | 12.65        | 23.31        | 17.21        |
| Tent          | 8.15         | 1.15         | 1.50         | 5.20         | 4.00         |
| EATA          | 10.20        | 1.95         | 2.10         | 8.50         | 5.69         |
| TLM           | 27.64        | 15.75        | 13.63        | 24.59        | 20.41        |
| <b>AT-TTA</b> | <b>30.95</b> | <b>18.72</b> | <b>23.45</b> | <b>33.80</b> | <b>26.73</b> |

Table 4: True Online TTA performance on DomainBench using Llama-3-8B. ROUGE-Lsum scores ( $\times 100$ ).

### 5.5 Hyperparameter Sensitivity Analysis

We analyze the impact of key hyperparameters on the trade-off between plasticity and stability. As shown in Figure 3, performance peaks at 34.73 with moderate regularization ( $\beta = 0.05$ ,  $\gamma_{\text{base}} = 1.0$ ). Excessive energy ( $\beta \geq 0.5$ ) induces overconfidence, while strong replay ( $\gamma_{\text{base}} \geq 5.0$ ) inhibits adaptation. This confirms that AT-TTA relies on synergistic optimization to prevent objective dominance.

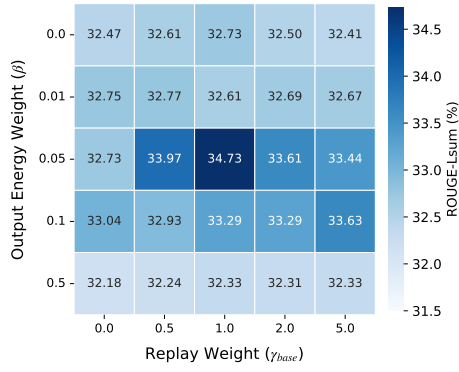


Figure 3: Sensitivity analysis of Output Energy Weight ( $\beta$ ) and Replay Weight ( $\gamma_{\text{base}}$ ) on Geography. Peak performance (dark blue) at balanced configurations confirms synergistic optimization.

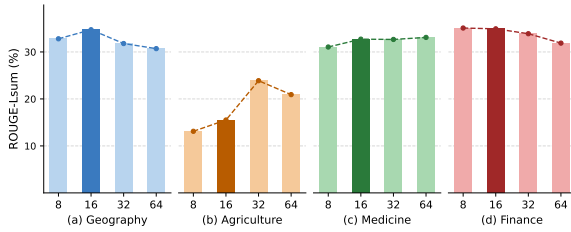


Figure 4: Sensitivity analysis of support set size  $K$  on Qwen2.5-7B. The optimal  $K$  varies across domains.

Figure 4 shows that the optimal  $K$  is domain-dependent, with numerical details provided in Table 20. Low-shift domains (e.g., Finance) peak at  $K = 8$ , where larger  $K$  over-regularizes, while high-shift domains (e.g., Agriculture) require more context ( $K = 32$ ). We adopt  $K = 16$  as a balanced default, which achieves near-peak performance on high-shift domains while maintaining robustness on low-shift ones.

## 5.6 Uncertainty Signal Analysis

Figure 5 compares length-normalized sequence energy and token-level entropy for correct and incorrect AT-TTA generations on the Medicine dataset using Qwen2.5-7B. We label an output as incorrect when its GPT-4o reference-answer correctness score is below 7. The entropy distributions exhibit substantial overlap, with a gap of only 0.09, whereas sequence energy yields a larger separation of 0.65. This result suggests that energy provides a more informative empirical reliability signal in this setting.

We further examine whether the adaptive replay weight follows the error-prone region identified by sequence energy. Outputs are grouped into energy quintiles, and the corresponding error rate

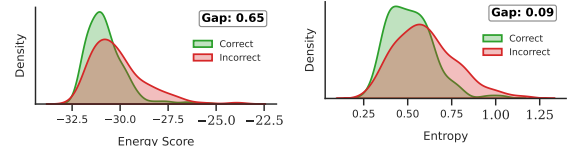


Figure 5: Distributions of length-normalized sequence energy and token-level entropy for correct and incorrect AT-TTA generations on Medicine using Qwen2.5-7B. Correctness is determined by a GPT-4o reference-answer correctness score threshold of 7.

and realized replay weight are measured within each group. Both quantities increase monotonically with energy, indicating that AT-TTA applies stronger anchor supervision in regions where errors occur more frequently. Meanwhile, Eq. 7 guarantees  $\gamma(x) \geq \gamma_{\text{base}}$ , preserving a baseline tether even for low-energy outputs. The complete protocol and results are provided in Appendix E.2.

## 5.7 General Reasoning Retention

To assess whether domain adaptation preserves the model’s prior reasoning capability, we evaluate AT-TTA on ReasoningBench under the evaluation protocol of TLM (Hu et al., 2025). Averaged across GSM8K, MetaMath, and LogiQA, AT-TTA obtains 0.7367 on Qwen2.5-7B, compared with 0.7253 for the base model and 0.7343 for TLM. On Llama-3-8B, the corresponding scores are 0.6670, 0.6357, and 0.6649, respectively. These results show no observed degradation relative to the base models and performance comparable to TLM. We therefore interpret ReasoningBench as evidence of reasoning retention rather than substantial reasoning improvement. Complete per-task results are provided in Appendix E.3.

## 5.8 Reference-Based Quality Evaluation

To complement lexical metrics (e.g., ROUGE and BLEU; see Appendix B), we use GPT-4o as an automatic judge across Medicine and Finance with Qwen2.5-7B and Llama-3-8B. Given a question, reference answer, and model response, the judge scores *Reference-Answer Correctness* and *Logical Reasoning* on a 1–10 scale (Appendix D.5). As shown in Table 5, AT-TTA obtains the highest scores in all reported settings, while Tent and EATA exhibit substantial quality degradation. Because correctness is assessed against the provided reference, this metric should not be interpreted as independent external fact verification. In a cross-evaluator validation on Medicine with Qwen2.5-7B, AT-TTA consistently ranks above TLM under

Claude Opus 4.7, Qwen-Max, and human evaluation, with human inter-annotator agreement of  $\kappa = 0.64$ ; the full score table and evaluation details are provided in Appendix E.5.

| Dataset  | Method               | Qwen2.5-7B |            | Llama-3-8B |            |
|----------|----------------------|------------|------------|------------|------------|
|          |                      | Ref. Corr. | Logic      | Ref. Corr. | Logic      |
| Medicine | Base Model           | 4.9        | 4.6        | 4.8        | 4.5        |
|          | Tent (Unsup.)        | 2.1        | 1.8        | 1.5        | 1.3        |
|          | EATA (Unsup.)        | 3.5        | 3.2        | 1.8        | 1.5        |
|          | TLM                  | 7.7        | 7.3        | 7.4        | 7.0        |
|          | <b>AT-TTA (Ours)</b> | <b>8.5</b> | <b>8.2</b> | <b>7.9</b> | <b>7.6</b> |
| Finance  | Base Model           | 6.5        | 6.2        | 6.2        | 5.8        |
|          | Tent (Unsup.)        | 3.8        | 3.5        | 1.9        | 1.6        |
|          | EATA (Unsup.)        | 5.8        | 5.5        | 3.2        | 2.9        |
|          | TLM                  | 8.0        | 7.7        | 7.8        | 7.5        |
|          | <b>AT-TTA (Ours)</b> | <b>8.6</b> | <b>8.3</b> | <b>8.3</b> | <b>8.0</b> |

Table 5: Reference-based evaluation (1–10) via GPT-4o on *Reference-Answer Correctness* (Ref. Corr.) and *Logical Reasoning* (Logic). Higher values are better.

## 5.9 Robustness to Augmentation Models

To examine whether AT-TTA depends on GPT-4 augmentation, we evaluate alternative augmentation backends, including Llama-3-8B and Qwen3.5-plus. As shown in Figure 6, Qwen3.5-plus produces more diverse and complex anchors than GPT-4, achieving the highest ROUGE-Lsum score of 25.69. These results suggest that AT-TTA scales with augmentation quality rather than relying on a specific proprietary model. Full results are provided in Appendix E.8.

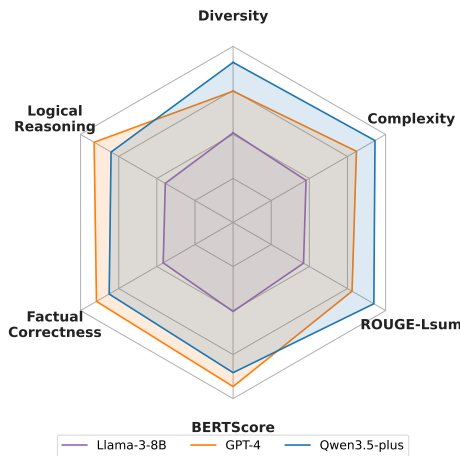


Figure 6: Comparison of augmentation models. Qwen3.5-plus yields richer anchors than GPT-4 in diversity and complexity while achieving higher ROUGE-Lsum.

## 5.10 Inference Efficiency

Despite online gradient updates, AT-TTA reduces per-sample latency from 10.65s to 6.91s for Qwen2.5-7B and from 3.10s to 2.60s for Llama-3-8B. As shown in Figure 7, the Semantic Anchor suppresses degenerative repetition and shortens generation length, effectively offsetting the update overhead. While Tent and EATA achieve lower latency, this reflects premature termination during model collapse rather than genuine efficiency, as evidenced by their degraded scores in Table 1. Full results are provided in Appendix E.9.

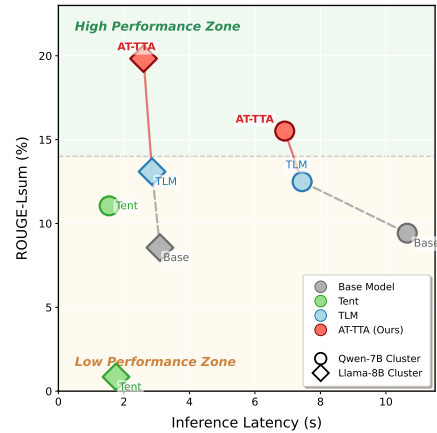


Figure 7: Efficiency–performance trade-off. AT-TTA reduces latency relative to the base model, whereas Tent gains speed only by sacrificing output quality.

## 6 Conclusion

In this paper, we presented AT-TTA, a dual-view framework for few-shot anchored test-time adaptation of LLMs in specialized domains. AT-TTA combines input-side perplexity alignment, output energy calibration, and a Semantic Anchor derived from a small labeled support set and its filtered augmentations, constraining test-time updates through Energy-Guided Adaptive Replay. This design balances plasticity and stability, mitigating confidence traps and model drift in entropy-based adaptation. Experiments on AdaptEval show strong performance across diverse domains, including under matched few-shot initialization, while maintaining competitive inference efficiency. These results demonstrate the effectiveness of few-shot anchored test-time adaptation for reliable LLM deployment under domain shift, suggesting its potential applicability to specialized domains such as finance, medicine, and agriculture.

## 7 Limitations

While AT-TTA achieves strong performance across diverse NLP tasks, two directions remain unexplored. First, we have not investigated extending AT-TTA to computer vision or multimodal settings. The current dual-view uncertainty signals are designed for autoregressive language models, and adapting them to different architectures and output spaces represents a natural avenue for future work. Second, we have not explored integrating AT-TTA with continual learning mechanisms. In scenarios where test distributions shift continuously over time, combining AT-TTA with dynamic adaptation strategies could further broaden its applicability to non-stationary environments.

## Acknowledgments

This work was supported in part by the Yangtze River Delta Science and Technology Innovation Community Joint Research Project under Grant YDZX20233100004031 and in part by the National Natural Science Foundation of China under Grant 72595845, Grant 72595840, and Grant 62502225.

## References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*, 25(70):1–53.
- 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. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. [A survey on RAG meeting LLMs: Towards retrieval-augmented large language models](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.
- Guhao Feng, Shengjie Luo, Kai Hua, Ge Zhang, Wenhao Huang, Di He, and Tianle Cai. 2026. [In-place test-time training](#). In *The Fourteenth International Conference on Learning Representations*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *arXiv preprint arXiv:2312.10997*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Jinwu Hu, Zitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. 2025. [Test-time learning for large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24823–24849. PMLR.
- Jonas Hübner, Sascha Bongni, Ido Hakimi, and Andreas Krause. 2025. [Efficiently learning at test-time: Active fine-tuning of LLMs](#). In *The Thirteenth International Conference on Learning Representations*.
- Matheus Thomas Kuska, Mirwaes Wahabzada, and Stefan Paulus. 2024. [AI for crop production—where can large language models \(LLMs\) provide substantial value?](#) *Computers and Electronics in Agriculture*, 221:108924.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

- Jian Liang, Ran He, and Tieniu Tan. 2025. [A comprehensive survey on test-time adaptation under distribution shifts](#). *International Journal of Computer Vision*, 133(1):31–64.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020a. [LogiQA: A challenge dataset for machine reading comprehension with logical reasoning](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3622–3628.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020b. [Energy-based out-of-distribution detection](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475.
- Llama Team. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. [#InsTag: Instruction tagging for analyzing supervised fine-tuning of large language models](#). In *International Conference on Learning Representations*.
- Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. [Efficient test-time model adaptation without forgetting](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16888–16905. PMLR.
- OpenAI. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620(7972):172–180.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. [Tent: Fully test-time adaptation by entropy minimization](#). In *International Conference on Learning Representations*.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhjanan Kam-badur, David Rosenberg, and Gideon Mann. 2023. [BloombergGPT: A large language model for finance](#). *arXiv preprint arXiv:2303.17564*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chuji Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [MetaMath: Bootstrap your own mathematical questions for large language models](#). In *International Conference on Learning Representations*.
- Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. 2025. [COME: Test-time adaption by conservatively minimizing entropy](#). In *The Thirteenth International Conference on Learning Representations*. OpenReview.net.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. [LlamaFactory: Unified efficient fine-tuning of 100+ language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410. Association for Computational Linguistics.

## A Details of Datasets

To ensure a rigorous evaluation of model adaptability, we use DomainBench and InstructionBench from AdaptEval (Hu et al., 2025). We use the seven 5,000-example subsets released with AdaptEval and apply the hold-out protocol described below. The evaluation framework consists of two primary categories: domain-specific reasoning and general instruction following.

### A.1 DomainBench

DomainBench is designed to assess model performance in specialized verticals that demand high factual precision and expert-level reasoning. It draws on four corpora spanning geography, agriculture, medicine, and finance:

**Geography.** GeoSignal<sup>1</sup> contains approximately 39.7k examples and combines a general instruction-alignment section with a professional Earth-science section. Its domain-specific tasks include named entity recognition, relation inference, fact verification, and question answering over specialized geoscience terminology.

**Agriculture.** Agriculture-QA<sup>2</sup> comprises approximately 22.6k Q&A pairs covering crop cultivation, livestock farming, soil management, pest control, and other practical agricultural topics. It evaluates whether models can transfer domain knowledge to farmer-oriented queries.

**Medicine.** GenMedGPT-5k-en<sup>3</sup> contains approximately 5.45k doctor-patient dialogues generated by ChatGPT. The interactions cover patient queries and corresponding medical responses or diagnoses, providing a testbed for adaptation to clinical dialogue.

**Finance.** Wealth-Alpaca-LoRA<sup>4</sup> contains approximately 44.3k examples combining general instructions from Alpaca, financial data from FiQA, and 1.3k custom examples generated with GPT-3.5. AdaptEval uses a randomly sampled subset to evaluate adaptation to this finance-oriented mixture.

<sup>1</sup><https://huggingface.co/datasets/daven3/geosignal>

<sup>2</sup><https://huggingface.co/datasets/KisanVaani/agriculture-qa-english-only>

<sup>3</sup><https://huggingface.co/datasets/wangrongsheng/GenMedGPT-5k-en>

### A.2 InstructionBench

InstructionBench serves as a comprehensive testbed for general instruction-following capabilities, aggregating datasets with varying degrees of complexity and noise to test robustness and generalization.

**Alpaca-GPT4.** Alpaca-GPT4<sup>5</sup> comprises 52k instruction-response pairs whose responses were generated by GPT-4. It covers diverse tasks such as question answering, explanation, summarization, and procedural guidance.

**Dolly.** Dolly-15k<sup>6</sup> consists of 15k human-authored instruction-response pairs spanning tasks such as brainstorming, question answering, summarization, classification, and information extraction.

**InstructionWild.** InstructionWild contains more than 50k bilingual Chinese-English instructions derived from real-world user-shared scenarios. Following AdaptEval, we use its released 5,000-example subset to evaluate robustness to diverse instruction-following tasks.

## B Evaluation Metrics

We evaluate our method with three families of metrics: lexical overlap, semantic similarity, and confidence calibration. Following established summarization evaluation practice, we use ROUGE-Lsum (Lin, 2004) as the primary lexical-overlap metric.

**Preprocessing.** We normalize non-standard Unicode characters, segment text into sentences using NLTK, and join sentences with newline separators. All ROUGE scores are computed with the rouge-score library with stemming enabled.

**ROUGE- $N$  (ROUGE-1/2).** Let  $G_n(x)$  be the multiset of  $n$ -grams extracted from text  $x$ , and let  $c_x(g)$  denote the count of  $n$ -gram  $g$  in  $x$ . We define the matched  $n$ -grams as

$$\text{match}_n(\hat{y}, y) = \sum_{g \in G_n(y)} \min(c_{\hat{y}}(g), c_y(g)). \quad (9)$$

Then ROUGE- $N$  recall, precision, and  $F_1$  are

$$R_n = \frac{\text{match}_n(\hat{y}, y)}{\sum_{g \in G_n(y)} c_y(g)}, \quad P_n = \frac{\text{match}_n(\hat{y}, y)}{\sum_{g \in G_n(\hat{y})} c_{\hat{y}}(g)}, \quad (10)$$

<sup>4</sup>[https://huggingface.co/datasets/gbharti/wealth-alpaca\\_lora](https://huggingface.co/datasets/gbharti/wealth-alpaca_lora)

<sup>5</sup>[https://huggingface.co/datasets/llamafactory/alpaca\\_gpt4\\_en](https://huggingface.co/datasets/llamafactory/alpaca_gpt4_en)

<sup>6</sup><https://huggingface.co/datasets/databricks/databricks-dolly-15k>

| Benchmark               | Dataset         | Raw Corpus | Anchor Size ( $ \mathcal{S}_{\text{anchor}} $ ) | Eval. Size ( $N_{\text{test}}$ ) |
|-------------------------|-----------------|------------|-------------------------------------------------|----------------------------------|
| <b>DomainBench</b>      | Agriculture     | 22.6k      | 64                                              | 4,900                            |
|                         | Geography       | 39.7k      | 64                                              | 4,900                            |
|                         | Medicine        | 5.5k       | 64                                              | 4,900                            |
|                         | Finance         | 44.3k      | 80                                              | 4,900                            |
| <b>InstructionBench</b> | Dolly           | 15.0k      | 75                                              | 4,900                            |
|                         | Alpaca-GPT4     | 52.0k      | 78                                              | 4,900                            |
|                         | InstructionWild | >50k       | 79                                              | 4,900                            |

Table 6: Detailed statistics of the datasets in AdaptEval. For each dataset, we randomly sample a subset of 5,000 instances. To ensure **zero data leakage**, we adopt a strict hold-out protocol: 100 instances are reserved as a candidate pool for sampling the **initial labeled support set** ( $K = 16$ ), while the remaining 4,900 instances form the unified Evaluation Set. **Note on Anchor Size:** The reported Anchor Size denotes the total count after data augmentation (i.e.,  $|\mathcal{S}_{\text{anchor}}| = |\mathcal{S}| + |\mathcal{S}_{\text{aug}}|$ ). While the initial seed size is fixed at  $K = 16$  for all datasets, the final anchor size varies (e.g., 80 for Finance vs. 64 for Agriculture) due to domain-specific valid generation yields and filtering criteria applied during the augmentation process.

$$F_n = \frac{2P_n R_n}{P_n + R_n}. \quad (11)$$

ROUGE-1 uses  $n = 1$  and ROUGE-2 uses  $n = 2$ .

**ROUGE-L.** Let  $L = \text{LCS}(\hat{y}, y)$  be the length of the longest common subsequence between the token sequences of  $\hat{y}$  and  $y$ . Let  $m = |\hat{y}|$  and  $n = |y|$  be sequence lengths. We compute

$$R_L = \frac{L}{n}, \quad P_L = \frac{L}{m}, \quad F_L = \frac{2P_L R_L}{P_L + R_L}. \quad (12)$$

**ROUGE-Lsum (Main Metric).** We split  $\hat{y}$  and  $y$  into sentences and apply newline separators, as required by the standard ROUGE-Lsum setup. Let  $\{y_i\}_{i=1}^S$  and  $\{\hat{y}_j\}_{j=1}^T$  be the reference and candidate sentence sets. We use a sentence-level LCS aggregation

$$L_{\text{sum}} = \sum_{i=1}^S \max_{1 \leq j \leq T} \text{LCS}(y_i, \hat{y}_j), \quad (13)$$

and define

$$R_{\text{sum}} = \frac{L_{\text{sum}}}{|y|}, \quad P_{\text{sum}} = \frac{L_{\text{sum}}}{|\hat{y}|}, \quad F_{\text{sum}} = \frac{2P_{\text{sum}} R_{\text{sum}}}{P_{\text{sum}} + R_{\text{sum}}}. \quad (14)$$

**BLEU.** We report sentence-level BLEU (Papineni et al., 2002) with modified  $n$ -gram precisions  $p_k$  and brevity penalty BP:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{k=1}^K w_k \log p_k\right), \quad (15)$$

$$\text{BP} = \begin{cases} 1, & m > n, \\ \exp(1 - n/m), & m \leq n, \end{cases} \quad (16)$$

where  $m = |\hat{y}|$  and  $n = |y|$ . We use NLTK BLEU with smoothing method 4 to avoid zero scores on short sequences.

**BERTScore.** We assess semantic similarity using BERTScore (Zhang et al., 2020) with the bert-base-multilingual-cased backbone. To standardize evaluation, both predictions and references are capped at 500 tokens and aligned to the same length prior to scoring.

### B.1 Calibration Metrics

We use calibration metrics to distinguish confident-but-incorrect generation from complete model collapse. For the  $i$ -th generated response, let  $p_{i,t}$  denote the predictive token distribution at generation step  $t$ , and let  $T_i$  be the number of tokens in the evaluated generation window. We first compute normalized predictive entropy as

$$\tilde{H}(p_{i,t}) = -\frac{1}{\log |\mathcal{V}|} \sum_{v \in \mathcal{V}} p_{i,t}(v) \log p_{i,t}(v), \quad (17)$$

where  $\mathcal{V}$  is the model vocabulary. The confidence assigned to the complete response is then

$$c_i = 1 - \frac{1}{T_i} \sum_{t=1}^{T_i} \tilde{H}(p_{i,t}). \quad (18)$$

Thus,  $c_i \in [0, 1]$ , with larger values indicating lower average predictive entropy. This confidence definition is kept separate from the sequence-energy signal used by AT-TTA.

We derive a binary reference-based correctness label from the GPT-4o judge score:  $a_i = 1$  when the reference-answer correctness score is at least 7, and  $a_i = 0$  otherwise. For Expected Calibration Error (ECE) (Guo et al., 2017), confidence scores are partitioned into  $M = 10$  equal-width

bins  $\{B_m\}_{m=1}^M$ , with empty bins omitted. We compute

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (19)$$

where  $\text{acc}(B_m) = |B_m|^{-1} \sum_{i \in B_m} a_i$  and  $\text{conf}(B_m) = |B_m|^{-1} \sum_{i \in B_m} c_i$ .

High-Confidence Error Rate (HCER) measures the error rate among the top-confidence quartile. Let  $q_{0.75}$  be the 75th percentile of the confidence scores and  $Q_{0.75} = \{i : c_i \geq q_{0.75}\}$ . We define

$$\text{HCER} = 1 - \frac{1}{|Q_{0.75}|} \sum_{i \in Q_{0.75}} a_i. \quad (20)$$

Lower ECE and HCER indicate better alignment between predictive confidence and reference-based correctness.

## C Baselines

To ensure a fair and rigorous comparison, we align the standard unsupervised configurations of the baseline methods with established protocols in recent LLM test-time adaptation literature (Hu et al., 2025); for matched-support comparisons, we initialize them from the same  $K = 16$  SFT checkpoint used by AT-TTA.

**Tent.** Originally designed for discriminative vision tasks, Tent adapts the model by minimizing the Shannon entropy of model predictions (Wang et al., 2021). For autoregressive LLMs, we calculate the entropy over the generated token sequence. Specifically, following the TLM protocol (Hu et al., 2025), we perform a forward pass to generate a sequence and minimize the average entropy of the first 80 generated tokens to update the model parameters via the AdamW optimizer.

**EATA.** EATA improves upon Tent by introducing a sample-efficient filtering mechanism and Fisher Information Matrix (FIM) regularization to prevent catastrophic forgetting (Niu et al., 2022). In our experiments, we adapt EATA to the generative setting using the same token-level entropy calculation as Tent. We set the entropy threshold margin  $E_0$  to 0.4 to filter out high-uncertainty samples. Regarding the optimization scope, we compute gradients over the first 80 generated tokens for standard models (e.g., 7B/8B). For larger models where memory is a bottleneck (e.g., Qwen2.5-14B), we reduce

this window to the first 30 tokens to avoid Out-of-Memory (OOM) errors, consistent with prior adaptation practices.

**COME.** COME replaces conventional entropy minimization with an uncertainty-aware conservative objective that avoids overconfidence on unreliable samples (Zhang et al., 2025). We adapt COME to autoregressive generation using the same output-token window, LoRA parameter scope, and initialization protocol as the other output-centric baselines.

**TLM.** TLM adapts models by minimizing the perplexity of unlabeled input queries (Hu et al., 2025). We adopt the official implementation, strictly adhering to its original perplexity-based sample selection strategy and learning rate configurations to ensure a rigorous comparison against this input-centric baseline.

## D Implementation Details

### D.1 Computational Setup

All experiments were conducted on a server equipped with eight NVIDIA A100 GPUs (80GB VRAM) running CUDA 12.1. Our implementation is based on PyTorch 2.5.1 integrated with the LLaMA-Factory library (Zheng et al., 2024). LoRA is applied to the query ( $W_q$ ) and value ( $W_v$ ) projection matrices with rank  $r = 8$ , following standard initialization (Hu et al., 2022). For inference, we use greedy decoding (temperature = 0) with a maximum of 512 new tokens per response.

**Stage 1 (Anchor Initialization)** fine-tunes the model on  $\mathcal{S}_{\text{anchor}}$  for 3 epochs using the AdamW optimizer with learning rate  $2 \times 10^{-5}$ , cosine learning rate scheduler, warmup ratio 0.1, and batch size 1.

**Stage 2 (Synergistic Adaptation)** performs one-pass online adaptation (1 epoch) over the test stream with learning rate  $5 \times 10^{-5}$ , cosine scheduler, warmup ratio 0.1, and batch size 1. Key method-specific hyperparameters are: output energy weight  $\beta = 0.05$ , base replay weight  $\gamma_{\text{base}} = 1.0$ , perplexity gating threshold  $\tau = 3$ , and input loss scaling factor  $\lambda = 0.1$ . Both stages use bf16 mixed-precision training.

### D.2 Support Set Selection

To construct the Semantic Anchor  $\mathcal{S}$  without compromising the integrity of the test evaluation, we adopt a strict Split-and-Hold-Out protocol. For

each domain dataset, we partition the raw corpus into two mutually exclusive subsets: a Reserved Candidate Pool and the Final Evaluation Stream  $\mathcal{D}_{test}$ .

Specifically, we randomly sample  $K$  labeled instances (default  $K = 16$ ) from the Reserved Candidate Pool to form the support set  $\mathcal{S}$ . These instances are treated as available few-shot exemplars, simulating a deployment scenario in which limited target-domain supervision is accessible. Crucially, these samples are strictly isolated from the evaluation stream. The final performance of all methods (including baselines) is reported on the identical  $\mathcal{D}_{test}$ , ensuring zero data leakage and a fair comparison.

### D.3 Augmentation Strategies and Prompts

To simulate diverse user behaviors and query complexities, we design an augmentation prompt that instructs the model to rewrite queries based on the original instruction and the golden answer. We define four distinct augmentation types (requirement) used in the prompt above:

- **Rephrasing:** Rephrase the original query with the same intention. This improves lexical diversity.
- **Complexity:** Increase the semantic complexity of the original query. This simulates expert-level or academic user queries.
- **Decomposition:** Decompose the original query into several sub-problems. This helps the model learn to handle multi-hop reasoning.
- **Constraint:** Add more conditional and constrained statements to the original query. This tests the model’s ability to follow complex instructions.

The exact prompt template used for this process is illustrated in Figure 8. The model is explicitly instructed to output only the rewritten query without additional conversational text.

To demonstrate the robustness and versatility of our augmentation pipeline across diverse domains, we present concrete examples from seven distinct datasets covering specialized vertical domains and general instruction following. Specific instances are detailed in the following tables:

- **Vertical Domains:** Geography (Table 21), Agriculture (Table 22), Medicine (Table 23), and Finance (Table 24).

#### Query Augmentation Prompt

You are an AI assistant helping me rewrite the query. I will give you the original query, input, answer and rewriting requirements. Please rewrite the query based on the following information. You only need to output the rewritten query without any additional text:

**Original Query:** {Instruction}

**Answer:** {Answer}

**Augmentation Requirements:**

{Augmented Requirements}

**New Queries:**

Figure 8: The prompt template used for data augmentation. The placeholders are filled with the specific instance data and the selected augmentation strategy requirement.

- **General Instructions:** Alpaca-GPT4 (Table 25), Dolly (Table 26), and InstructionWild (Table 27).

These examples illustrate how the model adapts query generation based on the specific ground-truth content and the constraints imposed by different augmentation strategies.

### D.4 Data Quality Filtering

To ensure that the augmented variations maintain semantic consistency with the original user intent, we implement a quality control mechanism based on Natural Language Inference (NLI).

We employ a pre-trained NLI model to evaluate the logical relationship within the query pair. Specifically, we designate the original query  $x$  as the *premise* and the augmented query  $x_{aug}$  as the *hypothesis*. The model accepts this pair as input and predicts a probability distribution over the label set  $\mathcal{C} = \{\text{ENTAILMENT}, \text{NEUTRAL}, \text{CONTRADICTION}\}$ :

$$P_{\phi}(c \mid x, x_{aug}) = \text{Softmax}(f_{\phi}(x, x_{aug})), \quad (21)$$

where  $f_{\phi}$  denotes the logits produced by the NLI encoder.

We determine the dominant semantic relationship  $\hat{c}$  by selecting the class with the maximum probability:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} P_{\phi}(c \mid x, x_{aug}). \quad (22)$$

Table 7: Comparison of experimental results on the Geography dataset of DomainBench.

| Method               | BERTScore $\uparrow$ | BLEU $\uparrow$ | Rouge-1 $\uparrow$ | Rouge-2 $\uparrow$ | Rouge-L $\uparrow$ |
|----------------------|----------------------|-----------------|--------------------|--------------------|--------------------|
| Qwen2.5-7B           | 0.6968               | 0.0766          | 0.2780             | 0.1034             | 0.2006             |
| Qwen2.5-7B + SFT     | 0.7019               | 0.0832          | 0.2892             | 0.1120             | 0.2106             |
| Tent                 | 0.7012               | 0.0815          | 0.2843             | 0.1102             | 0.2115             |
| EATA                 | 0.6984               | 0.0792          | 0.2611             | 0.1067             | 0.2054             |
| TLM                  | 0.7131               | 0.0998          | 0.3238             | 0.1354             | 0.2453             |
| <b>AT-TTA (Ours)</b> | <b>0.7325</b>        | <b>0.1218</b>   | <b>0.3742</b>      | <b>0.1711</b>      | <b>0.2957</b>      |
| Qwen2.5-14B          | 0.6999               | 0.0827          | 0.2909             | 0.1082             | 0.2092             |
| Qwen2.5-14B + SFT    | 0.7184               | 0.1096          | 0.3386             | 0.1395             | 0.2541             |
| Tent                 | 0.7085               | 0.0914          | 0.3056             | 0.1213             | 0.2241             |
| EATA                 | 0.7122               | 0.0945          | 0.3128             | 0.1284             | 0.2310             |
| TLM                  | 0.7290               | <b>0.1191</b>   | 0.3664             | 0.1602             | 0.2834             |
| <b>AT-TTA (Ours)</b> | <b>0.7341</b>        | 0.1185          | <b>0.3678</b>      | <b>0.1713</b>      | <b>0.2963</b>      |
| Llama-3-8B           | 0.6952               | 0.0717          | 0.2743             | 0.1041             | 0.1991             |
| Llama-3-8B + SFT     | 0.6993               | 0.0773          | 0.2892             | 0.1137             | 0.2142             |
| Tent                 | 0.6225               | 0.0512          | 0.1104             | 0.0485             | 0.1523             |
| EATA                 | 0.6312               | 0.0684          | 0.2621             | 0.0912             | 0.1847             |
| TLM                  | <b>0.7901</b>        | <b>0.1074</b>   | <b>0.3592</b>      | <b>0.1673</b>      | <b>0.2875</b>      |
| <b>AT-TTA (Ours)</b> | 0.7265               | 0.1022          | 0.3493             | 0.1648             | 0.2843             |

Table 8: Comparison of experimental results on the Agriculture dataset of DomainBench.

| Method               | BERTScore $\uparrow$ | BLEU $\uparrow$ | Rouge-1 $\uparrow$ | Rouge-2 $\uparrow$ | Rouge-L $\uparrow$ |
|----------------------|----------------------|-----------------|--------------------|--------------------|--------------------|
| Qwen2.5-7B           | 0.6560               | 0.0124          | 0.1016             | 0.0366             | 0.0732             |
| Qwen2.5-7B + SFT     | 0.6542               | 0.0132          | 0.1019             | 0.0371             | 0.0738             |
| Tent                 | 0.6582               | 0.0141          | 0.1145             | 0.0402             | 0.0812             |
| EATA                 | 0.6515               | 0.0133          | 0.1210             | 0.0435             | 0.0864             |
| TLM                  | 0.6524               | 0.0189          | 0.1372             | 0.0504             | 0.0998             |
| <b>AT-TTA (Ours)</b> | <b>0.6657</b>        | <b>0.0258</b>   | <b>0.1649</b>      | <b>0.0620</b>      | <b>0.1262</b>      |
| Qwen2.5-14B          | 0.6530               | 0.0146          | 0.1162             | 0.0412             | 0.0833             |
| Qwen2.5-14B + SFT    | 0.6715               | 0.0235          | 0.1477             | 0.0547             | 0.1091             |
| Tent                 | 0.6642               | 0.0185          | 0.1321             | 0.0498             | 0.0967             |
| EATA                 | 0.6691               | 0.0204          | 0.1415             | 0.0542             | 0.1023             |
| TLM                  | 0.6554               | 0.0213          | 0.1481             | 0.0556             | 0.1084             |
| <b>AT-TTA (Ours)</b> | <b>0.7044</b>        | <b>0.0436</b>   | <b>0.2485</b>      | <b>0.0883</b>      | <b>0.1952</b>      |
| Llama-3-8B           | 0.6667               | 0.0108          | 0.0927             | 0.0335             | 0.0680             |
| Llama-3-8B + SFT     | 0.6649               | 0.0133          | 0.1065             | 0.0381             | 0.0805             |
| Tent                 | 0.5614               | 0.0084          | 0.0102             | 0.0041             | 0.0104             |
| EATA                 | 0.5987               | 0.0072          | 0.0048             | 0.0002             | 0.0062             |
| TLM                  | 0.6547               | 0.0190          | 0.1462             | 0.0509             | 0.1106             |
| <b>AT-TTA (Ours)</b> | <b>0.6958</b>        | <b>0.0342</b>   | <b>0.2289</b>      | <b>0.0770</b>      | <b>0.1831</b>      |

To filter out noise while preserving diverse reasoning paths, we define the valid augmented set  $\mathcal{S}_{valid}$  by retaining only those samples that do not semantically contradict the source:

$$\mathcal{S}_{valid} = \{x_{aug} \in \mathcal{S}_{aug} \mid \hat{c} \neq \text{CONTRADICTION}\}. \quad (23)$$

This filtering strategy effectively removes rewritten queries that distort the core instruction, ensuring that the anchor remains reliable.

## D.5 Reference-Based Evaluation Prompt

To complement lexical metrics, we employ GPT-4o as an automatic judge to assess model outputs on two dimensions, each scored on a 1–10 integer

scale. The judge receives the original question, a reference answer, and the model’s prediction, and is instructed to score each dimension independently without referencing the other. The original prompt names the first dimension *Factual Correctness*; because the evaluator is given a reference answer rather than independent external evidence, we report this score as *Reference-Answer Correctness* in our analyses. The complete prompt template is shown in Figure 9.

Table 9: Comparison of experimental results on the Medicine dataset of DomainBench.

| Method               | BERTScore $\uparrow$ | BLEU $\uparrow$ | Rouge-1 $\uparrow$ | Rouge-2 $\uparrow$ | Rouge-L $\uparrow$ |
|----------------------|----------------------|-----------------|--------------------|--------------------|--------------------|
| Qwen2.5-7B           | 0.6548               | 0.0123          | 0.1363             | 0.0351             | 0.0856             |
| Qwen2.5-7B + SFT     | 0.7052               | 0.0584          | 0.2517             | 0.0957             | 0.1874             |
| Tent                 | 0.6221               | 0.0106          | 0.0645             | 0.0212             | 0.0624             |
| EATA                 | 0.6454               | 0.0112          | 0.1741             | 0.0445             | 0.1352             |
| TLM                  | 0.7187               | 0.0660          | 0.2778             | 0.1053             | 0.2077             |
| <b>AT-TTA (Ours)</b> | <b>0.7565</b>        | <b>0.1172</b>   | <b>0.3523</b>      | <b>0.1784</b>      | <b>0.3000</b>      |
| Qwen2.5-14B          | 0.6601               | 0.0148          | 0.1599             | 0.0985             | 0.1443             |
| Qwen2.5-14B + SFT    | 0.7351               | 0.0978          | 0.3160             | 0.1457             | 0.2612             |
| Tent                 | 0.6845               | 0.0167          | 0.2145             | 0.0721             | 0.1856             |
| EATA                 | 0.7114               | 0.0184          | 0.2356             | 0.0824             | 0.1951             |
| TLM                  | 0.7491               | 0.1056          | <b>0.3385</b>      | 0.1588             | 0.2824             |
| <b>AT-TTA (Ours)</b> | <b>0.7637</b>        | <b>0.1079</b>   | 0.3375             | <b>0.1833</b>      | <b>0.3056</b>      |
| Llama-3-8B           | 0.6627               | 0.0135          | 0.1397             | 0.0343             | 0.0910             |
| Llama-3-8B + SFT     | 0.7117               | 0.0647          | 0.2469             | 0.1071             | 0.2104             |
| Tent                 | 0.5742               | 0.0021          | 0.0154             | 0.0002             | 0.0145             |
| EATA                 | 0.5985               | 0.0035          | 0.0199             | 0.0008             | 0.0154             |
| TLM                  | 0.7036               | 0.0440          | 0.2565             | 0.0774             | 0.1801             |
| <b>AT-TTA (Ours)</b> | <b>0.7310</b>        | <b>0.0684</b>   | <b>0.2917</b>      | <b>0.1012</b>      | <b>0.2173</b>      |

Table 10: Comparison of experimental results on the Finance dataset of DomainBench.

| Method               | BERTScore $\uparrow$ | BLEU $\uparrow$ | Rouge-1 $\uparrow$ | Rouge-2 $\uparrow$ | Rouge-L $\uparrow$ |
|----------------------|----------------------|-----------------|--------------------|--------------------|--------------------|
| Qwen2.5-7B           | 0.6946               | 0.0568          | 0.2825             | 0.1047             | 0.1895             |
| Qwen2.5-7B + SFT     | 0.7063               | 0.0776          | 0.3198             | 0.1255             | 0.2242             |
| Tent                 | 0.6841               | 0.0424          | 0.2124             | 0.1045             | 0.1656             |
| EATA                 | 0.6912               | 0.0645          | 0.3312             | 0.1384             | 0.2341             |
| TLM                  | 0.7211               | 0.0987          | 0.3593             | 0.1505             | 0.2592             |
| <b>AT-TTA (Ours)</b> | <b>0.7310</b>        | <b>0.1122</b>   | <b>0.3785</b>      | <b>0.1624</b>      | <b>0.2855</b>      |
| Qwen2.5-14B          | 0.6940               | 0.0554          | 0.2864             | 0.1029             | 0.1893             |
| Qwen2.5-14B + SFT    | 0.7218               | 0.0966          | 0.3464             | 0.1440             | 0.2530             |
| Tent                 | 0.7102               | 0.0812          | 0.3241             | 0.1424             | 0.2312             |
| EATA                 | 0.7185               | 0.0724          | 0.3456             | 0.1412             | 0.2456             |
| TLM                  | 0.7256               | 0.1037          | 0.3718             | 0.1575             | 0.2683             |
| <b>AT-TTA (Ours)</b> | <b>0.7375</b>        | <b>0.1169</b>   | <b>0.3748</b>      | <b>0.1682</b>      | <b>0.2915</b>      |
| Llama-3-8B           | 0.6860               | 0.0409          | 0.2532             | 0.0874             | 0.1685             |
| Llama-3-8B + SFT     | 0.6901               | 0.0478          | 0.2621             | 0.0941             | 0.1787             |
| Tent                 | 0.6187               | 0.0092          | 0.0545             | 0.0056             | 0.0112             |
| EATA                 | 0.6552               | 0.0345          | 0.1312             | 0.0454             | 0.1196             |
| TLM                  | 0.7159               | 0.0939          | 0.3499             | 0.1445             | 0.2577             |
| <b>AT-TTA (Ours)</b> | <b>0.7195</b>        | <b>0.0991</b>   | <b>0.3517</b>      | <b>0.1498</b>      | <b>0.2652</b>      |

## E Additional Experimental Results

### E.1 Comparison with COME

We additionally compare AT-TTA with COME (Zhang et al., 2025), an uncertainty-aware entropy-minimization method designed to reduce overconfidence during test-time adaptation. Table 11 reports average ROUGE-Lsum scores on DomainBench and InstructionBench.

AT-TTA achieves higher average generation performance than the standard unsupervised COME configuration across both benchmarks and all three backbones. Because AT-TTA uses labeled support, this table is an end-to-end comparison across adaptation settings; matched-support results are re-

ported separately in Table 2. The comparison here is restricted to the generation benchmarks, while reasoning results are reported separately.

### E.2 Energy-Quantile Analysis

To examine the direction of the adaptive replay mechanism, we sort the AT-TTA test outputs on Medicine with Qwen2.5-7B by their length-normalized sequence energy and partition them into five equal-sized groups. Q1 and Q5 contain the lowest- and highest-energy outputs, respectively. Following the uncertainty analysis in Section 5.6, an output is treated as incorrect when its GPT-4o reference-answer correctness score is below 7. For each quintile, we report the error rate and the mean

| LLM-as-a-Judge Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |  |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
| <p>You are an expert evaluator assessing the quality of a model-generated response. You will be given a question, a reference answer, and a model response. Please evaluate the model response on the following two dimensions <b>independently</b>, using an integer score from <b>1</b> (very poor) to <b>10</b> (excellent).</p> <p><b>Factual Correctness (1–10):</b> Does the response contain accurate, verifiable facts consistent with the reference answer? Penalize hallucinations, factual errors, or unsupported claims.</p> <p><b>Logical Reasoning (1–10):</b> Is the reasoning process coherent and well-structured? Does the response arrive at conclusions through valid inferential steps?</p> <p><b>Question:</b> {Instruction}<br/> <b>Reference Answer:</b> {Reference}<br/> <b>Model Response:</b> {Prediction}</p> <p>Output your evaluation in the following format exactly, with no additional text:<br/> Factual Correctness: &lt;score&gt;<br/> Logical Reasoning: &lt;score&gt;</p> |  |

Figure 9: The original prompt template used for reference-based evaluation via GPT-4o. Placeholders {Instruction}, {Reference}, and {Prediction} are filled with the question, reference answer, and model-generated response respectively.

realized replay weight normalized by  $\gamma_{\text{base}}$ .

As shown in Table 12, the error rate increases monotonically from 0.08 in Q1 to 0.63 in Q5. The mean normalized replay weight follows the same trend, increasing from 1.05 in Q1 to 1.88 in Q5. Thus, the energy-dependent schedule strengthens anchor supervision in the empirically more error-prone high-energy region. Since  $\gamma(x)$  remains lower-bounded by  $\gamma_{\text{base}}$ , low-energy outputs continue to receive baseline anchor regularization.

### E.3 ReasoningBench Results

We evaluate reasoning retention on GSM8K, MetaMath, and LogiQA using the exact-match (EM) protocol adopted by TLM (Hu et al., 2025). The baseline results for the base models, Tent, EATA, COME, and TLM are taken from their reported evaluation, while AT-TTA is evaluated under the same protocol.

On Qwen2.5-7B, the strongest method varies across tasks: COME performs best on GSM8K, TLM on MetaMath, and AT-TTA on LogiQA. On Llama-3-8B, AT-TTA is close to TLM and obtains the highest score on all three tasks. More broadly, AT-TTA remains above the corresponding base model across the evaluated tasks, indicating

| Model       | Method        | DomainBench  | InstructionBench |
|-------------|---------------|--------------|------------------|
| Qwen2.5-7B  | COME          | 13.50        | 25.80            |
|             | TLM           | 25.11        | 37.73            |
|             | <b>AT-TTA</b> | <b>29.47</b> | <b>39.74</b>     |
| Qwen2.5-14B | COME          | 14.90        | 26.20            |
|             | TLM           | 28.05        | 37.96            |
|             | <b>AT-TTA</b> | <b>30.46</b> | <b>41.28</b>     |
| Llama-3-8B  | COME          | 1.70         | 7.50             |
|             | TLM           | 25.20        | 35.98            |
|             | <b>AT-TTA</b> | <b>27.54</b> | <b>39.21</b>     |

Table 11: Average ROUGE-Lsum performance on DomainBench and InstructionBench with COME as an additional baseline. Scores are multiplied by 100.

|                                       | Q1   | Q2   | Q3   | Q4   | Q5   |
|---------------------------------------|------|------|------|------|------|
| Error rate                            | 0.08 | 0.17 | 0.29 | 0.44 | 0.63 |
| Mean $\gamma(x)/\gamma_{\text{base}}$ | 1.05 | 1.18 | 1.42 | 1.68 | 1.88 |

Table 12: Energy-quantile analysis on Medicine using Qwen2.5-7B. Q1 and Q5 denote the lowest- and highest-energy groups, respectively.

that its domain adaptation gains do not come at the expense of the reasoning capabilities measured by ReasoningBench.

### E.4 Calibration and High-Confidence Errors

We evaluate confidence calibration on Medicine and Finance using Qwen2.5-7B and Llama-3-8B. Confidence is computed from normalized token-level predictive entropy, while binary correctness is derived from the GPT-4o reference-answer correctness score as defined in Appendix B.1. Sequence energy is not used as the confidence score in either metric.

As shown in Table 14, AT-TTA obtains the lowest ECE and HCER in all four model–dataset settings. Tent and EATA exhibit substantially higher HCER, indicating that a large fraction of their most confident outputs remain incorrect even when the model does not fully collapse. AT-TTA also improves over TLM on both metrics, suggesting closer alignment between token-level confidence and reference-based correctness. Since correctness labels are derived from a reference-based judge, these metrics evaluate calibration with respect to the provided references rather than independent external fact verification.

### E.5 Cross-Family and Human Judge Validation

To assess whether the primary result depends on GPT-4o as the evaluator, we re-evaluate the AT-

| Method        | Qwen2.5-7B    |               |               | Llama-3-8B    |               |               |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|               | GSM8K         | MetaMath      | LogiQA        | GSM8K         | MetaMath      | LogiQA        |
| Base          | 0.8378        | 0.7430        | 0.5952        | 0.7610        | 0.6912        | 0.4550        |
| Tent          | 0.8455        | 0.7412        | 0.5934        | 0.7578        | 0.6550        | 0.4378        |
| EATA          | 0.7098        | 0.0070        | 0.2172        | 0.0250        | 0.5454        | 0.2192        |
| COME          | <b>0.8556</b> | 0.7559        | 0.5908        | 0.7479        | 0.6460        | 0.2180        |
| TLM           | 0.8424        | <b>0.7560</b> | <u>0.6046</u> | <u>0.8074</u> | <u>0.7006</u> | <u>0.4868</u> |
| AT-TTA (Ours) | <u>0.8480</u> | 0.7550        | <b>0.6070</b> | <b>0.8100</b> | <b>0.7020</b> | <b>0.4890</b> |

Table 13: Exact-match accuracy on ReasoningBench. Results for the base models, Tent, EATA, COME, and TLM are taken from the TLM evaluation (Hu et al., 2025); AT-TTA is evaluated under the same protocol. **Bold** and underlined values denote the best and second-best results, respectively.

| Model and Dataset     | Metric | Base | Tent | EATA | TLM         | AT-TTA      |
|-----------------------|--------|------|------|------|-------------|-------------|
| Qwen2.5-7B · Medicine | ECE ↓  | 0.16 | 0.55 | 0.42 | <u>0.13</u> | <b>0.07</b> |
|                       | HCER ↓ | 0.45 | 0.88 | 0.72 | <u>0.28</u> | <b>0.16</b> |
| Qwen2.5-7B · Finance  | ECE ↓  | 0.12 | 0.48 | 0.30 | <u>0.11</u> | <b>0.06</b> |
|                       | HCER ↓ | 0.34 | 0.74 | 0.55 | <u>0.26</u> | <b>0.14</b> |
| Llama-3-8B · Medicine | ECE ↓  | 0.18 | 0.62 | 0.58 | <u>0.15</u> | <b>0.08</b> |
|                       | HCER ↓ | 0.48 | 0.92 | 0.88 | <u>0.30</u> | <b>0.18</b> |
| Llama-3-8B · Finance  | ECE ↓  | 0.14 | 0.60 | 0.45 | <u>0.12</u> | <b>0.07</b> |
|                       | HCER ↓ | 0.38 | 0.90 | 0.72 | <u>0.25</u> | <b>0.15</b> |

Table 14: Expected Calibration Error (ECE; 10 equal-width bins) and High-Confidence Error Rate (HCER; top-confidence quartile) on Medicine and Finance. Correctness is determined by a GPT-4o reference-answer correctness score threshold of 7. Lower values are better. **Bold** and underlined values denote the best and second-best results, respectively.

| Evaluator       | AT-TTA     | TLM |
|-----------------|------------|-----|
| GPT-4o          | <b>8.5</b> | 7.7 |
| Claude Opus 4.7 | <b>8.1</b> | 7.2 |
| Qwen-Max        | <b>8.3</b> | 7.5 |
| Human           | <b>8.0</b> | 7.1 |

Table 15: Reference-Answer Correctness (1–10) on Medicine with Qwen2.5-7B across automatic and human evaluators. Higher values are better.

TTA and TLM outputs using Claude Opus 4.7 and Qwen-Max. Both automatic evaluators receive the same question, reference answer, and model response, and apply the 1–10 Reference-Answer Correctness rubric in Appendix D.5. We additionally conduct a human evaluation on a subset of the outputs using the same rubric. Table 15 reports the Medicine results for Qwen2.5-7B; human inter-annotator agreement is  $\kappa = 0.64$ .

AT-TTA ranks above TLM under every evaluator, although the absolute scores vary. This consistency indicates that the comparative result in this setting is not specific to the original GPT-4o judge.

## E.6 Comprehensive Evaluation on Domain Benchmarks

We extend the analysis of Section 5.2 by detailing the performance on specific subsets of DomainBench: Geography, Agriculture, Medicine, and Finance. As shown in Tables 7–10, **AT-TTA** achieves strong results across model families and domains. The matched few-shot comparisons with the SFT variants indicate that these gains cannot be explained solely by exposure to the labeled support set. Specific observations for each domain are as follows:

(1) The proposed method demonstrates superior linguistic quality, particularly on the Qwen series. As shown in Table 7, AT-TTA achieves a 5.12% improvement in BERTScore compared to Qwen2.5-7B. We note that while the baseline TLM achieves higher metrics specifically on Llama-3-8B in this domain, AT-TTA maintains a robust advantage on Qwen models (7B/14B) and consistently outperforms both the original base model and the SFT baseline across all backbones, highlighting its generalization stability.

(2) As detailed in Table 8, AT-TTA achieves substantial gains across all metrics. For instance,

compared to the Qwen2.5-7B base model, AT-TTA achieves a 108.06% improvement in the BLEU metric. It also surpasses the few-shot SFT variant, indicating that the test-time mechanism captures specialized terminology beyond initialization alone. AT-TTA further obtains a higher ROUGE-1 score than the standard unsupervised TLM configuration.

(3) AT-TTA demonstrates exceptional adaptation capabilities in specialized professional domains. As reported in Table 9, compared to the Qwen2.5-7B base model, it achieves a relative improvement of 15.53% in BERTScore. Notably, AT-TTA also consistently outperforms both the strong TLM baseline and the SFT variant across both BERTScore and Rouge-L metrics.

(4) On the Finance dataset (Table 10), AT-TTA exhibits remarkable stability and precision. It achieves a relative improvement of 97.54% in the BLEU metric compared to Qwen2.5-7B. Even compared to the SFT baseline, AT-TTA yields a substantial margin, highlighting the effectiveness of the anchor-tethered mechanism in handling complex financial terminology and reducing hallucinations.

## E.7 Stability across Random Seeds

To further examine the robustness of AT-TTA under stochastic variations, we conduct a targeted multi-seed evaluation on representative domain-specific datasets. Specifically, we evaluate AT-TTA on Medicine and Agriculture using Llama-3-8B and Qwen2.5-7B, and repeat each experiment with five different random seeds. The randomness comes from support-set sampling, LoRA initialization, and data ordering. Following prior LLM-TTA protocols, we report this targeted stability analysis rather than repeating all model–dataset combinations in the main experiments.

As shown in Table 16, AT-TTA exhibits small standard deviations across different models and datasets. The multi-seed means remain close to the corresponding single-run scores reported in Table 1, with absolute differences no larger than 0.07. These results indicate that the observed gains are not caused by favorable random seeds, but remain stable under different support-set samples, LoRA initializations, and data orders.

## E.8 Impact of Different Augmentation Models

In the main text (Section 5.9), we primarily utilize GPT-4 to generate the augmented support set  $S_{\text{aug}}$ . To address potential concerns regarding reliance on a single proprietary augmentation model,

| Model      | Dataset     | Main  | Mean $\pm$ Std   | $ \Delta $ |
|------------|-------------|-------|------------------|------------|
| Llama-3-8B | Medicine    | 24.19 | $24.15 \pm 0.42$ | 0.04       |
| Llama-3-8B | Agriculture | 19.83 | $19.76 \pm 0.35$ | 0.07       |
| Qwen2.5-7B | Medicine    | 32.72 | $32.65 \pm 0.48$ | 0.07       |
| Qwen2.5-7B | Agriculture | 15.51 | $15.47 \pm 0.31$ | 0.04       |

Table 16: Multi-seed stability analysis of AT-TTA over five random seeds. “Main” denotes the corresponding single-run score in Table 1.  $|\Delta|$  is the absolute difference between the multi-seed mean and the main-table result.

we conduct an extended ablation study using alternative augmentation backends with different model families and capabilities. Specifically, we compare the generation quality of a smaller baseline (Llama-3-8B), GPT-4, and Qwen3.5-plus under the same instruction-preserving augmentation strategies on the Medicine dataset, adapted with Llama-3-8B. To provide a comprehensive evaluation, we report six metrics: data characteristics (Diversity and Complexity), downstream generation quality (ROUGE-Lsum and BERTScore), and reference-based evaluation scores (Reference-Answer Correctness and Logical Reasoning).

As shown in Table 17 (and visualized in the radar chart in the main text), the models present a trade-off. While Llama-3-8B serves as a viable baseline, it obtains lower reference-based quality scores. GPT-4 maintains a slight advantage in BERTScore and reference-based correctness and reasoning, whereas Qwen3.5-plus yields anchors with the highest semantic diversity (1.20 vs. 1.11) and syntactic complexity (2.37 vs. 2.23).

Consequently, this richer and more diverse supervision signal is particularly beneficial for Test-Time Adaptation, where broadly covering the target distribution is crucial. This translates directly into an improved final adaptation performance, with Qwen3.5-plus increasing the ROUGE-Lsum from 24.19 to 25.69. This experiment demonstrates that AT-TTA is robust across augmentation backends and that its efficacy scales positively with the structural richness of the augmented data, rather than being strictly tied to a specific proprietary LLM.

## E.9 Efficiency and Computational Cost

To assess the practical deployment feasibility of AT-TTA, we analyze the trade-off between adaptation performance and inference latency. Figure 7 displays this balance across methods, with detailed numerical results provided in Table 18.

Notably, AT-TTA accelerates inference relative

| Augmentation Model | Diversity $\uparrow$ | Complexity $\uparrow$ | ROUGE-Lsum $\uparrow$ | BERTScore $\uparrow$ | Ref. Corr. $\uparrow$ | Logical Reasoning $\uparrow$ |
|--------------------|----------------------|-----------------------|-----------------------|----------------------|-----------------------|------------------------------|
| Llama-3-8B         | 0.98                 | 1.85                  | 20.85                 | 0.6980               | 6.20                  | 6.00                         |
| GPT-4              | 1.11                 | 2.23                  | 24.19                 | <b>0.7385</b>        | <b>8.30</b>           | <b>8.10</b>                  |
| Qwen3.5-plus       | <b>1.20</b>          | <b>2.37</b>           | <b>25.69</b>          | 0.7310               | 7.90                  | 7.60                         |

Table 17: Detailed comparison of augmentation models on the Medicine dataset (adapted by Llama-3-8B). Qwen3.5-plus generates the most diverse and complex anchors and obtains the highest ROUGE-Lsum, whereas GPT-4 obtains the highest BERTScore and reference-based quality scores.

Table 18: Detailed performance comparison of efficiency and adaptation quality on the Agriculture dataset. Note that Latency is measured in seconds (s), and performance is reported in ROUGE-Lsum.

| Model      | Method               | Latency     | ROUGE-Lsum   |
|------------|----------------------|-------------|--------------|
| Qwen2.5-7B | Base                 | 10.65       | 9.42         |
|            | Tent                 | <b>1.55</b> | 11.05        |
|            | TLM                  | 7.44        | <u>12.49</u> |
|            | <b>Ours (AT-TTA)</b> | <u>6.91</u> | <b>15.51</b> |
| Llama-3-8B | Base                 | 3.10        | 8.56         |
|            | Tent                 | <b>1.76</b> | 0.85         |
|            | TLM                  | 2.86        | <u>13.09</u> |
|            | <b>Ours (AT-TTA)</b> | <u>2.60</u> | <b>19.83</b> |

to unadapted base models, reducing latency from 10.65s to 6.91s on Qwen2.5-7B and from 3.10s to 2.60s on Llama-3-8B. This counterintuitive efficiency gain—despite the additional overhead of gradient updates—stems from the regularization effect of the Semantic Anchor. Specifically, unadapted models in specialized domains frequently suffer from prolonged hallucinations or degenerative repetition, leading to unnecessarily long generation sequences. In contrast, AT-TTA’s energy-guided anchoring mechanism enforces more concise and focused reasoning, effectively reducing the average generation length and offsetting the computational cost of parameter updates.

Conversely, while methods such as Tent and EATA achieve lower latency, they do so at the cost of severe model collapse. As evidenced by premature generation termination and degraded output quality (see Section 5.2), these methods sacrifice performance for speed. AT-TTA, by contrast, achieves a favorable balance between strong generation performance and competitive inference efficiency relative to the base model.

Table 18 presents detailed latency and ROUGE-Lsum results on the Agriculture dataset.

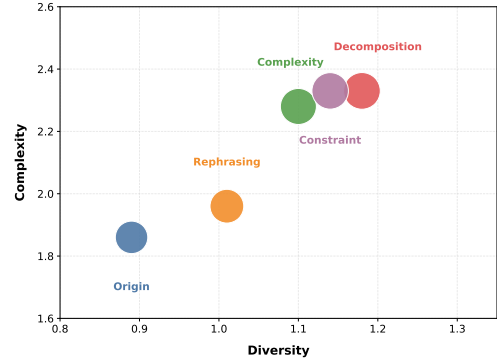


Figure 10: Visualization of augmentation strategies. The x-axis represents semantic diversity, the y-axis represents complexity, and the bubble size corresponds to the aggregated quality score.

## E.10 Discussion on support data augmentation quality

In Section 4.1, we introduced a data-augmented support set  $\mathcal{S}_{aug}$  to serve as a “Semantic Anchor” for mitigating model drift. To characterize the effectiveness of these augmented samples in constructing a robust anchor, we conduct a quantitative breakdown of different augmentation strategies. Specifically, we analyze the trade-off between semantic diversity and syntactic complexity to understand how different strategies contribute to the richness of the supervision signal. We visualize these properties in Figure 10.

**Metric Design.** Inspired by recent findings (Lu et al., 2024) highlighting the critical role of dataset complexity and diversity in model alignment, we employ the proposed InsTag framework to quantitatively evaluate our constructed anchors. Specifically, we aim to measure whether our augmentation strategies successfully broaden the semantic coverage and density of the support set compared to the original data.

Let  $\mathcal{D} = \{q_1, q_2, \dots, q_N\}$  denote a dataset containing  $N$  queries. For each query  $q_i$ , let  $T_i$  represent the set of normalized intention tags generated by InsTag. The metrics are defined as follows:

**Complexity.** We define complexity as the average

number of intention tags per query, reflecting the semantic density of the instructions:

$$\text{Complexity} = \frac{1}{N} \sum_{i=1}^N |T_i| \quad (24)$$

where  $|T_i|$  denotes the number of tags associated with the  $i$ -th query.

**Diversity.** We define diversity as the coverage of unique tags normalized by the dataset size, reflecting the breadth of semantic intentions:

$$\text{Diversity} = \frac{1}{N} \left| \bigcup_{i=1}^N T_i \right| \quad (25)$$

where  $\bigcup_{i=1}^N T_i$  represents the set of all unique normalized tags present in the dataset.

Table 19: Quantitative analysis of anchor quality metrics.

| Strategy      | Complexity | Diversity | Total |
|---------------|------------|-----------|-------|
| Origin        | 1.86       | 0.89      | 2.75  |
| Rephrasing    | 1.96       | 1.01      | 2.97  |
| Complexity    | 2.28       | 1.10      | 3.38  |
| Constraint    | 2.33       | 1.14      | 3.47  |
| Decomposition | 2.33       | 1.18      | 3.51  |

Table 20: Impact of different shot sizes on Domain-Bench.

| Shots | Geography    | Agriculture  | Medicine     | Finance      |
|-------|--------------|--------------|--------------|--------------|
| 8     | 32.80        | 13.11        | 31.05        | <b>35.10</b> |
| 16    | <b>34.73</b> | 15.51        | 32.72        | 34.92        |
| 32    | 31.79        | <b>23.88</b> | 32.66        | 33.89        |
| 64    | 30.71        | 20.92        | <b>33.09</b> | 31.88        |

**Quantitative Analysis.** Figure 10 visualizes the quality of the augmented anchors, showing that the augmentation strategies consistently enhance data complexity compared to the Origin baseline (Total Score: 2.75). Specifically, as detailed in Table 19, Complexity and Decomposition markedly boost both complexity and diversity scores. While surface-level methods, such as Rephrasing, offer only marginal gains, logic-heavy approaches like Constraint yield a substantial increase in complexity. Moreover, Decomposition achieves the highest Total Score (3.51) by balancing these two dimensions. Crucially, we observe that performance improves as the total quality score increases, confirming that anchor effectiveness is highly correlated with complexity and diversity.

| Example 1: Geography Dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Query</b>        | Write down a few tips concerning the usage of artificial intelligence in business.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Golden Answer</b>         | <ol style="list-style-type: none"> <li>1. Identify impact areas: Identify which tasks can benefit from automation (e.g., data analysis, customer service) and implement AI strategically.</li> <li>2. Focus on customer experience: Use AI for personalized recommendations and faster service.</li> <li>3. Invest in data quality: Ensure data is properly collected and managed to get accurate results.</li> <li>4. Understand limitations: Recognize potential biases and do not rely solely on AI without human oversight.</li> <li>5. Train employees: Communicate changes and provide training to prepare staff for working alongside AI.</li> <li>6. Stay updated: Keep business up to date with rapid technological advancements.</li> <li>7. Ethics and privacy: Address privacy concerns proactively and ensure AI systems are transparent and ethical.</li> </ol> |
| Augmented Variations         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| <b>Rephrasing</b>            | Provide some essential guidance on the effective implementation of artificial intelligence in business settings.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Complexity</b>            | Develop a comprehensive framework for strategically integrating artificial intelligence into business operations, emphasizing the balance between leveraging technological advantages and mitigating inherent risks, while addressing ethical and operational challenges.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>Decomposition</b>         | <ol style="list-style-type: none"> <li>1. How to identify areas where AI can have the most impact in business operations?</li> <li>2. What strategies can businesses use to enhance customer experience through AI applications?</li> <li>3. What steps should businesses take to ensure high-quality data for effective AI implementation?</li> <li>4. How can businesses recognize and manage the limitations and potential biases of AI technologies?</li> <li>5. What measures can be taken to train employees to work effectively with AI systems?</li> <li>6. How can businesses stay current with AI developments to remain competitive?</li> <li>7. What ethical considerations should businesses prioritize when implementing AI systems?</li> </ol>                                                                                                                 |
| <b>Constraint</b>            | Provide a list of AI implementation guidelines for businesses, subject to conditions such as adherence to industry-specific regulations, evaluation of data quality and integrity requirements, integration of ethical decision-making frameworks, allocation of resources for employee reskilling programs, consideration of operational scalability constraints, and alignment with organizational strategic goals.                                                                                                                                                                                                                                                                                                                                                                                                                                                         |

Table 21: Data augmentation example from the Geography dataset.

| Example 2: Agriculture Dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Query</b>          | What is the importance of drying seeds in relation to facilitating efficient storage and handling?                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Golden Answer</b>           | Dried seeds are easier to handle, store, and transport compared to seeds with higher moisture content. Lower moisture levels reduce the likelihood of seed clumping, mold formation, and damage during storage. Dry seeds also tend to have better flowability, making it easier to measure, package, and distribute them.                                                                                                                                                                                                                                                      |
| Augmented Variations           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <b>Rephrasing</b>              | Why is drying seeds important for enhancing storage and handling efficiency?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Complexity</b>              | What are the critical advantages of desiccating seeds in terms of enhancing their storability, minimizing degradation risks, and optimizing logistical efficiency through the reduction of hygroscopicity-related challenges?                                                                                                                                                                                                                                                                                                                                                   |
| <b>Decomposition</b>           | <ol style="list-style-type: none"> <li>1. How does reducing moisture content affect seed preservation during long-term storage?</li> <li>2. What physical changes occur in seeds after drying that improve their manageability?</li> <li>3. Why do lower moisture levels prevent issues like clumping, mold growth, and degradation?</li> <li>4. How does the flowability of dry seeds impact processes such as packaging, measuring, and distribution?</li> <li>5. What role does drying play in minimizing risks of seed damage during transportation and storage?</li> </ol> |
| <b>Constraint</b>              | What is the importance of drying seeds under controlled environmental conditions in relation to facilitating efficient storage and handling for agricultural crops, particularly when considering specific moisture thresholds, long-term stability, and resistance to microbial degradation?                                                                                                                                                                                                                                                                                   |

Table 22: Data augmentation example from the Agriculture dataset.

| Example 3: Medicine Dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Query Context</b>        | System: If you are a doctor, please answer the medical questions based on the patient's description.<br>User: Doctor, lately I have been feeling like my throat is tight and I've noticed a lump in my neck. Also, my neck seems to be more swollen than usual.                                                                                                                                                                                                                                                                                                                                  |
| <b>Golden Answer</b>        | These symptoms could be caused by a thyroid disease. I will have to conduct further tests to confirm a diagnosis, but it's likely that your thyroid gland is not functioning properly.                                                                                                                                                                                                                                                                                                                                                                                                           |
| Augmented Variations        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Rephrasing</b>           | Doctor, I've recently experienced a sensation of constriction in my throat and have noticed a palpable mass in my neck. Additionally, my neck has become noticeably enlarged.                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Complexity</b>           | Doctor, over the past few weeks, I've experienced a persistent sensation of tightness in my throat accompanied by intermittent pain when swallowing. Additionally, I've observed a palpable lump in my cervical region that appears to fluctuate in size throughout the day, along with generalized lymphadenopathy. My neck has also exhibited increased tenderness upon touch, and I've noticed subtle changes in my voice quality. Could these manifestations indicate a potential thyroid-related pathology or another underlying condition requiring immediate evaluation?                  |
| <b>Decomposition</b>        | <ol style="list-style-type: none"> <li>1. Identification of specific symptoms (throat tightness, neck lump, localized swelling).</li> <li>2. Association of symptoms with potential underlying pathologies (e.g., thyroid dysfunction, lymphadenopathy).</li> <li>3. Determination of clinical urgency and need for diagnostic investigation (imaging, lab tests).</li> <li>4. Differentiation between benign vs. malignant etiologies for the neck lump.</li> <li>5. Assessment of systemic involvement or related complications (e.g., airway obstruction, metabolic disturbances).</li> </ol> |
| <b>Constraint</b>           | Doctor, over the past two weeks, I have experienced persistent throat tightness and have noticed a painless, firm lump in my right neck. Additionally, there's noticeable swelling that hasn't improved with over-the-counter medications. I don't have a fever, but I've gained 5 pounds recently without changes in diet or exercise. Could this be related to thyroid dysfunction?                                                                                                                                                                                                            |

Table 23: Data augmentation example from the Medicine dataset.

| Example 4: Finance Dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Query</b>      | Produce a written summary of the movie “Jaws”.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Golden Answer</b>       | Jaws is a classic thriller about a giant white shark that terrorizes the small beach town of Amity Island. Police Chief Brody (Roy Scheider) is determined to protect his community from the terror that lurks in the waters, but he must overcome his own fears in order to do so. With the help of a daring oceanographer (Richard Dreyfuss) and a grizzled shark hunter (Robert Shaw), Brody sets out to outwit the giant predator. The suspenseful film has captivated audiences for decades and is considered one of the greats of Hollywood’s golden age. |
| Augmented Variations       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <b>Rephrasing</b>          | Generate a written summary of the movie “Jaws”.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <b>Complexity</b>          | Provide a comprehensive analysis of the narrative structure, character development, and thematic elements of the film Jaws, emphasizing its psychological underpinnings, the interplay between human fear and nature, and its seminal impact on the evolution of the thriller genre within Hollywood cinema.                                                                                                                                                                                                                                                    |
| <b>Decomposition</b>       | What is the genre of the movie Jaws?<br>What is the central plot of Jaws?<br>Who are the main characters in Jaws, and what roles do they play?<br>Where and when does the story of Jaws take place?<br>What is the primary conflict and resolution in Jaws?<br>What themes does Jaws explore, and how are they conveyed?<br>How has Jaws been received historically, and why is it considered significant?                                                                                                                                                      |
| <b>Constraint</b>          | Produce a concise written summary of the movie “Jaws” that highlights its impact on the thriller genre, includes the release year (1975), mentions the director (Steven Spielberg), and excludes any explicit plot details beyond the initial setup. The summary should be limited to three paragraphs, emphasizing the themes of fear and human resilience, and should not reference sequels or subsequent films.                                                                                                                                              |

Table 24: Data augmentation example from the Finance dataset.

| Example 5: Alpaca-GPT4 Dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Query</b>          | Suggest a strategy to increase understanding and engagement in the workplace.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Golden Answer</b>           | <p>One strategy is fostering a culture of two-way communication, active listening, and open feedback. This can be achieved through the following steps:</p> <ol style="list-style-type: none"> <li>1. Encourage regular check-ins: Schedule team meetings or one-on-ones to discuss progress and share ideas.</li> <li>2. Promote active listening: Create an environment where listening is encouraged and train staff to listen considerately.</li> <li>3. Provide constructive feedback: Encourage sharing positive or negative feedback to develop better relationships.</li> <li>4. Use visual aids: Use charts or diagrams to help comprehend complex ideas.</li> <li>5. Encourage continuous learning: Provide training opportunities and workshops to upgrade skills.</li> <li>6. Recognize and reward engagement: Acknowledge contributions to boost morale.</li> </ol> |
| Augmented Variations           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Rephrasing</b>              | Propose ways to enhance understanding and engagement in the workplace environment.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| <b>Complexity</b>              | Develop a multifaceted framework that synergistically integrates organizational structural reforms, psychological safety initiatives, and technology-driven collaboration tools to holistically enhance both cognitive comprehension and participatory engagement across hierarchical and cross-functional teams within a dynamic work environment.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Decomposition</b>           | <p>How can a workplace foster two-way communication to improve understanding and engagement?</p> <p>What strategies can promote active listening among employees?</p> <p>What methods are effective for encouraging constructive feedback in the workplace?</p> <p>How can visual aids be utilized to enhance comprehension of complex topics?</p> <p>What approaches support continuous learning and skill development for staff?</p> <p>How do recognition and reward systems contribute to increased employee engagement?</p>                                                                                                                                                                                                                                                                                                                                                 |
| <b>Constraint</b>              | Suggest a strategy to increase understanding and engagement in the workplace only if the organization has a hybrid workforce, provided that the strategy accommodates varying levels of digital literacy among employees, and ensuring that the approach aligns with existing compliance protocols while minimizing disruptions to daily operations.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |

Table 25: Data augmentation example from the Alpaca-GPT4 dataset.

| Example 6: Dolly Dataset |                                                                                                                                                                                                                                                                                                                                                                                                    |
|--------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Query</b>    | What are the generations of the Ford Bronco?                                                                                                                                                                                                                                                                                                                                                       |
| <b>Golden Answer</b>     | 1966-1977 Generation 1 (also considered EB or Early Broncos)<br>1978-1979 Generation 2<br>1980-1986 Generation 3 (also called bull nose)<br>1987-1991 Generation 4 (also called brick nose)<br>1992-1996 Generation 5<br>2021-Present Generation 6 (also called new bronco)                                                                                                                        |
| Augmented Variations     |                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>Rephrasing</b>        | What are the different generations of the Ford Bronco and their corresponding model years?                                                                                                                                                                                                                                                                                                         |
| <b>Complexity</b>        | What is the chronological breakdown of the Ford Bronco’s generational development, including distinct design evolutions, technical specifications, and market relevance across each major production phase from its debut to the latest iteration?                                                                                                                                                 |
| <b>Decomposition</b>     | What are the model years for each generation of the Ford Bronco?<br>What are the alternative names or classifications for each generation of the Ford Bronco?                                                                                                                                                                                                                                      |
| <b>Constraint</b>        | What are the generations of the Ford Bronco, specifically focusing on North American models produced between 1966 and 2021, including detailed production years, alternative naming conventions (e.g., “EB,” “bull nose,” “brick nose”), and notable design distinctions for each generation? Ensure the response excludes any information about international variants or post-2021 developments. |

Table 26: Data augmentation example from the Dolly dataset.

| Example 7: InstructionWild Dataset |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Original Query</b>              | Summarize the objectives of knowledge discovery and how it can be used to improve decision-making.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <b>Golden Answer</b>               | Knowledge Discovery (KD) is an interdisciplinary field of research that is concerned with extracting useful and meaningful information from large datasets. KD consists of multiple components such as data mining, machine learning, information retrieval, and natural language processing. KD is used to aid decision-making by extracting insights from data which can lead to better choices. The process involves extracting relevant knowledge from data sources and using it to identify opportunities, identify correlations, trends and patterns helpful for decision making. |
| Augmented Variations               |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <b>Rephrasing</b>                  | Outline the goals of knowledge discovery and discuss its application in supporting improved decision-making.                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Complexity</b>                  | Elaborate on the multifaceted objectives of Knowledge Discovery (KD) within the context of its interdisciplinary framework, emphasizing its methodological underpinnings and analytical processes, and explore its transformative potential in refining complex decision-making paradigms through the extraction and application of actionable insights from structured and unstructured data streams.                                                                                                                                                                                  |
| <b>Decomposition</b>               | What are the primary objectives of Knowledge Discovery (KD)?<br>What are the key components or methodologies involved in KD?<br>How does KD facilitate improved decision-making through data analysis?                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Constraint</b>                  | Summarize the objectives of knowledge discovery (KD) specifically within the context of analyzing structured versus unstructured datasets, and explain how it can improve decision-making in dynamic environments only if the data meets predefined quality thresholds and the extracted insights align with domain-specific regulatory frameworks. Additionally, clarify how KD addresses scalability challenges when applied to high-volume, real-time data streams in competitive markets.                                                                                           |

Table 27: Data augmentation example from the InstructionWild dataset.