DEAL-ICL: Robust In-Context Learning via Semantic Expansion, Alignment, and Dual-View Adaptation

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have demonstrated impressive In-Context Learning (ICL) capabilities, their performance remains highly sensitive to demonstration quality and test-time distribution shifts. Existing approaches primarily focus on optimizing demonstration retrieval or calibration during decoding, yet they leave the model parameters frozen during inference, limiting the model's ability to fundamentally adapt to unseen queries. To bridge this gap, we propose DEAL-ICL, a robust framework designed to enhance ICL through progressive adaptation. DEAL-ICL operates in three stages: (1) semantic expansion to enrich the demonstration pool, (2) ICL-aligned supervised fine-tuning to internalize the retrieval-augmented format, and (3) a novel Dual-View Test-Time Adaptation mechanism. During inference, we construct anchor and perturbed views of the input and leverage a geometric consistency objective to dynamically update model parameters. Extensive experiments on Llama3 and Qwen2 benchmarks demonstrate that DEAL-ICL achieves state-of-the-art performance. Notably, under a challenging random retrieval setting, our method consistently outperforms contrastive decoding baselines across various natural language understanding tasks by effectively mitigating pre-training biases.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive In-Context Learning (ICL) capabilities (Brown et al., 2020); however, their effectiveness remains highly sensitive to the quality of demonstrations (Zhao et al., 2021; Lu et al., 2022). Early research prioritized Demonstration Retrieval , with frameworks like OpenICL (Wu et al., 2023a) employing strategies such as BM25 and TopK to select optimal examples. Yet, subsequent studies revealed that retrieval alone is insufficient: even with high-quality demonstrations, models frequently ignore specific Input-Label Mappings within the context, relying instead on pre-training priors (Min et al., 2022b). To mitigate this, ICCD (Peng et al., 2025) introduced contrastive decoding to calibrate output probabilities. Despite these advancements, both retrieval optimization and decoding interventions share a fundamental limitation: model parameters remain frozen during inference, preventing the model from truly adapting to the specific distribution of test data.

To address the constraint of frozen parameters, we draw inspiration from Meta-ICL (Min et al., 2022a), which pioneered the use of Supervised Fine-Tuning (SFT) to explicitly internalize ICL capabilities. While this paradigm effectively teaches models to attend to demonstrations, it remains fundamentally an offline strategy: once training concludes, parameters are fixed, leaving the model static against unseen test queries and distribution shifts. To bridge the gap between this offline alignment and dynamic adaptation, Test-Time Adaptation (TTA) has emerged as a promising direction. Although early TTA methods like TENT (Wang et al., 2021) focused on vision, recent work such as TTRL (Zuo et al., 2025) has proven the feasibility of updating LLM parameters on unlabeled text. Building on theoretical insights regarding consistency from Self-Harmony (Wang et al., 2023, 2025), we explore a novel trajectory: using Meta-ICL-style SFT as a foundational alignment step, and further leveraging the consistency of test samples themselves as a self-supervised signal for dynamic parameter updates during inference.

To this end, we propose **DEAL-ICL**, an end-to-end framework for robust test-time adaptation. Specifically , DEAL-ICL moves beyond simple inference heuristics, offering a systematic solution comprising three progressive stages: (1) **Stage 1: Semantic Expansion.** We identify that standard retrieval pools often lack sufficient diversity to cover the semantic space of unseen test queries. To

mitigate this, we employ **consistency-aware data augmentation**. Drawing on the insight that consistency serves as a proxy for correctness (Wang et al., 2023; Xue et al., 2023), we utilize established augmentation techniques (Bayer et al., 2023) to filter and construct a high-density demonstration set, ensuring reliable context retrieval. (2) **Stage 2: ICL-Aligned SFT.** Acting as a bridge between pre-training and TTA, we fine-tune the model to internalize the utilization of these retrieved contexts, establishing a stable initialization for subsequent adaptation. (3) **Stage 3: Dual-View TTA.** During the test phase, we synthesize an "Anchor View" and a "Perturbed View," employing LoRA (Hu et al., 2022) to explicitly update parameters in real-time, thereby forcing the model to capture deep semantic invariances.

In summary, our contributions are three-fold: ❶ **Systematic Robust ICL Framework.** We propose DEAL-ICL, a systematic framework that seamlessly integrates semantic expansion with dual-view test-time adaptation. This effectively bridges the gap between offline internalization and online adaptation, addressing the limitations of frozen parameters in traditional ICL. ❷ **Novel Methodology.** Unlike retrieval-only methods, we introduce a self-supervised Dual-View TTA mechanism that constructs "Anchor" and "Perturbed" views during inference. By minimizing the consistency loss between these views, DEAL-ICL dynamically updates model parameters to capture semantic invariants. ❸ **Extensive Experiments & Analysis.** We achieve state-of-the-art (SOTA) performance on Llama3 (Team, 2024) and Qwen2 (Yang et al., 2024)series models. Notably, DEAL-ICL consistently outperforms ICCD even under the challenging Random Retrieval setting. Our analysis further confirms that both SFT alignment and test-time updates are crucial for these performance leaps.

2 Related Work

Since the emergence of GPT-3 (Brown et al., 2020), In-Context Learning (ICL) has become a dominant paradigm for utilizing LLMs. However, its performance is known to be highly volatile, suffering from sensitivity to demonstration ordering (Lu et al., 2022; Wu et al., 2023b) and label bias (Zhao et al., 2021). To mitigate this, previous works proposed calibration methods to adjust output distributions (Zhao et al., 2021; Zhang et al., 2024). In parallel to calibration, selecting high-quality demonstrations has been identified as a crucial factor (Peng et al., 2024). While early approaches relied on heuristics like BM25 (Robertson and Zaragoza, 2009), advanced methods now utilize embedding-based retrievers (Chen et al., 2024) or specialized policy networks (Wang et al., 2024). Despite these advancements, both calibration and retrieval strategies typically keep model parameters **frozen** during inference, limiting the model's ability to fundamentally align with the specific test distribution. To address this limitation, our work draws inspiration from Test-Time Adaptation (TTA) (Wang et al., 2021), which updates model parameters on-the-fly. While applying TTA to LLMs has historically faced stability issues (Xu et al., 2025), recent works like TTRL (Zuo et al., 2025) have demonstrated the feasibility of updating LLMs on unlabeled text. Distinct from prior arts, DEAL-ICL introduces a unified framework that not only enriches the retrieval pool via semantic expansion but also employs a dual-view adaptation mechanism. By treating the consistency between an *Anchor View* and a *Perturbed View* as supervision, we enable robust dynamic updates without deviating from the model's pre-trained distribution.

3 Methodology

3.1 Overview

We propose **DEAL-ICL**, a robust framework designed to bridge the gap between pre-trained models and downstream few-shot inference. As illustrated in Figure 1, the framework comprises three progressive stages: (1) **Semantic Expansion** (Stage 1), which alleviates demonstration sparsity via consistency-aware data augmentation; (2) **ICL-Aligned SFT** (Stage 2), which explicitly adapts the model to the retrieval-augmented paradigm; and (3) **Dual-View Test-Time Adaptation** (Stage 3), the core contribution of this work, which leverages dual-view consistency and a relative gain objective for dynamic online updates .

3.2 Stage 1: Semantic Expansion

Standard In-Context Learning (ICL) performance is often bottlenecked by the sparsity of high-quality demonstrations. To address this, we adopt a generation-based augmentation strategy to enrich the semantic coverage of the retrieval pool. Specifically, given a seed sample $(x, y) \in \mathcal{D}_{\text{seed}}$, we utilize an LLM to generate a diverse set of paraphrase candidates $\mathcal{X}'$. However, generated

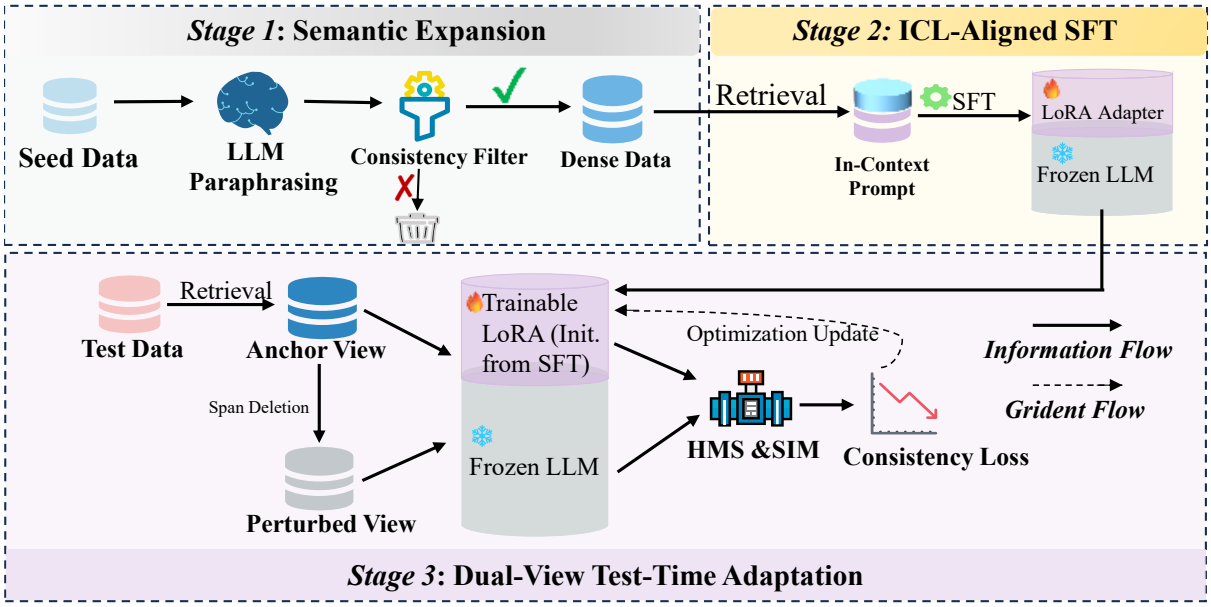


Figure 1: **Overview of the proposed fine-tuning framework combining data augmentation and consistency regularization.** The pipeline consists of three stages: (1) Utilizing a Paraphraser/T5 model to augment and filter the training data; (2) Performing Supervised Fine-Tuning (SFT) with LoRA on the generated dense data; (3) Optimizing robustness against retrieval noise by minimizing the consistency loss between the full context (Anchor View) and the incomplete context (Perturbed View) generated via Span Deletion.

paraphrases may suffer from semantic drift or ambiguity. Drawing inspiration from consistency voting strategies (Xue et al., 2023), we introduce a Stochastic Consistency Filter to ensure reliability. By leveraging Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016), we evaluate the stability of the model's predictions under perturbation. We execute T stochastic forward passes using a task-specific auxiliary classifier $\mathcal{C}_{\text{aux}}$. A prediction in the t-th pass is considered a *valid vote*, denoted as $\delta_t(\hat{x}, y) = 1$, only if it strictly satisfies three criteria: (1) **Correctness** (matches ground truth y); (2) **Confidence** (probability exceeds ξ_{conf}); and (3) **Discriminability** (margin between top-2 classes exceeds ξ_{margin}). We then aggregate these stochastic votes to compute a Consistency Score, representing the reliability of the candidate:

$$S_{\text{cons}}(\hat{x}, y) = \frac{1}{T} \sum_{t=1}^{T} \delta_t(\hat{x}, y). \quad (1)$$

Finally, we construct the expanded dense pool $\mathcal{D}_{\text{dense}}$ by merging the seed data with candidates that exceed the consistency threshold τ_{cons}:

$$\mathcal{D}_{\text{dense}} = \mathcal{D}_{\text{seed}} \cup \{(\hat{x}, y) \mid \hat{x} \in \mathcal{X}', S_{\text{cons}}(\hat{x}, y) \geq \tau_{\text{cons}}\}. \quad (2)$$

This rigorous filtering ensures the retrieval pool covers a broader semantic space while maintaining high label fidelity.

3.3 Stage 2: ICL-Aligned Supervised Fine-Tuning

While the semantic expansion in Stage 1 provides a rich source of demonstrations, pre-trained LLMs are not inherently optimized to leverage such context effectively. Standard Supervised Fine-Tuning (SFT) typically maps an input x directly to a label y, ignoring the contextual dependence on retrieved demonstrations. This leads to a significant *training-inference discrepancy* (Wei et al., 2023; Dong et al., 2024): during inference, the model is presented with a long sequence of examples it never learned to process during training, often resulting in the neglect of the retrieved context. To bridge this gap and maximize the utility of $\mathcal{D}_{\text{dense}}$, we implement an alignment strategy inspired by Meta-ICL (Min et al., 2022a). Our goal is to shift the model's focus from simple pattern recognition to analogy-based reasoning. Specifically, we construct retrieval-augmented training instances to simulate the few-shot inference process. For each training query (x, y), we employ a Dynamic Context Sampling strategy, drawing k demonstrations $C = \{(x_i, y_i)\}_{i=1}^{k}$ from the expanded dense pool $\mathcal{D}_{\text{dense}}$. This dynamic sampling not only aligns the training format with inference but also serves as a data augmentation technique to prevent overfitting. The optimization objective is to minimize

the negative log-likelihood of y conditioned on the augmented context:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}} \left[\log P_\theta(y \mid C \oplus x)\right], \tag{3}$$

where $\oplus$ denotes concatenation. This stage explicitly aligns the model's internal representations with the in-context learning paradigm, ensuring it learns to attend to and use the retrieved information.

3.4 Stage 3: Dual-View Test-Time Adaptation

To address distribution shifts during the inference phase, we propose an online test-time adaptation mechanism based on Dual-View Consistency. Unlike standard TTA methods that reset parameters for each instance, we adopt a continuous adaptation paradigm: the adapter parameters ϕ are updated in a streaming fashion across the test set to capture global distribution trends.

View Construction for Gradient Generation. Since test instances lack ground-truth labels, we cannot directly compute gradients for supervised updates. We construct two distinct views for each test query x_{test} under a "Teacher-Student" paradigm (Sohn et al., 2020):

- **Anchor View (v_a):** Contains the original query and standard retrieved demonstrations. It serves as the *Teacher*, providing stable pseudo-labels based on the model's current capability.
- **Perturbed View (v_p):** Generated via Span Deletion by randomly masking continuous token spans in the context. This view acts as the *Student*. The discrepancy between v_p and v_a drives the gradient backpropagation.

Online Adaptation Protocol. For each incoming test instance x_t, we perform a predict-then-adapt procedure. First, the model generates the prediction y_t using the current parameters $\theta_{\text{SFT}} + \phi_t$. Subsequently, if the sample passes the reliability check, we perform a single gradient step to update $\phi_t \to \phi_{t+1}$. This allows the model to progressively adapt to the test distribution and correct domain shifts for subsequent instances.

Reliability Filtering via Dual-Metric Gating. The primary risk in unsupervised online adaptation is the accumulation of noise from erroneous pseudo-labels (Pan et al., 2023). Since updates are persistent, we must ensure that only the most trustworthy predictions contribute to gradient updates.To achieve this, we propose a dual-metric mechanism that acts as a strict quality guard.

Algorithm 1 Overall Training and Inference Procedure of DEAL-ICL (Online Streaming Version)

Require: Training set $\mathcal{D}_{\text{train}}$, Test stream $\mathcal{D}_{\text{test}} = \{x_1, ..., x_N\}$, Pre-trained LLM $\mathcal{M}_\theta$, Context size k.
Ensure: Predicted labels $\{y_1, ..., y_N\}$.

// Stage 1: Semantic Expansion (Offline)
Initialize $\mathcal{D}_{\text{dense}} \leftarrow \mathcal{D}_{\text{train}}$
for each $(x, y) \in \mathcal{D}_{\text{train}}$ **do**
  Generate paraphrases $\mathcal{X}' = \text{LLM-Aug}(x)$
  Filter $\mathcal{X}'$ based on consistency score f using Eq. (1)
  Update $\mathcal{D}_{\text{dense}} \leftarrow \mathcal{D}_{\text{dense}} \cup \mathcal{X}'_{\text{filtered}}$
end for
// Stage 2: ICL-Aligned SFT (Offline)
Initialize $\theta_{\text{SFT}} \leftarrow \theta$
while not converged **do**
  Sample batch $\mathcal{B}$ from $\mathcal{D}_{\text{train}}$ with retrieved context from $\mathcal{D}_{\text{dense}}$
  Update θ_{SFT} to minimize SFT objective $\mathcal{L}_{\text{SFT}}$ via Eq. (3)
end while
// Stage 3: Dual-View Test-Time Adaptation (Online Streaming)
Initialize adapter ϕ on θ_{SFT}; Freeze reference $\theta_{\text{ref}} \leftarrow \theta_{\text{SFT}}$

Initialize prediction list $Y_{\text{pred}} \leftarrow \emptyset$
for each test instance x_t in $\mathcal{D}_{\text{test}}$ **do**
  Retrieve context $C_t = \text{Retrieve}(x_t, \mathcal{D}_{\text{dense}}, k)$
  Construct views v_a (Anchor) and v_p (Perturbed) {Code: build_noisy_text}
  // Step 3.1: Inference with current state
  Compute logits $P_{\theta_{\text{SFT}}+\phi}(\cdot \mid v_a)$
  Predict label $y_t = \arg\max_y P_{\theta_{\text{SFT}}+\phi}(y \mid v_a)$
  Add y_t to Y_{pred}
  // Step 3.2: Online Adaptation
  Compute pseudo-label $\hat{y}$ via HMS using Eq. (4)
  Compute gating mask $\mathbb{I}_{\text{update}}$ via Eq. (5)
  if $\mathbb{I}_{\text{update}} = 1$ **then**
    Compute Relative Gain $\mathcal{G}(v)$ for both views using Eq. (6)
    Calculate loss $\mathcal{L}_{\text{DV-TTA}}$ via Eq. (7)
    Update adapter parameters: $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}_{\text{DV-TTA}}$
  end if
end for
return Y_{pred}

First, we introduce the Harmonic Mean Score (HMS) to rigorously assess prediction reliability. Let $p_a = P_\theta(\hat{y} \mid v_a)$ and $p_p = P_\theta(\hat{y} \mid v_p)$ denote the prediction confidence of the pseudo-label $\hat{y}$ under the anchor and perturbed views, respectively. The HMS is defined as:

$$\text{HMS}(\hat{y}) = \frac{2 \cdot p_a \cdot p_p}{p_a + p_p + \epsilon}, \tag{4}$$

where ϵ is a small constant for numerical stability. Unlike the arithmetic mean, the harmonic mean is strictly dominated by the minimum value. This property allows us to use consistency as a proxy for correctness: if the model fluctuates significantly under perturbation (i.e., low p_p), the HMS drops sharply, effectively filtering out noise-sensitive samples.

Second, to further guarantee the stability of the

supervision signal, we employ Cosine Similarity to ensure **distributional alignment**. The final update mask is:

$$\mathbb{I}_{\text{update}} = \mathbb{I}\big(\text{HMS}(\hat{y}) > \tau_{\text{hms}}\big) \cdot \mathbb{I}\big(\cos(\mathbf{p}_a, \mathbf{p}_p) > \tau_{\cos}\big). \quad (5)$$

This gating mechanism functions as a **firewall**, preventing erroneous data from polluting the continuous self-training process.

Optimization via Relative Gain Consistency. Simply minimizing cross-entropy on pseudo-labels may bias the model towards its pre-existing knowledge. We optimize the Relative Gain for samples passing the filter. We define the gain function $\mathcal{G}(v)$ as:

$$\mathcal{G}(v) = \log P_\theta(\hat{y} \mid v) - \log P_{\theta_{\text{ref}}}(\hat{y} \mid v). \quad (6)$$

The final DV-TTA objective comprises a consistency term and an anchor term:

$$\mathcal{L}_{\text{DV-TTA}} = \underbrace{\big\|\mathcal{G}(v_p) - \text{sg}[\mathcal{G}(v_a)]\big\|_1}_{\text{Alignment Loss}} + \lambda \cdot \underbrace{\big\|\log P_\theta(\hat{y} \mid v_a) - \log P_{\theta_{\text{ref}}}(\hat{y} \mid v_a)\big\|_1}_{\text{Anchor Constraint}}. \quad (7)$$

The *Gain Consistency* forces the student view (v_p) to mimic the information gain of the teacher, while the Anchor Constraint prevents catastrophic forgetting during the online adaptation process.

4 Experiments

We evaluate DEAL-ICL on seven widely-used discriminative NLU tasks, casting classification as conditional text generation (see Table 5 for templates and verbalizers).

4.1 Experimental Setup

Datasets & Evaluation Protocol. We evaluate DEAL-ICL on seven NLU tasks, including subsets from the GLUE benchmark (Wang et al., 2019) and other standard datasets. Our evaluation covers Sentiment Analysis (SST-2, SST-5 (Socher et al., 2013), CR (Hu and Liu, 2004)), Subjectivity Classification (Subj (Pang and Lee, 2004)), News Classification (AG News (Zhang et al., 2015)), and Natural Language Inference (QNLI (Rajpurkar et al., 2016), MNLI (Williams et al., 2018)). All datasets are sourced via the Hugging Face library (Lhoest et al., 2021). To simulate low-resource scenarios, we sample few-shot training sets from the original data while reserving a subset for validation. Following ICCD (Peng et al., 2025), we report results on the official validation sets for MNLI and QNLI, and standard test sets for the other tasks.

Retrieval Strategy (Stress Test). Distinct from optimization-heavy retrieval methods, we adopt a minimalist Random Retrieval strategy ($k = 16$) using OpenICL (Wu et al., 2023a). This setup minimizes computational costs and serves as a **"Stress Test,"** ensuring that performance gains stem principally from our test-time adaptation mechanism rather than the quality of retrieved demonstrations.

Baselines. We benchmark DEAL-ICL against two primary baselines: **Random Retrieval (Standard ICL)**, which utilizes $k = 16$ randomly selected demonstrations to establish a fundamental performance lower bound; and **ICCD** (Peng et al., 2025), a state-of-the-art contrastive decoding approach. Additionally, for the Qwen2-7B setting, we include results from **TopK-L2D** (using Qwen2.5-7B-Instruct)(Jiang et al., 2025) as a strong external reference to contextualize our method's competitiveness against advanced retrieval-based strategies.

Implementation Details. We conduct experiments on the Llama-3 family (1B, 3B, 8B) (Team, 2024) and the Qwen2 family (0.5B, 1.5B, 7B) (Yang et al., 2024) using NVIDIA GPUs. Following Peng et al. (2025), we employ 16-shot ICL across all models. To ensure robustness, we repeat all experiments three times with different random seeds and report average accuracies.In Stage 2, we apply LoRA (Hu et al., 2022) ($r = 8, \alpha = 16, \text{dropout} = 0.1$) to the query and value projections, fine-tuning for 4 epochs with AdamW ($lr = 1 \times 10^{-5}$, batch size 2).In Stage 3 (Online DV-TTA), we perform test-time gradient updates on each instance without access to ground truth.Crucially, we adopt a reduced learning rate of 5×10^{-6} and an anchor loss weight of $\lambda = 0.01$ to maintain stability.The perturbed view is generated via Span Deletion ($\gamma = 0.25$). Finally, we enforce a strict gating mechanism ($\tau_{\text{hms/cos}} > 0.35$) with a maximum of $T = 5$ update steps per sample.

4.2 Main Results

Table 1 presents the performance comparison. Overall, DEAL-ICL consistently outperforms both the Random baseline and the state-of-the-art ICCD across all model families and scales.

Model	Method	SST2	CR	SST5	Subj	QNLI	MNLI	AGNews	Avg.
Llama3.2-1B	Random	89.8	83.0	43.7	72.8	53.5	36.6	83.3	66.1
	ICCD	91.1	83.7	43.3	83.0	**53.8**	39.2	84.1	68.3
	DEAL-ICL	**94.2**	**92.1**	**49.5**	**93.9**	51.6	**42.0**	**87.2**	**72.9 (+4.6)**
Llama3.2-3B	Random	93.7	87.2	46.2	86.0	54.2	56.9	86.4	72.9
	ICCD	94.0	88.1	46.5	92.1	**57.2**	57.0	86.9	74.6
	DEAL-ICL	**94.7**	**94.2**	**52.0**	**96.8**	56.2	**70.9**	**88.5**	**79.0 (+4.4)**
Llama3.1-8B	Random	**96.7**	92.3	48.0	94.0	60.3	65.3	86.7	77.6
	ICCD	96.5	93.2	49.3	96.1	65.4	67.5	87.6	79.4
	DEAL-ICL	95.1	**94.4**	**55.3**	**97.0**	**77.3**	**80.6**	**89.7**	**84.2 (+4.8)**
Qwen2-0.5B	Random	87.9	89.4	34.5	62.2	52.5	47.6	78.1	64.6
	ICCD	89.2	89.6	33.9	68.1	53.2	47.6	78.7	65.8
	DEAL-ICL	**91.9**	**91.2**	**42.0**	**84.6**	**62.3**	**56.0**	**85.0**	**73.3 (+7.5)**
Qwen2-1.5B	Random	**95.2**	91.0	49.0	72.3	60.2	61.8	76.7	72.3
	ICCD	95.1	**91.3**	48.3	81.5	61.8	65.2	79.1	74.6
	DEAL-ICL	94.6	91.2	**50.2**	**86.5**	**74.8**	**75.2**	**86.0**	**79.8 (+5.2)**
Qwen2-7B	Random	96.0	91.5	51.9	82.3	71.4	78.7	83.8	79.4
	ICCD	96.3	91.7	52.9	90.4	72.8	79.9	85.0	81.3
	TopK-L2D (Qwen2.5-7B-Instruct)†	***96.5***	***94.7***	*54.3*	*95.2*	***85.5***	*83.6*	*78.2*	*84.0*
	DEAL-ICL	95.2	93.9	**54.3**	**95.5**	85.2	**86.8**	**88.1**	**85.6 (+4.3)**

Table 1: Main results on NLU benchmark tasks. We report the accuracy (%) and the absolute improvement gap over the strong baseline ICCD (in parentheses). Note that for the Qwen2-7B setting, we include **L2D**† for comparison, which utilizes the significantly stronger **Qwen2.5-7B-Instruct** backbone. Despite this, our **DEAL-ICL** achieves consistent improvements across **various** model scales. **Red** indicates the best performance within each group.

Surpassing Strong Retrieval with Naive Sampling. The most compelling result is observed in the Qwen2-7B setting. While DEAL-ICL utilizes a larger number of shots ($k = 16$), it is well-established that increasing the few-shot count alone rarely compensates for the inherent performance gap between different model generations (e.g., Qwen2 vs. Qwen2.5-Instruct)(Brown et al., 2020). Remarkably, DEAL-ICL allows the base Qwen2-7B model to achieve an average accuracy of **85.6%**, surpassing *TopK-L2D* (**84.0%**), which employs $k = 8$ shots alongside the more powerful *Qwen2.5-7B-Instruct* and sophisticated retrieval algorithms. This suggests that the performance gains are not merely a result of the demonstration quantity, but are fundamentally driven by our DV-TTA mechanism's ability to refine the model's internal alignment on the fly, effectively bridging the capability gap between model versions.

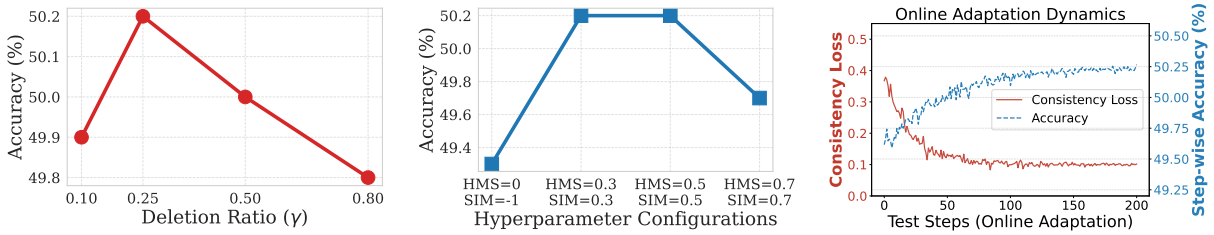
Robustness Across Model Scales. DEAL-ICL exhibits remarkable gains on smaller models. For instance, on Qwen2-0.5B, it boosts the average accuracy from 65.8% (ICCD) to 73.3%. This indicates that our dual-view self-consistency constraint is particularly beneficial for smaller LLMs, which typically struggle with context-following in few-shot scenarios.

Logical Reasoning Breakthrough. The advantage of DEAL-ICL is most pronounced in NLI tasks (QNLI and MNLI), where it outperforms ICCD by over 10% in several cases (e.g., +13.1% on MNLI for Llama3.1-8B). While random retrieval often provides noisy or irrelevant context, the DV-TTA mechanism forces the model to internalize the underlying logical relationship, effectively transforming a low-quality few-shot prompt into a high-fidelity reasoning signal.

4.3 Further Analysis

Ablation studies (Table 2) show that removing the DV-TTA module causes DEAL-ICL to degenerate into standard ICL-Aligned SFT, resulting in significant performance degradation. DV-TTA yields an average net gain of +0.3% $\sim$ +3.3%, indicating that Stage 2 fine-tuning alone is insufficient to handle test-time distribution shifts, while Stage 3's unsupervised adaptation effectively improves robustness at zero annotation cost. Further examination of Stage 3 gains relative to the ICL-Aligned SFT baseline reveals pronounced non-uniformity across tasks of varying cognitive complexity. On shallow pattern-matching tasks (e.g., SST-5), gains are modest (averaging +0.4%), as these tasks rely heavily on local keyword matching that Stage 2's static weights already capture well, leaving limited room for test-time optimization. In contrast, on deep logical reasoning tasks (e.g., QNLI), Stage 3 delivers substantial improvements (averaging +2.4%), representing the core strength of DEAL-ICL.

We attribute the greater benefits on reasoning tasks to parameter plasticity. Conventional SFT models



(a) Robustness to deletion ratio γ (b) Sensitivity of gating mechanism (c) Online adaptation dynamics

Figure 2: Sensitivity and robustness analysis of key components in Stage 3. (a) Performance under varying span deletion ratios γ (higher γ introduces more noise; even at $\gamma = 0.8$, performance remains strong at **49.8%**). (b) Impact of HMS and SIM thresholds (peak accuracy **50.2%** with stricter thresholds), validating the gating as a "semantic firewall". (c) Consistency loss (red, left axis) and step-wise accuracy (blue, right axis) show rapid convergence within 100 steps, demonstrating high data efficiency.

Model	Method	SST-5	QNLI	MNLI	Avg.
Llama3.2-1B	ICL-Aligned SFT	48.4	51.4	41.2	47.0
	DEAL-ICL	**49.5**	**51.6**	**42.0**	**47.7 (+0.7)**
Llama3.2-3B	ICL-Aligned SFT	51.0	55.6	69.7	58.8
	DEAL-ICL	**52.0**	**56.2**	**70.9**	**59.7 (+0.9)**
Llama3.1-8B	ICL-Aligned SFT	54.9	73.7	80.2	69.6
	DEAL-ICL	**55.3**	**77.3**	**80.6**	**71.1 (+1.5)**
Qwen2-0.5B	ICL-Aligned SFT	43.0	61.5	53.9	52.8
	DEAL-ICL	42.0	**62.3**	**56.0**	**53.4 (+0.6)**
Qwen2-1.5B	ICL-Aligned SFT	49.6	65.9	74.7	63.4
	DEAL-ICL	**50.2**	**74.8**	**75.2**	**66.7 (+3.3)**
Qwen2-7B	ICL-Aligned SFT	53.8	85.0	86.5	75.1
	DEAL-ICL	**54.3**	**85.2**	**86.8**	**75.4 (+0.3)**

Table 2: Ablation study on the impact of DV-TTA. We report accuracy (%) on SST-5, MNLI, and QNLI, along with the average score. DEAL-ICL achieves consistent improvements across different task types and model scales.

tend to overfit spurious correlations (McCoy et al., 2019) in training data, whereas DEAL-ICL's gradient updates during inference enable parameter space reshaping, thereby correcting logical biases and better adapting to the specific structure of the current context.

4.4 Further Analysis

In our DV-TTA framework, we conduct experiments using Qwen2-1.5B on the SST-5 dataset. We employ a Span Deletion strategy to construct the perturbed view. While essential for consistency loss computation, it introduces artificial noise. We analyze two critical components for robustness:

Robustness to Deletion Ratio (γ). As shown in Figure 2(a), DEAL-ICL exhibits strong robustness across $\gamma \in [0.1, 0.8]$. Even at the extreme $\gamma = 0.8$ (80% deletion), accuracy remains 49.8%, outperforming the ICL-Aligned SFT baseline (49.6%). This confirms our dual-view objective captures deep semantic invariants.

Necessity of Gating Mechanism (HMS & SIM). Figure 2(b) shows ablation of the dual gating mechanism. Removing gating (HMS=0, SIM=-1) drops accuracy to 49.3%, below the SFT baseline. Stricter thresholds in $[0.3, 0.5]$ effectively filter low-quality updates, achieving peak accuracy of 50.2%. This validates the gating as a crucial "semantic firewall" against negative transfer.

4.5 Adaptation Dynamics and Efficiency

As DV-TTA introduces inference-time gradient updates, we analyze its convergence behavior and computational overhead. Figure 2(c) illustrates a Step-wise Convergence pattern observed during the online process.

When performing the SST-5 task with Qwen2-1.5B, DEAL-ICL exhibits remarkable data efficiency: accuracy climbs from the 49.6% baseline and stabilizes at approximately 50.0% within the first 100 steps (only 5% of the test stream). This rapid adaptation suggests that our Stage 2 alignment successfully pre-positions parameters in a Critical Adaptation State, enabling swift fine-tuning within a concentrated parameter subspace rather than requiring costly global exploration.

To mitigate the 10× latency bottleneck of full TTA, we propose an Amortized Update Strategy. Leveraging the rapid convergence characteristic, we employ a *Warm-up* mechanism where gradient updates are restricted to the initial 15% of the test stream. Once performance stabilizes, parameters are frozen for the remaining inference. Experimental results demonstrate that this strategy preserves over 95% of the performance gains while reducing average latency to 2–3× that of SFT. Consequently, DEAL-ICL achieves an asymptotic complexity comparable to standard SFT, ensuring its feasibility for real-time streaming applications.

Table 3: Qualitative analysis on a representative sample from QNLI. We compare the prediction of the SFT Baseline and our DEAL-ICL. This case requires the model to identify the specific pathway to a religious goal.

Prompt Composition	
Instruction	Determine if the sentence contains the answer to the question. Answer with [entailment, not_entailment] only.
ICE (1/16)	**Question:** How is Nirvana achieved? **Sentence:** In Theravada Buddhism, the ultimate goal is the attainment of the sublime state of Nirvana, achieved by practicing the Noble Eightfold Path... **Label:** entailment
ICE (2-16)	*[... 15 other representative examples omitted for brevity ...]*
Test Query	**Question:** Where did Temüjin hide during his escape from the Tayichi'ud? **Sentence:** Temüjin's reputation also became widespread after his escape from the Tayichi'ud.
Gold Label	**not_entailment**
Model Response	
ICL-Aligned SFT	**entailment** (Incorrect)
DEAL-ICL	**not_entailment** (Correct)

4.6 Why L_1 Loss for Consistency?

We adopt the L_1 objective for the dual-view consistency constraint over KL-Divergence or L_2 loss for the following reasons:

Mathematical Applicability: KL-Divergence is restricted to probability distributions on a simplex. Our *Relative Gain* $\mathcal{G}(v)$, a scalar log-likelihood shift in $\mathbb{R}$, is not a density function and does not sum to unity. Thus, KL is mathematically ill-posed here, whereas L_1 naturally accommodates these unbounded real-valued shifts.

Robustness to Noise: Stochastic perturbations (e.g., span deletion) can introduce outliers with large semantic discrepancies. Unlike L_2 (MSE), where gradients scale linearly with error and risk destabilizing updates, L_1 provides a constant gradient magnitude. This robust regression mechanism effectively ensures stability against transient noise in the perturbed view.

Fine-grained Convergence: As teacher and student views align, L_2 gradients vanish quadratically, leading to optimization stagnation in small-error regimes. L_1 maintains constant pressure even as discrepancies approach zero, preventing premature convergence and forcing the model toward tighter and more precise alignment.

4.7 Case Study

To evaluate DEAL-ICL's efficacy in resolving logical conflicts, we present a QNLI case study in Table 3. The test query shows high lexical overlap (e.g., "Temüjin", "escape"), yet the sentence lacks the specific location required by the question.

While the SFT baseline succumbs to shallow keyword matching and erroneously predicts `entailment`, DEAL-ICL correctly identifies the relationship as `not_entailment`. This rectification is primarily due to our Dual-view Test-time Adaptation (DV-TTA) mechanism. By enforcing consistency between dual views and utilizing HMS gating to filter noisy gradients, the model undergoes localized parameter refinement. This process re-calibrates the attention mechanism, shifting focus from isolated keywords to deeper semantic verification. Consequently, DEAL-ICL effectively reshapes its decision boundary at test-time to capture precise logical relationships rather than relying on superficial word overlaps.

5 Conclusion

In this work, we presented **DEAL-ICL**, a framework designed to mitigate In-Context Learning (ICL) fragility regarding demonstration quality. By synergizing semantic expansion, aligned SFT, and our novel Dual-View Test-Time Adaptation (DV-TTA), DEAL-ICL enables models to internalize task patterns and perform dynamic self-alignment via geometric consistency. Extensive experiments on Llama-3 and Qwen2 demonstrate that our approach achieves state-of-the-art performance and exhibits remarkable resilience under suboptimal retrieval settings. Most notably, DEAL-ICL effectively decouples ICL's efficacy from a heavy reliance on sophisticated retrievers, successfully bridging the capability gap between different model generations. This provides a practical and scalable solution for real-world deployments where high-quality labeled data is unavailable. Our findings suggest that test-time adaptation is a computationally efficient supplement to traditional methods, enhancing model reliability across diverse tasks. This work contributes to the development of more resilient reasoning systems capable of maintaining performance in unpredictable environments.

6 Limitations

Despite the promising results, DEAL-ICL has the following limitations:

Inference Latency. A primary constraint of Test-Time Adaptation (TTA) methods is the computational overhead. Since DEAL-ICL involves **test-time gradient updates**, the inference speed is inevitably slower compared to standard frozen inference. Although we employ parameter-efficient techniques like LoRA to minimize memory usage, the backpropagation process still incurs additional time costs. Consequently, our current framework is best suited for scenarios where accuracy is prioritized over strict real-time latency.

Scope of Tasks. Our current evaluation primarily focuses on discriminative tasks (e.g., classification and reasoning selection) where consistency is easier to quantify. Extending the **Dual-View Consistency** mechanism to open-ended generation tasks (e.g., summarization or machine translation) remains challenging, as measuring semantic invariance in free-form text is non-trivial. We leave the adaptation of DEAL-ICL to generative tasks as a direction for future work.

References

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2023. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7):146:1–146:39.

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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 15159–15177. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1107–1128. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

Ye Jiang, Taihang Wang, Youzheng Liu, Yimin Wang, Yuhan Xia, and Yunfei Long. 2025. Learn to select: Exploring label distribution divergence for in-context demonstration selection in text classification. *Preprint*, arXiv:2511.10675.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. Metaicl: Learning to learn in

context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2791–2809. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1389–1403. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 271–278. ACL.

Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, Yancheng Yuan, and Dacheng Tao. 2025. Enhancing input-label mapping in in-context learning with contrastive decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 997–1004. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9090–9101. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 1752–1767. Association for Computational Linguistics.

Ru Wang, Wei Huang, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2025. Self-harmony: Learning to harmonize self-supervision and self-play in test-time reinforcement learning. *CoRR*, abs/2511.01191.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jerry W. Wei, Le Hou, Andrew K. Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. 2023. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 968–979. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

Zhenyu Wu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 2023a. Openicl: An open-source framework for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 489–498. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023b. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1423–1436. Association for Computational Linguistics.

Yijie Xu, Huizai Yao, Zhiyu Guo, Weiyu Guo, Pengteng Li, Aiwei Liu, Xuming Hu, and Hui Xiong. 2025. You only need 4 extra tokens: Synergistic test-time adaptation for llms. *CoRR*, abs/2510.10223.

Mingfeng Xue, Dayiheng Liu, Wenqiang Lei, Xingzhang Ren, Baosong Yang, Jun Xie, Yidan Zhang, Dezhong Peng, and Jiancheng Lv. 2023. Dynamic voting for efficient reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3085–3104, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Hanlin Zhang, Yifan Zhang, Yaodong Yu, Dhruv Madeka, Dean P. Foster, Eric P. Xing, Himabindu Lakkaraju, and Sham M. Kakade. 2024. A study on the calibration of in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6118–6136. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, and 1 others. 2025. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*.

A Dataset Statistics

Table 4 summarizes the statistics of the Natural Language Understanding (NLU) datasets used in our evaluation, including the task type, number of classes, and the size of data splits.

B Prompt Templates

We employ specific prompt templates and label mappings for each dataset to adapt the pre-trained model to downstream tasks. For multi-part datasets (MNLI and QNLI), we construct the input text by combining the components with appropriate formatting. Table 5 lists the exact input templates and verbalizers (label mappings) used in our experiments.

C Detailed Voting Example for Stage 1

C.1 Original Sample

Original Text:

> Money Funds Fell in Latest Week (AP)
> AP - Assets of the nation's retail money market mutual funds fell by $1.17 billion in the latest week to $849.98 trillion, the Investment Company Institute said Thursday.

Label: Business (Class 2)

C.2 Generated Paraphrases (N=10)

1. Retail Money Market Assets Decline - The Investment Company Institute reported Thursday that nation's retail money market mutual fund assets decreased by $1.17 billion, reaching $849.98 trillion this week.

2. Weekly Drop in Money Market Funds - According to Thursday's announcement from the Investment Company Institute, retail money market mutual fund assets fell to $849.98 trillion, down $1.17 billion.

3. Money Market Mutual Funds See Reduction - This week witnessed a $1.17 billion decrease in retail money market mutual fund assets,

Dataset	Task	# of Classes	Data Split
SST-2	Sentiment Classification	2	6920/872/1821
SST-5	Sentiment Classification	5	8544/1101/2210
CR	Sentiment Classification	2	3394/0/376
Subj	Subjectivity Analysis	2	8000/0/2000
AgNews	Topic Classification	4	120000/0/7600
MNLI	Natural Language Inference	3	392702/19647/19643
QNLI	Natural Language Inference	2	104743/5463/5463

Table 4: Details of the NLU datasets used in our experiments. The Data Split column denotes the number of samples in Train/Dev/Test sets respectively.

Dataset	Input Template	Label Mapping (Verbalizers)
SST-2	Review: *{text}* \n Sentiment:	0: negative, 1: positive
SST-5	Review: *{text}* \n Sentiment:	0: terrible, 1: bad, 2: okay, 3: good, 4: great
CR	Review: *{text}* \n Sentiment:	0: negative, 1: positive
Subj	Input: *{text}* \n Type:	0: objective, 1: subjective
AgNews	Input: *{text}* \n Type:	0: world, 1: sports, 2: business, 3: technology
MNLI	*{text}* \n Prediction:	0: entailment, 1: neutral, 2: contradiction
QNLI	*{text}* \n Answer:	0: entailment, 1: not_entailment

Table 5: Prompt templates and label mappings used in our experiments. *{text}* represents the input sequence from the dataset. For MNLI, we construct *{text}* as "Premise: {premise}\nHypothesis: {hypothesis}"; for QNLI, we construct *{text}* as "Question: {question}\nSentence: {sentence}". The '\n' symbol denotes a newline character. In few-shot settings, we prepend the instruction *"Choose one from [{label_list}]. Answer with one word only."*, where {*label_list*} is the comma-separated list of verbalized labels.

settling at $849.98 trillion, per the Investment Company Institute.

4. Investment Institute Reports Fund Decline - The nation's retail money market funds experienced a $1.17 billion reduction this week, with total assets at $849.98 trillion.

5. Latest Week Shows Money Fund Drop - Retail money market mutual fund holdings fell by $1.17 billion to reach $849.98 trillion, the Investment Company Institute disclosed.

6. Money Fund Assets Show Minor Fluctuation - This week retail market funds decreased by approximately one billion dollars. (Semantic drift)

7. Market Fund Reports Changes - Investment institutions stated that money markets have adjusted. (Information loss)

8. Thursday's Financial Report - Certain fund data showed variations. (Severe semantic drift)

9. Money-Related News - Financial markets have updates this week. (Label confusion)

10. Fund Information Update. (Severe error)

C.3 Consistency Voting Results

C.3.1 Voting Rules

For each paraphrase candidate $\hat{x}$, we conduct **T=20** stochastic forward passes using Monte Carlo Dropout. Each pass must satisfy three criteria: (1) **Correctness**: Predicted class = Ground truth label (Business); (2) **Confidence**: Maximum probability > 0.75; (3) **Discriminability**: Probability margin between top-2 classes > 0.15.
A pass is counted as a valid vote (✓) only if all three criteria are simultaneously satisfied; otherwise, it is marked as invalid ($\times$).

C.3.2 Detailed Voting Table

Table 8 presents the complete voting record for all 10 paraphrases across 20 Monte Carlo Dropout

Hyperparameter	Symbol	Value	Description
MC Passes	T	20	Stochastic forward passes
Confidence	ξ_{conf}	0.75	Min. prediction prob.
Margin	ξ_{margin}	0.15	Top-1/Top-2 gap
Acceptance	τ_{cons}	0.70	Min. valid vote ratio
Paraphrases	N	10	Generation count

Table 6: Key hyperparameters for Stage 1.

ID	Score	Rejection Reason
P6	0.55	Borderline; confidence drops below threshold
P7-P8	0.40-0.25	Semantic drift introduced ambiguity
P9-P10	0.15-0.05	Severe paraphrasing errors

Table 7: Analysis of rejected samples.

passes. Each cell indicates whether a stochastic forward pass yielded a valid vote (✓) or invalid vote (×) based on the three-criteria validation described above. The consistency score S_{cons} is computed as the fraction of valid votes (Eq. 1 in the main paper). Paraphrases with $S_{\text{cons}} \geq 0.70$ are accepted into the dense pool D_{dense}.

Several patterns emerge from this table: (1) High-quality paraphrases (P1-P3) achieve near-perfect consistency ($S_{\text{cons}} \geq 0.90$), demonstrating that semantic preservation leads to stable predictions under stochastic perturbation. (2) Borderline cases (P5) exactly meet the threshold, indicating that the filter successfully captures the transition between acceptable and unacceptable semantic variation. (3) Low-quality paraphrases (P8-P10) exhibit erratic voting patterns with consistency scores below 0.25, confirming severe semantic drift or grammatical errors.

C.4 Analysis of Rejected Samples

The consistency filter rejects 50% of generated paraphrases (P6-P10) in this example, demonstrating its effectiveness as a quality gate. Table 7 categorizes the rejection reasons into three tiers based on consistency scores.

Borderline Failures ($S_{\text{cons}} \approx 0.55$): Characterized by subtle semantic drift, such as replacing precise quantities with vague approximations. This introduces ambiguity, causing significant fluctuation in the classifier's confidence across stochastic evaluations.

Moderate Failures ($S_{\text{cons}} = 0.25$–0.40): Involve more substantial semantic damage, including the omission of critical factual attributions or the use of imprecise terminology, thereby failing to preserve the original's semantic integrity.

Severe Failures ($S_{\text{cons}} < 0.15$): Represent catastrophic paraphrasing errors where the generated text bears minimal resemblance to the source, discarding nearly all factual content and rendering the sample unusable.

C.4.1 Boundary Case Analysis: P5 vs. P6

This analysis illustrates the effectiveness of our filtering mechanism by comparing a borderline accepted case (P5) with a rejected case (P6).

- **P5 (Accepted, $S_{\text{cons}} = 0.70$):** The paraphrase (e.g., "...fell by $1.17 billion to $849.98 trillion...") **preserves all key factual elements** (precise figures, source attribution) while employing restructured syntax. Consequently, it achieved 14 valid votes, meeting the acceptance threshold.
- **P6 (Rejected, $S_{\text{cons}} = 0.55$):** The paraphrase (e.g., "...decreased by approximately one billion dollars.") exhibits critical flaws: semantic drift (precise quantification replaced by vagueness), information loss (omission of total assets and source), and imprecise terminology. These issues resulted in only 11 valid votes, failing the threshold.

This contrast demonstrates that the consistency filter functions as a *semantic firewall*, effectively rejecting paraphrases with factual imprecision or ambiguity while accepting those with mere structural variation.

C.5 Key Hyperparameter Settings for Stage 1

Table 6 summarizes the key hyperparameters for the semantic expansion stage, which were calibrated to balance quality, diversity, and computational cost.

- **MC Dropout Passes (T=20):** Provides sufficient statistical power for reliable consistency estimation. Fewer passes (<15) yield unstable scores, while more passes (>20) offer diminishing returns.
- **Confidence Threshold ($\xi_{\text{conf}} = 0.75$):** Ensures valid votes correspond to high-confidence predictions. A lower threshold (<0.6) admits unreliable predictions, while a higher one (>0.85) becomes overly restrictive.
- **Margin Threshold ($\xi_{\text{margin}} = 0.15$):** Enforces a clear probability gap between the top-two classes to prevent accepting ambiguous paraphrases where the classifier is confused.

ID	1-5					6-10					11-15					16-20					Valid	S	OK?
P1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	20	1.0	Y
P2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	19	0.95	Y
P3	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	18	0.9	Y
P4	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	×	16	0.8	Y
P5	✓	✓	✓	×	✓	✓	×	✓	✓	×	✓	×	✓	✓	×	✓	✓	×	✓	✓	14	0.7	Y
P6	✓	✓	×	✓	×	✓	×	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	✓	11	0.55	N
P7	✓	×	×	✓	×	×	✓	×	×	✓	×	✓	×	✓	×	✓	×	×	✓	×	8	0.4	N
P8	×	×	✓	×	×	×	×	✓	×	×	✓	×	×	×	✓	×	×	✓	×	×	5	0.25	N
P9	×	✓	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	✓	×	×	×	3	0.15	N
P10	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	1	0.05	N

Table 8: Detailed voting results across 20 Monte Carlo Dropout passes. ✓ denotes a valid vote (all three criteria satisfied), × denotes an invalid vote. Acceptance criterion: $S_{\text{cons}} \geq 0.70$ (at least 14 valid votes). Result: 5 out of 10 paraphrases passed (P1-P5). Column **S** represents consistency score, **OK?** indicates acceptance (Y=Yes, N=No).

- **Acceptance Threshold ($\tau_{\text{cons}} = 0.70$):** Requires a paraphrase to satisfy all criteria robustly across most dropout passes. This achieves a 70% acceptance rate, enabling a 7x data expansion while preserving high label fidelity.

- **Paraphrases per Sample (N=10):** Generating 10 candidates ensures sufficient diversity for the filter to select from, yielding 7 accepted paraphrases per seed sample after filtering.