Robust LLM Data Selection via Content-Invariant Gradient Attribution

Anonymous ACL submission

Abstract

Gradient-based data attribution has become an important tool for high-quality data selection in the instruction fine-tuning for large language models (LLMs), as it estimates the contribution of training examples through their influence on model optimization. However, we find that existing global gradient attribution methods suffer from surface-form bias: semantically equivalent instances with different formatting can produce highly inconsistent attribution scores, making data selection sensitive to superficial prompt variations. To address this issue, we propose Content-Invariant Gradient Attribution (CIGA), a robust framework for instruction data selection. CIGA first identifies layers whose gradients are both influential for attribution and stable under surface-form perturbations, thereby reducing formatting noise in gradient features. It then dynamically fuses these attribution signals with semantic representations of the input, combining optimization-level influence with content-level information. Extensive experiments across multiple benchmarks show that CIGA consistently improves data selection quality over strong attribution-based and similarity-based baselines while reducing attribution cost. Further analyses demonstrate that CIGA produces more stable rankings under format perturbations, offering a robust and efficient approach to data selection for LLM instruction tuning.

1 Introduction

In the instruction fine-tuning of large language models (LLMs), constructing a high-quality instruction dataset is a core challenge that determines the model's alignment capabilities and downstream task performance (Albalak et al., 2024). Traditional paradigms rely heavily on massive amounts of data; however, as model scale expands, automatically constructed vast data pools are inevitably filled with formatting errors and low-information redundant samples. Recent studies have shown that blindly increasing the volume of training data does not yield performance gains and can sometimes even lead to a degradation in instruction-following capabilities; conversely, utilizing an extremely small amount of curated, high-quality data (Xia et al., 2024) can stimulate superior model performance. Therefore, exploring automated and precise data selection methods has become an urgent need in the current LLM post-training phase.

Among existing data selection paradigms, gradient attribution-based methods (e.g., influence functions) (Xia et al., 2024) have demonstrated tremendous application value. By calculating the cosine similarity between training data and target validation data in the direction of parameter updates, these methods can measure the actual utility of the data for model optimization. However, this computational paradigm, which relies heavily on global gradients, exhibits a critical flaw: *Surface Form Bias*. Experiments reveal that when only the instruction format of the validation data is perturbed, existing gradient methods recall drastically different subsets of training data, as shown in Figure 2. This renders the selection results extremely fragile, lacking robustness and consistency. Our insight is that **the impact of truly high-quality training data on enhancing model capabilities should be driven by core semantics, rather than being constrained by surface-level instruction formats.**

In addition, there are data selection paradigms based on content similarity (Yang et al., 2025). These utilize pure content similarity metrics (e.g., semantic embeddings) to identify samples that semantically overlap with the validation data. However, effective data selection requires the ability to filter out data that hold enhancement value for the model (Jiang et al., 2025). Although similarity-based methods maintain consistency in surface content and are effective for tasks with prominent surface features, they are completely divorced from the model's actual training dynamics. Thus, they

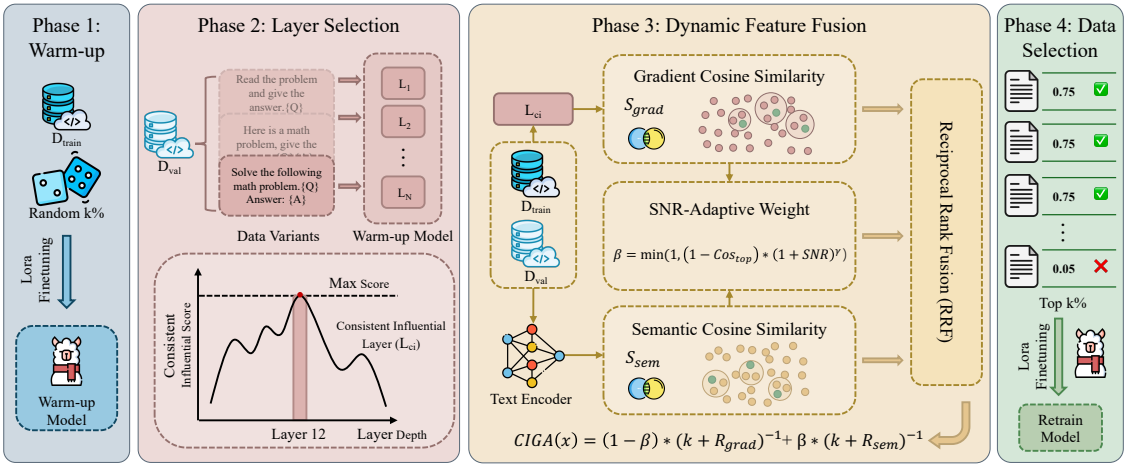

Figure 1: Overview of the CIGA framework. The overall architecture of this process mainly includes four stages: (1) warm-up training; (2) consistent influential layer selection, using semantically equivalent prompt variants to locate consistent influential layer (L_{ci}); (3) dynamic feature fusion, extracting gradient features as well as semantic features, and dynamically combining them through RRF; (4) final selection of the optimal Top k% data subset.

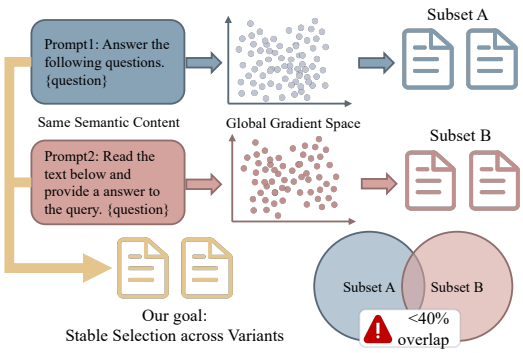

Figure 2: In gradient-based methods, prompt variations can lead to unstable data selection; semantically identical prompts map to disparate global gradients, yielding subsets with minimal overlap.

cannot guarantee that the selected data will practically facilitate model parameter updates.

To summarize, existing data selection methods fall into a dilemma: utility-aware gradient methods are unstable, while content-aware retrieval methods are disconnected from parameter dynamics. To break this impasse, we propose a framework, **Content-Invariant Gradient Attribution (CIGA)**. Prior research indicates that different network layers in LLMs play distinctly different roles in representation learning: intermediate layers often serve as the alignment hub for core semantics and task logic (Meng et al., 2022; Valeriani et al., 2023). Building on this, CIGA abandons global gradients. First, we dynamically locate the consistent influential layer across prompt variants of the same validation data, identifying the specific layer least perturbed by format variations yet exhibiting strong data influence. Computing similarity exclusively on this layer ensures robust gradient features. Finally, dynamic Reciprocal Rank Fusion (RRF) merges this gradient utility with semantic similarity, combining optimization-level influence with content-level information to select a high-quality data subset. Furthermore, unlike trajectory-based gradient methods that rely on multiple checkpoints, CIGA achieves superior performance using a single checkpoint, drastically reducing attribution costs.

The main contributions are as follows:

(i) Revealing surface form bias in data attribution: We systematically demonstrate the instability in global gradient attribution, highlighting the limitations of relying solely on global gradients.

(ii) Introducing a dynamic layer selection method: We introduce a dynamic layer selection mechanism to identify the content-invariant influential layer, reducing formatting noise and ensuring robustness while preserving influence.

(iii) Proposing the CIGA framework for effective data selection: We introduce the CIGA framework to integrate gradient-based attribution with semantic features. Extensive experiments demonstrate that this approach can select more valuable training data across various downstream tasks.

2 Related Work

2.1 Data Attribution

The core of data attribution lies in quantifying the true contribution of a single training sample to the model's prediction results or overall performance. Classical methods primarily rely on Influence Functions (IF) (Koh and Liang, 2017; Barshan et al., 2020), evaluating the parameter-level impact of removing a single data point by approximating the inverse of the Hessian matrix. To overcome the computational bottleneck in the era of LLM s, recent studies have proposed various efficient approximation frameworks (Schioppa et al., 2022; Park et al., 2023; Pruthi et al., 2020).

In addition to gradient-reliant technical pathways, another crucial direction is introducing Data Shapley based on cooperative game theory (Ghorbani and Zou, 2019; Fotouhi et al., 2024) to guarantee theoretical fairness in data value allocation. In-Run Data Shapley (Wang et al., 2025a) cleverly utilizes Taylor expansion within a single backward pass to achieve efficient approximate computation.

Regarding evaluation paradigms, the theoretical validity of data attribution algorithms typically requires rigorous verification through Leave-One-Out (LOO) retraining. However, due to the astronomically high computational cost of repeatedly retraining LLMs, current research also relies on utility evaluations based on downstream tasks. Numerous works indirectly confirm the reliability and guiding value of their attribution scores by examining the performance of attribution algorithms in practical application scenarios such as hallucination detection, fact tracing, and high-quality Data Selection(Chen et al., 2024; Wu et al., 2024).

2.2 Data Selection

In the instruction tuning of LLMs, data selection methods driven by gradients are dedicated to measuring the parameter-level utility of a single data for the target task. Regarding gradient-based methods for computing the maximum gradient of massive datasets (Nikdan et al., 2025), recent research has proposed optimization strategies. For example, some studies perform gradient feature clustering in the early stages of training (Wang et al., 2025b). Beyond computational efficiency, inherent noise similarly constrains the performance of gradient methods. The G-SNR method (Yuan et al., 2026) proposes using the degree of gradient decay and gradient variance of samples during model fine-tuning to reflect learning value and uncertainty, respectively. However, although these improvements have enhanced selection efficiency, they generally rely only on global gradient features and struggle to capture the true semantic relationships within the data, causing the selection results to lack robustness against variations in expression forms.

Another mainstream branch of research has shifted towards refined selection from the perspectives of internal model mechanisms and feature information theory (Chen et al., 2026; Duan et al., 2025; Dong et al., 2026). T-SHIRT (Fu et al., 2025) exclusively focuses on highly informative key tokens within the response sequence and introduces neighborhood perturbations to filter out redundant samples that rely on spurious lexical features. Meanwhile, methods based on the ICL single-shot loss difference (Jiang et al., 2025) have also been proven effective in quantifying the value of specific data in bridging model capability gaps.

In summary, the current fields of data selection face a dilemma: methods based on global gradients, while closely reflecting the true utility of parameter updates, are highly susceptible to surface form bias; whereas semantic methods based on internal representations, while focusing on the content, often remain detached from actual gradient dynamics.

3 Methods

The core workflow of CIGA comprises two primary stages: identifying the consistent influential layer guided by parameter update dynamics and dynamically fusing these gradient signals with surface semantic features.

3.1 Warm-up Fine-tuning

Directly extracting gradients from an unaligned pretrained model cannot accurately reflect the dynamics during training on the current candidate dataset (Xia et al., 2024). Therefore, before extracting gradient signals, we first randomly sample a $k\%$ subset from the full training set $\mathcal{D}_{train}$ to serve as the warm-up data $\mathcal{D}_{warmup}$.

By fine-tuning the base model M_θ on $\mathcal{D}_{warmup}$, we obtain the warmed-up model M_{θ^*}. This step ensures that the model successfully transitions past the initial phase of volatile optimization and enters a regime of relatively stable parameter adjustments. Consequently, this stability guarantees that the extracted gradients can faithfully and accurately reflect the true influence of the training data.

3.2 Consistent Influential Layer Selection

Different Transformer layers in LLMs have different roles when processing tasks. To identify layers that both reflect data influence and exhibit stability, we propose a consistent influential score based on the statistical properties of gradient norms.

For each data $x_i \in \mathcal{D}_{val}$, we construct m semantically equivalent prompt variants $\{p_j\}_{j=1}^{m}$. Feeding $p_j(x_i)$ into the warm-up model M_{θ^*}, we compute the gradient L_2 norm for the l-th layer:

$$n_{i,l}^{(j)} = \|\nabla_{\theta_l} \mathcal{L}(M_{\theta^*}(p_j(x_i)), y_i)\|_2 \quad (1)$$

We first calculate the mean $\mu_{i,l}$ and standard deviation $\sigma_{i,l}$ of $n_{i,l}^{(j)}$ across all m variants for each sample x_i. We then average these values across the entire validation set to obtain the layer's global μ_l and σ_l. Conceptually, μ_l reflects the magnitude of the data's impact on the parameters, representing its overall influence at layer l, while σ_l captures the instability introduced by different prompt variants. To eliminate the scale discrepancies between the μ_l and σ_l, we apply Z-score normalization to derive $\hat{\mu}_l$ and $\hat{\sigma}_l$. The final consistent influential score S_l^{ci} for layer l is defined as:

$$S_l^{ci} = \hat{\mu}_l - \alpha \cdot \hat{\sigma}_l \quad (2)$$

where α penalizes variance across variants.

The layer with the highest score is selected as the consistent influential layer, denoted as l_{ci}.

3.3 Dynamic Feature Fusion

After identifying the target layer l_{inv}, we compute the relevance between the training data x_{train} and the validation data x_{val} from two dimensions: parameter dynamics and semantics.

Grad-Score: We extract the gradients from L_{ci}, flattening and concatenating them into a unified gradient vector. For efficient storage, dimensionality reduction is applied in practical implementations to obtain a gradient vector G. We then calculate the cosine similarity, denoted as S_{grad}, to measure the consistency between the training data and the validation data in the direction of parameter updates.

Sem-Score: Using a external text embedding model, we extract the text representation E of the data. We then calculate the cosine similarity in the semantic representation space to reflect the surface expression similarity of the data, yielding S_{emb}.

Because gradient and semantic similarities have misaligned distributions, direct addition is unstable. Thus, we compute the final CIGA score via RRF. Let $R_{grad}(x)$ and $R_{sem}(x)$ be the rankings of the candidate data based on the Grad-Score and Sem-Score, respectively. The CIGA score is:

$$\text{CIGA}(x) = (1-\beta) \cdot \frac{1}{r + R_{grad}(x)} + \beta \cdot \frac{1}{r + R_{Sem}(x)} \quad (3)$$

where r is a smoothing constant. The dynamic weight $\beta \in [0, 1]$ controls the degree of reliance on semantic features during fusion.

Dynamic Fusion Ratio: A static combination ratio (e.g., $\beta = 0.5$) intrinsically assumes that gradient signals and semantic features are equally reliable across all tasks, which is empirically suboptimal. We posit that the reliance on semantic content should be dynamically adjusted based on the reliability of the gradient signals. To quantify this reliability, we introduce the gradient Signal-to-Noise Ratio (SNR).

Let $G_i^{(j)}$ denote the extracted gradient vector of the i-th data sample under the j-th prompt variant. We define noise (N) as the expected within-group gradient variance $\mathbb{E}_i \left[\text{Var}_j(G_i^{(j)})\right]$, which reflects the instability induced by diverse prompt formats for the same data. Meanwhile, we define the semantic signal (S) as the expected between-group gradient variance $\text{Var}_i \left(\mathbb{E}_j[G_i^{(j)}]\right)$, which captures the gradient distinction across distinct data samples, reflecting that the model's gradient features contain stronger data-discriminative signals.

The Gradient SNR is then defined as $SNR = S/(N + \epsilon)$, where ϵ is a small smoothing constant.

Higher SNR indicates that the model has a clear understanding of the underlying logic for the task, and the gradient direction is highly resistant to interference. In this case, the gradient features should be given higher confidence (reducing β).

Lower SNR indicates that the model's logical signal strength is weak and needs to rely on semantic content as a supplement. However, if at this time the top Sem-Score of the data are very high, it indicates that a large amount of content with very high surface-level similarity has been retrieved. In this case, relying too heavily on semantic content will reduce the diversity of the selected data.

Therefore, letting Sem^{Top} be the average of the top K semantic scores, the dynamic weight β for the validation set is defined as:

$$\beta = \min\left(1, \frac{1 - Sem^{Top}}{(1 + SNR)^{\gamma}}\right) \quad (4)$$

where γ is a hyperparameter that controls the SNR sensitivity, and the denominator $1 + SNR$ ensures numerical stability when the SNR is small.

3.4 Data Selection and Retraining

For any given validation data, we calculate the CIGA scores for all candidate samples in the full training set $\mathcal{D}_{train}$, and select the Top-$k\%$ high-value data with the highest scores. Finally, we retrain the original base model M_θ using this high-quality subset, which balances logical alignment and semantic diversity, to achieve alignment and enhancement of specific capabilities.

4 Experiment

4.1 Experimental Setup

Datasets. We construct a diverse instruction tuning training datasets including FLAN-v2 (Longpre et al., 2023), CoT (Wei et al., 2022), Dolly (Conover et al., 2023), OASST1 (Köpf et al., 2023), Alpaca-GPT4 (Peng et al., 2023), and GSM8K (Cobbe et al., 2021). We evaluate across five rigorous benchmarks: TyDiQA (Clark et al., 2020), BBH (Suzgun et al., 2023), GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2021), and HellaSwag (Zellers et al., 2019). Detailed dataset statistics are provided in Appendix A.1.

Models. We employ LLaMA-2-7B (Touvron et al., 2023) and Qwen3-4B-Base (Team, 2025) as our foundation models for both data attribution and subsequent instruction fine tuning.

Baselines. We compare CIGA against baselines below: (1) *Semantic Matching:* cosine similarity via dense retrieval (Reimers and Gurevych, 2019); (2) *Gradient Methods:* LESS (Xia et al., 2024) and TracIn (Pruthi et al., 2020); (3) *Information based Methods:* IFD (Li et al., 2024), T-SHIRT (Fu et al., 2025), and SelectIT (Liu et al., 2024). Detailed information are provided in Appendix A.2.

Evaluation Metrics. We assess the selected subsets via two dimensions: *Downstream Performance:* We report benchmark specific metrics (e.g., accuracy or exact match) after retraining models exclusively on the top $k\%$ selected data. *Robustness:* To quantify surface form bias, we measure the overlap rate (Jaccard similarity (Jaccard, 1912)) among top $k\%$ subsets retrieved across multiple semantic equivalent prompt variants.

Hyperparameter Settings. In the dynamic fusion calculation, the smoothing constant r defaults to the commonly used 60. Additionally, the data selection ratio k for all fine-tuning tasks is set to the Top 5% of the overall candidate dataset.

Implementation Details. We utilize Low-Rank Adaptation (LoRA) (Hu et al., 2021) to ensure efficient feature extraction and parameter updates (see Appendix A.4). Notably, replacing the multi checkpoint ensemble used in prior methods (e.g., 4 checkpoints in LESS's main experiments) with a single warmup checkpoint allows CIGA to accelerate the feature extraction process, yielding a 75% reduction in time cost. All core experiments, including model fine-tuning and gradient computation, were conducted on an Ubuntu Linux server equipped with 4 NVIDIA A100 GPUs (80GB).

4.2 Experimental Result

4.2.1 Downstream Performance Results

To comprehensively evaluate the quality of the data selected by the CIGA framework, we fine-tuned the base models on the Top-5% data subsets selected by different algorithms and tested their final performance across mainstream benchmarks. The experimental results are presented in the Table 1. "Full" represents training on the full dataset, while CIGA-G employs data selected exclusively via the Grad-Score without incorporating the Sem-Score.

Overall Performance Superiority: Compared to all gradient-based and heuristic data selection baselines, CIGA demonstrates the most robust and balanced comprehensive performance, achieving an average improvement of 2.11% on LLaMA-2 and 1.03% on Qwen3. Particularly on TyDiQA, CIGA achieves the highest overall scores. This proves the effectiveness of our selection method.

Validation of Semantic and Gradient Fusion: We observe that the Cos method demonstrates a substantial advantage on the GSM8K test set. This is likely because this dataset primarily consists of mathematical word problems, making their structural and semantic features particularly pronounced in the embedding space. Leveraging our dynamic feature fusion, the CIGA framework successfully incorporates the strengths of the Cos method in capturing semantic features, significantly outperforming traditional gradient-only methods on this task. More importantly, across other tasks where the Cos method underperforms, CIGA can adaptively rely more heavily on gradient feature.

Effectiveness Across Model Frameworks: We evaluated our framework on two distinct model families with different parameter scales: LLaMA-

2-7B and Qwen3-4B-Base. The results demonstrate that CIGA maintains strong generalizability across different architectures, consistently delivering positive gains over the random baseline for both models (+2.11% and +1.03%, respectively).

Approaching or Surpassing Full-Data Fine-Tuning: Comparing CIGA against full-data fine-tuning reveals a critical observation regarding model capacity. As a more capable base model, Qwen3-4B-Base is particularly sensitive to data quality. Fine-tuning it on the entire dataset actually degrades overall performance compared to random selection (-1.16%), indicating that blindly incorporating large volumes of data may undermine the original capabilities of powerful models. By effectively filtering out this noise, CIGA outperforms full-data fine-tuning. On LLaMA-2-7B, although using all data yields a slightly higher average score due to specific reasoning tasks like GSM8K, CIGA achieves highly competitive results with merely 1/20th of the computational budget. Furthermore, it even surpasses the full dataset on TyDiQA, MMLU, and HellaSwag. This strongly validates the "less is more" paradigm and highlights CIGA's efficiency.

4.2.2 Selection Consistency Across Variants

To verify whether our method overcomes the surface form bias inherent in gradient-based methods, we measured the Jaccard similarity of the Top-5% training subsets retrieved using different prompt variants for the same validation data.

As shown in Table 2, results demonstrate that CIGA exhibits significantly higher selection consistency, achieving an average overlap rate of 58.33% and 48.68%, which far exceeds LESS and TracIn. This strongly supports the following conclusions:

Superior robustness over gradient methods: CIGA outperforms global gradient methods across knowledge intensive (MMLU) and logical reasoning (BBH, GSM8K) tasks. This indicates that by targeting more stable layers and incorporating semantic content, CIGA can retrieve more consistent training data for validation data with the same content, thereby achieving more robust data selection.

Overcoming surface form bias: On the HellaSwag dataset, CIGA achieves an overlap rate (49.29% and 36.6%) that is approximately twice as high as that of LESS (22.64% and 18.47%). This sharp contrast reveals the inherent fragility of global gradient methods when subjected to varying prompt expressions. Ultimately, CIGA establishes that isolating format noise enables robust data attribution driven by semantics.

4.2.3 Ablation Experiment

Consistent Influential Layer Selection. To independently evaluate the contribution of the dynamic layer selection algorithm, we introduce the CIGA-G, which omits the Sem-Score and relies solely on the Grad-Score extracted from the consistent influential layer for data selection.

As demonstrated by the experimental results in Table 1, CIGA-G surpasses traditional global gradient baselines on most tasks, even when semantic features are excluded. For instance, on the TydiQA benchmark, CIGA-G achieves 55.27%, surpassing LESS's 54.65%. This phenomenon strongly demonstrates that traditional methods aggregating gradients across all layers often introduce unnecessary noise, whereas CIGA-G achieves highly precise data selection by isolating consistent and influential intermediate layers.

Semantic Feature Fusion. Furthermore, comparing CIGA-G with the complete CIGA framework demonstrates that dynamic feature fusion yields further improvements. This validates our hypothesis: while intermediate layer gradients already possess a certain degree of robustness, integrating semantic features effectively compensates for the limitations of pure gradient representations.

5 Discussion

5.1 Consistent Influential Score Analysis

To verify if the consistent influential score S_l^{ci} reflects practical utility of each Transformer layer, we compared it against downstream metrics of models trained on data selected via layer Grad-Score (sampled every three layers), as Figure 3 illustrates.

Precise Localization of CIGA: The global peak of the consistent influential score (approximately at Layer 15) corresponds to the absolute high-performance interval of the model's downstream performance. This proves that CIGA can locate the true data value layer inside the model, thereby eliminating part of the noise in subsequent gradient similarity calculations and ensuring an advantage in gradient features for high-quality data selection.

Semantic Dominance of Middle Layers: The figure shows that the best evaluation performance and the highest consistent influential scores are overlapped in the middle layers of the model. This phenomenon confirms our hypothesis in the Section 1 and the consensus in current LLM inter-

Table 1: Performance comparison across test datasets (%). Δ_{Rand} is the improvement over Random. Standard deviations over multiple runs are reported in parentheses. **Bold** and underline indicate the best selection result and performance surpassing "Full", respectively. Parentheses show standard deviations over three runs.

Method	LLaMA-2-7B						Qwen3-4B-Base					
	TydiQA	MMLU	BBH	GSM	Hella	Δ_{Rand}	TydiQA	MMLU	BBH	GSM	Hella	Δ_{Rand}
Reference Baselines												
Full	50.77 (0.3)	47.14 (0.2)	43.80 (0.8)	25.17 (1.0)	74.15 (0.3)	+2.48	66.34 (0.4)	72.89 (0.3)	68.52 (0.9)	78.09 (1.1)	70.40 (0.3)	-1.16
Random	49.54 (0.5)	45.51 (0.4)	**41.85** (0.6)	17.29 (1.6)	74.47 (0.4)	-	73.16 (0.5)	73.01 (0.4)	67.13 (1.3)	78.01 (0.9)	70.52 (0.4)	-
Heuristic & Semantic Methods												
Cos	48.31 (0.4)	46.06 (0.3)	40.93 (1.0)	**21.91** (0.7)	74.52 (0.3)	+0.62	72.23 (1.4)	73.23 (0.3)	67.41 (1.1)	78.85 (1.0)	70.69 (0.3)	+0.11
T-shirt	28.83 (0.8)	45.03 (0.4)	41.30 (1.2)	12.96 (1.6)	74.51 (0.4)	-5.20	58.90 (1.7)	73.11 (0.4)	69.54 (0.8)	60.50 (1.7)	70.61 (0.4)	-5.84
IFD	43.21 (0.6)	44.45 (0.5)	41.67 (0.7)	14.78 (0.5)	74.17 (0.4)	-2.08	71.58 (0.5)	72.86 (0.4)	67.87 (1.0)	79.15 (1.2)	70.44 (0.4)	+0.01
SelectIT	49.96 (0.7)	45.91 (0.4)	41.67 (0.5)	20.39 (0.6)	74.35 (0.3)	+0.72	71.37 (0.5)	73.21 (0.3)	67.87 (0.6)	77.03 (1.3)	**70.73** (0.3)	-0.33
Gradient-based Methods												
LESS	54.65 (0.5)	45.19 (0.8)	40.00 (1.1)	17.59 (1.2)	74.50 (0.4)	+0.66	73.38 (1.6)	**73.37** (0.4)	69.63 (0.5)	78.17 (1.6)	70.36 (0.4)	+0.61
TracIn	54.05 (1.2)	45.52 (0.5)	40.74 (1.2)	16.76 (1.1)	73.71 (0.6)	+0.43	72.05 (0.6)	72.67 (0.5)	68.33 (1.5)	78.32 (0.8)	69.44 (0.5)	-0.21
Ours												
CIGA-G	55.27 (0.3)	46.56 (0.3)	40.37 (1.0)	19.11 (0.9)	74.43 (0.3)	+1.42	**74.22** (0.3)	73.09 (0.3)	**70.28** (1.1)	78.32 (0.9)	70.38 (0.3)	+0.89
CIGA	**56.17** (0.2)	**47.46** (0.2)	41.39 (0.6)	19.64 (1.0)	**74.54** (0.2)	**+2.11**	74.10 (0.2)	73.05 (0.2)	**70.28** (0.7)	**79.30** (1.2)	70.29 (0.2)	**+1.03**

Table 2: Jaccard similarity (%) across prompt variants. Higher values reflect greater format robustness.

Test	LESS	TracIn	CIGA (ours)
LLaMA-2-7B			
TydiQA	41.40	36.75	**42.40**
MMLU	62.16	58.50	**76.04**
BBH	68.94	69.86	**71.92**
GSM8K	46.07	51.96	**52.00**
HellaSwag	22.64	28.24	**49.29**
Avg	48.24	49.06	**58.33**
Qwen3-4B-Base			
TydiQA	**62.75**	37.29	50.03
MMLU	57.03	55.49	**68.06**
BBH	48.62	50.59	**67.11**
GSM8K	13.28	17.77	**21.60**
HellaSwag	18.47	17.25	**36.60**
Avg	40.03	35.68	**48.68**

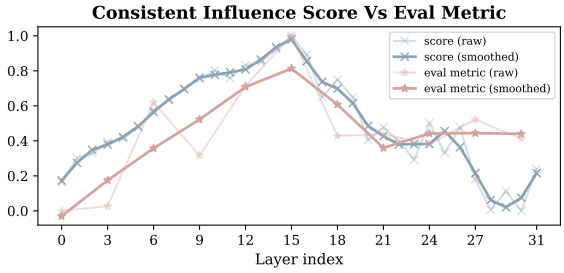

Figure 3: Correlation analysis between the consistent influence score and downstream evaluation metric.

pretability research: the shallow layers of the model are typically dedicated to processing low-level lexical and syntactic features, while the extremely deep layers are highly fitted to specific output formats and generation probability distributions. By contrast, the middle layers are more focused on semantic alignment, striking a balance between robustness and influence. They carry the core task-solving logic, and therefore achieve peak performance in both consistent influential scores and actual task performance.

5.2 Overcoming Surface Form Bias

To investigate the surface form bias in traditional gradient methods, we use t-SNE to visualize the gradient features of validation samples and their prompt variants. Figure 4 presents this 2D distribution for LLaMA-2-7B on the GSM8K task.

Observing global gradient methods reveals that, different prompt variants of the same validation data exhibit significant dispersion in the gradient space. This indicates that global gradient features are highly susceptible to superficial variations in the vocabulary or syntactic structure of the input text, leading to inconsistent parameter update directions for data sharing the same semantics.

In contrast, the gradients extracted from the consistent influential layer by the CIGA framework can partially mitigate the interference of such noise, even without incorporating semantic similarity. The variants of the same underlying sample are generally clustered much more tightly within the same neighborhood. This demonstrates that CIGA's strategy of exclusively extracting gradients from the intermediate layers effectively isolates the interference of surface forms, enabling gradient attribution to focus more accurately on the core logic and semantics underlying the data.

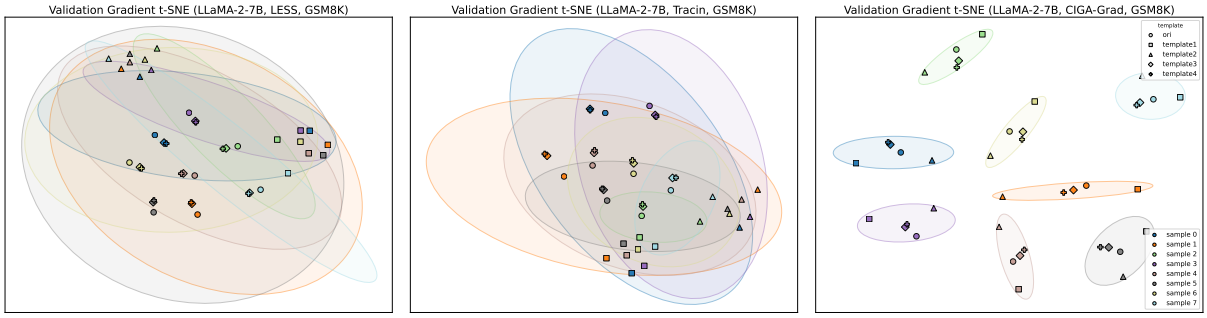

Figure 4: Comparison of gradient distributions via t-SNE.

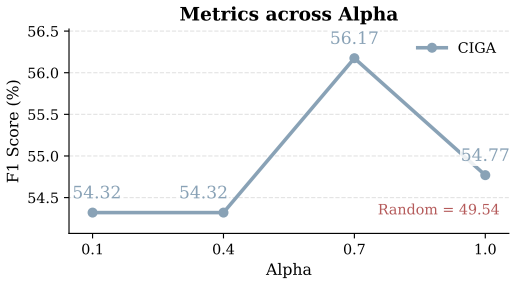

Figure 5: Sensitivity analysis of the variance penalty coefficient α in the CIGA layer selection mechanism.

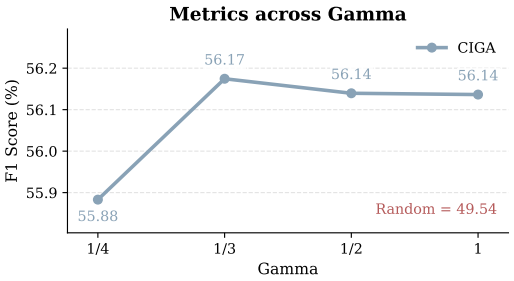

Figure 6: Sensitivity analysis of the semantic feature fusion weight γ during the comprehensive scoring stage.

5.3 Layer Selection Hyperparameter

In the CIGA layer selection mechanism, hyperparameter α penalizes layer gradient variance.

As shown in Figure 5, for small α (e.g., 0.1 and 0.4), performance stagnates around 54.3%. This indicates that an insufficient variance penalty causes layer scores to rely overly on the gradient mean, leading the algorithm to select influential but highly vulnerable layers disturbed by format variations. At $\alpha = 0.7$, the F1 score peaks near 56.2%. This moderate penalty optimally balances data influence and format robustness, allowing CIGA to precisely locate the consistent influential layer and filter out format noise. Conversely, increasing α to 1.0 degrades performance, as this excessive penalty erroneously discards genuinely influential layers over minor gradient fluctuations.

In summary, we adopt $\alpha = 0.7$ as the default setting in our experiments.

5.4 Feature Fusion Hyperparameter

The hyperparameter γ is used to adjust the fusion weight between gradient and semantic features.

When the value of γ is greater than 1/4, the metrics remain stable, maintaining an overall high performance range. This indicates that as long as a sufficiently strong semantic penalty is applied to samples with a high SNR, causing the algorithm to place more trust in gradient scores rather than blindly incorporating similar content, the model can achieve strong performance. However, reducing γ to $1/4$ degrades performance, as this excessively small γ value weakens the impact of high-confidence gradients and makes the RRF to introduce an excessive amount of semantic noise data.

Overall, we set γ to the stable value of 1/3.

6 Conclusion

In this paper, we delve into the limitations of traditional global gradient attribution methods in data selection tasks for LLMs, and reveal the interference of surface form bias in data value evaluation. To overcome this challenge, we propose the CIGA framework. By combining a layer selection algorithm with dynamic feature fusion, CIGA integrates gradient-based logic attribution signals with the underlying semantic alignment features, achieving a complement between the gradients significance and textual semantic features. Extensive benchmark experiments demonstrate that CIGA enhances high quality training data selection, consistently outperforming mainstream heuristic and gradient based baselines for instruction fine tuning.

Limitations

Although the CIGA framework demonstrates significant effectiveness and robustness in data selection for instruction fine-tuning, this study and the proposed method still exhibit the following limitations, which warrant further exploration and optimization in future work:

Inherent Computational Overhead: Despite CIGA substantially enhancing the noise robustness of gradients by precisely extracting consistent influential layers, as a gradient-based attribution framework, it inherently requires performing forward and backward passes across the entire candidate training corpus to compute and store the corresponding parameter gradients. When scaling to massive candidate data pools containing tens of millions of samples, this process incurs substantial temporal and computational overhead. Exploring methods to further compress gradient features or introduce efficient approximate computation techniques is crucial to alleviating this bottleneck.

Reliance on High-Quality Validation Sets: A core prerequisite for CIGA to accurately identify high-quality data is the utilization of a validation set from the target task to match core gradient and semantic features. While this mechanism endows the framework with robust task-specific adaptation capabilities, it also implies that for entirely novel target tasks lacking off-the-shelf validation sets, researchers or developers must incur additional costs to construct prior reference datasets. Investigating how to leverage unsupervised methods or rely on the model's intrinsic knowledge to evaluate the value of training data will be a primary focus of our future research.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. Relatif: Identifying explanatory training examples via relative influence. *CoRR*, abs/2003.11630.

Si Chen, Feiyang Kang, Ning Yu, and Ruoxi Jia. 2024. Fasttrack: Fast and accurate fact tracing for llms. *ArXiv*, abs/2404.15157.

Xin Chen, Junchao Wu, Shu Yang, Runzhe Zhan, Zeyu Wu, Min Yang, Shujian Huang, Lidia S. Chao, and Derek F. Wong. 2026. Neuron-aware data selection in instruction tuning for large language models.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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. *arXiv preprint arXiv:2110.14168*.

M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Victor Ye Dong, Kuan-Yun Lee, Jiamei Shuai, Shen Liu, Yi Liu, and Jian Jiao. 2026. Greedy information projection for llm data selection.

Feiyu Duan, Xuemiao Zhang, Sirui Wang, Haoran Que, Yuqi Liu, Wenge Rong, and Xunliang Cai. 2025. Enhancing llms via high-knowledge data selection. In *AAAI Conference on Artificial Intelligence*.

Milad Fotouhi, Mohammad Taha Bahadori, Oluwaseyi Feyisetan, Payman Arabshahi, and David Heckerman. 2024. Fast training dataset attribution via in-context learning. *CoRR*, abs/2408.11852.

Yanjun Fu, Faisal Hamman, and Sanghamitra Dutta. 2025. T-shirt: Token-selective hierarchical data selection for instruction tuning. *ArXiv*, abs/2506.01317.

Amirata Ghorbani and James Y. Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 2242–2251. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Tingyu Jiang, Shen Li, Yiyao Song, Lan Zhang, Hualei Zhu, Yuan Zhao, Xiaohang Xu, Kenjiro Taura, and Hao Henry Wang. 2025. Importance-aware data selection for efficient llm instruction tuning. In *AAAI Conference on Artificial Intelligence*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Proceedings of Machine Learning Research, pages 1885–1894. PMLR.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635, Mexico City, Mexico. Association for Computational Linguistics.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. SelectIT: Selective instruction tuning for LLMs via uncertainty-aware self-reflection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems 35*.

Mahdi Nikdan, Vincent Cohen-Addad, Dan Alistarh, and Vahab S. Mirrokni. 2025. Efficient data selection at scale via influence distillation. *ArXiv*, abs/2505.19051.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: attributing model behavior at scale. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 27074–27113. PMLR.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. 2022. Scaling up influence functions. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 8179–8186. AAAI Press.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. *ArXiv*, abs/2302.00294.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. 2025a. Data shapley in one training run. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zige Wang, Qi Zhu, Fei Mi, Minghui Xu, Ruochun Jin, and Wenjing Yang. 2025b. ClusterUCB: Efficient gradient-based data selection for targeted fine-tuning of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18867–18880, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kangxi Wu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2024. Enhancing training data attribution for large language models with fitting error consideration. In *Conference on Empirical Methods in Natural Language Processing*.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *ArXiv*, abs/2402.04333.

Jiaxi Yang, Wenlong Deng, Benlin Liu, Yangsibo Huang, James Zou, and Xiaoxiao Li. 2025. Gmvaluator: Similarity-based data valuation for generative models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhihang Yuan, Chen Yue, Longtao Huang, Litu Ou, and Lei Shi. 2026. Uncertainty-aware gradient signal-to-noise data selection for instruction tuning. *ArXiv*, abs/2601.13697.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Experimental Setup Details

A.1 Dataset Composition and Scale

To ensure our data selection framework is evaluated on a sufficiently diverse and massive candidate pool, our training corpus aggregates several widely adopted collections. Table 3 summarizes the approximate volume and primary domain of each dataset subset used in our experiments.

These diverse data sources collectively contribute to the foundation models' general instruction following capability, complex reasoning skills, and conversational alignment.

Regarding the licenses and intended use of these artifacts, we confirm that all public datasets and base models were utilized strictly for their intended academic research purposes (i.e., instruction tuning and model evaluation) in compliance with their original licenses. Upon acceptance, we will publicly release the CIGA codebase. This release is strictly intended to facilitate non-commercial academic research, ensuring full compatibility with the original access conditions and intended use of the source artifacts.

A.2 Detailed Baseline Descriptions

To comprehensively evaluate the effectiveness of the CIGA framework, we compare it against a diverse set of strong data selection baselines.

Semantic Matching Methods. This category relies on dense representation retrieval to find training samples semantically aligned with the validation set. We utilize the widely adopted all MiniLM-L6-v2 (Reimers and Gurevych, 2019) sentence transformer to encode both the training candidate pool and the validation sets into a shared embedding space. Data selection is subsequently performed by ranking the cosine similarity scores.

Gradient Methods. These methods attribute model performance directly to individual training samples by leveraging parameter gradients.

- **LESS** (Xia et al., 2024): A state of the art gradient attribution method that projects global parameter gradients into a lower dimensional space using random projections. It computes the influence score by aggregating gradient similarities across multiple saved training checkpoints. We followed LESS's methods for gradient computation and saving, but utilized only a single checkpoint; consequently, our efficiency is significantly higher than that of LESS.

- **TracIn** (Pruthi et al., 2020): A classic influence function approximation that traces the training trajectory. It estimates data utility by calculating the dot product of gradients between training and validation samples across multiple checkpoints, explicitly linking parameter updates to validation loss reduction.

Information based Methods. Instead of relying on heavy gradient computations, these methods select data based on intrinsic model feedback, such as perplexity, confidence, or generation difficulty during forward pass inference.

- **IFD** (Li et al., 2024): Instruction Following Difficulty identifies training samples that exhibit a high ratio of condition answer loss to direct language modeling loss. It operates on the premise that samples inherently difficult for the model to generate directly, yet coherent given the instruction, yield the highest training utility.

Table 3: Overview of the candidate training datasets, including their approximate volume and primary content focus.

Dataset	Volume	Primary Content Domain
FLAN v2 (Longpre et al., 2023)	100000	Zero shot and few shot NLP tasks
CoT (Wei et al., 2022)	100000	Step by step logical deduction
Alpaca GPT4 (Peng et al., 2023)	52002	High quality general instructions
Dolly (Conover et al., 2023)	15011	Human generated conversational records
OASST1 (Köpf et al., 2023)	55668	Multi turn dialogue sequences
GSM8K (Cobbe et al., 2021)	7473	Grade school mathematical reasoning

- **TSHIRT** (Fu et al., 2025): Evaluates instruction data quality by leveraging robust intrinsic metrics and model confidence scores to efficiently filter out noise and retain high quality instruction pairs.
- **SelectIT** (Liu et al., 2024): An advanced selection algorithm evaluating data based on instruction following difficulty and model confidence. It dynamically filters instruction tuning data by analyzing the foundation model's generation probabilities. Given the massive volume of our combined training corpus, we specifically reproduce the SelectIT model level score for the LLaMA-2-7B and LLaMA-2-13B architectures, discarding the LLaMA-2-70B score from the original paper.

A.3 Details of Prompt Variant Templates

To construct sufficient surface form variance during the feature extraction phase, we designed multiple sets of semantically equivalent prompt variants across various benchmark datasets. For each specific test task, five variant templates with different instruction expression formats (Template 1-5) are included. This appendix provides a detailed list of all the template formats used in our experiments.

In the templates below, placeholders are defined as follows: `{C}` represents the context or passage; `{Q}` represents the question or math problem; `{A}` represents the answer or ending continuation; `{S}` represents the specific subject in the MMLU task; and `{T}` represents the task instruction in the BBH task. Note that for the TyDiQA dataset, which is multilingual, we only list the English formats below; the corresponding translated versions were applied dynamically based on the language of the data in actual experiments.

A.4 Training Parameters

We fine-tune all models with LoRA. The LoRA rank is set to 128, the scaling factor alpha is set to 512, and the LoRA dropout rate is 0.1. Training is performed for one epoch with a maximum sequence length of 2048. We use AdamW with a learning rate of 2e-5, linear learning-rate scheduling, a warmup ratio of 0.03, and no weight decay.

B Hyperparameter Sensitivity Analysis

To more comprehensively evaluate the global impact of key hyperparameters in the CIGA framework on the final model performance, we supplement this appendix with the complete hyperparameter sensitivity experimental results on each test set, as shown in Table 5.

For the variance penalty coefficient α, we tested different values ranging from 0.1 to 1.0. For some coefficients, the selected layers were the same, and we reused the test results. The experiments indicate that when $\alpha = 0.7$, the model achieves the best or second-best performance on most tasks, such as TydiQA and MMLU, with the overall average score (47.74%) reaching the highest. Too small (e.g., 0.1) or too large (e.g., 1.0) α leads to an imbalanced penalty on variance fluctuations when locating content-invariant layers, thereby introducing additional formatting noise or losing effective information, which reduces the overall quality of data selection.

For the fusion ratio parameter γ, we compared performance under different scales ($\gamma \in \{1, 1/2, 1/3, 1/4\}$). The results show that when $\gamma = 1/3$, the gradient-based logical attribution signals and the underlying text semantic features achieve the best dynamic complementarity, reaching local optima on the TydiQA, MMLU, and HellaSwag tasks, and achieving the highest overall average score.

These detailed test results strongly validate the reasonableness of setting $\alpha = 0.7$ and $\gamma = 1/3$ as the default empirical configuration in our main experiments.

Table 4: Details of prompt variant templates across different benchmark datasets.

Dataset	Template	Content
TyDiQA	Template 0	Answer the following question based on the information in the given passage. Passage: {C}\nQuestion: {Q}\nAnswer: {A}
	Template 1	Read the text below and provide a answer to the query. Text: {C}\nQ: {Q}\nA: {A}
	Template 2	Given the background context, please extract the answer. Context: {C}\nQuery: {Q}\nResponse: {A}
	Template 3	Reference: {C}\nQuestion: {Q}\nResult: {A}
	Template 4	Answer the following questions. {C}\n{Q}\n{A}
MMLU	*Full Format*	{prefix}\n\n{Q}\nA. ...\nB. ...\nC. ...\nD. ...\nThe answer is: {A}
	Template 0	The following are multiple choice questions (with answers) about {S}.
	Template 1	Choose the correct answer (A, B, C, or D) for the question about {S}.
	Template 2	Multiple choice question on {S}:
	Template 3	Select the best option for the following question.
	Template 4	Answer the following question about {S}.
BBH	*Full Format*	{task_instruction}\n\n{Q}\nA: {A}
	Template 0	{T}
	Template 1	Read the instructions, then answer the question below.\n\n{T}
	Template 2	Below is a reasoning task. Answer the final question.\n\n{T}
	Template 3	Answer the following reasoning question.\n{T}
	Template 4	Answer the question below.
GSM8K	Template 0	Solve the following math problem.\n\n{Q}\nAnswer: {A}
	Template 1	Solve the math problem below.\n\n{Q}\nSolution: {A}
	Template 2	Here is a math problem, give the answer:\n\n{Q}\n {A}
	Template 3	Work through the problem and give the answer.\n\n{Q}\nAnswer: {A}
	Template 4	Read the problem and give the answer.\n\n{Q}\nAnswer: {A}
HellaSwag	Template 0	{C} {A}
	Template 1	Read the context and complete it with the best ending.\n\n{C} {A}
	Template 2	Write the ending that best continues the passage.\n\n{C} {A}
	Template 3	Write the most natural next sentence.\n\n{C} {A}
	Template 4	Write the most likely continuation for the description.\n\n{C} {A}

Table 5: Performance metrics (%) under different configurations of hyperparameter α and γ.

Hyperparameter	Value	TydiQA	MMLU	BBH	GSM8K	HellaSwag	Avg
Variance Penalty (α)	0.1	54.32	45.77	41.20	17.66	**74.62**	46.71
	0.4	54.32	**47.46**	41.20	17.66	74.54	47.04
	0.7	**56.17**	**47.46**	**41.20**	19.33	74.54	**47.74**
	1.0	54.77	**47.46**	37.78	**19.64**	74.54	46.84
Fusion Coefficient (γ)	1	56.14	47.45	**41.20**	19.18	74.47	47.69
	1/2	56.14	47.29	40.93	19.11	74.41	47.58
	1/3	**56.17**	**47.46**	**41.20**	19.33	**74.54**	**47.74**
	1/4	55.88	47.37	40.56	**19.41**	74.46	47.54

C Declaration of Generative AI Tool Usage

During the preparation of this manuscript, the authors utilized LLMs to assist with several auxiliary tasks. The primary applications of the AI tools included: Refining the grammar, phrasing, and flow of the English text, as well as translating non-English data samples, to improve overall readability. Structuring and formatting LaTeX tables for the clear presentation of the case study evaluations.

The generative AI tools were not used to conceptualize or generate the core scientific contributions, methodology, or theoretical claims of this work. All experimental designs, data analyses, and final scientific conclusions represent the original work of the authors. The authors have rigorously reviewed, edited, and verified all AI-assisted content and take full responsibility for the accuracy and integrity of this publication.

D Case Study

To intuitively illustrate the differences in data selection behaviors among various attribution methods, we present a detailed case study.

We present a detailed case study on the BBH dataset (Table 6). The evaluation instance tests logical reasoning via a step-by-step evaluation of a Boolean expression. As shown, CIGA retrieves a single-turn step-by-step reasoning sample, directly aligning with the logical deduction requirements of the target task. In contrast, baseline methods struggle with structural alignment: Cos selects an irrelevant instruction-following task about "steps", LESS retrieves a noisy multi-turn dialogue containing error corrections, and TracIn selects an answer that completely lacks the necessary reasoning steps.

For the MMLU dataset (Table 7), the evaluation instance tests domain-specific knowledge through a strict multiple-choice question format. As shown, Cos is distracted by dense textual features and retrieves an overwhelmingly long ethical story, while LESS selects a purely mathematical calculation. In contrast, CIGA captures the complex multiple-choice instruction paradigm.

For the GSM8K dataset (Table 8), the evaluation instance involves a mathematical word problem requiring step-by-step arithmetic deduction. As shown, Cos successfully retrieves an exact in-domain math problem via surface text similarity, while LESS selects a related numerical reasoning task. Crucially, CIGA retrieves a high-quality CoT reasoning sample that mirrors the step-by-step logical deduction required to solve the target math word problem.

For the TydiQA dataset (Table 9), the evaluation instance tests reading comprehension and factoid question answering based on a provided context. As shown, Cos incorrectly selects a translation task. TracIn retrieves an overly verbose reading comprehension task that misaligns with the concise nature of the target. Both LESS and CIGA successfully capture the core factoid QA paradigm.

For the HellaSwag dataset (Table 10), the evaluation instance tests commonsense reasoning and natural language continuation by requiring the model to select the most plausible ending to a given scenario. As shown, Cos retrieves an exact in-domain few-shot example based purely on surface formatting, while LESS selects a related commonsense reasoning passage. Remarkably, CIGA captures the underlying objective of natural, plausible short sentence generation.

Table 6: Top-1 selected training samples by different methods for the BBH evaluation task.

Method	Text (Validation Example / Selected Training Sample)
Eval Data	*BBH (Boolean Expressions)*: [User]: Evaluate the result of a random Boolean expression. Q: not ((not not True)) is [Assistant]: Let's think step by step. Remember that (i) expressions inside brackets are always evaluated first and that (ii) the order of operations from highest priority to lowest priority is "not", "and", "or", respectively. We first simplify this expression "Z" as follows... *[... Deduction steps truncated ...]* ...So the answer is False.
Cos	*dolly_2928*: [User]: Given these two sets of instructions, will you end up at the same end point if you started at the same beginning point? First instructions: 1) Take 2 steps forward... [Assistant]: Yes, you will end up at the same point for both sets of instructions...
LESS	*oasst1_15234*: [User]: If you have a 3 gallon cylindrical shaped jug and a 5 gallon... [Assistant]: The correct procedure to end up with exactly 4 gallons in the 5 gallon jug is: b... [User]: i dont understand, can you explain it to me like I was like 5? [Assistant]: My apologies if that was confusing. I have realized that my previous answer was incorrect...
TracIn	*oasst1_15229*: [User]: If you have a 3 gallon cylindrical shaped jug and a 5 gallon... [Assistant]: The correct answer is: **"e. None of the above"**.
CIGA	*oasst1_15233*: [User]: If you have a 3 gallon cylindrical shaped jug and a 5 gallon... [Assistant]: The correct procedure to end up with exactly 4 gallons in the 5 gallon jug is: b. Fill 3 gallon jug. Pour it all into the 5 gallon jug twice Here's the explanation: 1. Fill the 3 gallon jug with water. 2. Pour all 3 gallons from the 3 gallon jug into the 5 gallon jug...

Table 7: Top-1 selected training samples by different methods for the MMLU evaluation task.

Method	**Text (Validation Example / Selected Training Sample)**
Eval Data	*MMLU (Anatomy)*: [User]: What is the embryological origin of the hyoid bone? Options: A. The first pharyngeal arch B. The first and second pharyngeal arches C. The second pharyngeal arch D. The second and third pharyngeal arches [Assistant]: D
Cos	*flan_v2_83435*: [User]: Given a real-life anecdote of a complex ethical situation, identify who is wrong here. Classify each post into the following categories : AUTHOR [...], OTHER, EVERYBODY, and NOBODY. Example input: Pretty much was hanging out with two friends... Q: Buckle up, fam, because this is about to be a *ride,* whether you agree with me or not. So, I had two very good friends - my best friend, who we'll call Val, and another, who we'll call Todd... *[... Massive narrative text truncated ...]* ...I can't help but wonder if I was the asshole who over-reacted. A: [Assistant]: OTHER
LESS	*flan_v2_74587*: [User]: You will be given a definition of a task first, then some input of the task. In this task, you need to answer the given multiple-choice question on the general math. Classify your answers into 'a', 'b', 'c', 'd', and 'e'. Problem: ann and bob drive separately to a meeting . ann ' s average driving speed is greater than bob ' s avergae driving speed by one - third... what is the ratio r of the number of hours... Options: a) 8 : 3 , b) 3 : 2 , c) 4 : 3 , d) 2 : 3 , e) 3 : 8 Output: [Assistant]: b
TracIn	*flan_v2_40198*: [User]: OPTIONS: [i] momentum; [ii] power; [iii] force; [iv] work; Q: What is it called when a charge moves against a field? A: [Assistant]: [iv]
CIGA	*flan_v2_75125*: [User]: Given the task definition and input, reply with output. In this task, you will be presented with a question having multiple possible answers in Vietnamese language... Question: (Translated) Sam feels that you need to improve yourself. She says you should take time to gather your thoughts and clear your mind. Why does she want you to improve yourself? Options: A. learn new things B. feel better about yourself C. think D. practice yoga E. celebrate [Assistant]: B

Table 8: Top-1 selected training samples by different methods for the GSM8K evaluation task.

Method	Text (Validation Example / Selected Training Sample)
Eval Data	*GSM8K (Math Word Problem)*: [User]: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? [Assistant]: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. So the answer is 6. #### 6
Cos	*gsm8k_train_111*: [User]: Paul went to a shop to buy some groceries. He bought some bread for $2, butter for $3, and juice for two times the price of the bread. He had $15 for his shopping. How much money did Paul have left? [Assistant]: The price of the juice was 2 * 2 = «2*2=4»4. So in total he paid for his shopping 2 + 3 + 4 = «2+3+4=9»9. That means he had 15 - 9 = «15-9=6»6 left. #### 6
LESS	*cot_19506*: [User]: A train 360 m long is running at a speed of 45 km/hr. In what time will it pass a bridge 140 m long? Options: (A) 40 sec (B) 32 sec (C) 12 sec (D) 85 sec (E) 56 sec Stream of consciousness: [Assistant]: Answer: Option A Speed = 45 * 5/18 = 25/2 m/sec Total distance covered = 360 + 140 = 500 m Required time = 500 * 2/25 = 40 sec Thus, the answer is (A).
TracIn	*oasst1_15229*: [User]: If you have a 3 gallon cylindrical shaped jug and a 5 gallon cylindrical shaped jug with an unlimited supply of water (and no other containers), which is the correct procedure to end up with exactly 4 gallons in the 5 gallon jug? a. Fill 3 gallon jug... [Assistant]: The correct answer is: **"e. None of the above"**.
CIGA	*cot_30403*: [User]: Trish likes to go for walks every day. One day, she goes on a 1-mile walk. Every subsequent day, she doubles the distance she walked the previous day. On what day does she walk more than 10 times further than she did on the first day? Think carefully first, then make a decision: [Assistant]: Trish walks 1 mile on the first day, and since she walks double the previous day's amount each day that means she walks 1 * 2 = 2 miles the second day... *[... Deduction steps truncated ...]* ...As we're looking for the first day that her distance exceeds 10 times her original total, that means we're seeking a mile total of greater than or equal to 10 * 1 = 10 miles. Since she accomplished this on the 5th day of walking, the answer is 5. So, the answer is 5.

Table 9: Top-1 selected training samples by different methods for the TydiQA evaluation task.

Method	**Text (Validation Example / Selected Training Sample)**
Eval Data	*TydiQA (Reading Comprehension / Factoid QA)*: [User]: Context: Home Box Office (HBO) is an American premium cable and satellite television network that is owned by the namesake unit Home Box Office, Inc., a division of AT&T's WarnerMedia... Question: Who owns HBO? [Assistant]: Home Box Office
Cos	*flan_v2_35173*: [User]: [Q]: (Translated from Russian) These leaders - as well as those in Thailand... Translate this to English? [A]: These leaders – as well as those in Thailand... [Q]: (Translated from Russian) On March 6, 1990, the secretariat... Translate this to English? [A]: The Studios existed thanks to subsidies... [Q]: (Translated from Russian) If you have faced this problem, download Thessalonica 1.1... Translate this to English? [A]: [Assistant]: If you have faced this problem, download Thessalonica 1.1 prerelease...
LESS	*dolly_11483*: [User]: Plase answer the following request: Who wrote Les Miserable [Assistant]: Victor Hugo
TracIn	*flan_v2_40981*: [User]: Answer the following question: I am a movie director and I just received the following movie plot... Plot title: The Impossible Movie plot: Maria Bennett (Naomi Watts), her husband Henry (Ewan McGregor) and their three sons... *[... Massive plot text truncated ...]* ...Karl then... My question: How many children do Maria and Henry have? [Assistant]: Three.
CIGA	*flan_v2_139*: [User]: Answer the question...when was simple man by lynyrd skynyrd released?? [Assistant]: 1973

Table 10: Top-1 selected training samples by different methods for the HellaSwag evaluation task.

Method	Text (Validation Example / Selected Training Sample)
Eval Data	*HellaSwag (Commonsense Completion)*: [User]: A man talks in front of a house. A man climbs a ladder and looks at a roof. A man goes through a stack of papers. it Options: 0. is then announcin aluminum on a table. 1. shows the papers on the screen. 2. ends with credits and action figures on the screen attached to a fridge. 3. is then night time again. [Assistant]: shows the papers on the screen.
Cos	*flan_v2_10313*: [User]: context: A swimmer lifts his arms before and prepares for a dive on the... OPTIONS: - stands over the pool... - tramples and looks underwater. - surfaces and... **** *[... Multiple in-domain shots truncated ...]* context: How to bump a volleyball Get into position. You should stand with your legs about shoulder-width apart... OPTIONS: - Your hands should come together... - If your feet are wider... **** next sentence for the context: [Assistant]: Your hands should come together in the last moment before the ball comes...
LESS	*dolly_4340*: [User]: Given the reference text about the city of Cologne and its wildlife, can I expect to see any cats in Cologne? Cologne is the... *[... Wikipedia text truncated ...]* ...muted colours of the cityscape. [Assistant]: Although, the passage about the wildlife in the city of Cologne does not mention cats, it is possible to encounter cats in Cologne since it is very common for people in western cultures to own cats... The reason cats are not mentioned in the passage is that cats are not considered wildlife but pets.
TracIn	*cot_31937*: [User]: People standing in a straight line are wearing alternating colored shirts. The pattern of shirts begins with red, followed by green, blue, white and... If there are more than 5 people in line, which of the following cannot be the number of people in the line? Options: (A) 29 (B) 39 (C) 54 (D) 64 (E) 75 Stream of consciousness first, then make a decision: [Assistant]: The correct answer should not take the form: 5x + 4. Hence, it is Thus, the answer is (E).
CIGA	*flan_v2_32448*: [User]: question: Write a brief sentence. answer: A crowd is nearby. question: Write a brief sentence. answer: A woman using a cleaning product on her car. question: Write a brief sentence. answer: There are people in a pub. question: Write a brief sentence. answer: [Assistant]: Several girls gossip to each other.