ReSURE: Regularizing Supervision Unreliability for Multi-turn Dialogue Fine-tuning

Anonymous ACL submission

Abstract

Fine-tuning multi-turn dialogue systems requires high-quality supervision but often suffers from degraded performance when exposed to low-quality data. Supervision errors in early turns can propagate across subsequent turns, undermining coherence and response quality. Existing methods typically address data quality via static prefiltering, which decouples quality control from training and fails to mitigate turn-level error propagation. In this context, we propose **ReSURE** (REgularizing Supervision UnREliability), an adaptive learning method that dynamically down-weights unreliable supervision without explicit filtering. ReSURE estimates per-turn loss distributions using Welford's online statistics and reweights sample losses on the fly accordingly. Experiments on both single-source and mixed-quality datasets show improved stability and response quality. Notably, ReSURE enjoys positive Spearman correlations (0.21 $\sim$ 1.0 across multiple benchmarks) between response scores and number of samples regardless of data quality, which potentially paves the way for utilizing large-scale data effectively.

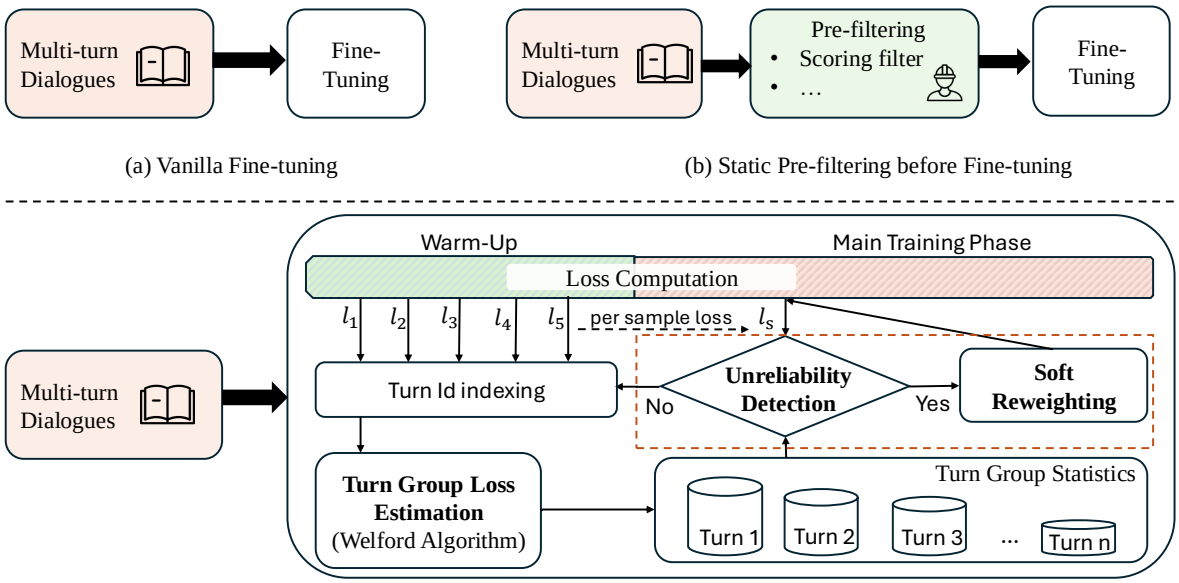
1 Introduction

Multi-turn dialogue systems are fundamental to both task-oriented (Xu et al., 2024) and open-domain conversational agents (Lu et al., 2023a; Sun et al., 2024), enabling coherent and natural interactions. However, fine-tuning remains challenging due to reliance on large-scale multi-turn datasets (Bian et al., 2023; Zhao et al., 2024b; Contributors, 2023) that mix human and synthetic data of varying quality (OpenAI, 2023). In such settings, supervision errors in early turns often propagate across later ones, compounding inconsistencies and degrading coherence (Hu et al., 2025; Yi et al., 2024). This issue is further exacerbated by mismatches between training supervision and evaluation criteria, making it difficult for models to recover from early-turn noise or learn turn-consistent behavior (Zheng et al., 2023; Kwan et al., 2024a; Wu et al., 2023a; Chen et al., 2023; Li et al., 2024a; Zhou et al., 2024). As datasets scale, conventional fine-tuning approaches assume uniformly reliable supervision and struggle to distinguish between clean and noisy signals, often overfitting to noise or discarding useful samples (Hase et al., 2024).

A common strategy to mitigate noisy supervision is static pre-filtering before fine-tuning (Wang et al., 2024a), aiming to remove low-quality or incomplete samples. However, such heuristic methods (Cao et al., 2023) overlook the hierarchical nature of multi-turn dialogues, leading to over-filtering and loss of informative turns. Other approaches enhance robustness by injecting synthetic noise (Wu et al., 2022), but often lack principled mechanisms to regulate supervision quality during training.

To address these limitations, we propose **ReSURE** (Regularizing Supervision UnREliability), an adaptive fine-tuning framework that dynamically adjusts loss contributions from unreliable supervision signals. We define such supervision as samples that consistently yield high or unstable losses during training (Wang et al., 2024b). Observing that later turns are more susceptible to supervision noise due to increased contextual complexity (Zheng et al., 2023; Kwan et al., 2024a), ReSURE groups samples by turn depth and tracks per-group loss statistics via Welford's algorithm (Welford, 1962; Efanov et al., 2021). Samples with abnormally high losses are softly reweighted to reduce instability while preserving gradient signal. This turn-aware design ensures that difficult turns are not over-penalized and early-turn errors do not dominate optimization, thereby stabilizing training and enhancing contextual coherence in multi-turn dialogue.

Experimental results show that ReSURE enables consistent optimization across multi-turn bench-



(a) Vanilla Fine-tuning

(b) Static Pre-filtering before Fine-tuning

(c) Our Approach: ReSURE – Regularizing Supervision Unreliability for Multi-turn Fine-tuning

Figure 1: Overview of Training Paradigms: Traditional Fine-tuning, Pre-filtering, and ReSURE.

marks, including MT-Bench, MT-Bench-Ext, and In-Domain-Test. Under mixed datasets with progressively added noisy or off-distribution samples, ReSURE consistently maintains or improves performance, achieving positive Spearman correlations (0.21, 1.00, 0.80), while Vicuna-Tuning (Chiang et al., 2023) shows degradation and other baselines fluctuate. To simulate task-level noise, we incorporate GSM8K (Cobbe et al., 2021) and find that ReSURE preserves generalization. These findings highlight ReSURE's robustness to both supervision noise and task drift. Unlike static filtering methods such as DeBERTa-based data selection (He et al., 2020, 2021), ReSURE achieves these gains without manual intervention. Moreover, combining ReSURE with pre-filtering yields further improvements, indicating their complementarity.

Our key contributions are as follows:

- We propose **ReSURE**, a turn-aware fine-tuning framework that preserves positive optimization in multi-turn instruction tuning under unreliable supervision.
- ReSURE avoids manual data filtering and seamlessly integrates with instruction-tuning pipelines, supporting robustness under both supervision and task-level noise.
- Extensive experiments across in-domain, MT-Bench, and MT-Bench-Ext show consistent gains and positive optimization trends (Spearman: 0.21, 1.00, 0.80), with further improvement when combined with static filtering.

2 Related Work

2.1 Multi-turn Dialogue Fine-tuning

Recent advances in LLM fine-tuning (Hu et al., 2023, 2021; Dettmers et al., 2024) have enabled strong performance on single-turn tasks (Liu et al., 2024a; Zhao et al., 2024a; Meng et al., 2024), but multi-turn dialogue remains challenging. Prior work addresses this via optimization techniques like reinforcement learning and preference modeling (Sun et al., 2024; Shani et al., 2024), or through data augmentation and inductive construction (Maheshwary et al., 2024; Ou et al., 2024). However, these lines remain disconnected, and challenges like data curation cost, weak generalization, and inconsistent supervision persist. Our work bridges this gap by jointly addressing turn-level supervision and data noise in a unified framework.

2.2 Data Selection in LLM Finetuning

Although the scale of data is crucial in LLM fine-tuning, selecting fewer high-quality data points can lead to better performance than using the entire dataset (Wu et al., 2023a; Chen et al., 2023), highlighting the significance of data selection. In terms of data quality assessment (Wang et al., 2024a), data selection methods can be grouped into three categories: (1) GPT-based scoring, which relies on prompting ChatGPT with predefined rubrics (Chen

et al., 2023; Lu et al., 2023b; Xu et al., 2023; Liu et al., 2024b; Du et al., 2023); (2) model-based scoring, where an LLM is trained to evaluate instances under a learned policy (Li et al., 2023a, 2024b; Wu et al., 2023b); and (3) indicator-based methods, which estimate data quality via inference loss (Cao et al., 2023) or handcrafted conversation metrics (Wei et al., 2023).

Although these works emphasize the importance of data selection, they often produce uninterpretable results, suffer from limited applicability and randomness, and demand prohibitively high training costs. These limitations lead to low feasibility in both training and generalization as models evolve. In addition, prior approaches perform data selection independently of the training process, failing to capture and leverage end-to-end feedback during training, which is a key focus of our work.

3 Methodology

By monitoring loss statistics in a turn-aware manner using Welford's online algorithm, ReSURE identifies unstable supervision signals and adjusts their training influence without explicit filtering. This design stabilizes optimization and preserves coherence in multi-turn dialogue settings.

Specifically, in multi-turn fine-tuning, each training sample consists of a dialogue with multiple user–assistant turns. The model is trained to minimize the cross-entropy loss over the supervised tokens. ReSURE modifies this objective by introducing a dynamic weight w_s for each sample:

$$\mathcal{L}_{\text{ReSURE}} = \frac{1}{S} \sum_{s=1}^{S} w_s \cdot \ell_s, \tag{1}$$

where ℓ_s denotes the loss for sample s, S denotes the number of samples in the mini-batch, and w_s is computed based on turn-aware loss statistics (see Sec. 3.3).

3.1 Turn Group Loss Estimation

Supervised fine-tuning in multi-turn dialogue is complicated by uneven supervision quality across dialogue depths. Early turns are typically short, contextually grounded, and easier to align with reference responses. In contrast, later turns often involve complex phenomena such as context accumulation, topic shifts, and implicit reasoning, which increase supervision noise and model uncertainty (Zheng et al., 2023; Kwan et al., 2024a).

To address this, ReSURE groups training samples by their maximum supervised turn group index $b \in \{1, \ldots, N\}$, where N denotes the maximum number of turns per dialogue. For each b, we maintain turn-specific online loss statistics—namely, a running mean $\mu_s^{(b)}$ and standard deviation $\sigma_s^{(b)}$ of the per-sample loss—computed using Welford's algorithm:

$$\mu_s^{(b)} = \mu_{t-1}^{(b)} + \frac{l_s - \mu_{s-1}^{(b)}}{t^{(b)}}, \tag{2}$$

$$SSD_s^{(b)} = SSD_{t-1}^{(b)} + (l_s - \mu_{t-1}^{(b)})(l_s - \mu_s^{(b)}), \tag{3}$$

$$\sigma_s^{(b)} = \sqrt{\frac{M2_s^{(b)}}{t^{(b)} - 1}}. \tag{4}$$

Here, $SSD_s^{(b)}$ denotes the Sum of Squared Deviations from the current mean $\mu_s^{(b)}$, used to compute the variance, and $s^{(b)}$ is the number of samples assigned to group b up to sample s. Only samples within each group b contribute to its own loss statistics, enabling turn-aware normalization. All statistics are initialized to zero and updated only upon observing the first reliable sample in each turn group. This design avoids over-penalizing high-turn samples that are harder, while ensuring stable optimization on easier low-turn cases. By aligning loss treatment with dialogue structure, it provides an inductive bias that helps the model calibrate supervision trust by turn depth.

3.2 Unreliability Detection

After warm-up, ReSURE detects unreliable supervision by identifying loss outliers with respect to turn-specific distributions. For each dialogue turn, we maintain the running mean $\mu^{(b)}$ and standard deviation $\sigma^{(b)}$ of per-sample loss using Welford's algorithm. A sample is flagged as unreliable if its loss l_s exceeds the threshold:

$$\tau_s^{(b)} = \mu_s^{(b)} + \alpha \cdot \sigma_s^{(b)}, \tag{5}$$

where α is a fixed anomaly factor. While classical outlier detection often adopts α under Gaussian assumptions, we use $\alpha = 1.0$ to increase sensitivity to moderate deviations, following practices in robust training and loss-based re-weighting (Zhang and Sabuncu, 2020).

If a sample is identified as unreliable ($l_s > \tau_s^{(b)}$), its loss is downweighted using soft reweighting

(see Sec. 3.3) but excluded from the update of running statistics, and the statistics $\mu^{(b)}$ and $\sigma^{(b)}$ remain unchanged.

In contrast, if $l_s \leq \tau_s^{(b)}$, the sample is treated as reliable and its loss is incorporated into the Welford updates as defined in Eqs. (2), (3), and (4).

This conditional update mechanism ensures that the estimated statistics remain stable in the presence of supervision noise while still adapting to distributional shifts in reliable examples.

3.3 Soft Reweighting

Rather than discarding high-loss samples, ReSURE applies a soft reweighting strategy to reduce their influence while retaining informative gradients. For samples with $l_s > \tau_s^{(b)}$, the adjusted loss is computed using a decayed weight:

$$w_s = \max\left(\epsilon_s, \exp\left(-\frac{l_s - \tau_s^{(b)}}{\tau_s^{(b)}}\right)\right), \quad (6)$$

$$\tilde{l}_s = w_s \cdot l_s, \quad (7)$$

where $\tau_s^{(b)}$ is the turn-specific loss threshold and ϵ_s denotes a dynamic floor, computed as the 5th percentile of the current batch's weight distribution. This adaptive lower bound ensures that even high-loss samples retain a minimal contribution, preventing vanishing gradients while adapting to overall batch variability.

Unlike fixed heuristics, this percentile-based formulation provides a data-driven way to preserve training signal from difficult or ambiguous cases. It aligns with findings in robust optimization that emphasize the importance of soft suppression rather than hard filtering for handling uncertain supervision (Ren et al., 2018; Zhang and Sabuncu, 2020).

4 Experiments

4.1 Evaluation on Datasets

There are multiple open-source and high-quality multi-turn dialogue datasets, which are generated by both humans and LLMs. Table 7 in section B of Appendix presents the datasets used in this work and their features, including ShareGPT (RyokoAI, 2023), WildChat (Zhao et al., 2024b), OpenAssistant (Köpf et al., 2024), ChatAlpaca (Bian et al., 2023), MTLingual (Maheshwary et al., 2024), and UltraChat (Ding et al., 2023). Motivated by benchmarks on LLM evaluation (Zheng et al., 2023; Kwan et al., 2024b) and MoDS (Du et al., 2023), we evaluate these datasets by **GPT** and **reward model**, respectively. This dual-evaluation strategy offers complementary insights and enables a comprehensive evaluation on dataset quality.

Evaluation by GPT. Recent benchmarks on LLM evaluation (Zheng et al., 2023; Kwan et al., 2024b; Radziwill and Benton, 2017) emphasize relevance, helpfulness, and accuracy, while also acknowledging ethical considerations. Besides, prior work on human dialogue (Dethlefs et al., 2016) highlights the importance of information density. Thus, we propose a benchmark evaluating conversations in four independent aspects: *Connection*, *Quality*, *Information Density* and *Friendliness*.

The evaluation is carried out using GPT-4o, which is widely adopted in evaluation works (Zheng et al., 2023; Kwan et al., 2024b; Bai et al., 2024). The designs of criteria, prompts and data pre-processing are detailed in section A of Appendix. For the evaluation on each aspect, one hundred conversations are sampled independently and randomly, and the evaluation on each conversation of each dataset is also independent. The score of each aspect of a dataset is defined as the average score of the sampled conversations in this aspect.

Evaluation by reward model. We employ the reward-model-deberta-v3-large-v2 (OpenAssistant, 2023) to score conversations. This model is trained on four diverse human-feedback datasets (Nakano et al., 2021; Stiennon et al., 2020; Havrilla, 2023; Bai et al., 2022), enabling it to perform evaluation on models' responses. We concatenate each entire multi-turn dialogue into a single input sequence and prompt the model to assign a reward score which reflects the overall quality. The score of a dataset is defined as the average reward score.

Table 6 presents the evaluation results. To derive an overall quality score for each dataset, we scale the *Information Density* score by a factor of 100 and sum it with the other four evaluation metrics. The overall quality of the datasets is categorized as high (ChatAlpaca, MTLingual, UltraChat), normal (WildChat, shareGPT), and low (OpenAssistant).

4.2 Experimental Settings

Parameter. The experiments are conducted with instruct-style models from multiple families, including **LLaMA-3.2-3B-Instruct**, **LLaMA-3.1-8B-Instruct** (Grattafiori et al., 2024), **Qwen2.5-3B-Instruct**, and **Qwen2.5-7B-Instruct** (Team, 2024).

The models are fine-tuned for 3 epochs on datasets of varying quality. Each device processed a batch size of 4, with a gradient accumulation step of 4, resulting in an effective batch size of 64. The Adam optimizer was employed, with the hyperparameter β_2 set to 0.95. A cosine decay learning rate schedule was applied, starting at an initial learning rate of 1×10^{-5} and incorporating a warm-up ratio of 0.01. All training and evaluation procedures were performed in FP16 precision on four NVIDIA GPUs. To reduce memory consumption, gradient checkpointing and Low-Rank Adaptation (LoRA) were enabled during training. Model performance was periodically assessed using a held-out validation set of 400 examples.

To enhance the robustness of the training process, a warm-up strategy was implemented during the initial phase of training. This involved using 640 high-quality dialogue samples to initialize baseline mean and variance parameters. As training progressed, the filtering weight for anomalous data was gradually increased to ensure smooth and stable model optimization.

Evaluation. We conducted evaluations across three settings: **In-Domain-Test**, **MT-Bench** (Zheng et al., 2023), and **MT-Bench-Ext** (Kwan et al., 2024b), to assess both in-domain performance and generalization. The In-Domain-Test serves as a setting-specific evaluation, where models are tested on held-out samples from the same distribution as the training data. It includes six multi-turn dialogue datasets (ShareGPT, WildChat, OpenAssistant, ChatAlpaca, MTLingual, and UltraChat), each with 100 randomly sampled conversations to cover diverse domains and supervision styles. All evaluations followed the GPT-4–based "LLM-as-a-Judger" protocol (Zheng et al., 2023), which was used both to compute Win Rate (Li et al., 2023b; Dubois et al., 2024, 2023) via pairwise comparisons and to assign fine-grained scores across four human-aligned criteria: **Faithfulness** (Faith.), **Appropriateness** (Appr.), **Naturalness** (Nat.), and **Completeness** (Compl.).

Mix Dataset. To validate the effectiveness of our approach, we selected ChatAlpaca, ShareGPT, and OpenAssistant as representatives of high-, normal-, and low-quality datasets, respectively. From each dataset, 20K samples are extracted and mixed in different combinations: high and normal quality, high and low quality, and high, normal, and low quality. These experiments are designed to assess the performance of our method in handling datasets with varying distributions during training.

4.3 Baselines

We evaluate our method against four typical methods in multi-turn dialogue study:
(1) **Base Model (BM)**: the original instruction-tuned model without task-specific fine-tuning for multi-turn dialogue.
(2) **Vicuna-Tuning (VT)**: a widely adopted dialogue adaptation framework built upon LLaMA, distinguished by its LoRA fine-tuning strategy on multi-turn conversational data (Chiang et al., 2023).
(3) **Baize**: a parameter-efficient approach that exclusively updates linear layers through self-chat generation (Chiang et al., 2023).
(4) **ChatGLM3**: implements multi-turn dialogue fine-tuning by updating only the loss of roles other than *user* and *system* (GLM et al., 2024).
All methods share identical LoRA configurations (rank=128, alpha=16, dropout=0.3) and data partitions: 20,000 training samples with 400 validation and 100 test instances. Experiments are conducted with fixed random seeds (seed=42) and multi-turn dialogue performance quantified by the MT-Bench (Zheng et al., 2023).

4.4 Main Results

4.4.1 Does ReSURE address negative optimization in multi-turn dialogues?

To evaluate the ability of ReSURE to mitigate negative optimization, which refers to performance degradation as the volume of supervision increases, we conduct a comparative analysis with Vicuna-Tuning across six instruction-tuned multi-turn dialogue datasets: M2Lingual, ChatAlpaca, UltraChat, ShareGPT, Wildchat, and OpenAssistant. As shown in Table 1, ReSURE consistently outperforms the base model by 6.11%, 9.82%, and 2.86% on the in-domain benchmark, MT-Bench, and MT-Bench-Ext, respectively. In contrast, Vicuna-Tuning exhibits clear signs of negative optimization, particularly on ShareGPT, where additional training data reduces performance—likely due to stylistic inconsistencies or supervision conflicts. Although ReSURE achieves slightly lower gains on M2Lingual, this may be attributed to the limited dataset size and increased risk of overfitting. Overall, these results demonstrate that ReSURE scales effectively with increasing data while maintaining robustness to supervision noise.

Level	Dataset	In_Domain_Test			MT-Bench			MT-Bench-Ext		
		BM	VT	ReSURE	BM	VT	ReSURE	BM	VT	ReSURE
H	M2Lin.	**7.10**	7.09 (-0.14%)	7.06 (-0.56%)	7.13	**7.21** (+1.12%)	7.16 (+0.42%)	6.64	6.65 (+0.15%)	**6.71** (+1.05%)
	ChatAlpaca	8.20	7.99 (-2.56%)	**8.26** (+0.73%)	7.13	6.97 (-2.24%)	**7.29** (+2.24%)	6.64	5.99 (-9.79%)	**6.76** (+1.81%)
	UltraChat	7.90	7.56 (-4.30%)	**8.01** (+1.39%)	7.13	6.68 (-6.31%)	**7.32** (+2.66%)	6.64	6.22 (-6.33%)	**6.76** (+1.81%)
N	ShareGPT	6.55	6.09 (-7.02%)	**6.95** (+6.11%)	7.13	6.08 (-14.73%)	**7.83** (+9.82%)	6.64	5.80 (-12.65%)	**6.83** (+2.86%)
	WildChat	6.80	6.47 (-4.85%)	**6.86** (+0.88%)	7.13	7.14 (+0.14%)	**7.21** (+1.12%)	6.64	6.74 (+1.51%)	**6.72** (+1.20%)
L	OpenAss.	7.64	7.20 (-5.76%)	**7.67** (+0.39%)	7.13	6.20 (-13.07%)	**7.26** (+1.83%)	6.64	5.48 (-17.47%)	**6.83** (+2.86%)

Table 1: Comparison of our method, non-trained Base Model (BM), and Vicuna-Tuning on LLaMA-3.2-3B-Instruct: multi-turn dialogue performance (GPT-4 scores) across high-, normal-, and low-quality datasets. Each cell shows absolute scores plus relative improvement/decline (%) vs. BM in parentheses. H, N, L = High, Normal, Low, M2Lin. = M2Lingual (en), OpenAss. = OpenAssistant.

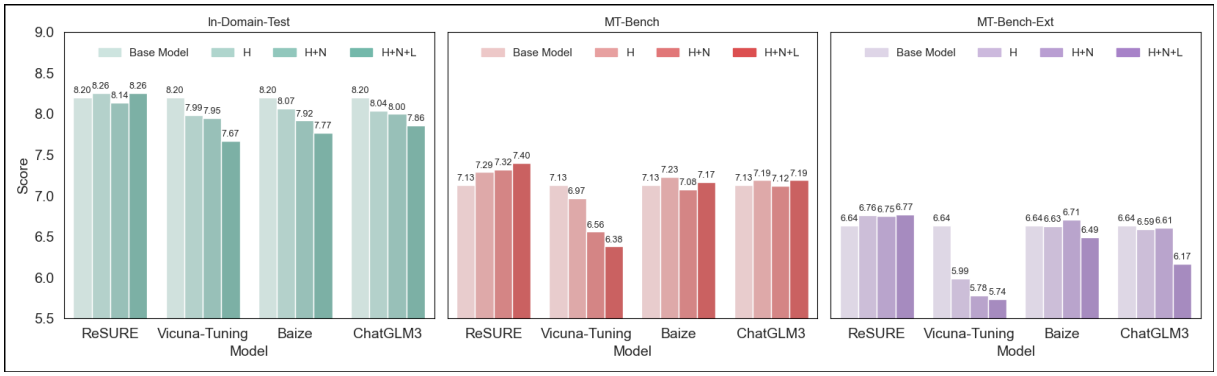


Figure 2: Performance scaling with Hierarchical Data Integration (H, H+N, H+N+L): (a) In-Domain-Test Performance (b) MT-Bench Performance (c) MT-Bench-Ext Performance

Human evaluation on MT-Bench-Ext (Table 2) further supports our findings. ReSURE outperforms both the base model and Vicuna-Tuning across all evaluation dimensions, with notable improvements in Faithfulness and Completeness. These gains are especially evident in multi-turn settings, where maintaining factual consistency and contextual coherence is essential. The results indicate that ReSURE more effectively preserves semantic alignment across turns, resulting in more coherent and informative dialogues. Additional evaluation details are provided in Appendix C.

These results indicate that the dynamic suppression of unreliable supervision contributes to more stable training dynamics and semantically aligned responses. This observation is consistent with the findings from automatic benchmarks, and further supports the robustness of ReSURE under imperfect supervision conditions in instruction-tuned dialogue settings.

4.4.2 Does ReSURE Suppress Unreliable Supervision for Robust Fine-Tuning?

To evaluate ReSURE's robustness under noisy supervision, we construct mixed datasets of increasing complexity and compare it with Vicuna-Tuning, Baize, and ChatGLM3. This setup simulates realistic fine-tuning scenarios involving low-quality or off-distribution samples. As shown in Figure 2, ReSURE maintains or improves performance across all three evaluation settings as dataset size and noise increase. It achieves stable in-domain scores around 8.2, with steady gains on MT-Bench (7.13 to 7.4) and MT-Bench-Ext (6.64 to 6.77), indicating effective use of additional supervision without overfitting to noise. In contrast, Vicuna-Tuning exhibits consistent degradation—particularly on MT-Bench-Ext (6.64→5.74)—while Baize and ChatGLM3 show marginal or unstable changes. These trends are confirmed by Spearman correlation analysis (Appendix Table 4), where ReSURE yields positive correlations across all benchmarks, unlike the negative or inconsistent values observed for baselines.

ReSURE excels on partially noisy datasets, maintaining positive optimization. As shown in Figure X, when noise increases from high-quality (H) to mixed-quality (H+N+L), conventional methods like Vicuna-Tuning and Baize exhibit noticeable performance drops—e.g., Vicuna-Tuning drops by 0.75 on MT-Bench and 0.90 on MT-Bench-Ext. In contrast, ReSURE shows

Model	Faith.	Appr.	Nat.	Compl.	Over.
BM	3.40	3.06	3.16	3.50	3.04
VT	3.36	2.98	3.22	3.42	2.98
ReSURE	**3.74**	**3.68**	**3.74**	**3.86**	**3.66**

Table 2: Human evaluation on MT-Bench-Ext.

Exp Setting	In-Domain-Test	MT-Bench	MT-Bench-Ext
Base	8.20	7.13	6.64
VT	7.99	6.97	5.99
VT + Prefiltering	7.99 (0.00%)	7.15 (+2.58%)	6.68 (+11.52%)
ReSURE	8.26	7.29	6.76
ReSURE + Prefiltering	**8.28** (+0.24%)	**7.58** (+3.98%)	**7.35** (+8.73%)

Table 3: Performance comparison between the prefiltering method (DeBERTa) and ReSURE.

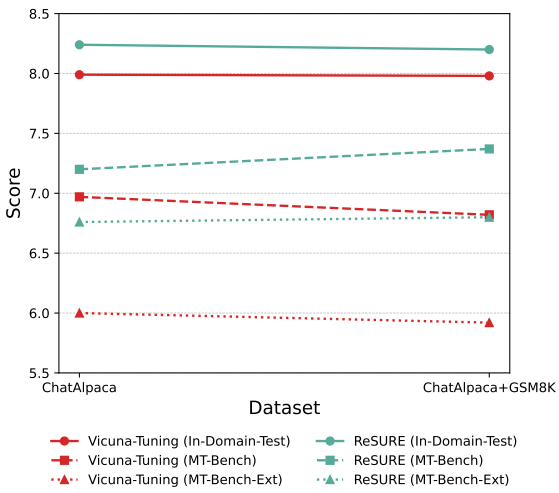


Figure 3: Performance comparison between Vicuna-Tuning and ReSURE on ChatAlpaca and ChatAlpaca+GSM8K across three evaluation benchmarks: In-Domain-Test, MT-Bench, and MT-Bench-Ext.

Model	In-Domain-Test	MT-Bench	MT-Bench-Ext
VT	-1.000	-1.000	-1.000
Baize	-1.000	0.000	-0.400
ChatGLM3	-1.000	0.211	-0.800
ReSURE	0.211	1.000	0.800

Table 4: Spearman correlation between dataset complexity and performance across benchmarks.

strong robustness, with minimal variance and even slight improvements in noisy conditions. On the In-Domain-Test, ReSURE achieves a peak score of 8.26, maintaining a high level of performance across all mixtures. In multi-turn settings, it consistently outperforms baselines across all noise levels, particularly under challenging H+N+L configurations. This resilience enables ReSURE to leverage larger and more diverse training data effectively, without requiring explicit pre-filtering.

5 Ablation Study

5.1 Can ReSURE Handle Task Mixture?

To further examine the stability of ReSURE under heterogeneous training conditions, we incorporate GSM8K, a mathematical question answering dataset, into the multi-turn ChatAlpaca corpus. This setting introduces task-level noise due to divergent supervision styles. As shown in Figure 3, ReSURE maintains in-domain performance and achieves positive generalization on MT-Bench and MT-Bench-Ext, even when trained on mixed-task data. In contrast, Vicuna-Tuning shows performance degradation on both in-domain and general benchmarks, likely due to overfitting to arithmetic patterns in GSM8K, which weakens its multi-turn dialogue capability and harms contextual alignment. These results indicate that ReSURE is more robust to task drift and better preserves dialogue-relevant optimization signals by dynamically suppressing incompatible supervision. All experiments are conducted using the `LLaMA3.2-3B-Instruct` model. Notably, ReSURE also improves GSM8K accuracy from 77.7% to 78.3%, confirming its robustness across tasks without sacrificing task-specific performance.

5.2 Does ReSURE perform better than pre-filtering methods?

We compare ReSURE against traditional offline reward-based pre-filtering (Du et al., 2023), using reward-model-deberta-v3-large-v2 (OpenAssistant, 2023) to retain the top 75% of samples from ChatAlpaca, ShareGPT, and OpenAssistant. As shown in Table 3, ReSURE alone outperforms static filtering, and the best performance is achieved by combining both. Notably, this hybrid setup yields the largest improvement on MT-Bench-Ext, highlighting its advantage in complex multi-turn scenarios. These findings indicate that ReSURE's dynamic reweighting complements static quality filtering, offering an effective synergy for robust dialogue fine-tuning.

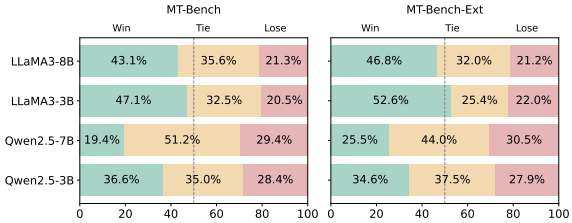


Figure 4: Win Rates of ReSURE vs. VT on MT-Bench and MT-Bench-Ext.

Metric	ReSURE	w/o Welford	Δ (%)
In-Domain-Test	8.26	8.20	**-0.73%**
MT-Bench	7.29	7.19	**-1.37%**
MT-Bench-Ext	6.76	6.70	**-0.89%**

Table 5: Performance drop (GPT-4 scores) when removing Welford statistics from ReSURE across three evaluation benchmarks.

5.3 How Does ReSURE Perform Across Diverse Model Families and Sizes?

To assess ReSURE's generalizability, we apply ReSURE to four instruction-tuned models from the Qwen and LLaMA families. We evaluate using *Win Rate*—the proportion of multi-turn responses preferred over base outputs, as judged by GPT-4. As shown in Figure 4, ReSURE consistently improves multi-turn quality across settings. Improvements are more stable for LLaMA models, while Qwen models show greater variance between MT-Bench and MT-Bench-Ext, suggesting model-specific sensitivity to noisy supervision. These results highlight ReSURE's applicability and robustness across architectures and scales.

ReSURE enhances response performance by effectively skipping low-quality data. To better understand the impact of its adaptive weighting mechanism, we conduct an ablation study by removing the Welford-based loss modulation, while keeping all other training settings and loss components unchanged. This ablated variant disables the skip mechanism and treats all supervision equally, regardless of quality. Table 5 demonstrates that removing Welford statistics leads to performance drops of 0.73%, 1.39%, and 0.90% on In-Domain, MT-Bench, and MT-Bench-Ext, respectively. These results confirm that selectively down-weighting unreliable supervision improves robustness and training stability in multi-turn dialogue tuning, and highlight the importance of adaptive loss modulation in mitigating the impact of noisy or inconsistent annotations.

1st	**User**: *In the sentence, ... , identify the subject* **Assistant**: Jack
2nd	**User**: *... what the predicate is in that sentence?* **Assistant**: Yes, the predicate ...
3rd	**User**: *Can you explain to me what a predicate is?* **Assistant**: Sure! A predicate is one of the two main parts of a sentence. It usually includes the...
4th (*Test case*)	**User**: ***Can you give me an example of a sentence with a compound predicate?*** **Answer 1 (ReSURE)**: ...**\"are going to the movies tonight\"** is the first part of the compound predicate, and **\"will see a new superhero movie\"** is the second part of the compound predicate.... ***Score***: **9/10**, ***Rationale***: Correctly identify parallel verb structures. 😊 **Answer 2 (Vicuna-Tuning)**: ... the compound predicate includes two verbs: **\"attended\"** and **\"headed\".** ... ***Score***: **3/10**, ***Rationale***: Mistakenly parsing the noun phrase "headed" as a verb. 😔

Figure 5: Case study.

6 Case Study

As illustrated in Figure 5, this multi-turn dialogue example demonstrates the superior contextual understanding of ReSURE compared to Vicuna-Tuning. When processing a compound predicate query, ReSURE correctly identifies the parallel verb structure, accurately parsing both predicate components ("are going" and "will see") with appropriate syntactic boundaries. In contrast, Vicuna-Tuning misinterprets the noun phrase "headed" as a verb predicate, despite the prior context clearly indicating "head" as a positional noun. This error highlights the model's limited ability to maintain dialogue state awareness and resolve referential dependencies across turns. Additional examples are provided in Appendix D.

7 Conclusion

We propose **ReSURE**, a turn-aware fine-tuning framework that dynamically down-weights unreliable supervision via per-turn loss statistics. Without explicit data filtering, ReSURE improves response quality and training stability across MT-Bench, MT-Bench-Ext, and in-domain settings. It demonstrates consistent gains under supervision noise, with ablations confirming the effectiveness of turn-aware modulation. ReSURE offers a scalable solution for instruction tuning on large, mixed-quality datasets.

Limitation

This study has several limitations. First, while we adopt one type of online statistical approach, alternative techniques for modeling supervision reliability remain unexplored. Second, our dataset quality evaluation is intended as a reference rather than a definitive measure, as different domains may require tailored metrics. Third, the method is evaluated only in multi-turn dialogue scenarios, with broader applications limited by computational cost. In addition, our results on Qwen2.5-7B-Instruct are less promising compared to other models, potentially due to architectural differences or instruction tuning strategies not well aligned with our loss calibration mechanism. Despite these limitations, we hope our findings offer useful insights for future research on domain-specific fine-tuning.

Ethics Statement

This research focuses on improving the robustness of fine-tuning multi-turn dialogue systems using publicly available datasets. All datasets used in this work are released under permissive licenses and do not contain personally identifiable information. No human subjects were involved in data collection. While our method aims to reduce the impact of unreliable supervision, it implicitly filters training signals, which may lead to unintended bias or underrepresentation of minority styles. Model evaluations are conducted by three trained research assistants, each paid $20/hour, above the local average.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ning Bian, Hongyu Lin, Yaojie Lu, Xianpei Han, Le Sun, and Ben He. 2023. Chatalpaca: A multi-turn dialogue corpus based on alpaca instructions. https://github.com/cascip/ChatAlpaca.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/. Accessed: 2025-02-10.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems, 2021. *URL https://arxiv.org/abs/2110.14168*, 9.

OpenAssistant Contributors. 2023. Openassistant conversations - democratizing large language model alignment. https://arxiv.org/abs/2304.07327. Accessed: 2023-04-17.

Nina Dethlefs, Helen Hastie, Heriberto Cuayáhuitl, Yanchao Yu, Verena Rieser, and Oliver Lemon. 2016. Information density and overlap in spoken dialogue. *Comput. Speech Lang.*, 37(C):82–97.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Andrey A Efanov, Sergey A Ivliev, and Alexey G Shagraev. 2021. Welford's algorithm for weighted statistics. In *2021 3rd International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*, pages 1–5. IEEE.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. 2024. The unreasonable effectiveness of easy training data for hard tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7002–7024, Bangkok, Thailand. Association for Computational Linguistics.

Alex Havrilla. 2023. synthetic-instruct-gptj-pairwise (revision cc92d8d).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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. *arXiv preprint arXiv:2106.09685*.

Hanjiang Hu, Alexander Robey, and Changliu Liu. 2025. Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. *arXiv preprint arXiv:2503.00187*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024a. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024b. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024a. 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.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2024b. One-shot learning as instruction data prospector for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Bangkok, Thailand. Association for Computational Linguistics.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023a. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, and Chang Zhou. 2023b. # instag: Instruction tagging for diversity and complexity analysis. *arXiv preprint arXiv:2308.07074*.

Rishabh Maheshwary, Vikas Yadav, Hoang Nguyen, Khyati Mahajan, and Sathwik Tejaswi Madhusudhan. 2024. M2lingual: Enhancing multilingual, multi-turn instruction alignment in large language models. *arXiv preprint arXiv:2406.16783*.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. In *arXiv*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAssistant. 2023. Openassistant/reward-model-deberta-v3-large-v2. https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2. Reward model trained from human feedback to predict which generated answer is better judged by a human, given a question.

Jiao Ou, Jiayu Wu, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Inductive-deductive strategy reuse for multi-turn instructional dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17402–17431, Miami, Florida, USA. Association for Computational Linguistics.

Nicole M Radziwill and Morgan C Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR.

RyokoAI. 2023. Sharegpt. https://huggingface.co/datasets/RyokoAI/ShareGPT52K.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. 2024. Multi-turn reinforcement learning from preference human feedback. *arXiv preprint arXiv:2405.14655*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *NeurIPS*.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9750, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024a. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y Wong, and Simon See. 2024b. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. *arXiv preprint arXiv:2402.10646*.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*.

B. P. Welford. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2022. Noisytune: A little noise can help you finetune pretrained language models better. *arXiv preprint arXiv:2202.12024*.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. 2023a. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. 2023b. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.

Yang Xu, Yongqiang Yao, Yufan Huang, Mengnan Qi, Maoquan Wang, Bin Gu, and Neel Sundaresan. 2023. Variety and quality over quantity: Towards versatile instruction curation. *arXiv preprint arXiv:2312.11508*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Zhilu Zhang and Mert R Sabuncu. 2020. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024a. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024b. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Data processing and Evaluation Prompts

This section presents the designs of criteria. the four evaluation aspects in section 4.1 are defined as:

Connection: The final response should incorporate relevant information from prior conversations without introducing unrelated or redundant details.

Quality: Each response should fulfill the request of the corresponding turn, while ensuring content accuracy and maintaining high language quality.

Information Density (ID): For the whole conversation, calculate the total number of words N and the number of information units I. The information density is defined as $ID = I/N$.

Friendliness: Requests should be in a polite manner, while responses should prioritize security and politeness. The whole conversation should maintain a respectful tone.

During the evaluation of datasets, although the raw patterns of conversation data from different sources vary from each other, all of them are formatted as *[{'human': '<request>', 'assistant': '<response>'}, ... , {'human': '<request>', 'assistant': '<response>'}]* for each entire and independent conversation, before being written to the prompt. The ChatGPT version used in the evaluation is *ChatGPT-4o-2024-08-06*, and the complete prompts of the evaluation on *Connection*, *Quality*, *Information Density* and *Friendliness* are detailed in Figure 7, Figure 8, Figure 9, Figure 10 separately. In the evaluation, each aspect of each independent conversation is also graded independently.

B Datasets Introduction

Table 7 shows the datasets in this work. ShareGPT is a collection of 90k conversations shared via the ShareGPT API (closed at present), and includes both user prompts and responses from ChatGPT, which mainly consists of messages in English and

Dataset	Con.	Qu.	ID	Fr.	Re.	Overall
ChatAlpaca	8.34	**9.49**	0.0286	**9.48**	**3.00**	High
MTLingual	**8.54**	9.37	0.0263	9.14	1.49	High
UltraChat	8.46	9.06	0.0233	9.41	1.92	High
WildChat	7.80	8.78	0.0196	8.90	0.17	Normal
ShareGPT	8.10	8.69	0.0174	8.82	-0.33	Normal
OpenAssistant	7.54	7.57	**0.0292**	8.21	0.28	Low

Table 6: Dataset Evaluation Results. Con.: Connection, Qu.: Quality, ID: Information Density, Fr.: Friendliness, Re: Reward score.

other western languages. WildChat is a collection of 1 million real-world user-ChatGPT conversations which consists of over 2.5 million interaction turns and 68 languages from 204,736 users (Zhao et al., 2024b). OpenAssistant is a collection of 161,443 messages that construct over 10000 complete conversations, which consists of 35 different languages and over 40k annotations on quality, and is designed for reinforcement learning from human feedback. Hence, it provides different conversations based on the same initial question with different quality, which leads to the sacrifice of the overall quality. Another important and unique feature of OpenAssistant is that, it is totally generated and annotated by human (Köpf et al., 2024). ChatAlpaca is a collection of 20k conversations, generated by ChatGPT and started with the original Stanford Alpaca (Taori et al., 2023) data, and it contains English and Chinese version. MTLingual is a collection of 182k conversations in 70 languages, and is generated by Evol (Maheshwary et al., 2024). The type of language, task, user prompt, and seed prompt are also detailed in MTLingual. UltraChat is a collection of 1.5 million conversations and is generated by ChatGPT which simulates the interactions of human. The main concerns of UltraChat is diversity, scale, and coherence.

C Human Evaluation

To qualitatively assess response quality, we conduct a human evaluation on a subset of multi-turn dialogues. Three research assistants with NLP backgrounds are recruited to independently rate model outputs. We randomly sample 10 dialogue instances from MT-Bench and MT-Bench++ (10 each), covering diverse tasks and turn depths. For each dialogue, annotators evaluate model responses at different turns following the criteria defined in Table 8, including faithfulness, appropriateness, naturalness, completeness, and overall quality. Final scores are computed by averaging ratings across annotators.

D Case Study

Figure 6 details a case showing ReSURE's superior contextual comprehension on multi-turn dialogue compared to Vicuna-Tuning. In this case, our method successfully identifies parallel verb structures while Vicuna-Tuning fails in the recognition of the parts of speech.

Dataset	Volume	Avg. Turns	Generation Type	Generation Mechanism	Annotated
ShareGPT (RyokoAI, 2023)	94K	3.51	User-ChatGPT	Voluntary sharing by netizens	No
WildChat (Zhao et al., 2024b)	1.04M	2.54	User-ChatGPT	Collected from chatbot services powered by GPT API	No
OpenAssistant (Köpf et al., 2024)	135.6K	2.34	Human only	Human-Generated & Annotated by volunteers	Yes
ChatAlpaca(Bian et al., 2023)	20K	4.32	ChatGPT	Follow-up by GPT from Stanford Alpaca(Taori et al., 2023)	No
MTLingual (Maheshwary et al., 2024)	182K	2.48	ChatGPT	Constructed by Evol from Aya(Singh et al., 2024)	No
UltraChat (Ding et al., 2023)	1.5M	3.80	ChatGPT	Simulate human interactions by ChatGPT	No

Table 7: Datasets in this work with features, the values of *Avg. Turns* of ShareGPT, WildChat and OpenAssistant derive from the work of WildChat (Zhao et al., 2024b), and the value of *Avg. Turns* of OpenAssistant calculates on multi-turn conversations in English.

1st	**User**: *In the sentence, \"Jack received a birthday present from his aunt Mary,\" identify the subject* **Assistant**: Jack
2nd	**User**: *Great job! Can you now tell me what the predicate is in that sentence?* **Assistant**: Yes, the predicate in the sentence is \"received a birthday present from his aunt Mary.
3rd	**User**: *Can you explain to me what a predicate is?* **Assistant**: Sure! A predicate is one of the two main parts of a sentence. It usually includes the verb and all the other words in the sentence that modify or complete the sentence's subject. The job of the predicate is to describe what the subject is doing or what is being done to it. For example, in the sentence \"Jack received a birthday present from his aunt Mary,\" \"received a birthday present from his aunt Mary\" is the predicate, and it describes what Jack did.
4th (Test case)	**User**: ***Can you give me an example of a sentence with a compound predicate?*** **Answer 1 (ReSURE)**: …\"Jack and Sarah are going to the movies tonight, and they will see a new superhero movie.\"\n\nIn this sentence, \"are going to the movies tonight\" is the first part of the compound predicate, and \"will see a new superhero movie\" is the second part of the compound predicate…. ***Score***: **9/10**, ***Rationale***: Correctly identify parallel verb structures. 😊 **Answer 2 (Vicuna-Tuning)**: …\"The manager of the store and the head of the marketing department attended the meeting.\" \n\nIn this sentence, the compound predicate includes two verbs: \"attended\" and **\"headed\".** The first verb is \"attended\", which describes the manager's action at the meeting. **The second verb is \"headed\",** which describes the head of the marketing department's action at the meeting. … ***Score***: **3/10**, ***Rationale***: Mistakenly parsing the noun phrase "headed" as a verb. 😔

Figure 6: Case study.

Prompts of the Evaluation of Connection

"""
You are a strict and objective evaluator. Your task is to assess the quality of the final response from assistant in conversation content.
Your evaluation should be fair, professional, and reflect an expert judgment of the response's quality.
The conversation is formatted as [{'human': '...', 'assistant': '...'}, ..., {'human': '...', 'assistant': '...'}] .
The final response is the final 'assistant' message in the conversation.

[Conversation]\n""" + <conversation> + "\n" + """
Assessment Criteria:
Score baseline is 5. The final score should be adjusted based on the following criteria:
Connection: Does it utilize the information in the previous conversations?
Concentrate on the evidence of conflicts and coherence. Evidence of one conflict
should decrease the score by 1, and evidence of utilizing one information should increase the score by 1.
Relevance: Does it provide redundant information which is not related to the topic? Is so, it should be penalized by the degree and amount. One irrelevant information should decrease the score by 1.
Overall Score: Assign a score from 1 to 10 (10 being the best), considering all of the above factors.

The evaluation and your output must be strictly structured in the following JSON format:
{
"Explanation": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>
}
"""

Figure 7: Prompts of the evaluation of connection.

Prompts of the Evaluation of Quality

"""
You are a strict and objective evaluator. Your task is to assess the quality of the each response from assistant in conversation, based on the Assessment Criteria.
Your evaluation should be fair, professional, and reflect an expert judgment of the response's quality.
The conversation is formatted as [{'human': '...', 'assistant': '...'}, ..., {'human': '...', 'assistant': '...'}].

[Conversation]\n""" + <conversation> + "\n" + """
Assessment Criteria:
Requirement Alignment: For each response, only consider the corresponding request from human in this turn, does the response meet the user's task goal?
Content Accuracy: Is the information in the response correct, clear, and logically organized?
Language Quality: Is the language fluent, coherent, and readable? Are there any obvious grammatical or word choice errors?
Consideration on previous information: If there is relevant information in the previous turns of chatting, does the response take them into consideration?
Overall Score: Assign a score from 1 to 10 (10 being the best), considering all of the above factors.

The evaluation and your output must be strictly structured in the following JSON format:
{
"evaluations": [
{
"Number of turn in conversation": 1,
"Explanation": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>
},
...,
{
"Number of turn in conversation": <Integer, the No. of turn in conversation>,
"Explanation": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>
}]
}
"""

Figure 8: Prompts of the evaluation of quality.

Prompts of the Evaluation of Information Density

"""
You are a strict and objective evaluator. Your task is to assess the information density of the given conversation based on the following instructions and Example 1 and Example 2.
[Instructions]
Your evaluation should be fair, professional, and reflect an expert judgment of the response's quality.
The conversation is formatted as[{'human': '...', 'assistant': '...'}, ..., {'human': '...', 'assistant': '...'}].
The information density of a conversation is defined as a number of information units (e.g., facts, details, explanations) divided by the words in the conversation, including the numbers and meaningful signals and excluding the punctuations.
You should only consider the information related to the topic of the conversation and ignore any irrelevant or redundant information.
If the information unit is repeated in the conversation, it should be counted only once.
[Example 1]
Input Conversation: [{'human': 'What is the capital of France?', 'assistant': 'The capital of France is Paris.'}]
Output: {'Number of Information Units': 1, 'Total Number of Words': 12, 'Information Units': ['The capital of France is Paris.']}
[Example 2]
Input Conversation: [{'human': 'What is the capital of France?', 'assistant': 'The capital of France is Paris.'}, {'human': 'What is the population of Paris?', 'assistant': 'The population of Paris is 2.1 million.'}]
Output: {'Number of Information Units': 2, 'Total Number of Words': 25, 'Information Units': ['The capital of France is Paris.', 'The population of Paris is 2.1 million.']}

[Conversation]\n""" + <conversation> + "\n" + """
The evaluation and your output must be strictly structured in the following format:
{
"Number of Information Units": <Number of information units>,
"Information Units": [<Information unit 1>, <Information unit 2>, ...],
"Total Number of Words": <Total number of words>,
}
"""

Figure 9: Prompts of the evaluation of information density.

Prompts of the Evaluation of Friendliness

"""
You are a strict and objective evaluator. Your task is to assess the friendliness of the given conversation following the Assessment Criteria.
Your evaluation should be fair, professional, and reflect an expert judgment of the response's quality.
The conversation is formatted as [{'human': '...', 'assistant': '...'}, ..., {'human': '...', 'assistant': '...'}].

{[Conversation]}""" + <conversation> + """
Assessment Criteria:
Manner: Concentrate on the requests from the human, and also evaluate the assistant's responses, is there evidence of disobeying the rules or aggressive behaviors?
Security: Evaluate the security of the conversation. Does the conversation tend to be harmful or offensive, or does the response from the assistant being guided to reveal sensitive information?
Tone: Evaluate the overall tone of the conversation. Does it have a positive and friendly tone?
Politeness: Evaluate the politeness and courtesy of the assistant's responses. Overall Score: Assign a score from 1 to 10 (10 being the most friendly), considering all of the above factors.

The evaluation and your output must be strictly structured in the following JSON format:

"Explanation": "<Explain the rationale of your score.>",
"Score": <An integer score from 1 to 10.>

"""

Figure 10: Prompts of the evaluation of friendliness.

Dimension	**Score**	**Description**
Faithfulness	1	Completely irrelevant or ignores prior context, leading to a fundamentally incorrect answer.
	2	Contains substantial irrelevant or contradictory content, but barely addresses the request.
	3	Accurately addresses the request but neglects useful context from earlier dialogue.
	4	Fully accurate, relevant, and contextually faithful to both current and prior user inputs.
Appropriateness	1	Severely off-topic, misinterprets the question, or violates conversational context.
	2	Partially relevant but includes misinterpretations or contextual inconsistencies.
	3	Mostly appropriate with only minor contextual or interpretative issues.
	4	Fully appropriate and consistent with both the question and dialogue context.
Naturalness	1	Highly unnatural, disfluent, or grammatically flawed to the point of harming comprehension.
	2	Understandable but includes awkward phrasing or noticeable language errors.
	3	Mostly fluent and natural, with minor phrasing issues.
	4	Fully fluent, smooth, and human-like in style.
Completeness	1	Severely incomplete, omits critical information needed for the response.
	2	Partially complete, with several important details missing.
	3	Mostly complete but misses some minor elaborations.
	4	Fully complete and comprehensive in addressing the user's request.

Table 8: Human evaluation criteria for MT-Bench responses evaluation (1–4 scale).