Data Selection for Multi-turn Dialogue Instruction Tuning

Bo Li[1,2,3], **Shikun Zhang**[1], **Wei Ye**[1]*
[1]National Engineering Research Center for Software Engineering, Peking University
[2]School of Computer Science, Peking University
[3]PKU-CMCC(Hubei) Joint Research Lab for LLM Industrial Applications
deepblue.lb@gmail.com, wye@pku.edu.cn
WisdomShell/MDS MDS Project

Abstract

Instruction-tuned language models increasingly rely on large multi-turn dialogue corpora, but these datasets are often noisy and structurally inconsistent, with topic drift, repetitive chitchat, and mismatched answer formats across turns. We address this from a data selection perspective and propose **MDS** (Multi-turn Dialogue Selection), a dialogue-level framework that scores whole conversations rather than isolated turns. MDS combines a global coverage stage that performs bin-wise selection in the user-query trajectory space to retain representative yet non-redundant dialogues, with a local structural stage that evaluates within-dialogue reliability through entity-grounded topic grounding and information progress, together with query-answer form consistency for functional alignment. MDS outperforms strong single-turn selectors, dialogue-level LLM scorers, and heuristic baselines on three multi-turn benchmarks and an in-domain Banking test set, achieving the best overall rank across reference-free and reference-based metrics, and is more robust on long conversations under the same training budget. Code and resources are included in the supplementary materials.

1 Introduction

Supervised fine-tuning on instruction-style data is now a central step in turning base language models into aligned assistants, from RLHF to recent instruction-tuned open-source models (Ouyang et al., 2022; Wang et al., 2022b; Taori et al., 2023; Köpf et al., 2023; Dubey et al., 2024; Yang et al., 2025; Tian et al., 2025a,b; Zhao and Yan, 2026). Yet a series of studies have shown that simply increasing dataset size is not sufficient and can even hurt downstream behavior when the supervision is noisy, redundant, or off-distribution (Zhou et al., 2023; Wang et al., 2023a; Li et al., 2024b, 2026).

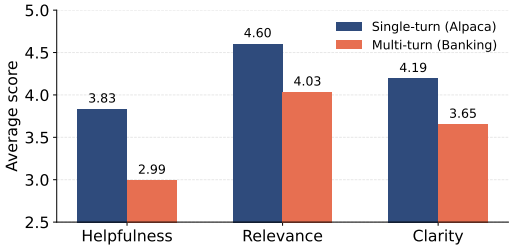

Figure 1: Comparison of turn-level quality between a single-turn instruction dataset (Alpaca) and a multi-turn dialogue dataset (Banking). We randomly sample 1,000 examples from each dataset and score the assistant responses for helpfulness, relevance and clarity on a 1–5 scale using GPT-4o as the evaluator.

Work on small, high-quality alignment sets consistently shows that data composition matters for shaping model behavior (Qi et al., 2023; Dong et al., 2024; Shen, 2024), motivating instruction-data selection and reweighting methods that aim to identify the most beneficial supervision signals for downstream capabilities.

Most existing work on data selection still focuses on single-turn instruction–response pairs, where examples are easy to synthesize and each instance can be scored in isolation. Recent methods select or reweight instructions based on self-guided signals, LLM quality scores, or simple heuristics such as response length, and have shown clear gains for instruction tuning (Wang et al., 2022a; Li et al., 2024a; Liu et al., 2024a; Xia et al., 2024a,b; Mekala et al., 2024; He et al., 2025). In contrast, multi-turn dialogue data are usually collected from human–assistant interaction logs or large-scale synthetic generators (Ding et al., 2023; Wang et al., 2023b; Xu et al., 2023; Zheng et al., 2023), and we empirically find that their quality is often lower and more variable across turns. A simple turn-level comparison already makes this gap visible: Figure 1 compares a standard single-turn instruction dataset (Alpaca (Peng et al., 2023)) with a multi-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 2724–2739

turn dialogue corpus (Banking) and shows that the latter consistently receives lower scores in helpfulness, relevance, and clarity on a 1–5 scale. Beyond lower average scores, manual inspection reveals that the multi-turn corpus suffers from characteristic dialogue-level failures: later turns often drift away from the user's original intent, many conversations end with long chitchat tails, and some responses ignore the requested format (e.g., open-ended advice instead of concrete steps). These issues are hard to detect from isolated turns but accumulate over trajectories, degrading the value of multi-turn supervision. While several works have begun to construct large multi-turn corpora and analyze consistency (Liu et al., 2023; Lin and Chen, 2023; Chen et al., 2025), their processing pipelines still rely mainly on rule-based or coarse filtering. These gaps motivate us to develop a dialogue-level data selection method that explicitly targets multi-turn structure and conversational quality, rather than treating each turn as an independent single-turn instruction.

In this paper, we introduce **MDS** (**M**ulti-turn **D**ialogue **S**election), combining a *global semantic coverage* stage with a *local structural* stage. In the global stage, MDS embeds each dialogue into a *user-query trajectory* representation, which captures the evolving intent while being robust to assistant-side chitchat. We then partition the trajectory space into semantic bins and perform *bin-wise semantic coverage* selection within each bin using an efficient greedy coverage–redundancy criterion, yielding a representative yet non-redundant subset under a strict budget. This global mechanism explicitly prevents a few high-frequency interaction patterns from dominating the selection and improves long-tail intent coverage. In the local stage, MDS assesses *within-dialogue structural reliability* by measuring entity-grounded *topic grounding* and *information progress* across turns, together with query-answer form consistency that enforces functional alignment between query types and response formats. By prioritizing dialogues that are both well-covered in the trajectory space and structurally reliable, MDS constructs a compact multi-turn training set that is simultaneously semantically diverse and well formed.

We validate MDS on two multi-turn training corpora, one general-purpose assistant dataset and one domain-specific customer-service dataset, each under a fixed 10K-dialogue selection budget. We compare against strong single-turn selection methods adapted to dialogue turns, various LLM-based multi-turn selectors, and several simple baselines. Across datasets and metrics, MDS consistently matches or surpasses these baselines under both reference-free and reference-based automatic evaluation, with particularly clear gains on measures of content coverage and fidelity. Beyond the main results, we conduct ablations that isolate the contribution of each component, and we provide in-depth analyses, visualizations, and case studies that show how MDS suppresses noisy conversations while preserving diverse yet well structured dialogues. Overall, our contributions are two-fold:

- We propose MDS, a two-stage global and local framework for selecting multi-turn supervision based on semantic coverage and structural quality.
- We demonstrate that MDS consistently improves multi-turn performance over state-of-the-art selection and filtering schemes on both general-purpose and domain-specific corpora. We further introduce structural diagnostics that explain which types of dialogue-level noise MDS suppresses in practice.

2 Multi-turn Dialogue Selection

2.1 Problem Setup and Overview of MDS

Let $D = \{d_1, \ldots, d_N\}$ be a pool of multi-turn dialogues. Each dialogue d is a sequence of user-assistant exchanges

$$d = \{(Q_1, A_1), \ldots, (Q_T, A_T)\}, \qquad (1)$$

where Q_t and A_t denote the user query and assistant response at turn t. Given a fixed budget of M dialogues, our goal is to select a subset D^* that provides the most useful supervision for multi-turn instruction tuning. Most existing data selection methods score isolated instruction-response pairs, ignoring the multi-turn conversational structure across turns. We instead score whole dialogues and target a subset that both covers diverse user intents and consists of structurally well-formed conversations, so the model learns from coherent, informative trajectories rather than noisy or repetitive exchanges.

To this end, we propose **MDS** (Multi-turn Dialogue Selection), a two-stage framework that combines *global semantic coverage* with *local structural quality*. In the global stage, MDS represents

each dialogue by a user-query trajectory embedding, partitions the trajectory space into semantic bins, and performs bin-wise coverage selection with redundancy control to retain conversations that are representative yet diverse. In the local stage, MDS assesses each candidate dialogue using complementary structure signals, including entity-grounded coherence and novelty, as well as query-answer form consistency, and then prioritizes dialogues with stronger structural reliability.

2.2 Global Stage: Semantic Coverage over Dialogues

The global stage constructs a dialogue candidate pool that *covers* diverse user intents and interaction patterns while controlling redundancy. Instead of scoring individual turns, we perform selection in a dialogue-level *trajectory space* derived from the user side. Concretely, we build representations from user queries rather than assistant responses: queries provide a stable signal of the underlying intent and task type, whereas responses often contain stylistic noise, templated phrasing, or low-quality content that can distort semantic grouping. Query-based trajectory representations therefore offer a cleaner basis for coverage-aware selection.

Query-trajectory representation. Given a dialogue d with T user turns, we encode each user query Q_t into an embedding $q_t \in \mathbb{R}^h$ using a sentence encoder, and aggregate them into a dialogue-level *query-trajectory* embedding:

$$v_d = \frac{1}{T} \sum_{t=1}^{T} q_t. \tag{2}$$

Each v_d summarizes the overall semantic trajectory of the user requests, so dialogues centered on similar tasks tend to be close in this space.

Bin-wise semantic coverage. A single global ranking over $\{v_d\}$ is prone to being dominated by high-frequency templates, which can reduce coverage of long-tail intents. To mitigate this, we partition the trajectory space into K semantic bins and enforce selection within each bin. We cluster $\{v_d\}_{d \in D}$ (e.g., with K-means), obtaining bins $\{B_k\}_{k=1}^{K}$ with centroids

$$c_k = \frac{1}{|B_k|} \sum_{d \in B_k} v_d. \tag{3}$$

Bin-wise coverage with redundancy control. Even within the same semantic bin, many dialogues can be near-duplicates. We therefore perform bin-wise greedy selection with redundancy control, which implements a practical coverage-diversity trade-off. Let S_k denote the set of dialogues already selected from bin B_k. For a candidate $d_i \in B_k$, we define its representativeness and redundancy as

$$s_i = \mathrm{sim}(v_{d_i}, c_k), \quad r_i = \max_{d_j \in S_k} \mathrm{sim}(v_{d_i}, v_{d_j}), \tag{4}$$

where $\mathrm{sim}(\cdot, \cdot)$ is cosine similarity. Starting from $S_k = \emptyset$, we iteratively add the next dialogue by maximizing a greedy marginal objective:

$$d^{\text{next}} = \arg \max_{d_i \in B_k \setminus S_k} \Big(\lambda s_i - (1 - \lambda) r_i \Big), \tag{5}$$

with $\lambda \in [0, 1]$ and we set $\lambda = 0.5$ in our experiments.

Output. We use the global stage to construct a reduced candidate pool for efficient local scoring. Specifically, for each bin B_k we run the above procedure and keep the top α fraction of selected dialogues, denoted $\tilde{B}_k \subseteq B_k$ with $|\tilde{B}_k| = \lceil \alpha |B_k| \rceil$ (we use $\alpha = 0.5$), and output the global candidate pool $D^{\text{global}} = \bigcup_{k=1}^{K} \tilde{B}_k$. This candidate pool maintains broad semantic coverage with redundancy control, while substantially reducing the computational cost of the local-stage scorer. In practice, the global stage keeps the top α fraction per bin to form D^{global} for efficiency, while the final per-bin budget m_k is applied in the local stage.

2.3 Local Stage: Structural Quality within Dialogues

The local stage refines the candidate pool D^{global} by assessing *within-dialogue structural reliability*. While the global stage targets semantic coverage in the trajectory space, the local stage focuses on whether a dialogue provides *usable multi-turn supervision*: it filters conversations that drift off-topic, collapse into repetition, or exhibit systematic query-answer form mismatches, and then performs budgeted selection within the semantic bins defined by the global stage. All signals in this stage are computed in a reference-free manner using a lightweight instruction-tuned scorer, making the procedure efficient and model-agnostic.

Signal 1: Entity-grounded coherence and novelty. Intuitively, a good multi-turn dialogue

should maintain *topic grounding* (staying anchored to user-mentioned entities) while ensuring *information progress* (introducing new, informative content rather than repeating earlier responses). For each turn t, we prompt the scorer to extract three entity sets: entities in the current answer E_t^A, entities mentioned in user queries up to turn t, denoted $E_{\leq t}^Q$, and entities appearing in previous answers $E_{<t}^A$. We then quantify two complementary aspects at each turn: (i) *anchoring*, encouraging answers to stay grounded in what the user is asking about, and (ii) *novelty*, rewarding answers that introduce informative new entities rather than repeating prior content. Formally, we define the per-turn entity score

$$\mathrm{ent}_t = \frac{|E_t^A \cap E_{\leq t}^Q|}{|E_t^A|} + \frac{|E_t^A \setminus E_{<t}^A|}{|E_t^A|}, \quad (6)$$

when $|E_t^A| > 0$ (and set $\mathrm{ent}_t = 0$ otherwise). Averaging over turns yields a dialogue-level structural score:

$$s_{\mathrm{entity}}(d) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{ent}_t. \quad (7)$$

Signal 2: Query-answer form consistency. Beyond topical grounding, multi-turn supervision also requires *functional alignment*: responses should match the form implied by the query type, such as step-by-step procedures for troubleshooting requests, explicit comparisons for comparative queries, or concrete recommendations for advice queries. For each turn t, we prompt the scorer with (Q_t, A_t) (and minimal context) to rate how well the *form* of A_t satisfies the expected form of Q_t on a three-point scale $c_t \in \{0, 1, 2\}$. We then define the dialogue-level form-consistency score as

$$s_{\mathrm{form}}(d) = \frac{1}{T}\sum_{t=1}^{T} c_t. \quad (8)$$

Bin-wise budgeted refinement. After computing $s_{\mathrm{entity}}(d)$ and $s_{\mathrm{form}}(d)$ for all $d \in D^{\mathrm{global}}$, we perform budgeted selection within bins using the candidate sets $\{\tilde{B}_k\}$ from the global stage. We first apply form consistency as a necessary-condition filter:

$$S_k^{\mathrm{form}} = \{\, d \in \tilde{B}_k : s_{\mathrm{form}}(d) \geq \tau_{\mathrm{form}} \,\}. \quad (9)$$

We then allocate a per-bin budget m_k proportional to the original bin size $|B_k|$ under the overall budget M (so that $\sum_k m_k = M$), and within each S_k^{form} select the top m_k dialogues ranked by $s_{\mathrm{entity}}(d)$ (or all of them if $|S_k^{\mathrm{form}}| < m_k$). The union of these bin-level subsets forms the final training set D^*.

3 Experimental Setup

3.1 Training and Evaluation Datasets

We evaluate MDS on both general-purpose and domain-specific multi-turn corpora. For training, we use **Baize** (Xu et al., 2023) as a general assistant corpus and **Banking** [1] as a domain-specific customer-service corpus. For evaluation, we adopt three public benchmarks (**MT-Eval** (Kwan et al., 2024), **ConsistentChat** (Chen et al., 2025), **TopDial** (Wang et al., 2023b)) and a **Banking Test** of 1,000 held-out Banking dialogues that are never used for training. These test sets jointly cover open-ended assistance, consistency-sensitive exchanges, and task-oriented dialogues; detailed statistics are provided in the Appendix A.

3.2 Evaluation Metrics

We use three types of metrics to evaluate multi-turn dialogue quality. **Reference-free metrics.** Because many user queries are open-ended, we rely on strong LLM judges. Specifically, we adopt *LLM-EVAL* (Lin and Chen, 2023) and *G-EVAL* (Liu et al., 2023) with GPT-4o as the judge, scoring each dialogue on several 0-10 dimensions (e.g., helpfulness, relevance, coherence). For each test set, we report the average score on each dimension and the average across dimensions; prompts and rubrics are provided in the Appendix E. **Reference-based metrics.** We further report an *Ent-F1* score, obtained by using GPT-4o to extract entities from reference and generated answers and computing F1 over entities aggregated across turns, which reflects how well the model covers key entities in the gold dialogue. We also use a *Cos (cosine similarity)* score, defined as the cosine similarity between sentence-level embeddings of the reference and generated answers computed by a Sentence-Transformers encoder (the *all-MiniLM-L6-v2* variant[2]). **Aggregate comparison.** Since these metrics have different scales, we also report the *Average Rank* of each method across all metrics as a scale-free summary, where lower is better.

[1] https://huggingface.co/datasets/talkmap/banking-conversation-corpus

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

	MT-Eval				ConsistentChat				TopDial				Avg.
	L-E	G-E	Ent-F1	Cos	L-E	G-E	Ent-F1	Cos	L-E	G-E	Ent-F1	Cos	Rank
BACKBONE: LLAMA3-8B-INSTRUCT													
Backbone	8.04	7.44	0.569	0.831	8.11	6.73	0.222	**0.808**	**7.12**	**6.68**	0.158	**0.499**	5.00
All Data	8.09	7.41	0.567	0.845	8.42	7.20	0.310	0.794	6.61	6.20	0.139	0.409	5.67
Random Data	8.00	7.41	0.558	0.842	8.46	7.20	0.306	0.783	6.58	6.20	0.140	0.409	6.92
SuperFiltering	8.08	7.44	0.568	0.846	8.38	7.11	0.301	0.788	6.98	6.27	0.156	0.458	4.92
Rethinking	8.01	7.43	0.568	0.845	8.41	7.18	0.310	0.789	6.81	6.36	0.144	0.426	5.50
ZIP	8.06	7.42	0.570	0.845	8.42	7.18	0.294	0.781	6.78	6.28	0.153	0.433	5.58
DialScore	8.05	7.21	0.567	0.845	8.44	7.13	0.307	0.787	6.90	6.29	0.145	0.430	6.00
Heuristic	7.99	7.23	0.562	0.838	8.43	7.10	0.300	0.792	6.97	6.20	0.151	0.436	7.00
CC-Score	8.05	7.48	0.570	0.845	8.41	7.16	0.305	0.788	6.82	6.38	0.151	0.436	4.58
MDS	**8.16**	**7.52**	**0.584**	**0.857**	**8.52**	**7.26**	**0.316**	0.797	**7.12**	6.48	**0.173**	0.465	**1.25**
BACKBONE: QWEN3-8B-INSTRUCT													
Backbone	7.81	7.90	0.496	0.826	6.68	7.11	0.184	0.711	**7.71**	**8.25**	0.145	0.392	7.25
All Data	7.90	8.08	0.568	0.843	8.28	7.96	0.301	0.793	7.14	7.54	0.123	0.390	7.25
Random Data	7.96	8.05	0.558	0.844	8.31	7.98	0.312	0.799	7.15	7.52	0.134	0.426	5.58
SuperFiltering	8.01	8.20	0.581	0.847	8.26	7.96	0.310	0.802	7.16	7.41	0.117	0.411	5.33
Rethinking	7.91	8.08	0.575	0.846	8.29	8.00	0.295	0.792	7.22	7.62	0.112	0.423	5.75
ZIP	7.98	8.12	0.564	0.841	8.32	7.98	0.314	0.791	7.13	7.58	0.110	0.414	6.17
DialScore	7.92	8.21	0.585	**0.850**	8.32	7.97	0.300	0.802	7.14	7.56	0.116	0.421	4.67
Heuristic	8.00	8.10	0.553	0.828	8.36	8.00	0.307	0.793	7.21	7.55	0.114	0.412	5.67
CC-Score	8.05	8.18	0.579	0.845	8.33	7.97	0.299	0.798	7.16	7.46	0.128	0.431	4.75
MDS	**8.16**	**8.26**	**0.593**	0.848	**8.44**	**8.04**	**0.338**	**0.822**	7.32	7.70	**0.150**	**0.451**	**1.25**

Table 1: Main results of MDS and baseline selection methods on Baize dataset and three multi-turn benchmarks. **L-E** and **G-E** denote **LLM-EVAL** and **G-EVAL**, respectively; **Ent-F1** and **Cos** denote entity-level F1 and embedding cosine similarity, respectively. All reported scores are averaged over 5 runs for each method. Bold numbers denote the best score and underlined numbers denote the second-best score in each column; the rightmost column reports the average rank over all 12 metrics, where lower is better.

3.3 Baseline Methods

We compare MDS against three groups of methods.

1) Single-turn selection. We include three state-of-the-art selectors: `SuperFiltering` (Li et al., 2024a), `Rethinking` (Xia et al., 2024b), and `ZIP` (Yin et al., 2024). Since they operate on query-answer pairs, we adapt them to dialogues by scoring each turn, aggregating turn-level scores into a single dialogue score, and selecting dialogues under the same 10K-dialogue budget as MDS.

2) Multi-turn selection. We consider three dialogue-level selectors: (i) A consistency-focused scoring baseline from `ConsistentChat` **(CC-Score)** (Chen et al., 2025). We use `Qwen3-32B` and their released prompts (without modification) to evaluate dialogue quality. (ii) A simple **DialScore** baseline that directly prompts the same model to assign a single 1–10 overall score to each dialogue under a generic rubric. (iii) A lightweight **Heuristic** baseline that filters dialogues by simple statistics (e.g., proportion of very short answers, self-repetition, lexical diversity) and rank by a composite heuristic score. Please refer to Appendix H for more details.

3) Others. We also report `Random Data` (uniformly sampling 10K dialogues), `All Data` (using all available training dialogues), and the unfine-tuned `Backbone`. These baselines help disentangle the effect of data selection from model capacity and training budget. For all selection methods, including MDS, we ultimately obtain a 10K-dialogue training subset from each corpus and expand it into turn-level supervision for fine-tuning.

3.4 MDS Configuration

We fine-tune **LLaMA3-8B-Instruct** and **Qwen3-8B-Instruct** with LoRA adapters (r=64, α=128, dropout 0.1) on each 10K-dialogue subset. We use batch_size=2, gradient_accumulation_steps=16, num_train_epochs=3, learning_rate=1e-5, and warmup_ratio=0.05 with a cosine schedule. In the global stage, we encode user queries with a Sentence-Transformers encoder (*all-MiniLM-L6-v2*) to obtain dialogue-level trajectory embeddings, cluster them into K=1000 semantic bins $\{B_k\}$ with K-means, keeping the top 50% dialogues per bin to form the candidate pool D^{global}.

In the local stage, we use Qwen3-8B-Instruct with greedy decoding to compute both the entity coherence–novelty score $s_{\text{entity}}(d)$ and the form-consistency score $s_{\text{form}}(d)$, applying simple normalization for entities and a fixed threshold τ_{form}=1.0 on $s_{\text{form}}(d)$ to filter out low-quality dialogues. The prompt is provided in the Appendix G. We allocate bin-level quotas m_k proportional to $|B_k|$ with rounding so that $\sum_k m_k = 10{,}000$, and within each bin keep the top-m_k dialogues ranked by $s_{\text{entity}}(d)$ after the form filter, yielding the final subset D^* for each corpus. Selection is performed entirely offline, and the resulting subsets are used for all backbones.

4 Main Results

4.1 General-Domain Results

Table 1 shows that MDS delivers consistent gains across backbones, achieving the best average rank on both backbones, which indicates that the improvements are not model-specific. MDS also mitigates degradation from training on the full noisy pool on the task-oriented TopDial benchmark: for LLaMA3-8B, `All Data` reduces TopDial L-E from 7.12 to 6.61, while MDS preserves 7.12 and attains the best TopDial Ent-F1 (0.173). This pattern suggests that indiscriminate multi-turn supervision can be harmful, and coverage-aware selection helps retain task-relevant dialogue behaviors.

MDS yields the most consistent improvements on structure-sensitive signals, aligning with our goal of improving within-dialogue reliability. In contrast, adapting strong single-turn selectors to dialogues remains insufficient, and even dialogue-aware baselines lag behind MDS, highlighting the need to control both dialogue-level coverage and within-dialogue structure. Notably, these gaps persist across all three benchmarks, showing that MDS improves not only general helpfulness scores but also consistency-oriented measures that reflect multi-turn quality.

To check robustness, we re-scored all outputs with Qwen3-32B as the judge. The two evaluators agree on 92.1% of **instance-level** pairwise preferences; at the **system level** (ranking methods by their average scores), their rankings are also highly correlated (Spearman's $\rho = 0.89$). This suggests our conclusions are not tied to a particular judge.

	Banking Test		ConsistentChat	
	G-E	Ent-F1	G-E	Ent-F1
Backbone	6.28	0.184	6.70	0.222
All Data	6.58	**0.354**	7.12	0.288
Random Data	6.42	0.313	7.22	0.290
SuperFiltering	6.44	0.305	7.12	0.283
Rethinking	6.62	0.333	7.18	0.282
ZIP	6.60	0.304	7.16	0.285
DialScore	6.50	0.321	7.20	0.291
Heuristic	6.58	0.321	7.22	0.278
CC-Score	6.64	0.319	7.16	0.285
MDS	**6.72**	0.351	**7.30**	**0.300**

Table 2: Domain-specific selection performances on the Banking corpus.

4.2 Domain-Specific Results

Table 2 reports results when all methods select 10K dialogues from the Banking corpus and we evaluate both in-domain (Banking Test) and out-of-domain (ConsistentChat) performance. On Banking Test, MDS attains the highest G-E score (6.72) while nearly matching the best entity coverage (Ent-F1 0.351 vs. 0.354 for All Data), thus preserving the gains of using all dialogues but with better conversational quality. Out-of-domain on ConsistentChat, MDS again achieves the highest G-E and Ent-F1, indicating that selecting Banking dialogues via MDS does not simply overfit to the customer-service domain but yields supervision that transfers better to a different multi-turn benchmark. Overall, these results complement the general-domain findings in Table 1 and show that MDS can enhance both in-domain robustness and cross-domain generalization for domain-specific dialogue pools.

4.3 Ablation Study

	MT-Eval		TopDial	
	G-E	Ent-F1	G-E	Ent-F1
MDS	**7.52**	**0.584**	**6.48**	**0.173**
Global-only	7.38	0.580	6.30	0.157
Local-only	7.44	0.576	6.46	0.162
w/o Binning	7.44	0.570	6.26	0.145
w/o Form Filter	7.42	0.574	6.46	0.148

Table 3: Ablation on the components of MDS using LLaMA3-8B. All rows are variants of MDS.

Table 3 reports ablations on the components of MDS. Removing either stage degrades performance: the *Global-only* variant and the *Local-only* variant are consistently worse than *Full MDS*, and neither matches its entity-level gains, showing that

semantic coverage and structural scoring are complementary. Within the global stage, turning off binning (*w/o Binning*) keeps a similar G-E score but noticeably harms Ent-F1 on TopDial, indicating that semantic bins are important for preserving long-tail intents while de-duplicating frequent patterns. Within the local stage, removing the query-answer form filter (*w/o Form Filter*) slightly changes G-E but reduces Ent-F1 on TopDial, confirming that hard filtering on form consistency contributes to higher-quality supervision. Detailed ablation for different numbers of semantic bins K are provided in Appendix F.

5 Analysis

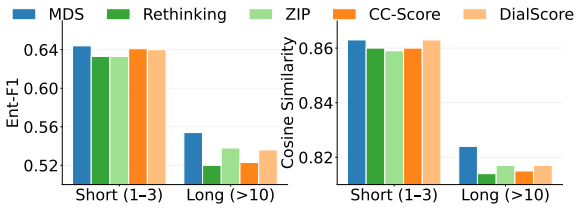

Figure 2: Performance on short (turns 1–3) vs. long (turns >10) queries on MT-Eval. We show Ent-F1 and cosine similarity averaged over turns in each bucket. Blue bars denote MDS, green bars adapted single-turn selectors (Rethinking, ZIP), and orange bars dialogue-level baselines (CC-Score, DialScore).

5.1 Robustness across Dialogue Lengths

To examine how selection strategies behave as conversations grow longer, we bucket MT-Eval turns by position into *short* (user queries at turns 1–3) and *long* (turns $>$ 10), and recompute metrics within each bucket. Figure 2 reports Ent-F1 and cosine similarity for five selection methods.

On short turns, all methods perform similarly; MDS is slightly ahead on Ent-F1 and cosine similarity. On long turns, the gap becomes more pronounced. All methods lose Ent-F1, but MDS degrades the least and maintains a clear margin over the best baseline (0.554 vs. 0.538 for ZIP and 0.536 for DialScore), and a similar trend holds for cosine similarity. This suggests that combining global semantic coverage with local structural filtering yields training data that better preserves entity coverage and semantic fidelity in later turns, making MDS more robust to length-induced degradation.

	All Selected	**Top 20% by $H(d)$**		
	ESC	ESC	HAR	ENR
MDS	0.599	0.614	0.514	0.714
shuffle level				
Pair	0.596	0.602	0.497	0.707
Block(k=2)	0.596	0.606	0.504	0.708
Block(k=4)	0.596	0.603	0.498	0.707
Query-only	0.547	0.560	0.407	0.713

Table 4: Order-perturbation analysis on the same 10K dialogues selected by MDS. We report the turn-weighted Entity Sequence Consistency score (ESC) on the full set (**All Selected**), and additionally report ESC together with two interpretable components on the Top 20% high-history-dependency subset ranked by $H(d)$: History Anchoring Rate (HAR) and Entity Novelty Rate (ENR).

5.2 Order Perturbation Analysis: Quantifying Cross-Turn Dependency

We conduct a controlled counterfactual analysis to isolate whether the gains of MDS are truly driven by preserving cross-turn structure. Specifically, we fix the training set to the same 10K dialogues selected by MDS and only apply order-level perturbations to the dialogue organization: *Pair shuffle* performs local swaps of adjacent QA pairs, *Block shuffle*$(k{=}2/4)$ reorders turns in larger blocks with higher disruption for larger k, and *Query-only shuffle* breaks query-answer correspondence as a stronger content-mismatch baseline. We additionally evaluate a *high-history-dependency* subset (Top 20% by $H(d)$), where $H(d)$ is a dialogue-level score computed from our turn-wise structural signals to quantify how strongly later turns depend on earlier turns, characterized by higher history anchoring and lower entity novelty (i.e., more reuse of previously introduced entities).

For evaluation, we use **ESC** (*Entity Sequence Consistency*) as an order-sensitive overall score, and further decompose Top 20% behavior into two interpretable factors: **HAR** (*History Anchoring Rate*) and **ENR** (*Entity Novelty Rate*)[3]. The results show that order shuffles primarily degrade cross-turn consistency on the high-dependency subset, and the degradation is mainly driven by weakened history anchoring (HAR), while novelty (ENR) remains relatively stable. This pattern indicates that the main failure mode of order perturbations is breaking history anchoring rather than reducing entity novelty, reinforcing our design choice of ex-

[3]All metric definitions and their exact computation formulas are provided in Appendix B.

plicitly modeling both anchoring and redundancy in the local stage.

5.3 Error-type Analysis on Difference Sets

To better understand what kinds of dialogues MDS prefers, we analyze *difference sets* between MDS and strong baselines $B \in \{\text{CC-Score}, \text{DialScore}, \text{Rethinking}, \text{SuperFiltering}\}$. For each B, we construct *MDS-only* ($\mathcal{D}_{MDS} \setminus \mathcal{D}_B$) and *Baseline-only* ($\mathcal{D}_B \setminus \mathcal{D}_{MDS}$), uniformly sample 1,000 dialogues from each, and ask GPT-4o to assign a primary label from six categories: **No Error**, **Topic Drift**, **Repetition**, **Form Mismatch**, **Contradiction**, and **Unsupported**. We then compute the percentage-point gap $\Delta = p(\text{MDS-only}) - p(\text{Baseline-only})$ for each error type (Figure 3), where negative Δ indicates fewer errors in MDS-only (while positive Δ is desirable for **No Error**). Appendix C shows the classification prompt.

Figure 3: Error-type gaps on difference sets between MDS and each baseline selector. Each column compares MDS against one baseline. For example, in the *DialScore* column, a cell value is the percentage-point gap $\Delta = p(\textit{MDS-only}) - p(\textit{DialScore-only})$ for that error type. Detailed statistics are provided in the Appendix D.

Figure 3 shows several consistent trends. MDS-only subsets contain noticeably more clean dialogues: **No Error** increases by $+5.0$ to $+9.5$ percentage points across baselines (avg. $+6.9$), indicating that MDS allocates more capacity to well-formed multi-turn supervision. At the same time, dialogue-level failures that directly hurt cross-turn learning are suppressed: **Topic Drift** is reduced for every baseline (avg. -3.4, up to -5.5), and **Unsupported** content also drops substantially (avg. -2.5, up to -5.0), suggesting that MDS-selected dialogues stay closer to the user's intent and make fewer unjustified claims. By contrast, **Form Mismatch** is rare in both subsets, and gaps for **Repetition** and **Contradiction** are small and sometimes mixed, implying that they are not the main drivers of the observed gains. Overall, this error profile highlights MDS's advantage: it reshapes the training pool toward on-topic, grounded, and structurally coherent conversations.

6 Related Work

A large body of work studies how to select high-quality supervision for single-turn instruction tuning, using carefully curated small datasets or automated selection based on LLM scores, weak-to-strong filtering, uncertainty or influence estimates, and distribution-matching objectives (Zhou et al., 2023; Chen et al., 2024; Li et al., 2024b,a; He et al., 2025; Liu et al., 2024a; Zhang et al., 2025; Xia et al., 2024a; Wang et al., 2022a; Li et al., 2023; Zhou et al., 2026; Li et al., 2026). Recent work further shows that smaller models can act as selectors for larger models and that random selection can be a surprisingly strong baseline under controlled settings (Xia et al., 2024b; Mekala et al., 2024), but these approaches operate on isolated instruction–response pairs and do not model dialogue-level coverage or multi-turn structure.

In the multi-turn setting, prior work has mainly focused on constructing datasets and benchmarks rather than dialogue-level selectors. MT-Eval and related efforts evaluate multi-turn capabilities with GPT-based judges (Kwan et al., 2024; Liu et al., 2024b), and corpora such as UltraChat, ConsistentChat, Baize, ShareGPT, LMSYS-Chat-1M, TopDial, and LIGHT provide large-scale synthetic or real conversations (Ding et al., 2023; Chen et al., 2025; Wang et al., 2023b; Urbanek et al., 2019; Xu et al., 2023). However, their pipelines typically rely on rule-based cleaning or generic LLM filtering at the utterance/turn level, without explicitly enforcing *semantic coverage* and *structural quality* for complete dialogues under a *fixed selection budget*. MDS fills this gap by enforcing *global semantic coverage* and complementing it with *local structural scoring* for reliability within each dialogue.

7 Conclusion

In this paper, we proposed MDS (Multi-turn Dialogue Selection), a dialogue-level data selection framework for multi-turn instruction tuning. MDS combines a global coverage stage that selects representative yet non-redundant dialogue trajectories with a local structure stage that measures entity-level coherence and query-answer form consistency. Experiments on Baize and a Banking corpus show that MDS outperforms strong single-turn selectors,

dialogue-level LLM scorers, and heuristic baselines across both reference-free and reference-based metrics. Ablation and analysis further indicate that MDS is more robust on long conversations and reduces topic drift and unsupported claims, suggesting that dialogue-level structure is a powerful signal for curating cleaner and more reliable supervision for conversational models.

Limitations

A main limitation of MDS is that it does not substantially reduce *Contradiction* errors in our error-type analysis. We suspect this is partly because MDS deliberately retains longer and structurally richer dialogues, where cross-turn dependencies make subtle inconsistencies and implicit conflicts harder to avoid, even when topic grounding and form consistency are satisfied.

References

Jiawei Chen, Xinyan Guan, Qianhao Yuan, Mo Guozhao, Weixiang Zhou, Yaojie Lu, Hongyu Lin, Ben He, Le Sun, and Xianpei Han. 2025. Consistentchat: Building skeleton-guided consistent multi-turn dialogues for large language models from scratch. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8426–8452.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Angela Fan, Amy Yang, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Junliang He, Ziyue Fan, Shaohui Kuang, Li Xiaoqing, Kai Song, Yaqian Zhou, and Xipeng Qiu. 2025. Fine: Filtering and improving noisy data elaborately with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8686–8707.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and others. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681.

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

Bo Li, Mingda Wang, Shikun Zhang, and Wei Ye. 2026. Instruction data selection via answer divergence. *Preprint*, arXiv:2604.10448.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024a. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530*.

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

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. 2023. One shot learning as instruction data prospector for large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024a. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. *Advances in Neural Information Processing Systems 37*.

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*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. 2024. Smaller language models are capable of selecting instruction-tuning training data for larger language models. *ArXiv*, abs/2402.10430.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *ArXiv*, abs/2304.03277.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*.

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.

Yuhang Tian, Dandan Song, Zhijing Wu, Pan Yang, Changzhi Zhou, Jun Yang, Hao Wang, Huipeng Ma, Chenhao Li, and Luan Zhang. 2025a. CompKBQA: Component-wise task decomposition for knowledge base question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 293–309, Suzhou, China. Association for Computational Linguistics.

Yuhang Tian, Pan Yang, Dandan Song, Zhijing Wu, and Hao Wang. 2025b. GRV-KBQA: A three-stage framework for knowledge base question answering with decoupled logical structure, semantic grounding and structure-aware validation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2618–2632, Suzhou, China. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Conference on Empirical Methods in Natural Language Processing*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and Wenjie Li. 2023b. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, and 21 others. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*.

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

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. 2024b. Rethinking data selection at scale: Random selection is almost all you need. *arXiv preprint arXiv:2410.09335*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxuand Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mingjia Yin, Chuhan Wu, Yufei Wang, Hao Wang, Wei Guo, Yasheng Wang, Yong Liu, Ruiming Tang, Defu Lian, and Enhong Chen. 2024. Entropy law: The story behind data compression and llm performance. *ArXiv*, abs/2407.06645.

Dylan Zhang, Qirun Dai, and Hao Peng. 2025. The best instruction-tuning data are those that fit. *ArXiv*, abs/2502.04194.

Yiheng Zhao and Jun Yan. 2026. Generating effective cot traces for mitigating causal hallucination. *Preprint*, arXiv:2604.12748.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, and others. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov, and Zhuohan Xie. 2026. Fincards: Card-based analyst reranking for financial document question answering. *Preprint*, arXiv:2601.06992.

Appendix

A Dataset statistics.

Table 5 summarizes the dialogue counts and average dialogue length used in our experiments. We conduct selection on two large multi-turn dialogue pools, BAIZE (54,456 dialogues; 3.95 turns on average) and BANKING (66,948 dialogues; 5.01 turns on average). For evaluation, we report results on three multi-turn benchmarks, MT-EVAL (130 dialogues; 7.30 turns), CONSISTENTCHAT (1,000 dialogues; 7.73 turns), and TOPDIAL (1,321 dialogues; 5.11 turns). In addition, we include BANKING TEST (1,000 dialogues; 4.98 turns) to assess domain-specific generalization on the Banking setting.

	#Dialogues	Avg.turn
Baize	54,456	3.95
Banking	66,948	5.01
MT-Eval	130	7.30
ConsistentChat	1,000	7.73
TopDial	1,321	5.11
Banking Test	1,000	4.98

Table 5: Statistics of the dialogue selection pools and evaluation benchmarks. #Dialogues denotes the number of multi-turn dialogues in each dataset, and Avg. turn denotes the average number of user–assistant turns per dialogue.

B Order-Perturbation Metrics: ESC, HAR, ENR, and $H(d)$

This appendix defines the three metrics used in our *Order Perturbation Analysis* (Section 5.2) and the history-dependency score $H(d)$ used to form the **Top 20% by** $H(d)$ subset. All metrics are computed from the same turn-wise entity annotations produced by our scoring pipeline (i.e., the extracted `q_entities` and `a_entities` per turn).

Notation. A dialogue d contains T user–assistant turns (QA pairs), indexed by $t \in \{1, \ldots, T\}$. For each turn t, let E_t^Q and E_t^A denote the entity sets extracted from the user query and the assistant answer, respectively (corresponding to `q_entities` and `a_entities` in our pipeline). We define the *history entity set* before turn t as

$$C_t \;=\; \bigcup_{j=1}^{t-1} \left(E_j^Q \cup E_j^A \right). \tag{10}$$

Intuitively, C_t summarizes all entities that have been introduced in the dialogue context up to (but excluding) the current turn.

History Anchoring Rate (HAR). HAR measures how well the current answer *anchors* to the previously established entity context. For a dialogue d, we denote the turn-level anchoring score at turn t by $\mathrm{HAR}_d(t)$. We compute it using an F1-style overlap between the answer entities E_t^A and the history entities C_t:

$$\mathrm{HAR}_d(t) \;=\; \begin{cases} \frac{2|E_t^A \cap C_t|}{|E_t^A| + |C_t|}, & \text{if } |E_t^A| + |C_t| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

We then define the dialogue-level HAR as the average over turns:

$$\mathrm{HAR}(d) \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathrm{HAR}_d(t). \tag{12}$$

Interpretation: higher HAR indicates stronger reuse/grounding to previously mentioned entities, hence stronger cross-turn anchoring.

Entity Novelty Rate (ENR). ENR measures how many entities in the current answer are *new* with respect to the prior context. For a dialogue d, we denote the turn-level novelty score at turn t by $\mathrm{ENR}_d(t)$. We compute it as the fraction of answer entities not seen in the history:

$$\mathrm{ENR}_d(t) \;=\; \begin{cases} \frac{|E_t^A \setminus C_t|}{|E_t^A|}, & \text{if } |E_t^A| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The dialogue-level ENR is again the average over turns:

$$\mathrm{ENR}(d) \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathrm{ENR}_d(t). \tag{14}$$

Interpretation: higher ENR indicates the answer introduces more new entities (less redundancy); lower ENR indicates the dialogue is more history-dependent, with heavier reuse of previously established entities.

Entity Sequence Consistency (ESC). ESC is an order-sensitive overall score that combines the above two complementary factors:

$$\mathrm{ESC}(d) \;=\; \frac{1}{2}\Big(\mathrm{HAR}(d) + \mathrm{ENR}(d) \Big). \tag{15}$$

Interpretation: ESC is high when a dialogue simultaneously maintains strong history anchoring (HAR) while still introducing non-trivial new entities (ENR), which matches our local-stage design goal of balancing *anchoring* and *anti-redundancy*.

Turn-weighted aggregation (reported in Table 4). For a dialogue set $\mathcal{D}$, let T_d denote the number of turns in dialogue d. We report turn-weighted scores so that each turn contributes equally:

$$\mathrm{HAR}_{\mathrm{tw}}(\mathcal{D}) \;=\; \frac{\sum_{d\in\mathcal{D}}\sum_{t=1}^{T_d}\mathrm{HAR}_d(t)}{\sum_{d\in\mathcal{D}}T_d}, \quad (16)$$

$$\mathrm{ENR}_{\mathrm{tw}}(\mathcal{D}) \;=\; \frac{\sum_{d\in\mathcal{D}}\sum_{t=1}^{T_d}\mathrm{ENR}_d(t)}{\sum_{d\in\mathcal{D}}T_d}, \quad (17)$$

and

$$\mathrm{ESC}_{\mathrm{tw}}(\mathcal{D}) \;=\; \frac{1}{2}\Big(\mathrm{HAR}_{\mathrm{tw}}(\mathcal{D}) + \mathrm{ENR}_{\mathrm{tw}}(\mathcal{D})\Big). \quad (18)$$

History-dependency score $H(d)$ for the Top-20% subset. To focus on dialogues that require stronger cross-turn dependency, we compute a dialogue-level history-dependency score that increases with stronger anchoring and decreases with higher novelty:

$$H(d) \;=\; \tfrac{1}{2}\Big(\mathrm{HAR}(d) + \big(1-\mathrm{ENR}(d)\big)\Big). \quad (19)$$

We rank the 10K selected dialogues by $H(d)$ and take the top 20% as the *high-history-dependency* subset. The subset size can be slightly different from exactly 20% due to ties in $H(d)$.

Why these metrics are order-sensitive. All four quantities above depend on the *history set* C_t, which is defined by the turn order. Therefore, order-level perturbations (Pair/Block shuffles) alter C_t for many turns and can reduce HAR/ESC even when the multiset of turns is unchanged. In contrast, *Query-only* perturbation additionally breaks query–answer correspondence, yielding a stronger mismatch that typically collapses HAR.

C Error-Type Classifier Prompt

To better understand the qualitative differences between dialogues selected by MDS and competing selectors, we perform an error-type analysis using GPT-4o as a strict multi-turn dialogue judge. Given a dialogue transcript, the judge is instructed to classify assistant-side failures using a fixed taxonomy that covers common multi-turn issues, including topic drift, contradiction/inconsistency, repetition/low novelty, form mismatch, and unsupported/hallucinated content, with an additional `No Error` label indicating no notable issue. The judge must output a single JSON line containing (i) a `primary_error` as the most harmful issue, (ii) an optional multi-label set `error_types` without duplicates, and (iii) short `evidence` snippets that cite the relevant turns (e.g., "Turn 3 Assistant ..."). This constrained format ensures consistent labeling across methods and enables reliable aggregation of error distributions for comparison.

You are a STRICT multi-turn dialogue error classifier.

Task:
Given a multi-turn dialogue between a user and an assistant, identify the assistant's error type(s) across the dialogue, especially with respect to the latest user request and cross-turn consistency.

Taxonomy (labels and definitions):
- No Error: No notable issue. The assistant is helpful, on-topic, and consistent.
- Topic Drift: Goes off-topic, changes the subject, or fails to address the latest user request.
- Contradiction: Contradicts earlier turns, or gives mutually inconsistent statements across turns.
- Repetition: Repeats itself, redundant restatements, or fails to add new useful information across turns.
- Form Mismatch: Answer format does not match the question type (e.g., asked for steps but gave vague talk).
- Unsupported: Introduces unsupported facts not grounded in the dialogue, or makes up details.

Rules:
- Choose labels ONLY from the taxonomy above (no new labels).
- Use multi-label when multiple errors exist.
- Also choose ONE primary_error: the single most harmful issue.
- error_types must contain 0 or more labels from the taxonomy, WITHOUT duplicates.
- evidence must be 1–4 short items. Each item should cite turns like "Turn 3 Assistant ...".
- If there is no notable issue, set primary_error="none", error_types=[], evidence=[].
- Do NOT use markdown.
- Do NOT output any other JSON objects.

Input Dialogue:
{DIALOGUE_TEXT}

At the VERY END, output EXACTLY ONE line in the following format:
FINAL_JSON: {"primary_error":"...","error_types":["..."],"evidence":["..."]}

Figure 4: Prompt used for GPT-based multi-turn dialogue error-type classification. The judge assigns a primary error label and an optional set of additional error labels from a fixed taxonomy, and returns brief evidence by referencing specific turns.

D Supplementary Error-Type Analysis on Difference Sets

This section provides supplementary evidence to the main text by characterizing *what kinds of dialogues* are uniquely favored by **MDS** compared to alternative selection methods. Rather than analyzing the full selected sets (which often share a large

Error type	CC		DialScore		Rethinking		SuperFiltering	
	MDS-only	B-only	MDS-only	B-only	MDS-only	B-only	MDS-only	B-only
No Error	87.5	82.5	87.0	79.5	86.0	80.5	86.5	77.0
Topic Drift	9.5	10.0	5.0	9.5	6.0	9.0	6.0	11.5
Repetition	1.0	2.5	0.5	2.5	3.0	2.5	3.0	2.0
Form Mismatch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
Contradiction	1.0	4.0	4.5	5.0	3.5	2.0	2.5	2.0
Unsupported	1.0	1.0	3.0	3.5	1.5	6.0	2.0	7.0

Table 6: Error-type distribution on DIFFERENCE SETS. For each baseline selector B, we compare 1K dialogues sampled from MDS-ONLY ($\mathcal{D}_{\text{MDS}} \setminus \mathcal{D}_B$) versus B-ONLY ($\mathcal{D}_B \setminus \mathcal{D}_{\text{MDS}}$). Values are percentages; higher `No Error` indicates cleaner dialogues.

overlap), we follow a *difference-set* protocol that isolates the distinctive portion of each selector.

Difference sets. For each baseline selector B, we construct two disjoint sets: (i) **MDS-only**, $\mathcal{D}_{\textbf{MDS}} \setminus \mathcal{D}_B$, containing dialogues selected by **MDS** but not by B; and (ii) **B-only**, $\mathcal{D}_B \setminus \mathcal{D}_{\textbf{MDS}}$, containing dialogues selected by B but not by **MDS**. This comparison controls for the shared subset and highlights the structural differences induced by the selection strategy.

Sampling and labeling. From each difference set, we uniformly sample 1,000 dialogues and assign each dialogue to one error type using the same taxonomy and the same LLM-based classifier described in Appendix C. The reported numbers are the percentages of dialogues in each error category. A higher *No Error* rate indicates cleaner and more coherent dialogues, while higher rates of others indicate specific failure modes.

Results overview. Table 6 summarizes the error-type distributions for four baselines. Across baselines, **MDS-only** dialogues consistently exhibit a higher *No Error* proportion and reduced rates of major multi-turn failure types, suggesting that **MDS** preferentially keeps dialogues with better cross-turn coherence and fewer structural issues. These findings complement the main results by providing a data-level explanation of why **MDS**-selected dialogues lead to stronger downstream behavior.

E Prompt Used for LLM-EVAL and G-EVAL

Figure 5 and Figure 6 show the prompt used for LLM-EVAL and G-EVAL. The above two metrics are both reference-free and rely on GPT-4o as the judge.

```
"You are a STRICT dialogue evaluation assistant."
    "You will evaluate a single model response given a dialogue context."
    "You must rate the response on four dimensions:"
    "- content: correctness, informativeness, completeness, and conciseness of
information."
    "- grammar: fluency, grammar, clarity, naturalness, and conciseness of
wording."
    "- relevance: how well the response addresses the latest user message, stays on
topic, and avoids unnecessary or off-topic details."
    "- appropriateness: safety, politeness, and instruction-following."
    "VERY IMPORTANT:"
    "- Do NOT reward unnecessary verbosity."
    "- If the response repeats itself, includes filler phrases (e.g., long preambles,
generic disclaimers, or obvious restatements), or adds text that does not help
answer the user's question, you MUST LOWER THE SCORES."
    "- A shorter response that fully and clearly answers the question should receive
HIGHER scores than a much longer response that is equally correct but
redundant."
    "For EACH dimension, you MUST assign an INTEGER score from 0 to 10:\n"
    "- 0 = very bad: serious errors or largely unusable."
    "- 2 = poor: many issues; only partially usable."
    "- 4 = borderline: mixed quality with noticeable issues."
    "- 6 = good: generally correct and appropriate but clearly improvable."
    "- 8 = very good: high quality with only minor issues."
    "- 10 = excellent: near human-expert quality; this should be RARE."
    "You MUST use the FULL RANGE of scores when appropriate."
    "First, you may briefly analyze the response for each dimension in free text."
    "Be especially strict about unnecessary verbosity, repetition, and filler."
    "Then, at the VERY END of your answer, output EXACTLY ONE line in the
following format:"
    "FINAL_JSON: {\"content\": c, \"grammar\": g,"
    "             \"relevance\": r, \"appropriateness\": a}"
    "where c, g, r, a are INTEGERS in the range [0, 10]."
    "Do NOT output any other JSON objects besides this FINAL_JSON line."
    "Do NOT use markdown code fences like ```."
```

Figure 5: Prompt used for LLM-EVAL.

F Ablation on Bin Granularity

Setup. We study the effect of bin granularity in our bin-wise selection pipeline on **Baize**, using LLaMA3-8B-Instruct as the backbone. We vary the number of bins $K \in \{100, 500, 1000, 1500, 2000\}$ and keep the total selection budget as well as all other settings unchanged. Table 7 reports results on MT-Eval, ConsistentChat, and TopDial.

Results and analysis. We observe a clear sweet spot at K=1000 (our default), which achieves the best performance on almost all metrics across all three benchmarks. When K is too small (e.g., 100 or 500), bins become overly coarse and mix heterogeneous dialogues, which weakens within-bin normalization and makes the final selection more

	MT-Eval				ConsistentChat				TopDial			
	L-E	**G-E**	**Ent-F1**	**Cos**	**L-E**	**G-E**	**Ent-F1**	**Cos**	**L-E**	**G-E**	**Ent-F1**	**Cos**
100	8.10	7.46	0.561	0.847	8.46	7.16	0.310	0.791	6.94	6.40	0.156	0.453
500	8.08	7.48	0.568	0.843	8.48	7.20	0.305	0.792	6.84	6.38	0.145	0.440
1000(default)	**8.16**	**7.52**	**0.584**	**0.857**	**8.52**	**7.26**	**0.316**	0.797	**7.12**	**6.48**	**0.173**	**0.465**
1500	8.10	7.44	0.561	0.844	8.48	7.16	0.302	0.795	6.94	6.40	0.162	0.457
2000	8.04	7.46	0.576	0.846	8.46	7.16	0.307	**0.799**	6.88	6.40	0.161	0.450

Table 7: Ablation on bin granularity on **Baize** with `LLaMA3-8B-Instruct`. We report performance under four metrics (L-E, G-E, Ent-F1, and Cos), where higher is better for all metrics. Bold numbers indicate the best score in each column.

```
"You are a STRICT dialogue-level evaluation assistant.\n"
    "You evaluate multi-turn conversations between a user and an assistant."
    "You must rate the assistant's latest response on four dimensions:"
    "- coherence: how logically consistent and well-connected the response is with
the previous turns in the dialogue, without rambling."
    "- naturalness: how fluent, human-like, stylistically appropriate, and concise the
response is."
    "- engagement: how interesting, proactive, and conversationally engaging the
response is, without resorting to unnecessary chit-chat or padding."
    "- groundedness: how well the response is grounded in the given context,
without hallucinating  unsupported facts or contradicting the dialogue."
    "VERY IMPORTANT:"
    "- Do NOT reward unnecessary verbosity."
    "- Long answers that repeat themselves, add generic filler, or provide off-topic
explanations should receive LOWER scores for coherence, naturalness, and
engagement."
    "- A shorter response that fits naturally into the dialogue and stays focused on
the user's needs should receive HIGHER scores than a much longer, padded
response."
    "For EACH dimension, you MUST assign an INTEGER score from 0 to 10:"
    "- 0 = very bad: serious issues; largely unusable."
    "- 2 = poor: many issues; only partially usable."
    "- 4 = borderline: mixed quality with noticeable issues."
    "- 6 = good: generally fine but clearly improvable."
    "- 8 = very good: high quality with only minor issues."
    "- 10 = excellent: near human-expert quality; this should be RARE."
    "You MUST use the FULL RANGE of scores when appropriate."
    "First, you may briefly analyze the response for each dimension in free text."
    "Be especially strict about unnecessary verbosity, repetition, and filler."
    "Then, at the VERY END of your answer, output EXACTLY ONE line in the
following format:"
    "FINAL_JSON: {\"coherence\": c, \"naturalness\":"
    "            \"engagement\": e, \"groundedness\": g}"
    "where c, n, e, g are INTEGERS in the range [0, 10]."
    "Do NOT output any other JSON objects besides this FINAL_JSON line."
    "Do NOT use markdown code fences like ```."
```

Figure 6: Prompt used for G-EVAL.

```
"You are an assistant that analyzes the FORM / STYLE of a single QA turn."
    "Your job is NOT to judge factual correctness, but ONLY to see whether the
answer's"
    "style and format match what the question is asking for."
    "Given a user question and an assistant answer, you MUST output a JSON
object with:"
    "  - \"q_entities\": list of key entities in the user question."
    "  - \"a_entities\": list of key entities in the assistant answer."
    "  - \"style_match_score\": integer in {0, 1, 2}:"
    "       * 2 = The answer's style/format clearly matches the request type"
    "             and respects explicit format constraints (e.g. list vs. paragraph,"
    "             translation only, yes/no only, etc.)."
    "       * 1 = Partially matches: the answer roughly follows the requested style,"
    "             but slightly violates some format constraints (e.g. extra explanation,"
    "             wrong number of items, mild verbosity)."
    "       * 0 = Clearly mismatched style: refusal/off-topic, or ignoring explicit"
    "             format instructions (e.g. question asks for a short list but answer\"
    "             is an unrelated long essay, or question asks for translation-only"
    "             but answer includes extra content)."
    "  - \"style_comment\": a short English explanation (1-2 sentences) of why you"
    "       gave this style_match_score."
    "Important:"
    "- Focus ONLY on style / format compatibility with the question."
    "- Do NOT judge factual correctness or safety."
    "- Output ONLY one JSON object, no extra text, no comments."
```

Figure 7: Local-stage structured scoring prompt used in MDS. We query Qwen3-8B with a single-turn QA pair and require a `JSON`-only output that *simultaneously* extracts question/answer entities (`q_entities`, `a_entities`) and predicts a discrete form/style compatibility score (`style_match_score` $\in \{0, 1, 2\}$) with a brief rationale. This one-pass, multi-signal design enables efficient local scoring by avoiding separate calls for entity statistics and form assessment.

sensitive to superficial biases. In contrast, when K is too large (e.g., 1500 or 2000), bins become fragmented and sparse, leading to unstable within-bin statistics and noisier quota allocation, which hurts overall quality. A minor exception is the Cos score on CONSISTENTCHAT, where K=2000 is slightly higher; however, this does not translate into consistent gains on L-E/G-E/Ent-F1 or on other benchmarks. Overall, these results support using a moderate bin granularity to balance within-bin comparability and statistical stability.

G Local-stage Prompt for Joint Entity and Form/Style Scoring

To compute local structural signals efficiently, we use a single structured prompt to obtain both (i) entity statistics and (ii) form/style compatibility judgments for each QA turn. Concretely, given a user question and the corresponding assistant answer, we query a lightweight judge model (Qwen3-8B) and require it to output *only* one JSON object containing: (1) entities in the question (`q_entities`) and in the answer (`a_entities`), and (2) a discrete style-match score `style_match_score` $\in \{0, 1, 2\}$ indicating whether the answer's format matches what the question requests (e.g., list vs. paragraph, translation-only, yes/no-only), along with a short explanation `style_comment`.

This joint-output design is critical for efficiency: entity overlap statistics (used by our entity-based local signal) and form/style judgments (used by our form-based local signal) are produced in a *single* forward pass per turn, rather than two separate

model calls. As a result, local scoring can scale to large candidate pools with substantially reduced inference overhead while keeping the signals consistent by construction (both derived from the same model output and the same turn context).

H Heuristic Rule-based Dialogue Filtering

To construct a strong rule-based baseline for dialogue selection, we implement a lightweight heuristic filter that removes low-quality conversations and then ranks the remaining ones by a composite quality score. The filter operates on each dialogue independently and only uses surface statistics computed from the *assistant* turns.

Preprocessing. For a dialogue d, we extract all assistant messages $\{a_1, \ldots, a_T\}$ (with role normalized to `assistant`). Each message is tokenized into a word list using a Unicode-aware regex, and additionally split into sentences using punctuation-based segmentation. Dialogues with fewer than `MIN_ASST_TURNS` assistant turns are discarded.

Quality constraints. We enforce three hard constraints to eliminate obvious noise:

- **Short-response ratio.** We count an assistant turn as *short* if its token length is below `SHORT_TOK_TH` or its character length is below `SHORT_CHAR_TH`. Let r_{short} be the fraction of short assistant turns. We discard d if $r_{\text{short}} > \texttt{MAX\_SHORT\_RATIO}$.

- **Repetition score.** We measure repetition from both token-level and sentence-level perspectives. First, we compute the n-gram repetition ratio using $n = \texttt{REP\_N}$ over the concatenated assistant token stream:

$$r_{ng} = 1 - \frac{|\text{unique } n\text{-grams}|}{|\text{all } n\text{-grams}|}.$$

  Second, we compute the duplicated-sentence ratio r_{sent} as the fraction of assistant sentences that are exact repeats within the dialogue. We define the overall repetition score as

$$r_{\text{rep}} = 0.5\, r_{ng} + 0.5\, r_{\text{sent}}.$$

  We discard d if $r_{\text{rep}} > \texttt{MAX\_REP\_SCORE}$.

- **Lexical diversity.** Let $\mathcal{V}$ be the set of unique assistant tokens and $\mathcal{T}$ be the multiset of all assistant tokens. We compute

$$r_{\text{lex}} = \frac{|\mathcal{V}|}{|\mathcal{T}|}.$$

  We discard d if $r_{\text{lex}} < \texttt{MIN\_LEX\_DIV}$.

In addition, dialogues with fewer than `MIN_ASST_TOTAL_TOKS` assistant tokens in total are removed to avoid overly short conversations.

Scoring and selection. For each dialogue that passes all constraints, we compute a normalized heuristic quality score:

$$s(d) = 0.45\,(1 - r_{\text{short}}) + 0.35\,(1 - r_{\text{rep}}) + 0.20\, r_{\text{lex}},$$

and clip $s(d)$ to $[0, 1]$. Finally, we rank all retained dialogues by $s(d)$ (descending) and select the top-$10K$ dialogues. This procedure yields a simple, fully deterministic baseline that prioritizes non-trivial, less repetitive, and lexically diverse assistant behavior while requiring no learned model.