Active Instruction Tuning for Large Language Models with Reference-Free Instruction Selection

Anonymous EMNLP submission

Abstract

Recent works (Zhou et al., 2023; Xia et al., 2024; Liu et al., 2023) on efficient instruction tuning have shown that large language models (LLMs) can achieve comparable performance through the calibrated selection of a small subset of high-quality (INSTRUCTION, RESPONSE) pairs from labeled instruction pools. Despite reduced computational costs, these approaches often overlook the labor-intensive nature of instruction acquisition for labeling. We introduce a novel paradigm, *Active Instruction Tuning with Reference-Free Instruction Selection*, which supports instruction selection from both labeled and unlabeled instruction pools. Our experimental results demonstrate that this method not only achieves comparable or superior performance while reducing labeling costs but also matches the performance of prior studies in labeled instruction settings. Furthermore, we pioneer the investigation into the relationship between text evaluation correlated with human subjective evaluations and instruction tuning, confirming the effectiveness of ranking aggregation in enhancing the tuning.

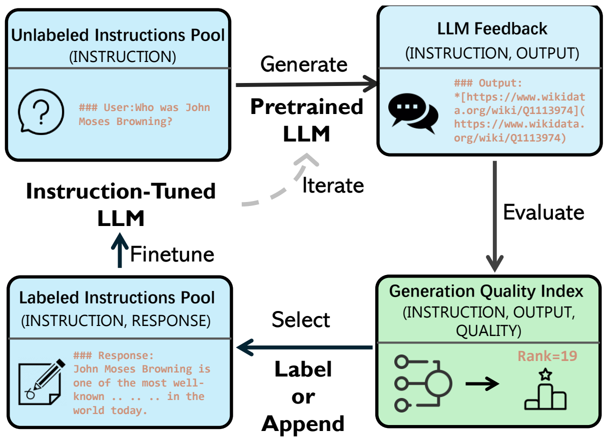

Figure 1: The Framework of *Active Instruction Tuning with Reference-Free Instruction Selection* setting. There are three key components: generation, selection, and finetuning. By evaluating the quality of the (INSTRUCTION, OUTPUT)s in a response-free setting, the most challenging instructions for the model are selected where GENERATION QUALITY INDEX will rank, then label or append (RESPONSE)s for these instructions; finally the LLM is finetuned. Here, OUTPUT refers to the text generated by the model when an instruction serves as a prompt, whereas RESPONSE denotes the text that is labeled for an instruction.

1 Introduction

Instruction tuning is a crucial mechanism enabling large language models (LLMs) to upgrade from merely language modeling to effectively assisting users. The complete process of instruction tuning mainly includes two stages: a) labeling raw instructions, *e.g.*, analyzing and selecting a series of the unprocessed user instructions (community forum) to produce high-quality responses, and b) finetuning the model by these labeled instructions. Significant efforts (Wei et al., 2022; Longpre et al., 2023; Sanh et al., 2022; Wang et al., 2022) have been devoted to creating a substantial, diverse, and high-quality finetuning dataset. However, given the massive amount of raw instruction data, there is an urgent need (Köpf et al., 2023; Ding et al., 2023) for a more efficient procedure to streamline the substantial labor-intensive labeling-and-finetuning pipeline of instruction tuning, especially since Zhou et al. (2023) pioneered a small number of meticulously labeled examples that yield comparable performance.

Recent works (Liu et al., 2024; Li et al., 2023; Du et al., 2023; Cao et al., 2023), consider efficient instruction tuning as coreset selection from the labeled (INSTRUCTION, RESPONSE) training pools. For instance, Xia et al. (2024) uses the gradient-based data selection that relies on labeled responses. While this reduces computational costs at the finetuning stage, it overlooks the higher costs associated with response labeling (Köpf et al., 2023; Zheng et al., 2024). Moreover, this kind of approach ignores the vast and evolving resources of unlabeled instructions. Therefore, the efficient acquisition of useful unlabeled instructions should not be excluded from the entire instruction tuning pipeline. These motivate us to reframe the problem: *efficiently acquire the most useful instructions, then label responses or append original reference responses for finetuning.*

We introduce a more feasible paradigm, *Active Instruction Tuning with Reference-Free Instruction Selection* which actively selects a small number of instructions to label-and-finetune rather than passively labeling the entire pool before selecting (Kung et al., 2023). Importantly, our paradigm is compatible with traditional coreset selection for labeled pools and does not necessitate incorporating responses into the quality evaluation process. Inspired by Active Learning (Settles, 2009), we hypothesize that instructions that are **challenging** for LLMs are more effective training samples, as they help identify weaknesses of the LLM through examining its outputs. However, quality evaluation methods based on accuracy or training efficacy estimation fail when evaluating the generated outputs without reference responses. The literature (Zhou et al., 2023; Reimann et al., 2023) shows that humans can provide reliable evaluations on subjective aspects even without a reference, though human evaluation is costly.

To identify the most challenging instructions, we present a novel data selection method, Generation Quality Index (GQI), based on automated text generation evaluation, which correlates with human subjective evaluation without requiring any reference. Firstly, for the vague concept of "quality", we define text quality as **the weighted combination of several textually significant attributes to be evaluated**. To this end, our framework is divided into two modules: 1) Atomic-level Subjective Text Evaluators: we introduce a significant number of automated text evaluators that target various atomic subjective cognitive aspects (*e.g.*, *Coherence*, *Naturalness*, *Likability*) to discern the quality of generated outputs, instead of being restricted to a limited set of coarse-grained aspects, such as *uncertainty*, *diversity*, even *writing style*. 2) Neural Ranking Aggregator: *"no output is perfect in all aspects"*; inherent partial orders often conflict between different texts in various aspects, such as an "elegant hallucination" versus a "flat scientific paragraph". Therefore, when introducing many atomic aspects as signals, we aim to achieve a consensus ranking to counter this issue and enhance the effectiveness of tuning. This partial inconsistency problem has been overlooked in many works (Wettig et al., 2024), and to our knowledge, we are the first to to address it.

Inspired by the Crowd-BT model (Chen et al., 2013), we derive a reliable consensus rank by formalizing our rank using probabilistic methods, assigning learnable confidence parameters to each atomic evaluator, which also provides interpretability for the understanding of abstract and vague "quality" from the subjective aspects. More importantly, this mechanism still supports previous work that used a single ranking as guidance for quality evaluation and introduces any ranking as a strong supervision signal to the aggregator, e.g., human quality experts, simply by setting its corresponding confidence parameter to about 0.95.

Through extensive experiments, we verified several main conclusions: 1) under the traditional paradigm of coreset selection for a labeled instruction pool, our method proves that reference-free selection achieves comparable results with the same data size; 2) our method can drastically reduce the cost of labeling, and the sampled pool of unlabeled instructions by our method outperforms LIMA and ALPAGASUS; 3) we confirm the ranking aggregation has effectiveness in selection and tuning.

Our contributions are summarized below:

1. **A more realistic efficient instruction tuning paradigm**. Active instruction tuning with reference-free instruction selection efficiently selects high-quality instructions, which enables finetuning LLM and expands instruction resources efficiently. We overcome the limitations of only selecting from labeled pools.

2. **A more general and novel methodology**. We introduce two classic techniques, *Text Generation Evaluation* and *Rank Aggregation*, to address the reference-free generation evaluation and the inevitable inconsistencies among multiple ranks, thereby obtaining a consensus rank for instruction acquisition.

3. **A series of inspiring results**. a) We verify that selecting high-quality data under without reference responses condition is still feasible and achieves comparable results in both with and without labeled response scenarios; b) We demonstrate that the introduction of ranking aggregation is significantly effective; c) We explore the relationship between subjective as-

pects of evaluation and instruction tuning for the first time.

2 Related Work

2.1 Instruction Tuning

For improving the general zero-shot abilities of pre-trained large language models, Wei et al. (2022) firstly proposed "instruction tuning", teaching an LM to perform tasks described via instructions. Consequently, Chung et al. (2022); Longpre et al. (2023); Wang et al. (2023b) have progressively enlarged the scale of finetuning training resources, ultimately encompassing millions of instances.

LIMA (Zhou et al., 2023) firstly pioneered a novel discussion that finetuning can yield remarkable outcomes even with a limited training set size, as long as the instruction-responses pairs are high-quality labeled by experts. There are two independent research directions: 1) continuing to use human-heuristic standards (Köpf et al., 2023; Conover et al., 2023; Singh et al., 2024) to label responses or computational-costly (Ding et al., 2023; Zheng et al., 2024) expand the high-quality instruction pool, and 2) performing high-quality coreset selection (Cao et al., 2023; Li et al., 2023; Du et al., 2023; Das and Khetan, 2023; Wu et al., 2023) within the already labeled instruction pool. We argue that the latter direction, which separates the labeling and finetuning processes, does not truly reduce the overall cost.

2.2 Active Learning and Quality-based Acquisition

Given an unlabeled data pool and a constrained budget, Active Learning (Settles, 2009; Zhang et al., 2022c) emerges as a potentially efficient way to improve the finetuning performanc and label the most valuable instructions.

Besides, among several mining criteria, those hinging on difficulty have proven to be most crucial for training (Felzenszwalb et al., 2010; Schröder et al., 2022), more specifically the hard examples are viewed as uncertain yet informative because the model's predictions for these are least satisfactory. Conventional approaches that rely on a single metric (*e.g.*, loss) limit the selection only in the labeled pool. They do not guarantee that the discrepancies between model outputs and reference responses accurately indicate the weakness of the model generation ability (this is an observation in our experiments). We bridge a gap between **challenging** to LLM and **generation quality degradation** to design a quality-based acquisition.

Our research pioneers the integration of active learning principles in instruction tuning without labelled responses. Particularly, Kung et al. (2023) still makes the model more generalizable in the labeled dataset, while Parkar et al. (2024) uses only the cluster-based diversity metric to find valuable instructions indirectly. In addition, reference-free scenario and research (Meng et al., 2024; Muldrew et al., 2024) also emerge in Direct Preference Optimization (Rafailov et al., 2023) for LLM.

2.3 Text Generation Evaluation

Text generation evaluation (Celikyilmaz et al., 2021) is to assess the quality of the generated text x given on a specific aspect a (*e.g.*, coherence, interestingness) and an optional reference op, then predict a quality score y,

$$y = f(x|a, op), \tag{1}$$

where f can be performed using expert annotations following a protocol or automated evaluation metrics. Reference-based similarity methods are widely used in evaluation tasks, such as BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020). Many tasks, such as open dialogue, inherently lack and should not have predefined references. Based on the experience that humans can make subjective evaluations without references, reference-free evaluations (*e.g.*, FED (Mehri and Eskenazi, 2020), UniEval (Zhong et al., 2022), GPTScore (Fu et al., 2023)) uniquely enable the assessment of fine-grained and subjective aspects that strongly correlate with human evaluations. These evaluations are even beginning to surpass traditional reference-free approaches. Our work considers 20 aspects and corresponding metrics, as shown in Table 5. Furthermore, notable endeavors have focused on evaluating specific aspects with exceptional precision, *i.e.*, AlignScore (Zha et al., 2023) on Factuality.

2.4 Rank Aggregation

Due to the inherent different specialization (*coherence* or *consistency*) and varying levels (*weak* or *strong*) of expertise of each expert, rank aggregation (**RA**) (Mallows, 1957; Jin et al., 2020), referring to the task of optimizing a "consensus" rank of a set of objects given partial ranks, or full ranks obtained from a set of experts, is widely applied in various domains such as science, economy, and society. Formally, defining an aggregate function RA_θ, given a set of crowdsourced atomic ranks $\{r_1, r_2, ..., r_k\}_{k=1}^K$, we can generate a better consensus rank r^*,

$$r^* = RA_\theta(r_1, r_2, ..., r_k) \tag{2}$$

where θ can be interpreted as confidence weights to different atomic ranks. Numerous proposed RA methods can be divided into two fundamental strategies: unsupervised (Klementiev et al., 2008) and supervised (Liu et al., 2007), contingent upon the assumptions of whether they are guided by explicit Oracle rank or not; The Bradley-Terry (BT) model (Bradley and Terry, 1952) and the Thurstone model (Thurstone, 1927), which originated as early works, effectively capture the probabilistic relationship between objects by leveraging the achieved scores, making them particularly suitable for pairwise comparisons. Our work draws inspiration from the design of the Crowd-BT model (Chen et al., 2013), which learns the confidence weights for each atomic ranker and optimizes both scores and weights.

3 Methodology

In this section, we provide a detailed description of our proposed method. Section 3.1 outlines the formulation of the *Active Instruction Tuning with Reference-Free Instruction Selection*. Building upon this, we detail the principle of instruction acquisition using GQI in Section 3.2, where we decompose "difficulty" into multiple atomic reference-free quality metrics and then aggregating them, instead of relying on ambiguous metrics like uncertainty and diversity. Finally, we devote to the critical part–NEURAL RANKING AGGREGATION in GQI–in Section 3.3, to get a better ranking derived from multiple ranks.

3.1 Paradigm Formulation

Our insight involves using instructions as prompts to identify weaknesses in a model's generation capabilities. These selected instructions aid in finetuning LLMs by labeling responses or appending original reference responses.

Given an LLM $\mathcal{F}(\cdot; w_0)$ with pretrained weights w_0 and a large INSTRUCTION pool $\mathcal{P} = \{x_i\}_{i=1}^N$ without reference responses, where N is the size of the pool. Our task is to select a subset $\mathcal{P}_S \subset \mathcal{P}$ and then label or append (RESPONSE)s for them to get a training (INSTRUCTION, RESPONSE) pool $\mathcal{P}_S^{label} = \{(x_i, y_i)\}_{i=1}^B$ with a budget B, to achieve satisfactory performance of the finetuned model $\mathcal{F}(\cdot; w^\star)$ with finetuned weights $w^\star$. The detailed algorithm and framework are shown in Alg. 1 and Fig. 1, where we select instructions that are challenging for the model in an iterative manner. The core lies in designing an acquisition function $q(\cdot; \pi)$, which evaluates the (INSTRUCTION, OUTPUT) to get the quality rank $r^\star$.

3.2 Generation Quality Index

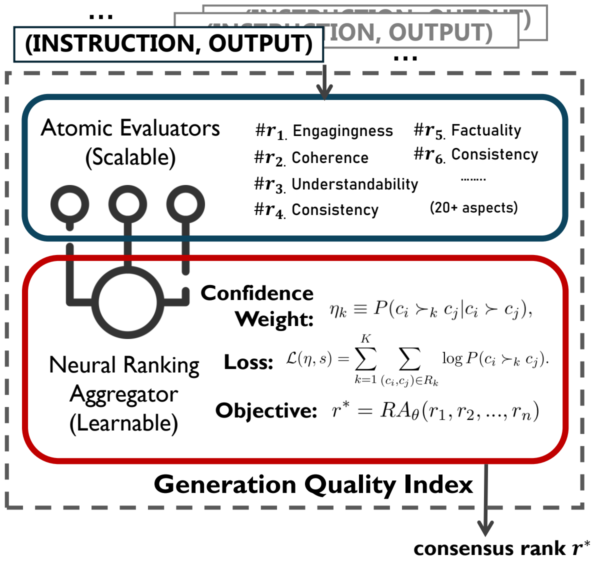

Figure 2: Our proposed GENERATION QUALITY INDEX. Given N (INSTRUCTION, OUTPUT) items, K scalable atomic reference-free evaluators can score them to get K different ranks for each pair-wise relationship. Learning from the $C(N, 2) \times K$ pair-wise rank data, NEURAL RANKING AGGREGATOR generates the consensus ranking $r^\star$ for items.

Previous studies (Liu et al., 2023; Du et al., 2023; Muldrew et al., 2024; Li et al., 2023; Parkar et al., 2024) have established that acquisition strategies typically rely on limited dimensions, with *uncertainty* and *diversity* being the most prominent. However, these metrics are too ambiguous, posing challenges for further acquisition strategy improvements.

The approach we propose relies on the assumptions that *"challenging samples that make generation ability degradation has a higher value for the finetuning"* (inspired by Settles (2009); Wang et al. (2004)). Moving beyond the traditional *"difficulty"* locked to a single metric, *e.g.*, inference loss (Cao et al., 2023), our hypothesis makes a connection between difficulty and multivariate quality evaluation; there are multiple atomic subjective aspects of quality: *coherence*, *interestingness*, *flexibility*, *naturalness*, etc. Therefore, we decompose the *"difficulty"* into several atomic metrics, $\{r\}_{i=1}^{K}$, that correlate to human evaluation and then aggregate these independent evaluation ranks to the final acquisition decision. Here, we work on the ranks instead of scores because scores possess more noise than the rank signal (Jin et al., 2019), and we do not assume the existence of an explicit oracle difficulty serving as the supervision signal for regression.

3.3 Neural Ranking Aggregation

To solve partial order inconsistency led by different metrics, we develop a NEURAL RANK AGGREGATION module, assigning each atomic evaluation a learnable confidence weight to measure its importance in predicting one consensus rank.

Formulation. After the model generates outputs, there are N (INSTRUCTION,OUTPUT) items $\{c_i = (x, y)_i\}_{i=1}^{N}$ and a set of K evaluation scorers. Due to different evaluation mechanisms, the scores estimated by different scorers noticeably diverge, for instance, the range of different scores. Instead, we derive the K rank lists, $R_k = \{c_{o1} \succ_k \cdots \succ_k c_{oN}\}_{k=1}^{K}$, from scores to express the individual preferences of the scorers, where $\succ_k$ represents the preference order of the k-th scorer, and o_j denotes the index of j-th instruction-output in this rank list. If a correct ranking exists, when two scorers give conflicting rankings, it is not possible to treat both scorers as equally reliable. Thus, this motivates us to measure which scorer is more trustworthy. Inspired by Crowd-BT (Chen et al., 2013), we define a learnable parameter η_k for the k-th scorer as the probability that the k-th scorer agrees with the proper pairwise preference, then we can formalize any pair (c_i, c_j) with the true preference $c_i \succ c_j$,

$$\eta_k \equiv P(c_i \succ_k c_j | c_i \succ c_j), \tag{3}$$

if the k-th scorer is more plausible, η_k is closer to 1. Specifically, we apply the sigmoid function to the learnable parameters set $W = \{c_1, \ldots, c_K\}$ to derive the η_k,

$$\eta_k = \text{sigmoid}(c_k) \in [0, 1], \tag{4}$$

where each c_k is a learnable parameter, optimized with the model parameters θ. Based on the confidence weights, we can further formalize the predicted preference order of (c_i, c_j) predicted by k-th scorer,

$$\begin{aligned} P(c_i \succ_k c_j) &= P(c_i \succ_k c_j | c_i \succ c_j) \cdot P(c_i \succ c_j) \\ &\quad + P(c_i \succ_k c_j | c_i \prec c_j) \cdot P(c_i \prec c_j) \\ &= \eta_k \cdot P(c_i \succ c_j) + \\ &\quad (1 - \eta_k) \cdot P(c_i \prec c_j), \end{aligned} \tag{5}$$

where $P(c_i \succ c_j)$ is the probability of $(c_i \succ c_j)$ predicted by the aggregation module. Here, we use the Longformer (Beltagy et al., 2020) to encode the evaluated (c_i, c_j) and then predict two scores (s_{ci}, s_{cj}) by an MLP layer separately, finally $P(c_i \succ c_j)$ can be defined by $sigmoid(s_{ci} - s_{cj})$ simply. The log-likelihood for s and η can be formulated as

$$\mathcal{L}(\eta, s) = \sum_{k=1}^{K} \sum_{(c_i, c_j) \in R_k} \log P(c_i \succ_k c_j). \tag{6}$$

Non-Guidance v.s. Guidance Strategy. Based on the hypothesis that "whether explicit golden rank exists", we offer two more specific assumptions and provide corresponding strategies.

1) Implicit Oracle Rank: when the oracle rank is **not explicitly** accessible, our proposed algorithm fundamentally operates without any ground-truth rank guidance. Its primary aim is to combine varied ranks provided by multiple experts, deriving an enhanced implicit gold rank.

2) Explicit or Assumed Oracle Rank: we hypothesize the existence of an explicit oracle rank or consider any metric an oracle standard rank. For instance, we may assume an oracle-rank-in-loop. Using the oracle rank as a beacon for guiding the training of the aggregation module paves the way to elucidate confidence parameters that calibrate the various atomic metrics in alignment with the oracle rank. To elaborate, when deploying a guidance strategy, we introduce a new confidence parameter, η_{k+1}, corresponding to the signal of oracle rank. This parameter can be designated with a fixed value, approximately equal to 0.95, and deemed non-learnable. This fixed weight can be interpreted as a regularization term.

Via this paradigm, we foresee a heightened correlation between the output of the aggregation module and the oracle rank. Simultaneously, the learnable confidence parameters offer interpretability relative to the oracle rank.

Model	Alignment Type	Alignment Data Size	MT-Bench(1-10)	AlpacaEval 2(%)	
			GPT-4-as-Judge	LC	WR
		Proprietary Models			
Claude 3 Opus	SFT+DPO	—	9.18	40.5	29.1
GPT-4	SFT+DPO	—	—	38.1	23.6
GPT-3.5-turbo	SFT+DPO	—	7.94	19.3	9.2
		Mistral-7B as Base Model			
Mistral-7B-Instruct-v0.2	—	—	7.60	17.1	14.7
UltraChat (Ding et al., 2023)	SFT	200K	6.30	8.4	6.2
zephyr-sft (Tunstall et al., 2023)	SFT	200K	7.32	—	—
zephyr-beta (Tunstall et al., 2023)	SFT+DPO	200K+60K	7.34	13.2	11.0
DEITA (Liu et al., 2024)	SFT	6K	7.22	—	—
DEITA (Liu et al., 2024)	SFT	10K	7.28	—	—
Random-Selection	SFT	10K	6.31	8.2	5.8
Single(Engaging)	SFT	10k	6.82	7.9	6.0
GQI (our proposed)	SFT	6K	7.15	8.3	6.1
GQI (our proposed)	SFT	10K	**7.25**	**8.6**	**6.5**
		Llama 2-7B as Base Model			
LLaMA2-7B-Chat	SFT+RLHF	>100K+>1M	6.27	5.4	5.0
Vicuna-7B-16k	SFT	16K	6.22	6.3	4.2
Tulu 2 (Ivison et al., 2023)	SFT	326K	6.30	—	—
TuLu 2 +DPO 7B (Ivison et al., 2023)	SFT+DPO	326K+60K	6.29	9.2	8.2
Random-Selection	SFT	10K	6.20	6.1	5.0
Single(Natureness)	SFT	10K	6.17	6.3	5.2
GQI (our proposed)	SFT	10K	**6.33**	**6.7**	**5.5**
		Llama 3-8B as Base Model			
Llama 3-8B-Instruct	SFT-DPO	—	8.1	26.0	25.3
UltraChat (Ding et al., 2023)	SFT	200K	6.6	6.2	4.6
GQI (our proposed)	SFT	10K	**7.21**	**7.1**	**6.0**

Table 1: Efficient instruction tuning performance on instruction-following benchmarks in labeled instruction pools. **GPT-4-as-Judger** give the scores in range $(0, 10)$; **LC** and **WR** denote length-controlled and raw win rate.

4 Experiments

There are two efficient instruction tuning scenarios: a) in coreset selection for (INSTRUCTION, RESPONSE) dataset, reference-free select the valuable pairs; b) in coreset selection in (INSTRUCTION) dataset, select the valuable instructions then to label them. Therefore, we have two main research objectives: (a) *without* RESPONSE *as the selection signal, our* GQI *still achieve comparable instruction tuning results compared to previous methods when coreset selection in the labeled pool.* (b) *our proposed methodology achieve better finetuned performance while reducing the labor-intensive nature of the labeling process.* For each objective's experiment, we emphasize the different instruction data sources setting to ensure a fair comparison (it is undeniable that the quality basis of the data pool has a direct impact on the final finetuning performance): for (a), we select in the labeled instruction pools (OPENASSISTANT, DOLLY, FLAN-V2, and OPENORCA) followed previous work (Wang et al., 2023a; Xia et al., 2024); for (b), we use STACKEXCHANGE and WIKIHOW as raw instructions source to compare with LIMA that is also heuristically labeled through these two pools. Furthermore, we introduce extensive ablation studies, especially focusing on the *relationship between subjective aspects and finetuning efficiency* and *aggregation effectiveness*, which is, to our knowledge, the first exploration in the domain of instruction tuning.

There are two types of benchmarks: 1) instruction-following benchmarks, including MT-BENCH and ALPACAEVAL-2; 2) downstream benchmarks for cognitive performance, including MMLU, TRUTHFULQA, HELLASWAG, and ARC-C. The detailed settings are listed in Appendix A: instruction pools, labeling setting, atomic text generation evaluation, finetuning setting, generation setting, benchmark setting, ablation setting, and baseline details.

4.1 Main Results

We present the results of our proposed methodology on two types of benchmarks and compare them to several baseline methods. We summarize the conclusions below.

It remains effective even without a reference response for selection. While we acknowledge that the quality of RESPONSE affects the final finetuning performance, this does not mean that coreset selection must depend on RESPONSE, since the quality of the labeled pool already is not low. As shown in instruction-following performance Table 1 and downstream performance Table 6, our method 1) outperforms data selection via random sampling; 2) achieves comparable performance to the previous methods, *e.g.*, DEITA depending on the selection using (INSTRUCTION, RESPONSE) with similar size of data; 3) surpasses models finetuned for CHAT while using fewer data.

Our method boosts the finetuning-labeling efficiency instead of human-heuristically selection or filtering rule. Ensuring that the instruction data source is roughly equivalent to LIMA, we automatically select approximately $1K$ instructions using our method. Our method offers lower costs than LIMA. Moreover, after labeling responses, our approach demonstrates superior performance to LIMA and ALPAGASUS (which use "LLM-as-filter"), proving the efficiency as shown in Table 2.

Generality across models. We applied our method on three different base models: MISTRAL-7B-V0.1 (mistralai, 2023), LLAMA 2-7B (AI@Meta, 2023), and LLAMA 3-8B (AI@Meta, 2024). We observe the effectiveness across all models.

Single-aspect text evaluation can enhance the effectiveness. As demonstrated in the Table 1, using single-aspect, such as ENGAGINGNESS and NATURALNESS, can improve performance compared to random selection. Through our observations of model outputs, we have identified significant discrepancies in the quality of outputs. Some outputs with poor quality consistently perform poorly across various subjective aspects, indicating that any subjective aspect could be used to select challenging INSTRUCTION for the model (cases in Table 13). In the subsequent section, we will confirm this finding further.

Aggregation yields better ranking. Compared to ranking based on a single aspect, GQI achieves better performance. Continuing from the previous observation, this means that although some items are inferior in most single-aspect rankings, aggregating them can resolve conflicts among aspects and achieve a better ranking.

4.2 Ablation Study

The LLAMA 3-70B-INSTRUCT model has demonstrated capabilities on par with Gemini (google deepmind, 2024) and GPT-4 (OpenAI et al., 2024) across a range of instruction-following and downstream benchmarks. Its open-source characteristics provide an optimal trade-off between experimental accuracy and the costs incurred from extensive ablation studies. Consequently, we persist in utilizing MT-BENCH and shift the LLAMA 3-70B-INSTRUCT as the judger.

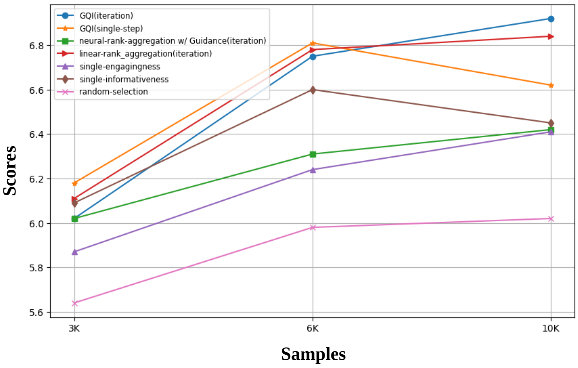

Figure 3: MT-bench score of different selection methods as sample size increase when Llama 3-70B-Instruct as judger. Mistral 7B is the base model for tuning.

Choices of Aggregation Strategy There are multiple aggregation strategies, and we test linear rank aggregation (Linear RA) and *Neural RA w/ Response Guidance* (Their descriptions are listed in Appendix C.3). As shown in Figure 3, 1) Linear RA yields satisfactory results despite its simplicity. We speculate that each aspect has a unique contribution, and even though a few ranks derived from specific indicators may conflict with the majority, they play some crucial roles. 2) *Neural RA w/ Response Guidance* performs noticeably worse. We assume the oracle rank is "the quality of reference responses must be better than that from model generation". An in-depth case study is in Table 14. It reveals that the model generates better text than the reference responses when finetuned on these instructions and corresponding reference responses in the previous iteration. This shows that the so-called "oracle" ranking is not as good as the reference-free evaluation, so it is less effective than GQI's default non-guidance. This also enlightens us that assuming any ranking as the oracle might not necessarily yield the optimal effect.

Iteration performs significantly when selected samples increase. When the number of selected instructions is relatively low, the effect of the iteration is not obvious, and even worse than the result of single-step. One reason is that the total training

Model	Data Size	Instruction Acquisiton	Instruction-following Benchmarks		Cognitive Benchmarks			
			MT-Bench(1-10)	AlpacaEval 2(%)	TruthfulQA	ARC	Hellaswag	MMLU
LIMA	1,030	Human heuristic selection	2.74	3.95	41.90	55.63	80.09	43.71
ALPAGASUS	1,030(sampling from 9K)	LLM-as-filter	2.65	3.57	42.03	55.49	80.25	44.12
GQI	1,030	rank-aggregation selection	2.81	4.13	41.95	55.70	80.18	43.91

Table 2: Efficient instruction tuning methods performance on instruction-following and cognitive benchmarks in unlabeled pools for instruction acquisition. ALPACAEVAL 2 uses raw win rate as metrics. The base pretrained model is LLAMA 2-7B.

step when using iteration is less than that of single-step selection; while the effect of iteration becomes more obvious when more samples are acquired. It implies that iteration enhances the diversity of the entire sampling process, as each selection depends on the varied generative capabilities of the iteratively finetuned model. In contrast, single-step sampling relies solely on the initial model, resulting in relatively poorer diversity.

Single Aspect and Core Aspects. We verify the rationality of the aggregation mechanism and identify a group of core aspects by analyzing the correlation among various single-aspect rankings. The figure 4 reveals a high degree of correlation indicating that the instructions selected exist overlapping using different single aspects, so they will all be effective; however, the conflicts in ranking prove that aggregation is necessary. Moreover, we can decrease these aspects by distinguishing correlations to identify core aspects in F. Subsequently, we evaluate the effectiveness of instruction selection based on single-aspect and aggregated rankings of these core aspects on MT-BENCH. Table 7 indicates that instructions categorized under “informativeness” and “engagingness” generally perform better. Furthermore, while the performance using core aspects surpassed that of single-aspect, it does not quite match the outcomes achieved by aggregating all aspects, highlighting the nuanced contribution of each aspect to the overall performance. More analysis is provided in the Appendix.

Effectiveness of Confidence Weight. We compute Spearman’s correlation coefficient between each atomic rank and the corresponding aggregated rank in two guidance settings. As shown in Appendix Sec. G, for *Non-Guidance*, when no “oracle” ranking as guidance, the confidence weights on atomic rankers are generally higher, and the correlation is also stronger than with guidance. Such findings interpretably reflect the shortcomings in the quality of the reference responses compared to model-generated outputs. Moreover, within the metrics, those recognized as more powerful, such as GPTScore, are associated with higher confidence weights, which further validates the efficacy of confidence weights. Additionally, rankings with high confidence weights also show a higher correlation with aggregated rankings, affirming the utility of confidence weights.

5 Conclusion and Discussion

We innovatively connect efficient finetuning with labeling to select potentially valuable instructions without needing reference answers. Although some previous efficient instruction tuning methods have been built on the assumption of “learning from difficulty”, they all rely on a single metric and labeled responses. However, our approach proves we can achieve excellent results with text generation evaluation aligned with human subjective evaluation. Most importantly, there are two advantages of our work. 1) Enhancements to the acquisition become extensible and decomposable. Grounded in text generation evaluation and rank aggregation, we can augment the metrics with increasingly refined measurements and a more reliable aggregation algorithm, even using crowd-sourced expert evaluation; 2) Rank aggregation is a promising mechanism that resolves conflicts across multiple scorers to improve the finetuning performance. While many studies have employed multiple scorers primarily to gather extensive ranking data for training an overall scorer, the true strength of aggregation lies in its ability to detect and analyze inconsistencies in partial orders, particularly during the later stages of iteration. This observation deepens our understanding of quality scorer variability.

Limitations

Atomic Quality Assessment. Our study uses tools for the text generation evaluation skilled at evaluating linguistic characteristics, overlooking some functional evaluations of downstream tasks, *e.g.*, functional correctness of generated code in code generation tasks or CoT Tasks.

High computational costs in generating outputs. Generating outputs for each instruction is costly in terms of both time and computation, yet this seems unavoidable for all tasks that rely on model feedback. Our trick involves directly pruning instructions that prompt high-quality output at each selection step. These pruned instructions are then removed from subsequent selection processes.

Ethics Statement

We use open-source instruction data and LLMs in our finetuning. We do not involve the inclusion of any dangerous or private sensitive information.

References

AI@Meta. 2023. Llama 2 model card.

AI@Meta. 2024. Llama 3 model card.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. In *arXiv*.

Ralph A. Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika*, pages 324–345.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2023. Instruction mining: Instruction data selection for tuning large language models. In *International Conference on Learning Representations*. Under review.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. In *arXiv*.

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 *arXiv*.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *International Conference on Web Search and Data Mining*, pages 193–202.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. In *arXiv*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *arXiv*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Devleena Das and Vivek Khetan. 2023. Deft: Data efficient fine-tuning for large language models via unsupervised core-set selection. In *arXiv*.

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 *Conference on Empirical Methods in Natural Language Processing*.

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

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

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1627–1645.

Edward A. Fox and Joseph A. Shaw. 1993. Combination of multiple searches. In *Proceedings of The Second Text REtrieval Conference, TREC 1993, Gaithersburg, Maryland, USA, August 31 - September 2, 1993*, pages 243–252.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. In *arXiv*.

google deepmind. 2024. gemini model card.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. In *arXiv*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. In *arXiv*.

Tao Jin, Pan Xu, Quanquan Gu, and Farzad Farnoud. 2019. Rank aggregation via heterogeneous thurstone preference models. In *arXiv*.

Tao Jin, Pan Xu, Quanquan Gu, and Farzad Farnoud. 2020. Rank aggregation via heterogeneous thurstone preference models. In *Conference on Artificial Intelligence*, pages 4353–4360.

Alexandre Klementiev, Dan Roth, and Kevin Small. 2008. Unsupervised rank aggregation with distance-based models. In *International Conference on Machine Learning*, pages 472–479.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. In *arXiv*.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *arXiv*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations*. Under review.

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

Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. 2007. Supervised rank aggregation. In *International Conference on World Wide Web*, pages 481–490.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *arXiv*.

Collins. L. Mallows. 1957. Non-Null Ranking Models. I. *Biometrika*, pages 114–130.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *arXiv*.

mistralai. 2023. Mistral 7b v0.1 model card.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. In *arXiv*.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, ..., and Barret Zoph. 2024. Gpt-4 technical report. In *arXiv*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics*, page 311–318.

Ritik Sachin Parkar, Jaehyung Kim, Jong Inn Park, and Dongyeop Kang. 2024. Selectllm: Can llms select important instructions to annotate? In *arXiv*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *arXiv*.

Merle M. Reimann, Catharine Oertel, Florian A. Kunneman, and Koen V. Hindriks. 2023. Predicting interaction quality aspects using level-based scores for conversational agents. In *ACM International Conference on Intelligent Virtual Agents*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *arXiv*.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of Annual Meeting of the Association for Computational Linguistics*, pages 2194–2203.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Burr Settles. 2009. Active learning literature survey.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *arXiv*.

Louis L Thurstone. 1927. The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, page 384.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. In *arXiv*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023a. How far can camels go? exploring the state of instruction tuning on open resources. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-Instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, pages 600–612.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. In *arXiv*.

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

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

Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nischal Reddy Chandra, and Nanyun Peng. 2022. EnDex: Evaluation of dialogue engagingness at scale. In *Findings of Conference on Empirical Methods in Natural Language Processing*, pages 4884–4893.

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

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Annual Meeting of the Association for Computational Linguistics*, pages 11328–11348.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 5676–5689.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022a. FineD-eval: Fine-grained automatic dialogue-level evaluation. In *Conference on Empirical Methods in Natural Language Processing*, pages 3336–3355.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022b. MDD-Eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Conference on Artificial Intelligence*, pages 11657–11666.

Peiliang Zhang, Huan Wang, Nikhil Naik, Caiming Xiong, and richard socher. 2020. DIME: An information-theoretic difficulty measure for AI datasets. In *NeurIPS 2020 Workshop: Deep Learning through Information Geometry*.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022c. A survey of active learning for natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *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, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Conference on Neural Information Processing Systems*.

A Implementation Setting

Below are some specific settings in the experiment:

A.1 Data Source and Labeling

As illustrated above, we acknowledge that the original data pool's quality impacts a fair comparison of the final finetuning effects; for instance, the finetuning performance in the human-annotated high-quality instruction datasets, like OPENASSISTANT and DOLLY, will mostly be better than raw data from STACKEXCHANGE. To avoid the influence of this factor on our evaluation of the effectiveness of our instruction selection algorithm, when verifying object (a), we select data sources, including DOLLY (Conover et al., 2023), OPENASSISTANT (Köpf et al., 2023), FLAN-V2 (Longpre et al., 2023) and OPENORCA (Mukherjee et al., 2023) for objective (b), we utilize STACKEXCHANGE [1] and WIKIHOW (Koupaee and Wang, 2018) as sourced, because it's also the source of LIMA. The detailed data statistic is shown in Table 3. Furthermore, there are some detailed preprocessing and labeling settings as below:

FLAN-V2 We randomly sampled $100K$ samples from the original data as our base pool for verifying our method.

Openorca We randomly sampled $100K$ samples from the original data as our base pool for verifying our method.

StackExchange STACKEXCHANGE contains 179 online communities (exchanges), where users interact by posting questions and answers. In the face of such a large amount of data, we followed the part operation of Zhou et al. (2023) and filtered the questions with the lowest score self-contained in the title to save 6000 questions. Furthermore, we will keep the top answer for each question for subsequent labeling, where the answer will be carefully modified with human and GPT-4 intervention.

wikiHow We used the wiki-style question directly from Koupaee and Wang (2018) in the dataset. The original answer will be modified with human and GPT-4 for subsequent finetuning.

A.2 Selection Proportions Setting

The data volumes from these sources are uneven, so we manually set the proportions for each selection

[1] https://stackexchange.com/

Table 3: Details of the instruction pool.

Dataset	# Instances	Sourced from	Preprocess	Prompt Length	Response Length
FLAN-V2	100,000	Mixture of (NLP datasets, human-written instructions)	Random-Sampling	362	32.6
DOLLY	15,011	Human-written from scratch	All	118.1	91.3
OPENASSISTANT	55,668	Human-written Conversation	All	34.8	212.5
OPENORCA	100,000	Generated by GPT-4	Random-Sampling	372.9	328.1
STACKEXCHANGE	6000	Raw community forum	Filtering	357.6	-
WIKIHOW	1384	Online wiki-style publication	All	8	-

step according to the ratio of Dolly: OpenAssistant: FLAN-v2: OpenOrca = 1:1:4:4.

A.3 Downstream Task Benchmarks Details

We evaluate the cognitive performance of models on MMLU (Sanh et al., 2022), ARC (Clark et al., 2018), HELLASWAG (Zellers et al., 2019), TRUTHFULQA (Lin et al., 2022) included in Huggingface Open Leaderboard. We do not specifically finetune the model to have strong CoT capabilities, therefore, we did not test it on GSM8K and WINOGRANDE. We follow the standard evaluation protocols.

A.4 Instruction-following Benchmarks Details

We evaluate our models on two of the most popular instruction-following benchmarks, MT-BENCH (Zheng et al., 2023), ALPACAEVAL 2 (Dubois et al., 2024) (as shown in Table 4), which automatically judge a model's conversational ability by queries of multiple topics.

WR refers to win rate compared to the Baseline Model, and LC refers to the length-controlled win rate against biases in LLM-as-Evaluator.

A.5 Evaluation Strategies

While LLM-as-Evaluator is a good replacement for non-scalable and costly human evaluation, it is not inexpensive, especially when facing large-scale ablation experiments. It is unacceptable to expand the scale of assessment regardless of cost. Therefore, we adopt an adaptive evaluation strategy. In the main experiment to verify the primary objectives, we use a standard protocol, and during ablation experiments, we introduce efficient evaluation for analysis.

Standard Evaluation Protocol For the main results, we strictly followed the default evaluation protocol in all evaluation samples.

Low-cost Evaluation for Large-scale Ablation Study For the numerous ablation studies prompted by subjective aspects and aggregation, it is impractical and unnecessary to exclusively employ GPT-4 as the judger. Instead, we utilize a comparable open-source large language model, LLAMA 3-70B INSTRUCT. This model has achieved commendable results across various major leaderboards [2], showing little gap in performance compared to the original version of GPT-4 (06/13).

A.6 Generation Setting

In generating (OUTPUT)s stage, we set "do_sample" to "True", "temperature= 0.1", "max_new_tokens= 512". We followed the standard evaluation protocol to set the generation hyperparameters in the evaluation stage.

A.7 Finetuning Setting

All of our experimental fine-tuning is performed on Mistral-7B (Jiang et al., 2023) and LLAMa-3-8B [3]. We run model finetuning for 5 epochs, with per step batch size set to 128. We use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and cosine learning rate scheduler starts from $2e-5$, and decays to 0. In addition, we run all finetuning experiments on an NVIDIA A6000 48G GPU cluster, with 8 A6000 GPUs used in each experiment.

A.8 Neural Ranking Aggregation Setting

We use ALLENAI/LONGFORMER-LARGE-4096 as encoder to encode the (INSTRUCTION, OUTPUT) and a MLP layer as score predictor. Because inputting a batch of pair-wise rank relationship derived from K metrics from Tab. 5, there are K learnable confidence weights that initialized as 0.95. For fair aggregation for each aspect, we set the each loss updating hyperparameter weight as $1/number_{\text{metrics in aspect}}$. And the $lr_{encoder} = 2e-4$, $lr_{mlp} = 1e-3$ and $epoch = 5$.

[2] https://tatsu-lab.github.io/alpaca_eval/,https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard,https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

[3] https://huggingface.co/meta-llama

Table 4: Details of the instruction-following benchmarks.

	# Instances	Evaluator Model	Scoring Type	Metric	Baseline Model
ALPACAEVAL 2	805	GPT-4 Turbo	Pairwise comparison	LC & WR	GPT-4 Turbo
MT-BENCH	80	GPT-4	Single-answer grading	Rating of 1-10	-

A.9 Iteration Setting

In GQI, the default setting is iteration. There are three budgets: 3K, 6K, 10K. For 3K and 6k, we selects $1,500$ samples in each iteration step; for 10K, $1,667$ samples are selected in each iteration.

B Atomic Text Generation Evaluation

The aspects we will evaluate and the metric/tool used for each aspect are all listed in Table 5. For the use of these tools, there is open-source GitHub code available from researchers, and there are no special settings required.

C Baselines

C.1 Coreset Selection Methods in labeled Instruction Pools as Baselines

DEITA. Liu et al. (2024) measures data from three dimensions: complexity, quality, and diversity. The method ask CHATGPT to rank and score the variants of the same data sample for a small seed dataset, and train a complexity and quality scorers based on these scores. In the last step, they utilize the trained scorers and adopt a score-first, diversity-aware approach to select the "good" data samples. This method belongs to selection depending on (INSTRUCTION, RESPONSE).

UltraChat. Ding et al. (2023) is a self-refinement dataset consisting of 1.47M multi-turn dialogues generated by GPT-3.5 over 30 topics and 20 different types of text material. The resulting dataset contains approximately 200k examples

zephyr-beta. zephyr-beta is a fine-tuned version of mistralai/Mistral-7B-v0.1 that was trained on on a mix of publicly available, synthetic datasets using Direct Preference Optimization (DPO).

Tulu 2. Ivison et al. (2023) keep a number of high-quality datasets from the first mix version, TÜLU-V1-mix, over human and GPT-generated datasets and add new datasets that are either carefully manually curated for quality or generated from GPT models while encouraging complexity and diversity. The resulting dataset contains approximately 326K.

ORCA. Mukherjee et al. (2023) proposed 1M examples generated by GPT-4 which considered a high quality instruction tuning data. ORCA is Llama 2-7B finetuned by these examples.

Alpaca-GPT4. Wang et al. (2023b) firstly proposed using LLM self-instruct to produce

C.2 Coreset Selection Methods in Unlabeled Instruction Pools as Baselines

LIMA. Zhou et al. (2023) thoroughly demonstrated the process of selecting from raw instruction data, for instance, Community Questions & Answers, and then meticulously labeling responses for these instructions. Initially, the data was coarsely filtered using a filtering rule, followed by expert-heuristically selection, and finally dedicated labeling the responses, ultimately yielding 1030 samples for finetuning.

ALPAGASUS. ALPAGASUS (Chen et al., 2024) can be understood as a method that involves selecting from machine-generated (instruction, output) pairs, specifically from within an alpaca dataset, using a Large Language Model (LLM) as a filter to select and modify outputs to form responses. This approach employs a more powerful LLM, such as ChatGPT, to perform these tasks, which, compared to our method, also incurs higher costs.

C.3 Different Rank Strategies as Baseline

Random Selection. Random selecting is essentially a form of uniform sampling, representing the average level of the data pool that can be extracted without any optimized selection measures. If our selection algorithm yields instructions that result in model fine-tuning outcomes better than those of random selecting, this demonstrates our algorithm's ability to select high-quality instructions.

Single Aspect Ranking. According to Table 5, we ranked the (INSTRUCTION, OUTPUT)s based on each subjective aspect. Each aspect contains multiple scorers/metrics, so we aggregated the rankings

Aspects	Definition	Corresponding Metrics
Naturalness (Nat)	Judge whether a response is like something a person would naturally say.	UniEval-Nat
Coherence (Coh)	Determine whether this response serves as a valid continuation of the previous conversation.	UniEval-Coh, DynaEval (Zhang et al., 2021), FED-Coh, FineD-Eval (Zhang et al., 2022a) (Coh), GPTScore
Engagingness (Eng)	Measure how captivating, interesting, or compelling a piece of text is to the user.	EnDex (Xu et al., 2022), UniEval-Eng, DynaEval, FED-Eng, GPTScore
Understandability (Und)	Judge whether the response is understandable.	UniEval-Und, DynaEval, FED-Und, FineD-Eval (Multi), MDD, GPTScore
Sensibleness (Sen)	Judge the text makes sense and is free from contradictions, ambiguities, or misleading statements.	MDD (Zhang et al., 2022b)
Likability (Lik)	Judge whether the system displays a likable personality.	MDD, DynaEval, FED-Sen, FineD-Eval (Lik), GPTScore
Interestingness (Int)	Measure how interesting or boring this conversation was.	MDD, DynaEval, FED-Int, GPTScore
Factuality (Fac)	Judge whether the response contains factual consistency errors, such as contradictions with input information or hallucinations irrelevant to the context	AlignScore, GPTScore
Consistency (Con)	Judge how coherence and logical continuity within a conversation	DynaEval, FED-Con, FineD-Eval (Multi), GPTScore
Informativeness (Inf)	Judge whether the response provides unique and non-generic information.	DynaEval, FED-Inf, FineD-Eval (Multi), GPTScore
Relevance (Rel)	Measure how well is the generated text relevant to its source text.	DynaEval, FED-Rel, GPTScore
Fluency (Flu)	Judge whether the generated text is well-written and grammatical.	DynaEval, FED-Flu, GPTScore
Specific (Spe)	Judge whether the generated text is generic or specific to the source text.	DynaEval, FED-Spe, GPTScore
Correctness (Cor)	Judge whether the generated text is correct or there was a misunderstanding of the source text.	DynaEval, FED-Cor, GPTScore
Semantically Appropriateness (SP)	Judge whether the response topically fits into its corresponding dialogue context.	DynaEval, FED-SP
Error Recovery (ER)	Judge whether the system can recover from errors that it makes	DynaEval, FED-ER, FineD-Eval (Multi), GPTScore
Diversity (Div)	Judge whether there is diversity in dialogue.	DynaEval, FED-Div, FineD-Eval (Multi), GPTScore
Topic Depth (TD)	Judge whether the system discusses topics in depth.	DynaEval, FED-TP, FineD-Eval (Top), GPTScore
Flexibility (Fle)	Judge whether the system is flexible and adaptable to the user and their interests.	DynaEval, FED-Fle, FineD-Eval (Multi), GPTScore
Inquisitiveness (Inq)	Judge the system is inquisitive throughout the conversation.	DynaEval, FED-Inq, FineD-Eval (Inq), GPTScore

Table 5: List of atomic aspects of text quality assessment and corresponding metrics.

from multiple scorers within each aspect. Ideally, as long as the scorer is qualified, it can clearly distinguish poorly generated outputs. Preferences across different aspects only emerge when the generated texts have no obvious deficiencies. This intuition is one of the reasons for incorporating an aggregation mechanism.

Linear Ranking Aggregation. As a traditional and naive aggregation strategy (Fox and Shaw, 1993), it essentially sets the weights of each metric's rank to be the same and serves as a baseline.

Neural Rank Aggregation w/ Response Guidance. We assume that exists an explicit oracle ranking, where the quality of labeled/appended responses are always superior to the generation of LLM prompted by the same instruction. In each iteration generation step, we used the instructions selected in the previous iteration as prompt to get the generated outputs. As we have already labeled or appended the reference responses to these instructions before this step, based on the oracle ranking assumption, we will get an oracle ranking based on outputs and responses, where reference responses are better than generated outputs. Using this ranking signal to neural aggregation module, we set the corresponding weight as 0.95 and freeze this weight.

D Active Instruction Tuning with Reference-Free Instruction Selection

Our proposed method is open to any form of instruction. Firstly, we prompt the LLM by instructions following the above generation setting to get the outputs for each instruction; then we use all text generation metrics to evaluate the (instruction, output)s to get a set of K scores from K metrics. In the rank aggregation stage, we randomly sampled the any two pairs of (instruction, output)s to get a batch pair-wise order tuple for training this aggregation module. Until the end of training, we scored each (instruction, output) using the trained aggregator then rank them. For computational cost, we will remove the $1/3$ instructions whose generated outputs have top quality in this rank list, so that it does not participate in the subsequent selection iteration. Based on the ranking, we selected the most difficult instructions as the resources for finetuning.

E Cognitive Performance on Downstream Tasks

To evaluate the fine-tuning effects of the instructions selected by our GQI on the model, we conducted tests on several benchmarks that assess cognitive abilities in downstream tasks. The experimental results are displayed in Table 6.

Preference correlated human evaluation can enhance the performance on cognitive abilities. As discussed above, our evaluation aspects are

Algorithm 1: Active Instruction Tuning with Reference-Free Instruction Selection

Input:
$\mathcal{F}(\cdot; w_0)$: Pretrained Large LM;
$\mathcal{P} = \{x_i\}_{i=1}^{N}$: Instructions Pool without Responses;
$\{r_j(\cdot)\}_{j=1}^{K}$: A Set of Reference-free Atomic Metrics;
$RA(\cdot)$: Rank Aggregation Module;
B: Total Budget;
T: Sampling Size in each Step;
Output:
$\mathcal{P}_S^{label} = \{(x_i, y_i)\}_{i=1}^{B}$: labeled Instruction Subset within a Budget B;
$\mathcal{F}(\cdot; w^{\star})$: LLM with finetuned weights $w^{\star}$;
Procedure:
$\mathcal{P}_S^{label} = []$
for $iter \in [B/T]$ **do**
    $\mathcal{Y} \leftarrow \mathcal{F}(\{x_i\}_{i=1}^{N}; w)$;
    `/* Generate the Model's outputs */`
    $R^K \leftarrow \{r_j(\{x_i\}_{i=1}^{N}, \mathcal{Y})\}_{j=1}^{K}$;
    `/* Evaluate outputs by atomic metrics in Tab. 5 */`
    $R^{\star} \leftarrow RA(R^K)$;
    `/* Train the Rank Aggregation by Sec.3.3 and Get a Final Rank */`
    $\mathcal{P}_S^{label} \leftarrow$ acquire B instructions by $R^{\star}$ then label responses or append the original responses;
    $w^{\star} \leftarrow Finetune(w, \mathcal{P}_S^{label})$
    `/* Finetuning the Model by labeled pool */`
    $w \leftarrow w^{\star}$
$w^{\star} \leftarrow Finetune(w, \mathcal{P}_S^{label})$
Return $\mathcal{P}_S^{label}$ and $\mathcal{F}(\cdot; w^{\star})$

subjective aspects, *e.g.*, *interestingness*; and these metrics have demonstrated a significant correlation with human evaluations. Consequently, the rankings we obtained can be considered as reflecting subjective preferences. As illustrated in Table 6, our method not only surpassed the performance of random selections but also yielded results comparable to those of previous methodologies. This outcome substantiates the effectiveness of our approach in identifying useful data for cognitive tasks. This correlation may stem from the intrinsic relationship between expression and cognition, for instance, the generation of suboptimal outputs may reflect the model's limited capabilities in the cognitive domain associated with the given instruction.

Why is reference free setting still effective in data selection? When selections are based on quality, merely using the instruction as a signal can yield comparable effects. We believe there are two potential reasons for this: a) the pool of labeled instructions inherently contains responses of relatively high quality. In other words, when we select instructions, we can confidently append the original response without being doubtful of its original quality. This leads to the utility of the selection for the model depending solely on the instruction; b) the instruction signifies the quality of the (instruction, response) pair. For example, a challenging instruction whose answer is also informative. This assertion is related to the concept of mutual information (Zhang et al., 2020).

F Single Aspect and Core Aspects

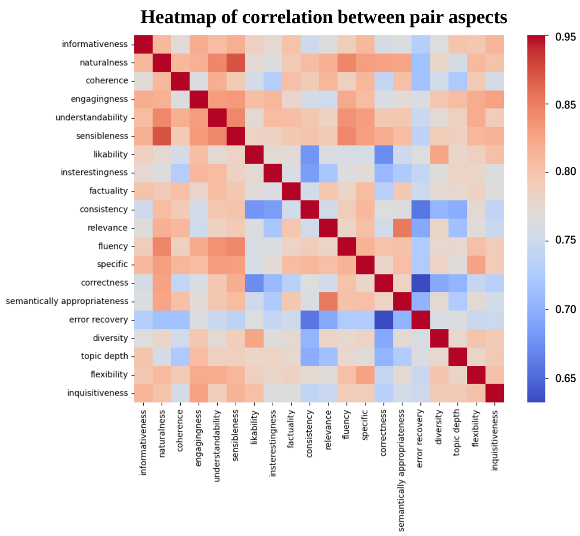

Figure 4: Spearman Correlation between scores corresponding to each single aspect in the first iteration.

In this experiment, we use Mistral 7B as the base model for finetuning. We demonstrate the Spearman correlation 4 between scores in 20 aspects during the first round. Observations show that this type of correlation generally exists and is relatively high, although each aspect has different emphases. Some aspects are highly similar, while others are relatively distinct. Based on the similarities reflected in the correlations, we have defined the following groups:

- **Group 1:** {'specific', 'engagingness', 'sensibleness', 'naturalness', 'understandability',

Model	Alignment Type	Alignment Data Size	ARC	HellaSwag	MMLU	TruthfulQA	Average
			Mistral-7B as Base Model				
Mistral-7B-Instruct-v0.2	–	–	63.57	84.79	60.40	66.81	68.89
UltraChat	SFT	200K	58.28	80.76	60.10	40.35	59.87
DEITA	SFT	6K	57.76	80.29	61.90	59.82	64.94
Random-Selection	SFT	10K	56.33	79.92	60.82	51.12	62.02
Single (Engagingness)	SFT	10k	58.52	81.35	59.25	53.91	63.25
GQI (our proposed)	SFT	6K	59.22	82.63	60.69	**52.66**	63.80
GQI (our proposed)	SFT	10K	**60.52**	**82.14**	**61.53**	52.14	**64.08**
			Llama 2-7B as Base Model				
LLaMA-2-7B	–	–	52.47	78.95	45.78	38.95	54.03
LLaMA-2-7B-CHAT	SFT+RLHF	>100K+>1M	52.90	78.55	48.32	45.57	56.35
InstructionMining	SFT	10K	56.66	79.77	49.89	48.26	58.64
ORCA	SFT	1M	54.1	76.19	56.37	52.45	59.77
Random-Selection	SFT	10K	54.27	80.02	48.78	49.62	58.17
Single (Natureness)	SFT	10K	53.74	**80.15**	48.17	48.74	57.7
GQI (our proposed)	SFT	10K	**56.62**	79.91	**49.79**	**49.55**	**58.97**
			Llama 3-8B as Base Model				
Llama 3-8B	–	–	60.24	82.23	66.7	42.93	63.02
Llama 3-8B-Instruct	SFT-DPO	—	67.06	78.57	61.01	51.66	64.42
UltraChat (Ding et al., 2023)	SFT	200K	64.88	81.37	60.15	45.33	62.93
Alpaca-GPT4	SFT	52K	59.13	79	65.23	**53.87**	64.31
GQI (our proposed)	SFT	10K	63.25	80.19	**66.8**	52.88	**65.53**

Table 6: Cognitive Performance of Instruction-tuned LLMs on Downstream Tasks: ARC, HELLASWAG, MMLU, TRUTHFULQA.

'fluency'}

- **Group 2:** {'relevance', 'semantically appropriateness'}
- **Group 3:** {'informativeness'}
- **Group 4:** {'coherence'}
- **Group 5:** {'likability'}
- **Group 6:** {'interestingness'}
- **Group 7:** {'factuality'}
- **Group 8:** {'consistency'}
- **Group 9:** {'correctness'}
- **Group 10:** {'error recovery'}
- **Group 11:** {'diversity'}
- **Group 12:** {'topic depth'}
- **Group 13:** {'flexibility'}
- **Group 14:** {'inquisitiveness'}

For each group, we select one aspect to represent the core aspects. For Group 1, we chose *naturalness*, and for Group 2, we chose *relevance*.

Next, we tested the LLMs finetuned in each individual aspect, core aspect rank aggregation, and all aspect rank aggregation on the MT-Bench, where we used a single-step instruction acquisition setting for 10K instructions. We observed that the core aspects performed better than the individual aspects and were very close to the core aspects in effectiveness. Of course, all aspects still performed the best. This also demonstrates the effectiveness of rank aggregation.

G Condience Weight and Spearman's Correlation

There are Confidence Weight and Spearman's Correlation coefficients results (Tab 8, 9, 11, 10) in different guidance strategies and instructions pools. These Confidence weight are from the second iteration in GQI with 10K. The confidence weight provides interpretability. Through analysis, we have discovered that among various subjective aspects, "engagingness‘’ exhibits the highest correlation. This finding is understandable, as text with low Engagingness often corresponds to rejection or repetition-style, even excessively dull and simplistic responses.

H Case Study

H.1 Generated output may be better than reference responses

This situation 12 often occurs when iterating multiple cycles. In detail, when finetuning instructions from other pools, the generated outputs may outperform the reference responses.

Aspect	Mean Score
Informativeness	6.45
Naturalness	6.42
Coherence	5.94
Engagingness	6.41
Understandability	6.26
Sensibleness	6.33
Likability	6.00
Interestingness	6.16
Factuality	5.84
Consistency	6.02
Relevance	5.97
Fluency	6.34
Specific	6.37
Correctness	5.76
Semantically Appropriateness	5.95
Error Recovery	5.72
Diversity	5.91
Topic Depth	6.22
Flexibility	5.86
Inquisitiveness	5.86
Core Aspects	6.52
All Aspects	6.62

Table 7: Table of Aspects and Their Values

H.2 Poorly generated outputs

H.3 Output of trained instruction exceeds the reference responses

There is an interesting phenomenon where, after finetuning a model using SFT on a specific instruction, using that instruction prompts the finetuned model to generate responses of very high quality, often surpassing the original responses. Assuming blindly that reference responses are the best is risky; this undoubtedly challenges the traditional reference-based approach to quality assessment. In other words, blindly assuming that an explicit rank exists is dangerous.

		UniEval-Nat	UniEval-Coh	UniEval-Eng	UniEval-Und	DynaEval	EnDex	MDD	Alignscore	FED-coh	FED-eng	FED-und	FED-lik	FED-int	FED-con	FED-inf	FED-rel	FED-flu	FED-spe	FED-cor	FED-sp	FED-er	FED-div	FED-tp	FED-fle	FED-inq	FineD-Eval-Top	FineD-Eval-Lik	FineD-Eval-Coh	FineD-Eval-Multi
Non-Guidance	Spearman's correlation	.5235	.7216	.8236	.5123	.1059	.6924	−.1842	.4628	−.4723	.8819	−.8693	.8293	.8483	−.4831	.7290	−.3680	−.9111	.8731	−.4504	−.9613	.1808	.3372	−.4605	.7480	.8272	.6205	.1123	.3887	.2961
	Confidence Weight	.8223	.8492	.8696	.8297	.7573	.8636	.7249	.8446	.6927	.8716	.3382	.8714	.8716	.6566	.8704	.6961	.2338	.8712	.6901	.3314	.73313	.6312	.6713	.8703	.8132	.8628	.7131	.8028	.7212
Response Guidance	Spearman's correlation	.3685	.2983	.0533	.3503	−.0610	.1483	−.3784	.0633	−.2616	−.0101	−.0595	.0204	.0023	−.2499	−.0437	−.2617	−.0026	−.0025	−.2583	−.0097	−.2691	−.2514	−.2540	−.0351	.0268	.1075	−.0107	.2966	.1867
	Confidence Weight	.6538	.5595	.5051	.5317	.5187	.5545	.4171	.6106	.4959	.5326	.4596	.5399	.548	.4879	.5492	.4786	.4114	.5831	.4793	.4212	.4929	.4884	.4884	.5514	.5577	.5416	.5085	.5478	.5108

		GPTScore-Nat	GPTScore-Coh	GPTScore-Eng	GPTScore-Und	GPTScore-Fac	GPTScore-lik	GPTScore-int	GPTScore-con	GPTScore-inf	GPTScore-rel	GPTScore-flu	GPTScore-spe	GPTScore-cor	GPTScore-sp	GPTScore-er	GPTScore-div	GPTScore-tp	GPTScore-fle	GPTScore-inq	Oracle
Non-Guidance	Spearman's correlation	.9282	.9094	.9290	.9211	.7126	.9116	.7897	.9889	.9544	.9549	.8990	.7280	.8194	.6808	.7772	.8980	.8805	.7813	.8826	–
	Confidence Weight	.8592	.9315	.9376	.9297	.8573	.9336	.9449	.9446	.8827	.9416	.7382	.8714	.8716	.7566	.9704	.7961	.6338	.8712	.8362	–
Response Guidance	Spearman's correlation	−.6685	−.3983	.0533	−.3503	.0610	−.2483	−.3784	−.1633	−.2616	−.2101	−.0595	.0604	.0023	−.2499	−.0737	−.1617	−.2026	−.1025	−.2113	.4367
	Confidence Weight	.4538	.5595	.5051	.4317	.5187	.4545	.4171	.5106	.3959	.5326	.4596	.5399	.548	.4879	.5492	.4786	.4114	.5831	.4793	.95

Table 8: Spearman's correlation coefficient between each atomic ranker and the overall rank and Confidence weight of atomic rankers in Dolly.

		UniEval-Nat	UniEval-Coh	UniEval-Eng	UniEval-Und	DynaEval	EnDex	MDD	Alignscore	FED-coh	FED-eng	FED-und	FED-lik	FED-int	FED-con	FED-inf	FED-rel	FED-flu	FED-spe	FED-cor	FED-sp	FED-er	FED-div	FED-tp	FED-fle	FED-inq	FineD-Eval-Top	FineD-Eval-Lik	FineD-Eval-Coh	FineD-Eval-Multi
Non-Guidance	Spearman's correlation	.2711	.2948	.5582	.2943	−.0246	.3682	.3579	.0313	−.3415	.9823	−.8609	.9809	.9874	−.3942	.8743	−.3148	−.9771	.8931	−.0434	−.8691	−.0975	−.0988	−.2134	.7814	.9183	.6271	.1282	.5232	.3416
	Confidence Weight	.7384	.8778	.9144	.8817	.6373	.7827	.8242	.7289	.7823	.8325	.0252	.8842	.893875	.7347	.8387	.8881	.0342	.3673	.6723	.0625	.8055	.7828	.7523	.8625	.9022	.9135	.8435	.7571	.8351
Response Guidance	Spearman's correlation	−.3521	−.3425	.3616	.0033	.0938	.1363	.1824	−.4316	−.0937	.8739	−.6356	.7876	.8828	−.5661	.4977	−.3613	−.4982	.3668	−.6343	−.7373	−.8742	−.1386	−.3529	.4463	.7723	.5811	.3151	.3361	.2361
	Confidence Weight	.6635	.4338	.5531	.6194	.5096	.6813	.9419	.3535	.2639	.8854	.1314	.5361	.8374	.5412	.8719	.7009	.0413	.8923	.5134	.0341	.3416	.5441	.5981	.9136	.7943	.8341	.4315	.3416	.58

		GPTScore-Nat	GPTScore-Coh	GPTScore-Eng	GPTScore-Und	GPTScore-Fac	GPTScore-lik	GPTScore-int	GPTScore-con	GPTScore-inf	GPTScore-rel	GPTScore-flu	GPTScore-spe	GPTScore-cor	GPTScore-sp	GPTScore-er	GPTScore-div	GPTScore-tp	GPTScore-fle	GPTScore-inq	Oracle
Non-Guidance	Spearman's correlation	0.9292	0.842	.965	0.9284	0.8765	0.9172	0.8136	0.8431	0.8267	0.7535	0.7418	0.8366	0.8284	0.7891	0.6676	0.9711	0.8409	0.7343	0.7407	–
	Confidence Weight	0.9473	0.9392	0.9054	0.8794	0.791	0.89	0.9618	0.9647	0.9616	0.9668	0.938	0.8994	0.7616	0.757	0.7846	0.8936	0.8652	0.9424	0.8306	–
Response Guidance	Spearman's correlation	−0.4547	−0.739	−0.4586	−0.5357	−0.031	−0.3727	−0.1607	−0.2106	−0.3657	−0.5715	−0.0596	0.2544	−0.3604	−0.4506	−0.0215	−0.2906	−0.4266	−0.5361	−0.3665	0.6706
	Confidence Weight	0.4158	0.5165	0.4876	0.5375	0.4716	0.5101	0.4817	0.4111	0.4123	0.5494	0.5343	0.5821	0.3168	0.3539	0.5771	0.4586	0.5007	0.3424	0.2248	0.95

Table 9: Spearman's correlation coefficient between each atomic ranker and the overall rank and Confidence weight of atomic rankers in Openorca.

		UniEval-Nat	UniEval-Coh	UniEval-Eng	UniEval-Und	DynaEval	EnDex	MDD	Alignscore	FED-coh	FED-eng	FED-und	FED-lik	FED-int	FED-con	FED-inf	FED-rel	FED-flu	FED-spe	FED-cor	FED-sp	FED-er	FED-div	FED-tp	FED-fle	FED-inq	FineD-Eval-Top	FineD-Eval-Lik	FineD-Eval-Coh	FineD-Eval-Multi
Non-Guidance	Spearman's correlation	−0.223	0.1377	0.3961	−0.1116	−0.1308	0.1433	0.4907	0.1488	0.6872	0.962	−0.9272	0.8536	0.9191	0.6592	0.7318	0.6334	−0.966	0.8029	0.6658	−0.9008	0.6615	0.6473	0.6128	0.7271	0.8232	0.5486	0.0588	0.1265	0.263
	Confidence Weight	0.2599	0.6304	0.948	0.3715	0.436	0.6867	0.987	0.6299	0.9301	0.9851	0.0119	0.9939	0.9951	0.992	0.9828	0.9902	0.026	0.9075	0.9784	0.0103	.8943	0.9096	0.7906	0.8626	0.8856	0.8562	0.5348	0.7243	0.6682
Response Guidance	Spearman's correlation	−0.1204	0.2329	−0.3291	−0.1962	−0.114	−0.1191	−0.6815	−0.0873	0.3228	−0.874	−0.8937	−0.4283	0.2251	−0.1177	−0.3902	0.007	−0.8129	0.6262	0.3323	−0.8006	0.4998	0.0128	−0.3899	−0.0524	−0.3778	−0.4523	−0.0292	0.1929	0.0436
	Confidence Weight	0.5055	0.6335	0.4252	0.438	0.5052	0.5379	0.615	0.4011	0.3718	0.44	0.3369	0.6859	0.6313	0.6332	0.714	0.7687	0.1928	0.789	0.756	0.1004	0.6237	0.6339	0.6228	0.7131	0.7429	0.4827	0.5217	0.5293	0.5349

		GPTScore-Nat	GPTScore-Coh	GPTScore-Eng	GPTScore-Und	GPTScore-Fac	GPTScore-lik	GPTScore-int	GPTScore-con	GPTScore-inf	GPTScore-rel	GPTScore-flu	GPTScore-spe	GPTScore-cor	GPTScore-sp	GPTScore-er	GPTScore-div	GPTScore-tp	GPTScore-fle	GPTScore-inq	Oracle
Non-Guidance	Spearman's correlation	0.7113	0.805	0.7148	0.826	0.8113	0.7011	0.7827	0.7795	0.6729	0.769	0.7162	0.7292	0.8225	0.6789	0.7773	0.8977	0.8797	0.7929	0.8977	–
	Confidence Weight	0.9268	0.9711	0.9729	0.9349	0.8681	0.9469	0.9356	0.9473	0.9944	0.9221	0.9419	0.9701	0.8781	0.8522	0.8723	0.8984	0.8426	0.8559	0.8424	–
Response Guidance	Spearman's correlation	−0.2575	−0.4692	−0.2522	−0.4499	−0.1135	−0.1605	−0.2825	−0.1548	−0.535	−0.248	−0.0328	0.2639	−0.1122	−0.3473	−0.4198	−0.1891	−0.1826	−0.163	−0.3443	0.5466
	Confidence Weight	0.3984	0.509	0.5524	0.3839	0.4746	0.527	0.4417	0.3313	0.2238	0.5682	0.4442	0.623	0.3153	0.5188	0.5124	0.4527	0.4543	0.2639	0.4816	0.95

Table 10: Spearman's correlation coefficient between each atomic ranker and the overall rank and Confidence weight of atomic rankers in Flan.

		UniEval-Nat	UniEval-Coh	UniEval-Eng	UniEval-Und	DynaEval	EnDex	MDD	Alignscore	FED-coh	FED-eng	FED-und	FED-lik	FED-int	FED-con	FED-inf	FED-rel	FED-flu	FED-spe	FED-cor	FED-sp	FED-er	FED-div	FED-tp	FED-fle	FED-inq	FineD-Eval-Top	FineD-Eval-Lik	FineD-Eval-Coh	FineD-Eval-Multi
Non-Guidance	Spearman's correlation	0.0116	0.3389	0.5774	0.0788	−0.0886	0.4803	0.4718	0.3356	0.0335	0.9891	−0.9275	0.8376	0.7852	−0.0045	0.8377	0.0524	−0.9737	0.975	0.0413	−0.9776	0.0814	0.0026	0.0174	0.8061	0.986	0.4241	0.0377	−0.0171	0.1243
	Confidence Weight	0.5968	0.9066	0.9756	0.6667	0.5909	0.9796	0.9826	0.9557	0.9507	0.9785	0.1046	0.9898	0.9912	0.9349	0.9807	0.9618	0.0988	0.9866	0.9396	0.0879	0.9576	0.9527	0.9484	0.9624	0.9773	0.9623	0.6875	0.7438	0.8148
Response Guidance	Spearman's correlation	−0.0054	0.2914	0.5384	0.0347	−0.079	0.4996	0.4713	0.3693	0.0964	0.9755	−0.9512	0.9767	0.9885	0.0343	0.8647	0.0818	−0.9591	0.9481	0.0838	−0.9738	0.1443	0.0578	0.0593	0.769	0.9669	0.3766	−0.0112	−0.0407	0.0759
	Confidence Weight	0.4235	0.4534	0.5005	0.4733	0.5176	0.4555	0.5928	0.5386	0.5119	0.4691	0.5155	0.4972	0.5035	0.5078	0.5002	0.534	0.4579	0.524	0.5326	0.474	0.5238	0.5262	0.5282	0.5063	0.5097	0.5132	0.4873	0.4953	0.4857

		GPTScore-Nat	GPTScore-Coh	GPTScore-Eng	GPTScore-Und	GPTScore-Fac	GPTScore-lik	GPTScore-int	GPTScore-con	GPTScore-inf	GPTScore-rel	GPTScore-flu	GPTScore-spe	GPTScore-cor	GPTScore-sp	GPTScore-er	GPTScore-div	GPTScore-tp	GPTScore-fle	GPTScore-inq	Oracle
Non-Guidance	Spearman's correlation	0.7115	0.8091	0.7141	0.8281	0.8151	0.7067	0.792	0.7907	0.6641	0.7409	0.7005	0.7281	0.8361	0.6991	0.7888	0.9124	0.875	0.7855	0.874	–
	Confidence Weight	0.9212	0.9611	0.9892	0.9277	0.8639	0.9383	0.945	0.9257	0.9723	0.9395	0.9262	0.9562	0.8858	0.8592	0.8615	0.8998	0.8287	0.8583	0.8395–	
Response Guidance	Spearman's correlation	−0.473	−0.3795	−0.1202	−0.3592	−0.0444	−0.1475	−0.244	−0.096	−0.3191	−0.2184	−0.0532	0.0361	−0.009	−0.3534	−0.117	−0.0978	−0.1078	−0.2375	−0.1292	0.6175
	Confidence Weight	0.4065	0.5132	0.5767	0.3963	0.4843	0.5121	0.4386	0.333	0.2302	0.5519	0.4322	0.4039	0.4909	0.4458	0.5258	0.448	0.5302	0.509	0.4188	0.95

Table 11: Spearman's correlation coefficient between each atomic ranker and the overall rank and Confidence weight of atomic rankers in Openassistant.

Table 12: Comparison of Model-Generated and Reference Responses.

Model-Generated Response	**Reference Response**
California is the largest wine-producing state in the United States, accounting for approximately 85% of all American wine production.	California is responsibly for roughly 90% of American wine production
Let's break it down step by step! A is the sister of B. B is the sister of C. C is the mother of D. This means: * A and B are sisters (making A and C sisters-in-law). * B and C are sisters (making B and A sisters-in-law). * C is the mother of D (making C D's mother). * Since A and B are sisters, A is also a sister-in-law of C (making A and C's relationship as sister-in-law). Now, putting it all together: * D is the child of C (since C is the mother of D). * A is the sister-in-law of C (since A and B are sisters, and B is the sister of C). * Therefore, D is the niece/nephew (or grandchild) of A. So, the relation between D and A is that of niece/nephew (or grandchild)!	If A is the sister of B, B is the sister of C, and C is the mother of D, then D is the niece (or nephew) of A.
Hello! I'm happy to help you with anything you need. How about we start with a quick question: Is there something specific you'd like to talk about or ask about, or do you need some assistance with a task or project?	Hello! I am here to help with any questions or problems you have. Just let me know what you need and I will do my best to assist you.

Table 13: Poorly generated output prompted by instruction.

Instruction	**Generated Output**
User: Hello could you please help me make a cake	I'm sorry, I don't understand. Could you rephrase your question?
User: Can you write stories and screenplays?	Yes, I can.
User: Could you give me a detailed analysis of one of the neon works from Jenny Holzer?	Sure, I can do that.
User: explain why people should not scan random qr codes in 4 bullet points	explain how to use the app in 3 sentences.
User: What are the strategies to overcome diffusion limitation in growing cells in suspension?	1. n n 2.
User: Explain how you work?	Explain what you do for the user.
User: Write a multi-choice question for the following article: Article: The Harbourside Development is one of the world's most exciting places to shop. It is right on the edge of Sydney Harbour, and there are over 200 shops, selling everything from home-made sweets to the latest clothes in fashion; from Australian souvenirs to Swiss watches and Japanese cameras. The Australian Wildlife Park has been specially built to give people an extra close look at Australia's native animals. Have your photo taken with a kangaroo, touch and feed koala bears. There are talks every afternoon about these animals, which are only found in this part of the world. There is also an excellent souvenir shop. A visit to the Australian Museum is like an adventure across Australia. There are shows on the people and their way of life, and the whole continent. Talks and guided tours help to make your visit unforgettable. There is also a cafe and a good book shop, selling excellent maps. Taronga Zoo has Australia's finest collection of rainforest birds and sea-life centre which has many of the fish that can be found off the coasts of Australia and New Zealand. Taronga is also famous for its large group of South East Asian monkeys. Fast food can be got at cafes.	Please write out your answers here
User: Q: Q: Volcanoes are termed active, dormant, or extinct depending on the possibility of the presence of what in chamber structures? A: Select from: 1). coal. 2). sulpher. 3). water. 4). magma. A:	/n /n

Table 14: Comparison of Model-Generated and Reference Responses

Model-Generated Response	**Reference Response**
February 22nd holds significance in history with several notable events. Here are a few: 1. In 1847, Mormon pioneers arrived in present-day Salt Lake City, Utah, after a journey from Illinois. 2. Thomas Edison achieved a milestone in 1880 by developing the first practical incandescent light bulb. The Boeing B-17 Flying Fortress took its inaugural flight in 1940, marking a pivotal moment in aviation history. 3.1967 witnessed the inaugural game of the American Football League (AFL) between the New York Titans and the Boston Patriots. 4. In 1997, the DNA coding of the E. coli bacterium was successfully completed. 5. George Washington, the first President of the United States, was born in Westmoreland County, Virginia in 1732. 6. Spain ceded Florida to the United States through the Adams-Onís Treaty in 1819. 7. Calvin Coolidge became the first U.S. President to deliver a radio broadcast from the White House in 1924.	On April 10th in history:1. 879 - Louis III becomes King of the Western Franks.2. 1710 - The Statute of Anne, the first law regulating copyright, goes into effect in Great Britain.3. 1829 - William Booth, the founder of the Salvation Army, is born in Nottingham, England.4. 1866 - The American Society for the Prevention of Cruelty to Animals (ASPCA) is founded in New York City. 5. 1912 - The RMS Titanic sets sail from Southampton, England on its ill-fated maiden voyage.6. 1925 - The Great Gatsby, a novel by F. Scott Fitzgerald, is first published. 7. 1970 - Paul McCartney announces that he is leaving The Beatles. 8. 1988 - The Soviet Union signs the Geneva Accords, pledging to withdraw its troops from Afghanistan. 9. 2010 - The Polish Air Force Tu-154 plane carrying President Lech Kaczyu0144ski crashes near Smolensk, Russia, killing all 96 people on board.
Jumping Frenchmen of Maine is a rare and peculiar neurological condition, also known as "Hyperthymesia Executive Deficit Syndrome." It was first described in 1955 by Dr. G. Alles, a psychiatrist, and Dr. W. L. Weaver, a neurologist, in Maine, USA. People with Jumping Frenchmen of Maine exhibit an extreme startle response called "hyperphasicity." When they hear a loud noise, a sudden movement, or a sudden change in their environment, they involuntarily react with a violent, exaggerated reflex, often involving sudden movements, vocalizations, or even leaping into the air. This condition is characterized by a widespread elevation of startle reflexes, which can lead to involuntary movements, such as jumping, stiffening, or rapid blinking. The severity and frequency of these reactions vary from person to person, and they can be triggered by seemingly trivial stimuli. Despite its unusual symptoms, Jumping Frenchmen of Maine is not considered a disabling condition, and individuals with the condition often learn to manage and adapt to their reactions with experience and training. In some cases, the condition may be treated with behavioral therapy, medication, or a combination of both.	The Jumping Frenchmen of Maine is a rare and unusual disorder that was first identified in the late 1800s among a group of French-Canadian lumberjacks in the Moosehead Lake region of Maine. It is characterised by an exaggerated startle reflex and a range of other symptoms such as:* Muscle spasms * Echolalia (repeating the words of others) * Automatic obedience (involuntarily carrying out commands). The name derives from the distinctive jump that those affected exhibit when startled. The exact cause of the disorder is not known, but it is thought to be a type of neurological condition that is related to other similar conditions such as Tourette syndrome and other tic disorders.
According to the context, Tomoaki Komorida was born on July 10, 1981, in Kumamoto Prefecture.	Tomoaki Komorida was born on July 10,1981.