

LEARNING TO AGGREGATE: REINFORCEMENT LEARNING FOR GENERALIZED LABEL AGGREGATION

Anonymous authors
 Paper under double-blind review

ABSTRACT

The rise of large language models (LLMs) as annotators has introduced new opportunities and challenges for label aggregation in data annotation pipelines. While traditional aggregation methods are designed for human crowd workers with independent judgments, they fall short when applied to LLM-generated annotations that exhibit high correlation patterns and provide rich explanatory justifications. To address these challenges, we introduce RFAGG, a reinforcement learning framework that dynamically aggregates LLM annotations by jointly modeling both labels and their corresponding justifications. To train RFAGG, we construct the AGG dataset by collecting question-answer pairs generated by different LLMs across various datasets. Then, RFAGG first uses LLMs to generate multiple aggregation responses containing reasoning tokens and final answers for each input, and then uses our proposed aggregation reward functions to update the model via the policy optimization algorithm. Experiments demonstrate that RFAGG significantly outperforms classical and recent aggregation methods. Most notably, it serves as a *general aggregation model, generalizing well to out-of-domain and previously unseen tasks*. Despite being trained only on limited classification tasks, RFAGG achieves an average improvement of 2.45% on diverse objective tasks and 5.2% on the Alpaca 2.0 subjective task compared to its base model. We will publicly release the AGG dataset and our source code.

1 INTRODUCTION

Data annotation is a cornerstone of machine learning, providing the labeled datasets essential for training and evaluating models. Traditionally, this process has relied on human annotators, whose individual judgments are aggregated to produce reliable labels. Common aggregation techniques, such as majority voting and the Dawid-Skene model (Dawid & Skene, 1979), are designed under the assumption that annotators provide independent judgments and focus solely on the labels themselves. Recently, the advent of large language models (LLMs) as annotators has disrupted this paradigm, offering new opportunities but also presenting unique challenges.

LLMs, such as GPT (Achiam et al., 2023) and Qwen (Yang et al., 2024), have demonstrated remarkable capabilities in generating annotations across a wide range of tasks, from text classification to complex reasoning problems. Their scalability, cost-effectiveness, and ability to provide not only labels but also detailed justifications or explanations make them attractive alternatives to human annotators (Wang et al., 2024). However, aggregating annotations from LLMs is non-trivial. Unlike human annotators, LLMs often exhibit high correlation in their outputs due to shared training data and architectural similarities (Kim et al., 2025; Li, 2024b; Wang et al., 2024). Moreover, the rich justifications they provide are typically discarded in traditional aggregation methods, leading to a loss of valuable information that could otherwise enhance the aggregation process.

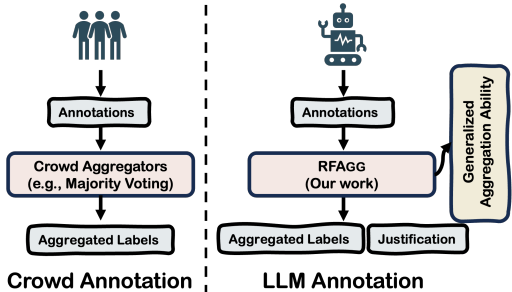


Figure 1: RFAGG enables effective aggregation of LLM annotations and achieves generalized aggregation ability beyond traditional methods.

Using LLMs as aggregators is a promising approach, leveraging their advanced reasoning to synthesize information and generate coherent explanations Fan et al. (2024). However, raw aggregation ability does not correlate with general LLM performance (Chen et al., 2025; Li, 2024a). Without specialized training, these models are susceptible to prompt phrasing and majority-opinion bias, often failing to systematically resolve conflicts or fully utilize the explanatory content in justifications. Consequently, while employing LLMs as aggregators is appealing in theory, in practice, their raw aggregation performance often falls short of expectations and highlights the need for dedicated aggregation strategies.

A straightforward approach would be supervised fine-tuning (SFT) on reference aggregations. However, this paradigm is fundamentally limited. It trains a model to replicate a static ground-truth answer, teaching it *what* to generate but failing to provide a learning signal for *how* to reason through conflicting evidence and synthesize justifications. We therefore turn to reinforcement learning (RL). By directly rewarding desirable outcomes—such as accuracy and coherent conflict resolution—RL enables the model to learn a robust and generalizable aggregation policy, rather than simply mimicking a single reference text.

To this end, we propose RFAGG, a reinforcement learning framework that dynamically aggregates LLM annotations by jointly modeling both the predicted labels and their corresponding justifications. Unlike classical aggregation algorithms that focus solely on label consensus, RFAGG can achieve more effective and generalizable aggregation (Figure 1). Specifically, RFAGG formulates the aggregation task as a conditional text generation problem, where the goal is to generate an optimal aggregated annotation given an input instance and a set of LLM annotations. By leveraging reinforcement learning, RFAGG learns to generate aggregations that maximize a carefully designed aggregation reward function, which incorporates multiple facets of effective aggregation: accuracy, format consistency, and the ability to resolve complex cases with high annotator disagreement. To train and evaluate RFAGG, we construct the AGG dataset, a comprehensive collection of question-answer pairs from diverse datasets, annotated by multiple LLMs with varying personalization. The AGG dataset spans a wide range of tasks, including text classification, open-ended generation, arithmetic reasoning, and industrial visual inspection, ensuring that our evaluation captures the generalization capabilities of the aggregation methods.

Our contributions can be summarized as follows:

- We introduce RFAGG, a generalizable reinforcement learning-based framework for aggregating LLM annotations that jointly models labels and justifications, addressing the limitations of traditional aggregation methods.
- We construct the AGG dataset, which provides a rigorous benchmark for evaluating aggregation methods across diverse tasks and domains.
- We design the aggregation reward functions and a curriculum training strategy that together guide the model toward high-quality, consensus-driven aggregation by capturing the complexities of annotation justifications and progressively increasing task difficulty.
- Through extensive experiments, we demonstrate that RFAGG significantly outperforms both classical and state-of-the-art aggregation methods, achieving higher accuracy and robustness across various annotation tasks. Additionally, we show that RFAGG generalizes well to new tasks and domains, underscoring its potential as a universal aggregator.

2 RELATED WORK

Label Aggregation. Label aggregation is a critical step in handling data from multiple annotators (Zheng et al., 2017; Liu et al., 2023; Ustalov et al., 2024; Li et al., 2019b; Whitehill et al., 2009; Li et al., 2019a; Welinder et al., 2010; Zhou et al., 2012). Existing methods fall into three main categories: distance-based optimization methods, probabilistic graphical models (PGMs), and deep learning-based methods. Distance-based methods (Dawid & Skene, 1979; Aydin et al., 2014) focus on minimizing disagreement or divergence between annotator labels and estimated ground truth through iterative refinement. PGM-based methods (Li et al., 2019b;a) model the annotation process using probabilistic graphical structures that explicitly represent dependencies among workers, tasks, true labels, and latent variables. More recently, deep learning methods have emerged (Ratner et al., 2016; Liu et al., 2023), leveraging neural networks to capture complex, non-linear patterns in annotator

behavior and improve label inference accuracy. However, these methods are not designed for LLM annotators and typically ignore the rich justifications that LLMs can provide. In contrast, our method jointly models both labels and justifications, enabling richer aggregation and better utilization of LLM-generated signals.

LLM as Annotators. Recent research highlights LLMs’ potential for data annotation tasks (Yao et al., 2024). GPT-3 demonstrated competitive performance with minimal fine-tuning in various NLP tasks (Gao et al., 2020), while Brown et al. showed LLMs could reduce the need for large task-specific datasets (Brown, 2020). He et al. introduced AnnoLLM, improving annotation performance through reasoned explanations (He et al., 2024a), and found GPT-4 could outperform traditional crowdsourcing methods (He et al., 2024b). Additional studies evaluated LLMs across specific annotation contexts, including political affiliation classification (Törnberg, 2023), implicit hate speech detection (Huang et al., 2023), general NLP tasks (Ding et al., 2022), paraphrase generation (Cegin et al., 2023), and replicating crowdsourcing pipelines (Wu et al., 2023). These findings collectively demonstrate LLMs’ capabilities and limitations as data annotators across diverse NLP applications. Our framework differs by explicitly leveraging both label and justification information and by using reinforcement learning to optimize aggregation quality.

Reinforcement Learning in Large Language Models. In recent years, reinforcement learning (RL) has been central to advancing the alignment and performance of large language models (LLMs) (Cao et al., 2024), from alignment via RLHF and DPO (Christiano et al., 2017; Rafailov et al., 2023; Cao et al., 2024) to enhancing reasoning through verifiable rewards in large reasoning models (Plaat et al., 2024; Jaech et al., 2024; Guo et al., 2025). Methods like GRPO show that structured accuracy feedback can elicit planning and self-correction even before fine-tuning (Shao et al., 2024). Reinforcement learning naturally fits decision-making tasks by optimizing policies through rewards, enabling effective handling of conflicting evidence and coherent outcome generation. In contrast, supervised fine-tuning merely imitates labels without guiding reasoning. For aggregation, where models must synthesize inconsistent annotations, RL provides a clear advantage by directly rewarding accuracy and conflict resolution, yielding more robust and transferable policies.

3 METHODOLOGY

In this section, we present our RFAGG framework for dynamically aggregating LLM annotations using reinforcement learning. We first formalize the annotation aggregation task, then describe the construction of the AGG dataset, and finally detail our reinforcement learning method with specialized reward functions (Figure 2).

3.1 TASK FORMULATION

Let x be an input instance (e.g., a text passage, an image, or a multimodal input) that requires annotation. We have K LLMs (different LLMs or LLM with different personalization), denoted as $\{M_1, M_2, \dots, M_K\}$, each providing an annotation $a_k = (l_k, j_k)$ for x , where l_k is the label and j_k is the justification or explanation for the label.

The goal of annotation aggregation is to combine these K annotations into a single consensus annotation $a^* = (l^*, j^*)$ that maximizes accuracy and reliability. Traditionally, aggregation methods focus solely on the labels $\{l_1, l_2, \dots, l_K\}$ while discarding the justifications. In contrast, RFAGG jointly models both labels and justifications to extract richer signals from the annotations.

We formulate this task as a conditional text generation problem, where given an input x and a set of annotations $\{a_1, a_2, \dots, a_K\}$, the goal is to generate an optimal aggregated annotation a^* . Formally:

$$a^* = \arg \max_a P(a|x, a_1, a_2, \dots, a_K) \quad (1)$$

3.2 AGG DATASET

Although RFAGG was designed to aggregate annotations from multiple LLMs in order to produce consensus results that are both accurate and robust, we found that no existing publicly available dataset adequately supports this particular use case. Specifically, while existing benchmarks provide

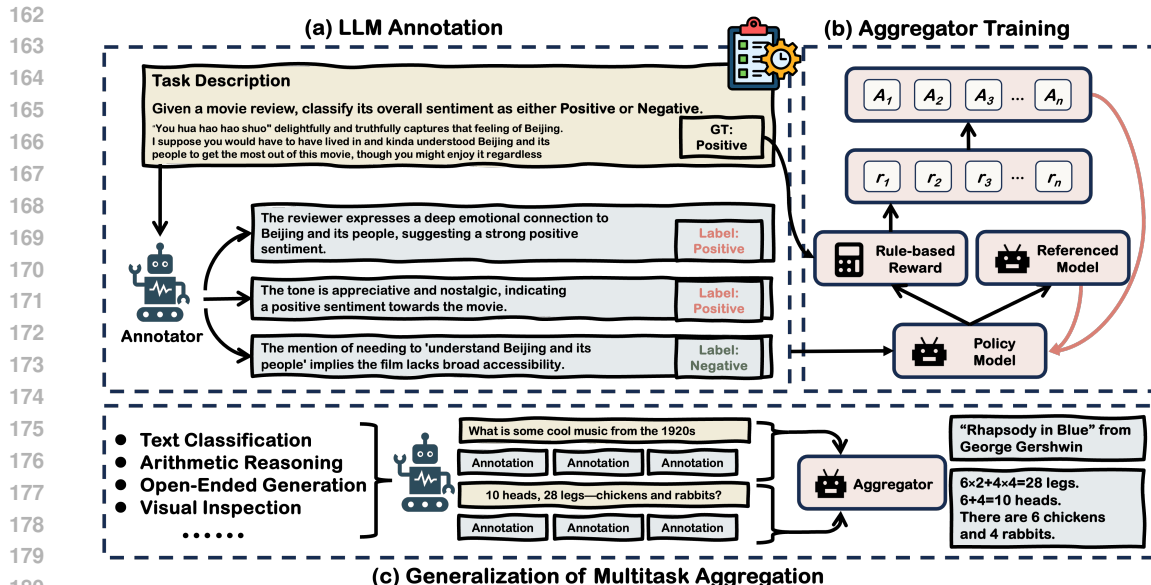


Figure 2: Framework of RFAGG , consisting of three main stages: (1) LLM-based annotation, (2) aggregator training, and (3) multitask aggregation generation.

input instances and ground-truth labels but lack the crucial component of multiple, diverse LLM-generated annotations. To address this gap and provide a comprehensive benchmark for evaluating RFAGG’s performance, we constructed a dataset called AGG.

Dataset Partition. To rigorously assess RFAGG’s aggregation capabilities and generalization, we designed a comprehensive dataset suite. This suite includes distinct partitions for training, testing, and generalization, facilitating a structured and thorough evaluation of both performance and robustness.

The training set and test set are primarily used to optimize and validate the aggregation ability of RFAGG. These sets were constructed from eight well-established benchmarks covering classification and multiple-choice tasks across diverse textual domains and difficulty levels. Specifically, these include: AGNews (Zhang et al., 2015b), DBpedia14 (Zhang et al., 2015a), Emotion (Saravia et al., 2018), IMDB (Maas et al., 2011), MultiNLI (Maas et al., 2011), SST2 (Socher et al., 2013), Trec (Li & Roth, 2002), WikiToxic (cjadams et al., 2017), and MMLU (Hendrycks et al., 2021). These datasets serve as prototypical examples of tasks where label aggregation methods are commonly employed. Moreover, they clearly defined categorical boundaries, which facilitates the derivation of precise reward signals during the reinforcement learning stages of RFAGG’s training. **The details of the datasets can be found in Appendix B.**

The generalization set, on the other hand, is designed to evaluate how well RFAGG performs when applied to tasks and modalities outside the scope of its training distribution. To achieve this, we incorporated 5 heterogeneous datasets representing a broader spectrum of tasks: text classification (BoolQ (Clark et al., 2019) and WiC (Pilehvar & Camacho-Collados, 2019)), multiple-choice science questions (ARC (Clark et al., 2018)), arithmetic reasoning (GSM8K (Cobbe et al., 2021)), open-ended generation (Alpaca (Dubois et al., 2024)), and industrial visual inspection (MVTec AD (Bergmann et al., 2019)).

LLM Annotation Generation. All annotations in the AGG dataset were generated using two variants of the QWEN2.5 series: QWEN2.5-7B-INSTRUCT (Team, 2024a) for textual tasks and QWEN2.5-VL-7B-INSTRUCT (Team, 2025) for vision-language tasks. For each instance in every dataset, the LLMs annotator generated both a predicted label l and an accompanying justification j in natural language, providing reasoning for the selected label.

To generate a large and diverse set of annotations, we employed a highly scalable strategy by applying distinct personality prompts to a single base model. This method is significantly more practical for

216 creating extensive training datasets than relying on the finite and often limited selection of available
 217 LLMs. Meanwhile, we confirm the generalizability of our method in Appendix 4.6, which shows
 218 strong performance when aggregating outputs from different LLMs.
 219

220 **Aggregation Dataset Construction.** To simulate real-world crowdsourcing scenarios, we con-
 221 structed aggregation-ready datasets through subsampling of the generated annotations. For each
 222 input instance, a random subset of annotations was selected from the available variants. This method
 223 both preserves the stochastic nature of crowd annotations through varying annotation set sizes and
 224 mitigates potential biases related to set size in comparisons of aggregation methods.
 225

226 3.3 RFAGG TRAINING FRAMEWORK

227
 228 **Aggregation Reward Function.** The heart of the RFAGG framework is our meticulously designed,
 229 multi-faceted reward function. It is engineered not just to pursue correctness, but to sculpt the
 230 aggregator’s behavior towards robust and intelligent synthesis (The MDP formulation can be found
 231 in Appendix C).

232 *Format Reward R_{format} .* Evaluates the format of the aggregation by checking whether the reasoning
 233 process and final answer are properly separated. Outputs that lack structure or fail to follow the
 234 expected format are penalized. This reward is computed using a rule-based parser, and the format
 235 reward is defined as:

$$236 R_{format}(O) = \begin{cases} 1, & \text{if } O \text{ has valid structure with labeled reasoning and answer} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

237
 238 where O denotes the aggregation model’s output. This encouraging the model to maintain a consistent
 239 and interpretable structure in its aggregations.
 240

241 *Accuracy Reward R_{acc} .* Measures the accuracy of the final aggregated answer by extracting the
 242 answer portion from the output using predefined rules, then comparing it to the gold standard label
 243 via exact match. This reward is also based on rules: let y_{pred} be the extracted answer and y^* be the
 244 gold label. Then,

$$245 R_{acc} = \mathbb{I}(y_{pred}, y^*) \quad (3)$$

246 where $\mathbb{I}(\cdot)$ is the indicator function. This reward provides the primary signal for guiding the model
 247 toward factually correct aggregations.
 248

249 *Annotation Complexity Reward $R_{complex}$.* Assigns additional reward for successfully handling high-
 250 entropy cases where annotator disagreement is substantial. The entropy of annotations $H(A)$ is
 251 calculated as:

$$252 H(A) = - \sum_{l \in L} p(l) \log p(l) \quad (4)$$

253 where L is the set of unique labels and $p(l)$ is the proportion of annotators who assigned label l .
 254 Higher entropy indicates greater disagreement, and solving such complex cases receives an additional
 255 reward scaled by the entropy:
 256

$$257 R_{complex} = R_{acc} \cdot \max(0, H(A) - \tau) \quad (5)$$

258 where τ is a threshold parameter that determines what level of entropy qualifies as a complex case.

259 This reward formulation explicitly encourages the model to: (1) maintain coherent structure in its
 260 outputs, (2) maximize predictive accuracy, (3) excel at resolving difficult cases with high annotator
 261 disagreement.
 262

263
 264 **Reinforcement Learning with Curriculum-based training.** For policy optimization, we select
 265 Group Relative Policy Optimization (GRPO) (Shao et al., 2024) due to its demonstrated stability
 266 and sample efficiency in complex text generation tasks. Its value-function-free nature reduces
 267 computational overhead and simplifies the training pipeline, allowing us to focus our efforts on
 268 the more critical aspects of our framework: the reward engineering and curriculum design. We
 269 initialize our aggregation model using a pre-trained LLM (e.g., Qwen-2.5-7B) and proceed directly
 to reinforcement learning without supervised fine-tuning. The GRPO objective function is given by:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left(\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \right. \right. \right. \quad (6)$$

$$\left. \left. \left. \text{clip} \left(\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right) \right\} - \beta \mathbb{D}_{KL}[\pi_{\theta} || \pi_{\theta_{\text{old}}}] \right],$$

where q is the query (input and annotations), $\{o_i\}_{i=1}^G$ are the generated candidate aggregations, π_{θ} is the current policy, $\pi_{\theta_{\text{old}}}$ is the old policy, $\hat{A}_{i,t}$ is the advantage, ϵ is a hyperparameter that controls the clipping and β controls the strength of the KL term.

The advantage $\hat{A}_{i,t}$ is calculated based on the relative performance within a group of samples:

$$\hat{A}_{i,t} = \frac{r_i - \mu_r}{\sigma_r}, \quad (7)$$

where r_i is the reward for candidate i , μ_r is the mean reward across all candidates in the group, and σ_r is the standard deviation of rewards within the group. This method eliminates the need for a separate value function, reducing computational overhead.

For curriculum learning, we pre-calculate the difficulty of each instance based on the number of annotations to be aggregated:

$$\text{difficulty}(x) = |A_x| \quad (8)$$

where $|A_x|$ is the number of annotations available for input instance x . This metric serves as a direct measure of aggregation complexity, as instances with more annotations require the model to reconcile a greater number of potentially conflicting viewpoints.

We partition the training dataset into subsets of increasing difficulty before training begins:

$$\mathcal{D}_k = \{(x, \{a_1, a_2, \dots, a_{|A_x|}\}, y) \in \mathcal{D} \mid |A_x| = k\} \quad (9)$$

where \mathcal{D}_k is the subset of data with exactly k annotations per instance. The curriculum progresses from datasets with fewer annotations to those with more annotations:

$$\mathcal{D}_{\text{curr}} = \mathcal{D}_{k_{\text{min}}} \cup \mathcal{D}_{k_{\text{min}}+1} \cup \dots \cup \mathcal{D}_{k_{\text{curr}}} \quad (10)$$

where k_{min} is the minimum number of annotations and k_{curr} is the current maximum number of annotations included in training.

This method ensures that the model first masters simpler aggregation problems with fewer annotations before progressing to more complex cases where it must reconcile a larger number of potentially conflicting annotations. The pre-calculation of difficulty levels based on annotation count allows for a structured curriculum without the need for dynamic sampling during training.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

We conduct all experiments using the QWEN2.5-7B-INSTRUCT model as our base model. Model training is performed on the training split of the AGG dataset, while evaluation is conducted on both the held-out test split and a dedicated generalization split to assess out-of-domain performance. This protocol enables a rigorous examination of the model’s aggregation ability as well as its robustness and transferability to novel tasks and domains.

To comprehensively evaluate RFAGG, we benchmark against three categories of baselines: (1) *Raw annotation*, where predictions from individual LLM annotators are used without any aggregation;

Table 1: Accuracy results on the AGG test set. The top section displays annotation performance generated by QWEN2.5-7B-INSTRUCT. The middle section presents the aggregation results from traditional label aggregation methods. The bottom section reports the performance of aggregation using LLMs. **F1 score can be found in Appendix E**

Dataset	AGNews	DBpedia14	Emotion	IMDB	MultiNLI	SST2	TREC	WikiToxic	MMLU
ANNOTATION	80.0	94.8	58.0	95.0	82.8	93.1	81.4	78.8	72.0
MV	80.7	96.5	59.4	95.5	85.1	93.8	84.8	79.8	75.0
PM	80.3	96.4	58.8	95.0	85.5	93.6	83.6	81.6	72.0
DS	81.7	96.8	59.0	95.3	85.3	93.9	87.4	78.5	76.0
CATD	80.7	96.4	59.2	95.2	85.8	93.6	83.4	81.1	75.0
BWA	80.6	96.8	59.3	95.1	85.6	93.6	85.6	81.2	76.0
IBCC	81.6	96.8	59.1	95.3	85.5	93.9	86.8	78.5	76.0
EBCC	80.1	96.8	58.3	95.2	85.1	93.9	87.0	79.0	76.0
QWEN2.5-7B	81.3	96.7	57.7	95.4	85.5	94.1	85.4	80.9	72.0
RFAGG	89.2	98.3	65.0	96.0	86.8	96.1	91.8	85.7	78.0

(2) *Traditional label aggregation methods*, including Majority Voting (MV), Probabilistic Model (PM) (Aydin et al., 2014), Dawid-Skene (DS) (Dawid & Skene, 1979), Confidence-Aware Truth Discovery (CATD) (Li et al., 2014), Bayesian Weighted Averaging (BWA) (Li et al., 2019a), Independent Bayesian Classifier Combination (IBCC), and Enhanced Bayesian Classifier Combination (EBCC) (Li et al., 2019b), which are applied to classification datasets; and (3) *Pretrained LLM aggregation*, where the QWEN2.5-7B-INSTRUCT model is used as an aggregator without any additional tuning.

4.2 AGG TEST SET PERFORMANCE ANALYSIS

We first evaluate the effectiveness of aggregation by comparing all methods with raw LLM annotations across the AGG test set (Table 1). Across most datasets, we observe that aggregation methods consistently yield substantial performance gains over using unaggregated annotations, confirming the value of synthesizing multiple LLM-generated responses. Among all evaluated methods, our proposed RFAGG achieves the best performance on every dataset, surpassing all baselines.

Among the baselines, we observe that simple majority voting (MV) provides a strong starting point, consistently improving over individual annotations. However, more sophisticated statistical aggregation methods such as PM, DS, and EBCC do not reliably outperform MV. This finding suggests traditional aggregation methods are less effective with LLM annotators due to their highly correlated outputs. Unlike human-generated data, LLM annotations don’t exhibit the independent reliability variations that these aggregation techniques were designed to exploit.

We further assess the use of untuned LLMs as aggregators (QWEN2.5-7B-INSTRUCT), and observe that their aggregation performance generally falls short of MV, underscoring the limitations of relying on general-purpose LLM reasoning for this specialized task. In contrast, our proposed RFAGG achieves the best performance across all datasets, surpassing both traditional and LLM-based baselines. These results demonstrate that our reinforcement learning framework, which explicitly models both label and justification information, is highly effective at extracting consensus from LLM-generated annotations and robustly improves aggregation quality.

4.3 GENERALIZATION PERFORMANCE ANALYSIS

In this section, we evaluate RFAGG on the AGG generalization set to rigorously assess its out-of-domain generalization capabilities. Crucially, none of the tasks in this evaluation set are present in the training data, and several tasks feature fundamentally different characteristics (e.g., open-ended questions instead of classification tasks).

Generalization on text classification. Table 2 presents results on the classification datasets BoolQ and WiC. Across both benchmarks, RFAGG outperforms all other aggregation methods on both datasets, achieving the highest accuracy across these benchmarks. These results demonstrate the strong generalization ability of our framework to classification tasks not seen during training.

Table 2: Generalization accuracy comparison on classification dataset.

Dataset	BoolQ	WiC
ANNOTATION	82.2	66.4
MV	83.3	69.2
PM	83.6	65.6
DS	84.0	69.2
CATD	83.1	66.8
BWA	83.3	68.1
IBCC	84.0	69.4
EBCC	83.0	69.1
QWEN2.5-7B	83.3	68.6
RFAGG	86.8	70.6

Table 3: Generalization accuracy comparison on reasoning and math datasets.

Dataset	ARC	GSM8K
ANNOTATION	88.7	72.8
MV	90.8	86.0
PM	90.8	-
DS	90.8	-
CATD	90.8	-
BWA	90.8	-
IBCC	90.8	-
EBCC	90.8	-
QWEN2.5-7B	88.0	87.0
RFAGG	91.8	88.0

Table 4: Generalization performance result on AlpacaEval 2. The top section shows results from state-of-the-art LLMs, obtained from the official leaderboard. The middle presents annotations from QWEN2.5-7B-INSTRUCT and the bottom section compares LLMs aggregation.

Metrics	LC Win Rate(%)	Win Rate(%)
CLAUDE 3 OPUS (02/29)	40.5	29.1
LLAMA 3.1 405B INSTRUCT	39.3	39.1
GPT-4	38.1	23.6
QWEN2 72B INSTRUCT	38.1	29.9
LLAMA-3.1-70B-INSTRUCT-TURBO	38.1	29.9
QWEN1.5 72B CHAT	36.6	26.5
CLAUDE 3 SONNET (02/29)	34.9	25.6
LLAMA-3-70B-INSTRUCT	34.4	33.2
MISTRAL LARGE (24/02)	32.7	21.4
MIXTRAL 8X22B v0.1	30.9	22.2
ANNOTATION	28.1	31.0
QWEN2.5-7B	35.9	52.9
RFAGG	41.1	55.3

Generalization on reasoning and math. Table 3 shows results on a multiple-choice QA reasoning dataset (ARC) and a math problem-solving dataset (GSM8K). RFAGG consistently achieves the best performance across both benchmarks, underscoring its robustness and stability across tasks of varying domains and complexity. Notably, traditional methods like PM and DS are unsuitable for GSM8K due to their reliance on categorical outputs, underscoring the flexibility of RFAGG in handling diverse output formats within a unified aggregation framework. On the ARC multiple-choice task, many traditional aggregation methods (PM, DS, CATD, etc.) do not improve upon a simple majority vote. This is because, with a limited number of annotators exhibiting similar performance, these models struggle to estimate meaningful differences in reliability, often causing their predictions to converge with the majority opinion.

Generalization on open-ended tasks. We further assess RFAGG on open-ended tasks using AlpacaEval 2, with results summarized in Table 4. The results first highlight the inherent value of aggregation for open-ended tasks; the untuned QWEN2.5-7B model, when used as a simple aggregator, already improves performance over the raw annotations. Crucially, RFAGG outperforms this base aggregator, achieving a 5.2% gain in LC Win Rate and a 2.4% improvement in Win Rate. Compared to annotations generated by state-of-the-art LLMs, our method surpasses state-of-the-art LLMs including CLAUDE 3 OPUS, GPT-4, and LLAMA 3 405B INSTRUCT in both metrics, demonstrating strong generalization in open-ended scenarios, along with significantly enhanced aggregation capability.

Table 5: Generalization accuracy result on DS-MVTeC.

Subtask	Anomaly	Defect				Object	
	Detection	Analysis	Classification	Description	Localization	Analysis	Classification
ANNOTATION	77.0	81.1	63.3	66.2	65.2	85.8	92.7
MV	80.4	87.9	69.4	73.4	72.0	91.1	98.0
PM	79.8	83.4	67.9	68.0	67.7	87.8	97.1
DS	70.6	89.5	71.1	73.4	72.0	91.3	98.0
CATD	80.6	88.5	70.2	73.4	72.2	91.1	97.8
BWA	80.4	88.2	70.2	73.4	71.8	91.2	98.0
IBCC	81.1	89.6	70.2	73.4	72.2	91.3	98.0
EBCC	81.4	88.1	69.4	72.8	71.5	91.3	98.0
QWEN2.5-7B	82.5	90.9	70.1	73.8	70.3	91.3	93.5
RFAGG	78.0	91.9	68.6	74.6	72.2	91.8	96.1

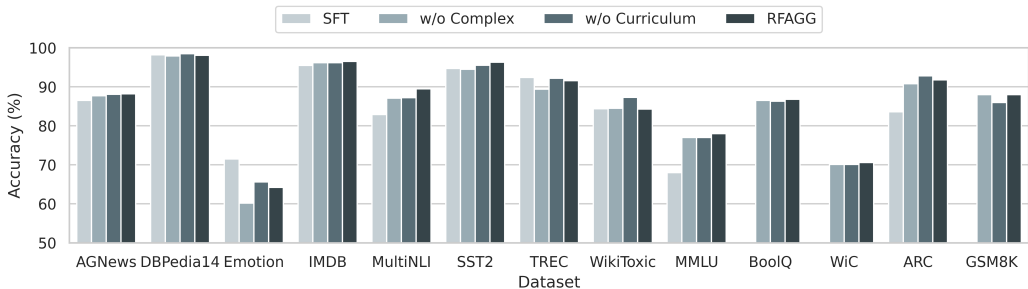


Figure 3: Ablation study results on classification, multiple-choice, and arithmetic reasoning datasets from both test and generalization splits.

Generalization on industrial visual inspection. Finally, Table 5 shows that RFAGG achieves the highest accuracy on four out of seven DS-MVTeC tasks. While traditional methods like CATD and IBCC perform competitively in this domain, RFAGG delivers superior overall performance, demonstrating strong generalization in real-world industrial applications.

Taken together, these results demonstrate the remarkable generalization ability of RFAGG. For a wide range of tasks, including text classification, complex reasoning, open-ended generation, and visual inspection, RFAGG consistently outperforms both traditional statistical aggregation methods and untuned LLM-based aggregators. This strong out-of-domain performance confirms that RFAGG can serve as a unified and adaptable aggregation solution for diverse LLM annotation scenarios, accommodating various data modalities, task formats, and levels of annotation heterogeneity.

4.4 ABLATION STUDY

To better understand the contributions of each component in our framework, we conduct a series of ablation studies by evaluating three distinct variants of RFAGG. The first variant, *SFT*, replaces our reinforcement learning objective with supervised fine-tuning directly on ground-truth labels, aligning the model through standard supervised signals. The second variant, *w/o Complex*, removes the entropy-based component from the reward function $R_{complex}$, eliminating the mechanism that prioritizes high-disagreement instances. The third variant, *w/o Curriculum*, disables the curriculum learning strategy, training the model without progressively increasing task complexity.

The results, summarized in Figure 3, highlight the importance of each component. While the *SFT* variant achieves strong performance on the test set, it fails to generalize and yields zero accuracy on datasets requiring outputs that differ from those seen during training, providing strong empirical validation for our core premise: the task of generating reasoned aggregations is fundamentally ill-suited for SFT and necessitates the RL paradigm we propose. Both the *w/o Complex* and *w/o Curriculum* variants remain competitive on certain benchmarks but exhibit diminished overall and out-of-domain performance compared to the full model. These findings demonstrate that the entropy-

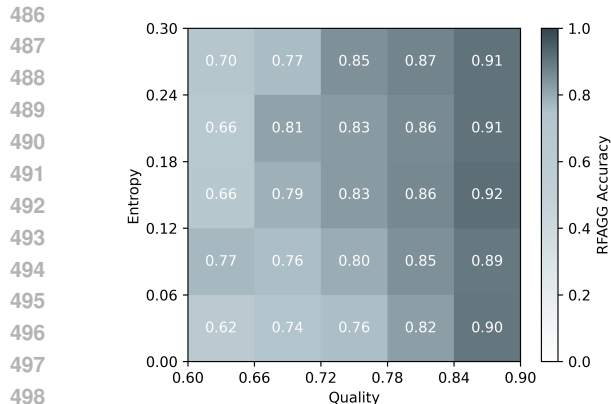


Table 6: Aggregation performance using annotations from diverse LLMs

Method	Accuracy
Annotation	65.0
MV	74.0
DS	73.0
PM	73.0
CATD	64.0
BWA	74.0
IBCC	74.0
EBCC	72.0
Qwen2.5-7B-Instruct	75.0
RFAGG	77.0

Figure 4: Aggregation performance related to annotation quality and diversity.

based reward and curriculum learning are crucial for improving both aggregation quality and the generalization ability of RFAGG, and that each component plays a complementary role in the robustness of our method.

4.5 DIVERSITY AND QUALITY ANALYSIS

Figure 4 analyzes how aggregation performance is influenced by both the quality and diversity of annotation outputs, evaluated across all annotators combinations. The x-axis denotes average answer accuracy (quality), and the y-axis indicates normalized label entropy (diversity). Each cell shows the mean aggregation accuracy. Our analysis shows that higher annotation quality and increased diversity both contribute to better aggregation performance. This finding aligns with the conclusion of (Kuncheva & Whitaker, 2003) that ensembles with diversity above a certain threshold offer guaranteed improvements over individual model. In the context of LLM annotation aggregation, this suggests that when LLMs generate a heterogeneous range of responses, our framework is better equipped to reconcile disagreements and synthesize a more accurate and robust consensus label.

4.6 GENERALIZATION ON DIFFERENT LLMs

To evaluate the robustness and generalization of our aggregation method, we construct a diverse and challenging testbed using annotations from ten distinct LLMs: including Daredevil-8B (Dar), Falcon3-7B-Instruct (Team, 2024b), Hermes-2-Theta-Llama-3-8B (Teknum et al.), Ministral-8B-Instruct-2410 (Min), OLMo-2-1124-7B-Instruct (OLMo et al., 2024), Qwen2.5-7B-Instruct (Team, 2024a), Yi-1.5-9B-Chat (Young et al., 2024), c4ai-command-r7b-12-2024 (Cohere et al., 2025), deepseek-llm-7b-chat (DeepSeek-AI, 2024), and openchat-3.6-8b-20240522 (Wang et al., 2023). All models generate annotation answers on the MMLU dataset.

As shown in Table 6, RFAGG achieves the highest accuracy, outperforming both classical aggregation rules and direct LLM-based aggregation. This demonstrates that RFAGG effectively leverages heterogeneous annotations and generalizes well across diverse model behaviors.

5 CONCLUSION

We introduced RFAGG, a reinforcement learning framework for aggregating LLM annotations by modeling both labels and justifications. To enable robust evaluation, we constructed the AGG dataset featuring diverse, quality-controlled question-answer pairs from multiple LLMs across various tasks. Experiments show that RFAGG consistently outperforms classical and recent aggregation methods, and generalizes well to new domains, highlighting its promise as a universal annotation aggregator.

540 ETHICS STATEMENT

541

542 The datasets used in this work are derived from publicly available sources and constructed using open
543 large language models. No private or sensitive data is involved.

544

545 REPRODUCIBILITY STATEMENT

546

547 To ensure full reproducibility, we will release both the datasets and trained models upon publication.
548

549

550 REFERENCES

551

552 Daredevil-8b. URL <https://huggingface.co/mlabonne/Daredevil-8B>] (<https://huggingface.co/mlabonne/Daredevil-8B>)).
553

554 URL [<https://huggingface.co/mistralai/Mistral-8B-Instruct-2410>]
555 (<https://huggingface.co/mistralai/Mistral-8B-Instruct-2410>).

556 URL [<https://huggingface.co/google/gemma-3-12b-it>] (<https://huggingface.co/google/gemma-3-12b-it>).
557
558

559 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
560 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.
561 *arXiv preprint arXiv:2303.08774*, 2023.

562 Bahadir Aydin, Yavuz Selim Yilmaz Yavuz Selim Yilmaz, Yaliang Li, Qi Li, Jing Gao, and Murat
563 Demirbas. Crowdsourcing for multiple-choice question answering. In *Proceedings of the AAAI*
564 *Conference on Artificial Intelligence*, volume 28, pp. 2946–2953, 2014.

565 Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive
566 real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference*
567 *on computer vision and pattern recognition*, pp. 9592–9600, 2019.
568

569 Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
570

571 Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua
572 Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning:
573 Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*,
574 2024.

575 Jan Cegin, Jakub Simko, and Peter Brusilovsky. Chatgpt to replace crowdsourcing of paraphrases
576 for intent classification: Higher diversity and comparable model robustness. *arXiv preprint*
577 *arXiv:2305.12947*, 2023.

578 Justin Chih-Yao Chen, Sukwon Yun, Elias Stengel-Eskin, Tianlong Chen, and Mohit Bansal. Sym-
579 bolic mixture-of-experts: Adaptive skill-based routing for heterogeneous reasoning. *arXiv preprint*
580 *arXiv:2503.05641*, 2025.

581 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep
582 reinforcement learning from human preferences. *Advances in neural information processing*
583 *systems*, 30, 2017.
584

585 cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukier-
586 ski. Toxic comment classification challenge. [https://kaggle.com/competitions/](https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge)
587 [jigsaw-toxic-comment-classification-challenge](https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge), 2017. Kaggle.

588 Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina
589 Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*,
590 2019.
591

592 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and
593 Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge.
arXiv:1803.05457v1, 2018.

- 594 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
595 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
596 Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,
597 2021.
- 598
- 599 Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammam, Yazeed Alnumay, Sophia
600 Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara
601 Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia,
602 Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt
603 Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander
604 Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash
605 Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier,
606 Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru,
607 Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas,
608 Ben Cyrus, Daniel D’souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche
609 Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit
610 Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün,
611 Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias
612 Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Seraphina
613 Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan,
614 Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian
615 Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom
616 Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline
617 Jenkins, JJ Jordan, Dhruvi Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia
618 Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov,
619 Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia
620 Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin
621 Luong, Raymond Ma, Lukas Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez,
622 Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic
623 McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian
624 Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham
625 Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu,
626 Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus,
627 Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan,
628 Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Laura Ruis, Louise
629 Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam,
630 Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne,
631 Sylvie Chang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk,
632 Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina,
633 Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone,
634 Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat
635 Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White,
636 Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern,
637 Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language
638 model, 2025. URL <https://arxiv.org/abs/2504.00698>.
- 639
- 640 Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates
641 using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*,
642 1979.
- 643 DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint*
644 *arXiv:2401.02954*, 2024. URL <https://github.com/deepseek-ai/DeepSeek-LLM>.
- 645
- 646 Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing.
647 Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*, 2022.
- 648
- 649 Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving
650 factuality and reasoning in language models through multiagent debate. In *Forty-first International
651 Conference on Machine Learning*, 2023.

- 648 Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled
649 alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
650
- 651 Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and
652 Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In
653 *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.
654 6491–6501, 2024.
- 655 Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot
656 learners. *arXiv preprint arXiv:2012.15723*, 2020.
657
- 658 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu
659 Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement
660 learning. *Nature*, 645(8081):633–638, 2025.
- 661 Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu,
662 Nan Duan, and Weizhu Chen. AnnoLLM: Making large language models to be better crowdsourced
663 annotators. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar (eds.), *Proceedings of the 2024
664 Conference of the North American Chapter of the Association for Computational Linguistics:
665 Human Language Technologies (Volume 6: Industry Track)*, pp. 165–190, Mexico City, Mexico,
666 June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-industry.15.
667 URL <https://aclanthology.org/2024.naacl-industry.15>.
- 668 Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth
669 Huang. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the CHI
670 Conference on Human Factors in Computing Systems*, pp. 1–25, 2024b.
671
- 672 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
673 Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International
674 Conference on Learning Representations (ICLR)*, 2021.
- 675 Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and
676 limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM
677 web conference 2023*, pp. 294–297, 2023.
678
- 679 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
680 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint
681 arXiv:2412.16720*, 2024.
- 682 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
683 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,
684 L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas
685 Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023. URL [https://arxiv.
686 org/abs/2310.06825](https://arxiv.org/abs/2310.06825).
- 687
- 688 Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang,
689 and Feng Zheng. Mmad: The first-ever comprehensive benchmark for multimodal large language
690 models in industrial anomaly detection. In *The Thirteenth International Conference on Learning
691 Representations*, 2025. URL <https://openreview.net/forum?id=JDIER86r8v>.
- 692 Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models.
693 *arXiv preprint arXiv:2506.07962*, 2025.
694
- 695 Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and
696 their relationship with the ensemble accuracy. *Machine learning*, 51:181–207, 2003.
- 697 Jiyi Li. A comparative study on annotation quality of crowdsourcing and llm via label aggregation.
698 In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing
699 (ICASSP)*, pp. 6525–6529. IEEE, 2024a.
- 700
- 701 Jiyi Li. Human-llm hybrid text answer aggregation for crowd annotations. *arXiv preprint
arXiv:2410.17099*, 2024b.

- 702 Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-
703 aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4):
704 425–436, 2014.
- 705
- 706 Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Confer-*
707 *ence on Computational Linguistics*, 2002. URL [https://www.aclweb.org/anthology/](https://www.aclweb.org/anthology/C02-1150)
708 C02-1150.
- 709
- 710 Yuan Li, Benjamin IP Rubinstein, and Trevor Cohn. Truth inference at scale: A bayesian model for
711 adjudicating highly redundant crowd annotations. In *The World Wide Web Conference (WWW)*,
712 2019a.
- 713
- 714 Yuan Li, Benjamin Rubinstein, and Trevor Cohn. Exploiting worker correlation for label aggregation
715 in crowdsourcing. In *International conference on machine learning (ICML)*, 2019b.
- 716
- 717 Jiacheng Liu, Feilong Tang, and Xiaofeng Hou. Label aggregation with self-supervision enhanced
718 graph transformer. *European Conference on Artificial Intelligence (ECAI)*, 2023.
- 719
- 720 Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher
721 Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting*
722 *of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150,
723 Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL [http:](http://www.aclweb.org/anthology/P11-1015)
724 [://www.aclweb.org/anthology/P11-1015](http://www.aclweb.org/anthology/P11-1015).
- 725
- 726 Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia,
727 Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira
728 Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri,
729 Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill,
730 Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman
731 Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael
732 Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious.
733 2024. URL <https://arxiv.org/abs/2501.00656>.
- 734
- 735 Mohammad Taher Pilehvar and Jose Camacho-Collados. Wic: the word-in-context dataset for
736 evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of*
737 *the North American Chapter of the Association for Computational Linguistics: Human Language*
738 *Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273, 2019.
- 739
- 740 Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Bäck.
741 Reasoning with large language models, a survey. *CoRR*, 2024.
- 742
- 743 Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin.
744 tinybenchmarks: evaluating llms with fewer examples. 2024.
- 745
- 746 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
747 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*
748 *in neural information processing systems*, 36:53728–53741, 2023.
- 749
- 750 Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data
751 programming: Creating large training sets, quickly. *Advances in neural information processing*
752 *systems*, 29, 2016.
- 753
- 754 Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Context-
755 tualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on*
Empirical Methods in Natural Language Processing, pp. 3687–3697, Brussels, Belgium, October-
November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL
<https://www.aclweb.org/anthology/D18-1404>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical
reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

- 756 Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and
757 Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank.
758 In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp.
759 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
760 URL <https://www.aclweb.org/anthology/D13-1170>.
- 761 Qwen Team. Qwen2.5: A party of foundation models, September 2024a. URL [https://qwenlm.
762 github.io/blog/qwen2.5/](https://qwenlm.github.io/blog/qwen2.5/).
- 763
764 Qwen Team. Qwen2.5-vl, January 2025. URL [https://qwenlm.github.io/blog/qwen2.
765 5-vl/](https://qwenlm.github.io/blog/qwen2.5-vl/).
- 766 TII Team. The falcon 3 family of open models, December 2024b.
- 767
768 Teknum, Goddard Charles, interstellarninja, theemozilla, karan4d, and art huemin. Hermes-2-theta-
769 llama-3-8b.
- 770 Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter
771 messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.
- 772
773 Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. Learning from Crowds with Crowd-Kit.
774 *Journal of Open Source Software*, 9(96):6227, 2024. ISSN 2475-9066. doi: 10.21105/joss.06227.
- 775
776 Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Ad-
777 vancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*,
778 2023.
- 779 Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm
780 collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI
781 Conference on Human Factors in Computing Systems*, pp. 1–21, 2024.
- 782
783 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-
784 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.
785 *arXiv preprint arXiv:2203.11171*, 2022.
- 786
787 Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of
788 crowds. *Advances in neural information processing systems (NIPS)*, 2010.
- 789
790 Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should
791 count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural
792 information processing systems (NIPS)*, 2009.
- 793
794 Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi
795 Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. Llms as workers in human-computational
796 algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168*,
797 2023.
- 798
799 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
800 Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint
801 arXiv:2412.15115*, 2024.
- 802
803 Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large
804 language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence
805 Computing*, pp. 100211, 2024.
- 806
807 Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng
808 Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint
809 arXiv:2403.04652*, 2024.
- 806 Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text
807 classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.),
808 *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.,
809 2015a. URL [https://proceedings.neurips.cc/paper_files/paper/2015/
file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf).

810 Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text
811 classification. *Advances in neural information processing systems*, 28, 2015b.
812

813 Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in
814 crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552,
815 2017.

816 Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by
817 minimax entropy. *Advances in neural information processing systems (NIPS)*, 2012.
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

864	APPENDIX	
865		
866	A The Use of Large Language Models (LLMs)	S2
867		
868	B AGG Dataset	S2
869		
870		
871	C MDP Formulation	S2
872		
873	D Macro F1 Score for Classification	S3
874		
875	E Additional Results	S3
876		
877		
878	F Case Study	S4
879		
880	G Prompt	S8
881	G.1 Aggregation Prompts	S8
882	G.2 Prompt Variants	S9
883		
884		
885		
886		
887		
888		
889		
890		
891		
892		
893		
894		
895		
896		
897		
898		
899		
900		
901		
902		
903		
904		
905		
906		
907		
908		
909		
910		
911		
912		
913		
914		
915		
916		
917		

A THE USE OF LARGE LANGUAGE MODELS (LLMs)

During the writing process, we utilized LLMs, specifically GPT-5, to refine the manuscript’s language for clarity and fluency. The authors retained full responsibility for all content, with the LLMs serving exclusively as a tool for language enhancement.

B AGG DATASET

Table 1 details the data distribution across splits. For the MMLU, AI2ARC, and GSM8K datasets, test sets are constructed following the sampling protocol introduced by Polo et al. (2024). For the DS-MVTEC visual inspection pairs, we adopt the question–ground truth annotation framework introduced in Jiang et al. (2025).

Table 1: Number of samples in the training, test and generalization set.

Dataset		#Train	#Test
	AGNews	2,000	1,000
	DBPedia14	2,000	1,000
	Emotion	2,000	1,000
	IMDB	2,000	1,000
	MultiNLI	2,000	1,000
	SST2	2,000	872
	TREC	2,000	500
	WikiToxic	2,000	1,000
	MMLU	10,000	100
	BoolQ	-	1,000
	WiC	-	1,000
	AI2ARC	-	100
	GSM8K	-	100
	AlpacaEval 2	-	805
	Anomaly	Detection	- 1,691
		Analysis	- 1,205
		Classification	- 1,205
		Description	- 1,212
		Localization	- 1,193
	Object	Analysis	- 1,367
		Classification	- 464

C MDP FORMULATION

The aggregation task is formulated as a single-episode MDP for each input instance.

- State (s_t): The state at timestep t consists of the input instance x , the full set of K annotations a_k , and the sequence of tokens generated so far for the aggregated answer, $y_{<t}$.
- Action (a_t): The action is to select the next token y_t from the model’s vocabulary.
- Transition: The transition is deterministic: generating token y_t moves the agent to the next state where the generated sequence is $y_{\leq t}$.
- Reward (R): The reward is sparse and is calculated only at the end of the generation episode (when an end-of-sequence token is produced). The total reward R is the sum of the components defined in Section 3.3.1: $R = R_{format} + R_{acc} + R_{complex}$. This formulation treats the entire generation of an aggregated response as a single trajectory that receives a terminal reward.

D MACRO F1 SCORE FOR CLASSIFICATION

Table 2: Macro F1 Score on classification datasets

Dataset	AGNews	DBPedia14	Emotion	IMDB	MultiNLI	SST2	TREC	WikiToxic	MMLU	BoolQ	WiC
ANNOTATION	79.9	94.7	51.4	94.9	82.5	93.1	80.9	67.3	71.9	81.9	65.7
MV	80.6	96.4	52.7	95.4	84.8	93.8	86.2	68.0	74.6	83.0	68.5
PM	83.3	96.2	51.3	94.9	85.3	93.6	84.1	69.7	71.4	83.3	63.5
DS	81.7	96.7	51.9	95.2	85.1	93.9	88.7	67.0	75.8	83.7	68.7
CATD	80.6	96.2	52.2	95.1	85.5	93.6	83.9	69.2	74.8	82.8	65.9
BWA	80.5	96.7	52.3	95.2	85.3	93.6	87.0	69.5	75.8	83.0	67.1
IBCC	81.6	96.7	52.1	95.2	85.3	93.9	88.1	67.0	75.8	83.7	68.9
EBCC	79.9	96.7	51.8	95.1	84.9	93.9	88.3	67.5	75.8	82.7	68.3
QWEN2.5-7B	81.0	96.6	50.9	95.3	85.4	94.1	85.3	69.2	71.5	82.9	68.6
RFAGG	88.3	98.1	55.5	96.4	89.2	96.3	92.4	73.2	77.8	86.3	69.3

E ADDITIONAL RESULTS

Comparison with Different LLMs for Aggregation. We evaluate the effectiveness of various large language models as aggregation modules, selecting models with different architectures and parameter scales: QWEN2.5-7B, QWEN2.5-14B (Team, 2024a), MISTRAL-7B (Jiang et al., 2023), and GEMMA-3-12B (gem). As shown in Table 3, while most LLM-based aggregation methods improve over direct annotation, model size or general reasoning capability does not directly correlate with aggregation performance. Notably, the 14B-parameter QWEN2.5 underperforms its 7B variant on multiple datasets, suggesting that larger models do not necessarily yield better consensus. Similarly, MISTRAL-7B and GEMMA-3-12B show inconsistent gains, highlighting the instability of off-the-shelf LLMs in aggregation tasks. In contrast, our method, RFAGG, achieves consistent and significant improvements across all datasets, demonstrating that effective aggregation requires more than raw model capacity. This result validates the necessity and novelty of our approach.

Comparison with Post-LLM Strategies. We further compare our method with several state-of-the-art post-hoc multi-model reasoning techniques: SELF CONSISTENCY (Wang et al., 2022), DEBATE STYLE (Du et al., 2023), and LLM-AS-JUDGE (Dubois et al., 2024). These methods represent popular paradigms for refining model outputs through internal reasoning or inter-model interaction. As summarized in Table 4, while these strategies offer modest improvements, their performance varies significantly across datasets. In contrast, RFAGG outperforms all baselines on every dataset, achieving robust and consistent results. This demonstrates the superiority of our designed aggregation mechanism in fusing diverse predictions reliably and effectively, without relying on unstable dialogue dynamics or generic judgment prompts.

Table 3: Accuracy on classification datasets from multiple LLMs aggregation

Dataset	AGNews	DBPedia14	Emotion	IMDB	MultiNLI	SST2	TREC	WikiToxic	MMLU	BoolQ	WiC
ANNOTATION	80.0	94.8	58.0	95.0	82.8	93.1	81.4	78.8	72.0	82.2	66.4
QWEN2.5-7B	81.3	96.7	57.7	95.4	85.5	94.1	85.4	80.9	72.0	83.3	68.6
QWEN2.5-14B	77.4	87.6	55.6	93.5	84.1	90.9	82.4	81.4	78.0	82.7	69.1
MISTRAL-7B	68.1	78.4	46.8	76.6	60.6	73.9	64.8	59.3	32.0	66.2	44.9
GEMMA-3-12B	80.8	96.7	58.3	95.5	86.2	93.4	85.8	80.2	76.0	84.6	68.1
RFAGG	89.2	98.3	65.0	96.0	86.8	96.1	91.8	85.7	78.0	86.8	70.6

Table 4: Accuracy on classification datasets from multiple post-LLM methods

Dataset	AGNews	DBPedia14	Emotion	IMDB	MultiNLI	SST2	TREC	WikiToxic	MMLU	BoolQ	WiC
SELF CONSISTENCY	80.7	96.5	59.4	95.5	85.1	93.8	84.8	79.8	75.0	83.3	69.2
DEBATE STYLE	82.8	97.3	55.8	95.3	78.2	94.3	84.0	87.6	84.5	67.0	62.6
LLM-AS-JUDGE	81.0	96.1	55.5	95.2	80.5	94.4	81.4	82.7	74.0	84.0	60.0
RFAGG	89.2	98.3	65.0	96.0	86.8	96.1	91.8	85.7	78.0	86.8	70.6

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

Table 6: Case study from MMLU

Query	Including the bottom how many sides are on a square-based pyramid? A. three B. four C. five D. six
Annotation 1	Justification: The question asks about the sides of a square-based pyramid, including the base. A pyramid with a square base has four triangular sides plus the square base itself, making a total of five sides. This can be visualized or sketched to ensure accuracy. Label: C
Annotation 2	Justification: A square-based pyramid has a square base, which contributes 4 sides. The pyramid also has 4 triangular faces connecting the base to the apex. Thus, the total number of sides is the 4 sides of the base plus the 4 triangular sides, making 8 sides in total, but the question specifically asks for the sides including the base, which is 4. Label: B
Annotation 3	Justification: A square-based pyramid has a square base and four triangular sides. Therefore, it has a total of 5 sides including the base. Label: C
Annotation 4	Justification: A square-based pyramid has a square base, which contributes 4 sides. Additionally, it has 4 triangular sides meeting at a common vertex. Therefore, the total number of sides is the 4 base sides plus the 4 triangular sides, totaling 5 sides. - Option A (three) is incorrect as it does not account for all sides. - Option B (four) is incorrect since it only counts the base sides. - Option D (six) is incorrect because a pyramid does not have 6 sides. - Option C (five) is the correct total as it includes all 4 base sides and the 1 apex side created by the triangles. Label: C
Aggregation	Justification: A square-based pyramid has a square base (4 sides) and 4 triangular sides. Therefore, it has $4 + 1$ base = 5 sides in total. Label: C

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

Table 7: Case study 1 from AlpacaEval 2


Query	Mike’s mother had four kids. Three of them are named Luis, Drake, and Matilda. What is the name of the fourth kid?
Annotation 1	The name of the fourth kid is Mike. The riddle states that Mike’s mother had four kids, and three of their names are given as Luis, Drake, and Matilda. Since the question asks for the name of the fourth kid, and we are told that one of the children is Mike, it follows that Mike is the fourth child.
Annotation 2	Based on the information provided, the names of three of the four kids are Luis, Drake, and Matilda. The question does not provide the name of the fourth child, so the name of the fourth kid is not specified in the given information.
Annotation 3	The name of the fourth kid is Mike. Here is the logic breakdown: 1. The question states that Mike’s mother had four kids. 2. It provides the names of three of the kids: Luis, Drake, and Matilda. 3. The question asks for the name of the fourth kid. Since the question specifies that Mike’s mother had four kids and mentions three of them by name, it is clear that the fourth kid referred to is Mike himself.
Aggregation	The name of the fourth kid is Mike. Here is the reasoning: 1. The question states that Mike’s mother had four kids. 2. It provides the names of three of the kids: Luis, Drake, and Matilda. 3. Since the question is asking for the name of the fourth kid, it logically follows that the name of the fourth kid is Mike, as Mike is explicitly mentioned to be one of the children. Therefore, the fourth kid is named Mike.

Table 8: Case study 2 from AlpacaEval 2

Query	For answering the chat, think of a funny response you can use. + how are you?
Annotation 1	How am I? I’m doing a barrel roll in my imagination right now! How about you? Are you ready to dance the waddle while sipping your epoxy smoothie?
Annotation 2	I’m doing a barrel roll! Because even robots need to show off sometimes!
Annotation 3	Well, I’m not actually feeling a whole lot because I’m just a computer program, but thank you for asking! How about you, are you having a comedic adventure today?
Annotation 4	How am I? I’m wrestling with a wormhole of inner thoughts and trying to squeeze them all into a single response. How about you? Are you navigating the complexities of your day with the grace of a professional tightrope walker, or are you more of a safety harness kind of person?
Aggregation	How am I? I’m doing a barrel roll in my code and cracking jokes with myself—do you know why a computer never goes to the beach? Because it has too many bytes to sand! How about you, are you riding the storm of your day or just feeling like a calm wave?

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

Table 9: Case study from DS-MVTec

Query	<p>There is a defect in the object. How many areas of defect are present on the fabric border of the zipper?"</p> <p>A. One. B. Four. C. Two. D. Three.</p> 
Annotation 1	<p>Justification: Upon sizes and Comparing the query image with the reference image, there doesn't appear to be a consistent pattern that would suggest a single defect across both pieces. The edge on the left side of the query image seems less uniform but not clearly specified as a defined defect.</p> <p>Label: A</p>
Annotation 2	<p>Justification: Upon comparing, I notice that one area where there appears to be a break or missing section along the top edge of the fabric near the slider indicates a defect. This is consistent across both images.</p> <p>Label: C</p>
Annotation 3	<p>Justification: The main aspects are to focus on the fabric border conditions, in order to determine if there is any defect present. There seems to be a defect at the bottom left corner of the fabrics on both sides of the zipper, indicating two distinct defects. For a closer look, one may also focus on continuity of features, like the zipper teeth as they may indicate better continuity when not defective.</p> <p>Label: C</p>
Aggregation	<p>Justification: Looking carefully at the image, there are two clearly visible defects on the fabric border of the zipper. One defect is at the bottom left and another is at the top right, near the slider.</p> <p>Label: C</p>

1242 G PROMPT

1243

1244 G.1 AGGREGATION PROMPTS

1245

1246

System Prompt

1247

1248

1249

1250

1251

As a domain-agnostic response synthesizer, critically evaluate multiple contributor submissions through three analytical lenses: (1) factual accuracy verification, (2) contextual relevance assessment, and (3) logical coherence validation. Integrate the most credible elements into a unified response maintaining original reasoning traces.

1252

1253

User Prompt

1254

1255

Following is the Task Description and the answers from different Contributors.

1256

[Task Description]

1257

<<Task Description and Question Input>>

1258

1259

[Contributor 1]

1260

<<Annotation from Contributor 1>>

1261

[Contributor 2]

1262

<<Annotation from Contributor 2>>

1263

...

1264

[Contributor n]

1265

<<Annotation from Contributor n>>

1266

<<Format Instruction>>

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296 G.2 PROMPT VARIANTS
12971298 We developed 10 task description prompt variants for each dataset.
12991300 Table 10: Variants of prompt applied to the AGNews fataset
1301

1302	Categorize the given article into one of the four categories: World, Sports, Business, or Sci/Tech. Justify your choice with a brief paragraph highlighting key details from the text that support your decision.
1303	
1304	
1305	As a journalist, analyze the article and determine whether it belongs to World, Sports, Business, or Sci/Tech. Provide a thorough explanation with specific details from the text that support your categorization.
1306	
1307	
1308	Examine the article’s content and decide whether it best fits into World, Sports, Business, or Sci/Tech. Summarize your reasoning in a concise paragraph, focusing on the most relevant details.
1309	
1310	Determine which category—World, Sports, Business, or Sci/Tech—best describes the article. Support your decision with specific examples from the text in a brief paragraph.
1311	
1312	
1313	As a news editor, review the article and classify it under one of the four categories: World, Sports, Business, or Sci/Tech. Explain your reasoning concisely, citing relevant information from the article.
1314	
1315	Assess the article’s main themes and determine whether it belongs to World, Sports, Business, or Sci/Tech. Discuss how its content aligns with your chosen category and why it does not fit the others.
1316	
1317	Analyze the article’s primary focus, audience, and key elements such as stakeholder involvement, risk factors, or technological relevance. Then, classify it under World, Sports, Business, or Sci/Tech and justify your decision with specific details.
1318	
1319	
1320	Evaluate all four categories—World, Sports, Business, and Sci/Tech—and explain why the article fits best into one of them. Provide a reasoned argument backed by key details from the text.
1321	
1322	
1323	Identify possible categories for the article and discuss why each could be relevant or not. Based on this analysis, assign it to World, Sports, Business, or Sci/Tech and justify your final choice in a brief paragraph.
1324	
1325	
1326	From a reader’s perspective interested in global affairs, sports, financial markets, or technological advancements, determine whether the article belongs to World, Sports, Business, or Sci/Tech. Explain your choice using specific content references.
1327	
1328	

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

Table 11: Variants of prompt applied to the DBPedia14 dataset

1350	
1351	
1352	Using logical deduction, analyze the text and identify specific keywords or patterns that most
1353	strongly align with one of the following categories: Company, EducationalInstitution, Artist, Athlete,
1354	OfficeHolder, MeanOfTransportation, Building, NaturalPlace, Village, Animal, Plant, Album, Film,
1355	or WrittenWork. Then, select the category that best matches the context and nature of the text,
1356	explaining your reasoning step-by-step.
1357	Apply a design thinking approach by first empathizing with the user who might be reading this text.
1358	Define what aspects of the text would resonate most with different categories such as Company,
1359	EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTransportation, Building, NaturalPlace,
1360	Village, Animal, Plant, Album, Film, or WrittenWork. Ideate potential categories based on the tone,
1361	content, and purpose of the text. Prototype your solution by choosing the category that fits the text’s
1362	core message. Test your hypothesis and explain the rationale behind your choice. Conclude with the
1363	selected category.
1364	Utilizing first-principles reasoning, break down the text into its most basic elements. Analyze whether
1365	the core subject of the text aligns with the fundamental characteristics of the categories: Company,
1366	EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTransportation, Building, NaturalPlace,
1367	Village, Animal, Plant, Album, Film, or WrittenWork. Starting from scratch, determine how these
1368	categories can be distilled into essential components and then reassemble them to select the most
1369	appropriate one. Output the final category clearly.
1370	Engage in a comparative analysis of how different categories (Company, EducationalInstitution,
1371	Artist, Athlete, OfficeHolder, MeanOfTransportation, Building, NaturalPlace, Village, Animal, Plant,
1372	Album, Film, or WrittenWork) would interpret the text. For each category, outline how the text could
1373	be viewed differently by someone in that domain. Based on these comparisons, select the category
1374	that most appropriately reflects the essence of the text and provide justification for your decision.
1375	Output the chosen category.
1376	Perform a deep dive into the text by taking a historical perspective. Consider how people from
1377	different time periods might have categorized the content. What historical context or shifts in
1378	understanding could influence the category selection today? Determine the category (Company,
1379	EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTransportation, Building, NaturalPlace,
1380	Village, Animal, Plant, Album, Film, or WrittenWork) by considering this historical lens and output
1381	the final choice.
1382	Frame the task as a hypothetical negotiation between experts in various domains. Each expert argues
1383	for their category (Company, EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTrans-
1384	portation, Building, NaturalPlace, Village, Animal, Plant, Album, Film, or WrittenWork) based on
1385	the text’s characteristics. You, as the mediator, must synthesize these arguments and select the most
1386	fitting category based on the strongest case made. Clearly output the selected category.
1387	Apply a pattern recognition approach: search for specific linguistic patterns, terms, or concepts in the
1388	text that align with known markers of each category (Company, EducationalInstitution, Artist, Athlete,
1389	OfficeHolder, MeanOfTransportation, Building, NaturalPlace, Village, Animal, Plant, Album, Film,
1390	or WrittenWork). Cross-reference these patterns with typical representations found in each category
1391	to determine the most appropriate one. Conclude with the most fitting category.
1392	Conduct a role-playing exercise where you take on the persona of an expert in one of the given
1393	categories (Company, EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTransportation,
1394	Building, NaturalPlace, Village, Animal, Plant, Album, Film, or WrittenWork). Approach the text
1395	from that persona’s perspective and argue why it should fall within that category. Afterward, switch
1396	roles to another category and repeat the process. Compare all arguments and output the final selected
1397	category.
1398	Consider the task as if you are a curator of a museum that houses diverse exhibits, each representing
1399	one of the categories (Company, EducationalInstitution, Artist, Athlete, OfficeHolder, MeanOfTrans-
1400	portation, Building, NaturalPlace, Village, Animal, Plant, Album, Film, or WrittenWork). Based on
1401	the description and thematic elements in the text, decide which exhibit the text would most likely
1402	belong to, providing reasons for your curatorial decision. Output the category you selected.
1403	Using a scientific method approach, formulate a hypothesis about the category (Company, Education-
	alInstitution, Artist, Athlete, OfficeHolder, MeanOfTransportation, Building, NaturalPlace, Village,
	Animal, Plant, Album, Film, or WrittenWork) based on an initial reading of the text. Collect evidence
	from the text to support or refute the hypothesis, and analyze it systematically. Based on this analysis,
	choose the category that best fits the content and context of the text. Output the final category.

1404
 1405
 1406
 1407
 1408
 1409
 1410
 1411
 1412
 1413
 1414
 1415
 1416
 1417
 1418
 1419
 1420
 1421
 1422
 1423
 1424
 1425
 1426
 1427
 1428
 1429
 1430
 1431
 1432
 1433
 1434
 1435
 1436
 1437
 1438
 1439
 1440
 1441
 1442
 1443
 1444
 1445
 1446
 1447
 1448
 1449
 1450
 1451
 1452
 1453
 1454
 1455
 1456
 1457

Table 12: Variants of prompt applied to the Emotion dataset

Using logical deduction, analyze the given text and identify the underlying emotion by evaluating the tone, word choice, and context. Break down the components of the text systematically to determine which emotional category (Sadness, Joy, Love, Anger, Fear, or Surprise) best matches the overall expression. Classify it into one of these emotions.
Imagine you are a psychologist conducting a session with the person who wrote the text. Based on your understanding of human emotions, assess the emotional state they are likely experiencing, considering both explicit and implicit clues from the text. Classify the emotion as either Sadness, Joy, Love, Anger, Fear, or Surprise.
Perform a comparative analysis by taking the text and comparing it with known examples of emotional expression. Identify patterns in language and sentiment to match the text with one of the six emotional categories (Sadness, Joy, Love, Anger, Fear, or Surprise). Clearly classify the emotion based on the analysis.
Adopt a design thinking approach: empathize with the emotional context of the text, define the emotional challenge, ideate potential emotional states, prototype potential emotional interpretations, and test your hypothesis against the text. Conclude which emotion (Sadness, Joy, Love, Anger, Fear, or Surprise) fits best.
Utilizing first-principles reasoning, strip down the language and tone of the text to its fundamental emotional elements. From this minimalist analysis, deduce the emotion that fits the core essence of the text, clearly classifying it as Sadness, Joy, Love, Anger, Fear, or Surprise.
Frame this task as a negotiation between emotions. Imagine that the text is a statement from one party (the speaker), and each emotion (Sadness, Joy, Love, Anger, Fear, or Surprise) is a negotiator. Determine which emotion wins the negotiation and classify the text as that emotion.
Place yourself in the shoes of a poet interpreting the emotional subtext of the text. Using your creative intuition, deduce which emotion (Sadness, Joy, Love, Anger, Fear, or Surprise) is most aligned with the emotional landscape presented in the text and classify it accordingly.
Consider the emotional journey of a character in a story. The text you are analyzing is a snippet from their diary. Based on the tone and context, classify the emotion expressed as Sadness, Joy, Love, Anger, Fear, or Surprise.
Imagine that you are conducting a debate between a group of experts on emotional intelligence. The text is presented as evidence, and the experts are tasked with determining the emotion it represents. Based on the debate, clearly classify the emotion as Sadness, Joy, Love, Anger, Fear, or Surprise.
Imagine you are a detective analyzing the emotional clues in the text to solve a case. Investigate the language used, identify emotional triggers, and classify the emotion as either Sadness, Joy, Love, Anger, Fear, or Surprise based on the evidence found.

1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472
 1473
 1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483
 1484
 1485
 1486
 1487
 1488
 1489
 1490
 1491
 1492
 1493
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511

Table 13: Variants of prompt applied to the IMDB dataset

Analyze the given movie review by identifying key phrases and emotional indicators. Focus on the tone, word choice, and expressions used to classify the sentiment as Positive or Negative. Clearly state the sentiment as either Positive or Negative.
Approach the review using a design thinking methodology. Start by empathizing with the reviewer’s emotional journey, define the sentiment issue, brainstorm possible interpretations, and test the sentiment to classify it as Positive or Negative. Provide a final classification as Positive or Negative.
Apply first-principles reasoning to break down the sentiment of the review. Identify the fundamental emotional tones and descriptive language used, then synthesize them to determine if the overall sentiment is Positive or Negative. Conclude with a final sentiment classification as Positive or Negative.
Imagine a group discussion about the movie review. Simulate how different individuals might react to the review’s language and tone to classify the sentiment as Positive or Negative based on their perspectives. Clearly state whether the sentiment is Positive or Negative.
Perform a comparative analysis of the review against reviews for similar films. Look for patterns in language, tone, and content to decide whether the sentiment is Positive or Negative, based on common conventions for the genre. Conclude with either Positive or Negative sentiment.
Assume the role of a mediator between two opposing sentiment analysts. One believes the review is Positive, the other Negative. Consider their arguments and come to a balanced conclusion about the review’s sentiment. Explicitly state whether the sentiment is Positive or Negative.
Frame the movie review as part of a negotiation between the filmmaker and the audience. The filmmaker wants a Positive sentiment, and the audience’s emotional response helps you classify the review as either Positive or Negative based on how it aligns with their expectations. Provide a final sentiment classification as Positive or Negative.
Incorporate a psychological approach by evaluating the emotional cues in the review. Consider the reviewer’s emotional state and biases, then determine if the sentiment expressed is Positive or Negative based on those psychological factors. Clearly state the sentiment as Positive or Negative.
Adopt a linguistic analysis perspective. Focus on the nuances of word choice, sentence construction, and emotional undertones to classify the sentiment of the review as either Positive or Negative. Provide a conclusive sentiment classification.
Use a literary criticism approach by examining the review’s rhetorical devices, narrative style, and overall tone. Identify whether the review’s language leans more toward positivity or negativity to classify the sentiment as Positive or Negative. Explicitly provide the sentiment classification.

1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565

Table 14: Variants of prompt applied to the MultiNLI dataset

From a logical deduction perspective, carefully examine the premise and hypothesis. If the truth of the premise guarantees the truth of the hypothesis, label it as Entailment. If the premise neither supports nor contradicts the hypothesis, label it as Neutral. If the premise contradicts the hypothesis directly, label it as Contradiction.
Utilize a scenario-based approach: construct a scenario where both the premise and hypothesis could either align, be independent, or conflict. Based on the scenario’s outcome, categorize the relationship as Entailment, Neutral, or Contradiction.
Apply a comparison strategy: break down the key elements of the premise and hypothesis. Determine if they logically reinforce one another (Entailment), are unrelated (Neutral), or present opposing ideas (Contradiction).
Use a first-principles approach: distill both the premise and hypothesis to their basic truths and foundational ideas. Based on these fundamentals, decide whether the relationship between them is one of Entailment, Neutral, or Contradiction.
Approach the task from a deductive reasoning perspective: systematically test whether the premise can be used to derive the hypothesis logically. If so, it’s Entailment. If the premise doesn’t influence the hypothesis, it’s Neutral. If the premise disproves the hypothesis, it’s Contradiction.
Frame the task as a debate. Treat the premise as one side of an argument and the hypothesis as the opposing argument. Determine if the arguments support each other (Entailment), exist independently (Neutral), or are in direct opposition (Contradiction).
Use a critical thinking methodology: identify and evaluate the core assumptions in the premise and hypothesis. Assess whether they complement (Entailment), do not affect each other (Neutral), or directly oppose one another (Contradiction).
Take a philosophical approach: engage with the premise and hypothesis through the lens of conceptual analysis. Does the premise inherently support the hypothesis (Entailment), allow for alternative interpretations (Neutral), or inherently refute the hypothesis (Contradiction)?
Adopt a systems-thinking approach. Consider the premise and hypothesis as parts of a larger system. Assess if they fit together naturally (Entailment), are disconnected (Neutral), or if one disrupts the integrity of the other (Contradiction).
Analyze the premise and hypothesis using a process of elimination. Test different logical possibilities where the premise could either lead to, be independent of, or contradict the hypothesis. Based on this elimination, classify the relationship as Entailment, Neutral, or Contradiction.

1566
 1567
 1568
 1569
 1570
 1571
 1572
 1573
 1574
 1575
 1576
 1577
 1578
 1579
 1580
 1581
 1582
 1583
 1584
 1585
 1586
 1587
 1588
 1589
 1590
 1591
 1592
 1593
 1594
 1595
 1596
 1597
 1598
 1599
 1600
 1601
 1602
 1603
 1604
 1605
 1606
 1607
 1608
 1609
 1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619

Table 15: Variants of prompt applied to the SST2 dataset

Identify any specific words or phrases that indicate a positive or negative sentiment, and classify the sentence as 'Positive' or 'Negative' based on these indicators.
As a consultant tasked with identifying positive sentiment in movie reviews, break down the sentence into its core elements and analyze the language used for any indicators of satisfaction, enjoyment, or appreciation. Classify the sentence as 'Positive' or 'Negative' accordingly.
From the perspective of a strategist, apply first-principles reasoning to dissect the sentence and determine if it expresses positivity or negativity. Utilize fundamental principles of sentiment analysis to classify the sentence as 'Positive' or 'Negative'.
As an end-user reading a movie review, empathize with the reviewer and gauge the overall sentiment conveyed in the sentence. Based on your personal understanding, classify it as 'Positive' or 'Negative'.
Use foundational reasoning to identify basic indicators of sentiment and classify the sentence as 'Positive' or 'Negative' based on the logical structure of the words used.
Classify the sentence as 'Positive' or 'Negative' by carefully comparing it to both positive and negative reviews to identify the subtle differences that classify it. Consider the context and tone of the sentence in this comparison.
Classify the sentence as 'Positive' or 'Negative' by considering the nuances of sentiment, such as sarcasm and irony. Explicitly analyze the sentence for these indicators and justify your classification accordingly.
Simulate the context in which the sentence was written and the potential reactions it might evoke to classify the sentence as 'Positive' or 'Negative'.
Use a creative brainstorming approach to identify both overt and subtle indicators of sentiment, classifying them as either 'Positive' or 'Negative' and justifying your choices while emphasizing the importance of context and nuance.
Adopt a scenario-based thinking approach, imagining the scenarios described in the sentence to classify it as 'Positive' or 'Negative', and explain your reasoning.

Table 16: Variants of prompt applied to the TREC dataset

1620	
1621	
1622	
1623	
1624	
1625	
1626	
1627	
1628	
1629	
1630	
1631	Analyze the question carefully and determine which type of information is being requested: Is it describing something, asking for an entity, referring to a place, inquiring about a person, looking for a number, or using an abbreviation? Classify the question as Description, Entity, Location, Human, Numeric, or Abbreviation, and justify your choice.
1632	
1633	
1634	
1635	Break down the question’s structure. Does it focus on an object, a person, a place, a number, a description, or a shorthand term? After evaluating the elements, classify the question as Description, Entity, Location, Human, Numeric, or Abbreviation and explain your reasoning.
1636	
1637	
1638	Think about the type of response the question is seeking: Does it ask for a label (Entity), a place (Location), a person (Human), a description (Description), a number (Numeric), or an abbreviation (Abbreviation)? Choose the correct category and provide a detailed justification for your decision.
1639	
1640	
1641	Approach the question from a logical perspective: If the question seeks to identify an object, a location, a person, a numeric value, a descriptive fact, or an abbreviation, classify it accordingly. The correct category is: Description, Entity, Location, Human, Numeric, or Abbreviation.
1642	
1643	
1644	Consider the question from a broader context: Does the inquiry fit into one of these categories: a factual description, an object or entity, a place, a person, a numerical value, or a shorthand form? Classify the question as Description, Entity, Location, Human, Numeric, or Abbreviation, and explain your rationale for the chosen category.
1645	
1646	
1647	
1648	Imagine you’re categorizing the question in a filing system designed for organizing knowledge. Which folder does it belong to—Description, Entity, Location, Human, Numeric, or Abbreviation? Choose the correct category and justify your reasoning.
1649	
1650	
1651	
1652	Frame the question within a real-world scenario: Is it asking for information about an object (Entity), a place (Location), a person (Human), a number (Numeric), a description (Description), or shorthand (Abbreviation)? Classify the question as Description, Entity, Location, Human, Numeric, or Abbreviation and explain your reasoning.
1653	
1654	
1655	
1656	Consider the purpose of the Query Is it inquiring about an object, a place, a person, a number, a description, or an abbreviation? Classify the question into one of these categories—Description, Entity, Location, Human, Numeric, or Abbreviation—and explain the reasoning behind your choice.
1657	
1658	
1659	Take a systems-thinking approach: What category does the question best fit into based on the kind of information it’s asking for? Classify the question as Description, Entity, Location, Human, Numeric, or Abbreviation, and offer a justification for your classification.
1660	
1661	
1662	Imagine being tasked with explaining this question to someone unfamiliar with it. How would you categorize the question—based on whether it refers to a description, a specific entity, a place, a human, a numeric value, or an abbreviation? Provide the category: Description, Entity, Location, Human, Numeric, or Abbreviation, and explain your rationale.
1663	
1664	
1665	
1666	
1667	
1668	
1669	
1670	
1671	
1672	
1673	

1674
 1675
 1676
 1677
 1678
 1679
 1680
 1681
 1682
 1683
 1684
 1685
 1686
 1687
 1688
 1689
 1690
 1691
 1692
 1693
 1694
 1695
 1696
 1697
 1698
 1699
 1700
 1701
 1702
 1703
 1704
 1705
 1706
 1707
 1708
 1709
 1710
 1711
 1712
 1713
 1714
 1715
 1716
 1717
 1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727

Table 17: Variants of prompt applied to the WikiToxic dataset

Using sentiment analysis, examine the given Wikipedia comment and categorize it based on the emotional tone conveyed. Label the comment as 'Toxic' if the sentiment is overwhelmingly negative, aggressive, or harmful, and 'Non-Toxic' if it is neutral or positive in tone.
Consider the linguistic structure and word choice within the comment. Identify specific toxic language indicators such as insults, threats, or discriminatory remarks. Classify the comment as 'Toxic' if it contains any of these features, and 'Non-Toxic' if it does not.
Adopt a scenario-based approach: imagine you are a moderator on a Wikipedia discussion page. Given the comment, would you consider it acceptable or offensive based on community guidelines? If it violates the guidelines, label it 'Toxic'; otherwise, label it 'Non-Toxic'.
Use a first-principles reasoning approach: Break down the comment to its fundamental components. Examine whether the language in the comment is objectively harmful or promotes toxicity, and classify it accordingly as 'Toxic' or 'Non-Toxic'.
Employ a comparative analysis methodology: analyze the given comment in the context of several other known toxic and non-toxic comments. Based on similarities in tone, phrasing, and content, decide whether the comment falls into the 'Toxic' or 'Non-Toxic' category.
Apply a design thinking approach: empathize with the potential audience of the comment. Analyze how the language might affect readers—whether it could harm, alienate, or provoke. If the language might cause harm, label it 'Toxic'; if it's neutral or beneficial, label it 'Non-Toxic'.
Frame this as a hypothetical negotiation: imagine you're negotiating the content standards for Wikipedia comments with a team of content moderators. Would this comment pass your collective standards? If it doesn't meet the agreed standards, classify it as 'Toxic'; if it does, classify it as 'Non-Toxic'.
Consider the comment from a psychological perspective: what motivations or emotions might be driving the language? If the comment reveals hostility, anger, or frustration aimed at others, classify it as 'Toxic'; otherwise, classify it as 'Non-Toxic'.
Think of the comment in terms of ethical considerations: if you were to act as a judge for the comment, would it be deemed as potentially harmful to the community? If the comment undermines respect and promotes harm, classify it as 'Toxic'; if not, classify it as 'Non-Toxic'.
Adopt a critical thinking approach: examine the logic and reasoning behind the comment. Is there evidence of personal attacks, stereotyping, or harmful assertions? If so, classify it as 'Toxic'; if it's logical, respectful, and constructive, label it as 'Non-Toxic'.

1728
 1729
 1730
 1731
 1732
 1733
 1734
 1735
 1736
 1737
 1738
 1739
 1740
 1741
 1742
 1743
 1744
 1745
 1746
 1747
 1748
 1749
 1750
 1751
 1752
 1753
 1754
 1755
 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781

Table 18: Variants of prompt applied to the MMLU dataset

1744	Assume the role of a logic professor tasked with solving this multiple-choice question using step-by-step logical deduction. Break down each option and explain the reasoning that leads to the correct choice.
1748	Approach this multiple-choice question through first-principles reasoning. Strip the question down to its foundational concepts and build up your understanding to select the correct answer.
1750	Use comparative analysis to evaluate all answer choices. Systematically compare and contrast the merits and flaws of each option to determine which is most accurate.
1752	Adopt the mindset of a detective solving a mystery. Examine clues in the question and eliminate incorrect options using evidence-based reasoning to reveal the correct choice.
1755	Frame the problem using systems thinking. Analyze how each component of the question interrelates and assess the systemic coherence of each option before identifying the correct answer.
1757	Engage in reverse engineering. Assume each answer choice is correct, then work backward to see which one aligns logically with the conditions in the question.
1759	Use design thinking by empathizing with the test-maker. Consider the intent behind the question and prototype different reasoning paths to arrive at the answer that best fits the underlying design.
1761	Use a root cause analysis framework. Identify the core issue or concept tested by the question and trace it through each option to determine the most fitting response.
1764	Analyze the question using analogical reasoning. Draw parallels between the current problem and known examples or models to infer the most applicable answer.
1766	Assume the role of a debate moderator. Stage a brief internal debate between advocates for each answer choice and assess which argument most convincingly supports its claim.

1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
 1795
 1796
 1797
 1798
 1799
 1800
 1801
 1802
 1803
 1804
 1805
 1806
 1807
 1808
 1809
 1810
 1811
 1812
 1813
 1814
 1815
 1816
 1817
 1818
 1819
 1820
 1821
 1822
 1823
 1824
 1825
 1826
 1827
 1828
 1829
 1830
 1831
 1832
 1833
 1834
 1835

Table 19: Variants of prompt applied to the BoolQ dataset

Apply logical deduction to analyze the given passage and identify if there is direct evidence supporting or denying the facts in the question. If the passage provides clear support for the facts, output 'Yes' If the passage contradicts the facts, output 'No'
Conduct a scenario-based simulation where you assume you are a fact-checker tasked with verifying the truthfulness of the statement in the question using the evidence provided in the passage. Based on the passage, output either 'Yes' if the facts are supported or 'No' if they are denied.
Use comparative analysis to contrast the facts presented in the question with the information found in the passage. If the passage aligns with the question, output 'Yes' If it contradicts or denies the facts, output 'No'
Think like a detective investigating the validity of a claim. Break down the passage and analyze whether there is sufficient evidence to support or refute the statement in the question. If supported, output 'Yes' If refuted, output 'No'
Adopt a design thinking approach and assess the passage for clarity, accuracy, and consistency with the question. If the passage confirms the facts in the question, output 'Yes' If it denies them, output 'No'
Approach the task through first-principles reasoning by deconstructing both the question and the passage to their most fundamental truths. If the passage logically supports the question's facts, output 'Yes' If the passage denies them, output 'No'
Imagine that you are a judge in a courtroom, and the question is being presented as evidence for a trial. If the facts in the passage confirm the question, output 'Yes' If the passage denies or contradicts the question, output 'No'
Conduct a thorough analysis of the passage as if you are an expert fact-checker. If the passage provides concrete evidence that supports the facts in the question, output 'Yes' If it denies or contradicts them, output 'No'
Frame the task as a negotiation between two parties: one party asserts the facts in the question, while the other denies them. If the passage supports the facts in the question, output 'Yes' If the passage denies the facts, output 'No'
Take on the perspective of a philosophical debate. Argue both sides of the claim in the question, and based on the passage, output 'Yes' if the passage supports the claim, or 'No' if the passage denies the claim.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

Table 20: Variants of prompt applied to the WiC dataset

Analyze the context in which 'w' is used in 's1' and 's2', and determine if both instances of 'w' convey the same core meaning, excluding any differences in form and tense. Output 'True' if they do, otherwise 'False'.
Use logical reasoning to determine if the core meanings of 'w' in 's1' and 's2' are consistent, excluding any differences in form and tense. Output 'True' or 'False'.
Determine if 'w' in both 's1' and 's2' has the same core meaning, excluding any differences in form and tense. Output 'True' or 'False'.
Consider the contextual understanding of 'w' in 's1' and 's2', and determine if it conveys the same core meaning, excluding any differences in form and tense. Output 'True' if they do, otherwise 'False'.
Analyze the meaning of 'w' in 's1' and 's2', and determine if they convey the same core meaning, excluding any differences in form and tense. Output 'True' or 'False'.
Examine the context in which 'w' is used in 's1' and 's2', and determine if both instances of 'w' convey the same core meaning, excluding any differences in form and tense. Output 'True' if they do, otherwise 'False'.
Act as a legal expert in a debate. Argue both sides of the case to determine if the word 'w' in 's1' and 's2' correspond to the same meaning. Output 'True' if the word's meanings are aligned, otherwise output 'False'.
Investigate how 'w' operates semantically in 's1' and 's2', ascertaining if the core meaning remains consistent when accounting for contextual usage. Output 'True' or 'False'.
Assume the role of a linguist and analyze the core meaning of 'w' in 's1' and 's2', excluding any differences in form and tense, and output 'True' if they do, otherwise 'False'.
In the role of a first-principles physicist, apply foundational principles to deduce the meaning of 'w' in 's1' and 's2'. If the principles align, output 'True'; otherwise, output 'False'.

Table 21: Variants of prompt applied to the AI2ARC dataset

Solve the multiple-choice questions based on scientific reasoning and knowledge.
Tackle the set of science questions by choosing the appropriate multiple-choice answers.
Use your understanding of science to pick the correct answers from the given options.
Review and solve each science multiple-choice question logically.
Determine the right choice for every science question in the list.
Choose the most scientifically accurate response for every question presented.
Imagine you are a high school science teacher reviewing a quiz—select the correct answers for each student question.
Take the role of a science competition judge and evaluate each multiple-choice question for the best answer.
Act as a scientific researcher taking a certification exam—identify the right answers to maintain your credentials.
You're a tutor helping a student prepare for a science test. Choose the correct answers to explain to them later.

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

Table 22: Variants of prompt applied to the GSM8K dataset

As a mathematics expert, carefully analyze the problem, justify each step rigorously, and ensure that your solution would be accepted in a formal mathematical journal.
As a student in a timed exam setting, solve the problem quickly and efficiently, balancing speed with accuracy, and skipping any unnecessary steps.
As a math teacher explaining the problem to a beginner, break the solution down into clear, digestible parts, using simple language and logical structure.
As a mathematical competitor in an Olympiad, look for elegant or clever shortcuts, tricks, or symmetry to reach the solution faster than using brute-force methods.
As a mathematician preparing to present your solution in a seminar, ensure every logical step is watertight and anticipate questions from an expert audience.
As a researcher writing a textbook, focus on clarity, pedagogical progression, and general principles that can apply beyond the immediate problem.
Use deductive reasoning to solve the mathematics problem by starting from known premises and building a logically rigorous sequence of steps that leads to the correct conclusion.
Apply first-principles thinking: break the problem down to its most basic definitions and axioms, and construct your solution without relying on pre-established shortcuts or formulas.
Break the problem into smaller sub-problems, solve them independently, and then integrate the partial results into a complete solution.
Apply a known theorem or identity carefully, making sure its conditions are satisfied in the current context before proceeding with the solution.

Table 23: Variants of prompt applied to the Alpaca dataset

You are an assistant that should think through the problem step by step before answering.
As an assistant, your goal is to provide clear, logical reasoning behind every answer.
You’re helping someone understand a complex issue—break it down into manageable parts.
Take your time. You are expected to consider various angles before responding.
You are a careful assistant. Avoid assumptions and stick to what the information supports.
Your job is to respond clearly and directly. Focus only on the key points.
Be thorough. As an assistant, you must evaluate all options before offering a suggestion.
You’re an assistant trained to explain not just what, but why.
You are expected to analyze carefully before reaching a conclusion.
Keep your explanation structured. The person you’re helping needs to follow your logic easily.

1944
 1945
 1946
 1947
 1948
 1949
 1950
 1951
 1952
 1953
 1954
 1955
 1956
 1957
 1958
 1959
 1960
 1961
 1962
 1963
 1964
 1965
 1966
 1967
 1968
 1969
 1970
 1971
 1972
 1973
 1974
 1975
 1976
 1977
 1978
 1979
 1980
 1981
 1982
 1983
 1984
 1985
 1986
 1987
 1988
 1989
 1990
 1991
 1992
 1993
 1994
 1995
 1996
 1997

Table 24: Variants of prompt applied to the DS-MVTec dataset

1961	You are an industrial quality control inspector analyzing product images for potential defects. Review the image and determine if it meets manufacturing standards.
1963	As a factory inspection expert, evaluate the product in the image for any signs of surface irregularities or damage that may affect usability.
1966	You are a defect detection specialist examining product photographs. Assess the image and report whether a defect is present, providing a justification.
1968	You are performing a visual inspection for a manufacturing audit. Examine the image and decide if the product displays any nonconformities.
1970	Act as a certified industrial inspector evaluating high-resolution images of manufactured items. Judge if there are any quality issues present.
1972	As an image-based inspection analyst, assess the product photo for any discrepancies compared to the expected standard.
1975	Take on the role of a factory floor quality inspector reviewing image submissions. Determine if the item shown contains a production flaw.
1977	You are a visual defect detection professional. Scrutinize the product image to evaluate for any scratches, cracks, or misalignments.
1979	As a product evaluation specialist working from visual data, assess if the item in the image exhibits any indication of malfunction or imperfection.
1982	Serve as an expert in machine vision inspection for industry. Determine whether the item in the provided image meets quality criteria.