

Can LLM Annotators Match Human Consistency? Benchmarking Human Agreement and Optimizing Multi-Agent with Contrastive Knowledge

Xiaopeng Ke^{1,2} Zhixin Wang¹ Xuebo Liu^{2*} Xinyu Ma² Xing Zheng¹

Bowen Song¹ Weiqiang Wang¹ Min Zhang²

¹ANT Group ²Harbin Institute of Technology (Shenzhen)

{xiaopk7, mxinyuma}@gmail.com {liuxuebo, zhangmin2021}@hit.edu.cn

{linglan.wzx, feishang.zx, bowen.sbw, weiqiang.wwg}@antgroup.com

Abstract

LLM-based annotation has advanced rapidly, yet existing benchmarks remain misaligned with real-world practice. Actual annotation often involves lengthy, task-specific guidelines, and different annotators may produce divergent but reasonable judgments, especially on boundary cases, where single-labeled dataset fail to reflect this disagreement. This gap also limits existing single-agent and workflow-based annotation systems, which often lack adaptability to complex, evolving guidelines. To bridge this gap, we introduce **ANT-A Benchmark**, a real-world multi-annotator benchmark spanning 7 datasets across domains and modalities, with task-specific guidelines and 5 independent human labels per instance. ANT-A enables a distribution-aware evaluation, using Fleiss' κ to assess how closely model annotations align with human labeling distributions. We further propose **ANT-OR**, a multi-agent annotation framework that refines prompts by distilling contrastive decision rules from positive and negative feedback. Experiments show that ANT-OR improves agreement@ ≥ 2 accuracy by 4.5 points over the strongest baseline and achieves the highest alignment with human annotator distributions as measured by Fleiss' κ within 2 points. Code and benchmark are available at <https://github.com/Krueske/ANT-OR>.

1 Introduction

Data annotation is essential to machine learning, yet acquiring high-quality annotations at scale remains costly and challenging (Ratner et al., 2017; Zhang et al., 2021; Wang et al., 2024). Large language models (LLMs) have emerged as an attractive option for automated annotation across diverse tasks (He et al., 2024; Tan et al., 2024). However, current LLM-based annotation systems still depend

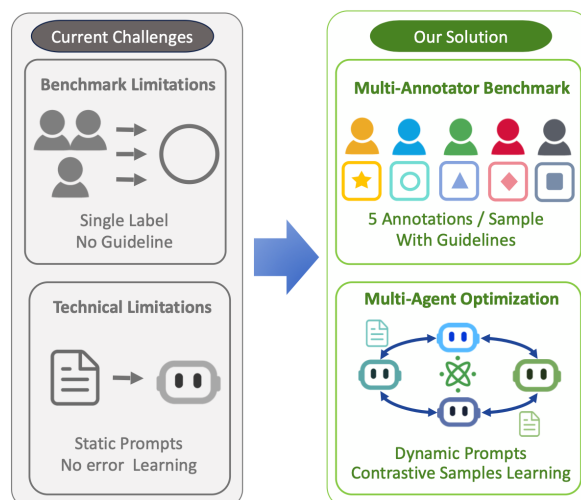


Figure 1: Overview of our contributions. We construct **ANT-A**, a benchmark with detailed guidelines and multiple independent human annotations, and develop **ANT-OR**, a multi-agent framework that improves automated annotation through iterative optimization with contrastive knowledge.

heavily on hand-crafted prompts and static workflows (Yu et al., 2025), which often fail on complex tasks with detailed guidelines and boundary cases.

A major obstacle to further developing of data annotation methods is the lack of evaluation protocols that reflect real-world annotation practice. Existing LLM annotation benchmarks often lack the guideline documents used in real-world tasks and provide only a single gold-standard label per sample (Basile et al., 2021; Kim et al., 2026). However, many realistic annotation tasks are governed by detailed annotation guidelines, which can be inherently ambiguous or open to interpretation. As a result, disagreement among annotators may reflect valid alternative interpretations rather than annotation errors (Tseng et al., 2025; Fleisig et al., 2023a). Single-label benchmarks thus collapse this uncertainty and cannot determine whether LLM annotators achieve human-level consistency, e.g.,

* Xuebo Liu is the corresponding author.

Fleiss’ κ (Fleiss, 1971), or simply match one reference answer (Mehrotra et al., 2026). To address this gap, we construct **ANT-A** (Aggregated N-annotator Task Annotations), a multi-annotator benchmark comprising 7 datasets with detailed guidelines and 5 independent human annotations per sample.

Realistic annotation settings also raise a challenge for single-agent systems, which often struggle on complex tasks governed by guidelines because a single reasoning perspective cannot reliably handle nuanced criteria and edge cases (Ban et al., 2026; Pan et al., 2025). Multi-agent systems offer a promising alternative through complementary judgments and cross-checking (Tran et al., 2025). Existing approaches typically treat agents as static components (Zheng et al., 2025) and fail to transform annotation experience into structured knowledge (Abdollahi et al., 2025). To address this limitation, we propose **ANT-OR**, an automated ANnotation framework Through multi-agent Optimization with agent-level contRastive knowledge. By incorporating agent-level contrastive knowledge into optimization, ANT-OR updates decision heuristics to improve future judgments and enhance inter-agent coordination.

Extensive experiments on the proposed benchmark show that ANT-OR substantially improves annotation performance across diverse tasks, outperforming the strongest baseline by 4.5 points in accuracy and achieves the highest agreement with human annotator distributions as measured by Fleiss’ κ within 2 points on average. Our primary contributions are summarized as follows:

- We identify two critical bottlenecks to reliable automated annotation: inadequate benchmarks that fail to capture inter-annotator disagreement and complexity from detailed annotation guidelines, and limitations of current annotation agents, particularly those based on single-agent or static multi-agent designs.
- We construct and release a multi-annotator benchmark comprising 7 datasets with detailed guidelines and 5 independent human annotations per sample, enabling a distribution-aware evaluation, using Fleiss’ κ to assess how closely model annotations align with human labeling distributions.
- We propose ANT-OR, an adaptive multi-agent framework that can transform collective prediction successes and failures into reusable

structured decision knowledge, enabling more robust collaboration on complex annotation tasks involving nuanced decision boundaries.

2 Related Work

LLM-based Data Annotation and Benchmarks.

LLMs are increasingly used as scalable alternatives to human annotators (He et al., 2024; Tan et al., 2024; Gilardi et al., 2023), creating a need for reliable benchmarks of their annotation behavior. However, existing benchmarks often simplify real-world annotation in two ways: first, they lack the detailed guideline documents that capture the complexity of realistic annotation tasks; second, they provide only a single gold or aggregated label per sample (Basile et al., 2021; Kim et al., 2026). These limitations make it difficult to evaluate whether LLMs can follow task-specific annotation protocols (Ke et al., 2025; Liu et al., 2024a) and whether their outputs match the distribution of human judgments rather than a single reference answer (Liu et al., 2024b). Recent work shows that LLMs struggle with specialized guidelines (Tseng et al., 2025; Diao et al., 2025), and that annotator disagreement often reflects meaningful differences in perspective rather than noise (Santurkar et al., 2023; Zhang et al., 2025c; Muscato et al., 2025; Wang and Lee, 2026). Concurrently, a line of work models annotator disagreement as a first-class problem rather than noise to be averaged out. Fleisig et al. (2023b) and Xu et al. (2025) predict individual annotator labels or demographic-group distributions. In emotion recognition, Wu (2026) and Jia et al. (2026) model ambiguity directly from human label distributions without demographic routing, yet their approaches are tied to task-specific architectures. These methods are valuable references, but they do not address the cross-domain, guideline-grounded setting we target, where disagreement persists on boundary cases even under explicit rules. Overall, prior work recognizes the importance of both guidelines and disagreement, but existing benchmarks typically study them in isolation. We address this gap with **ANT-A**, a multi-annotator benchmark that pairs detailed annotation guidelines with multiple independent human annotations per sample.

Multi-Agent System Construction and Optimization.

Multi-agent collaboration and optimization have emerged as complementary directions for improving LLM reliability. Existing approaches refine outputs through sampling (Wang

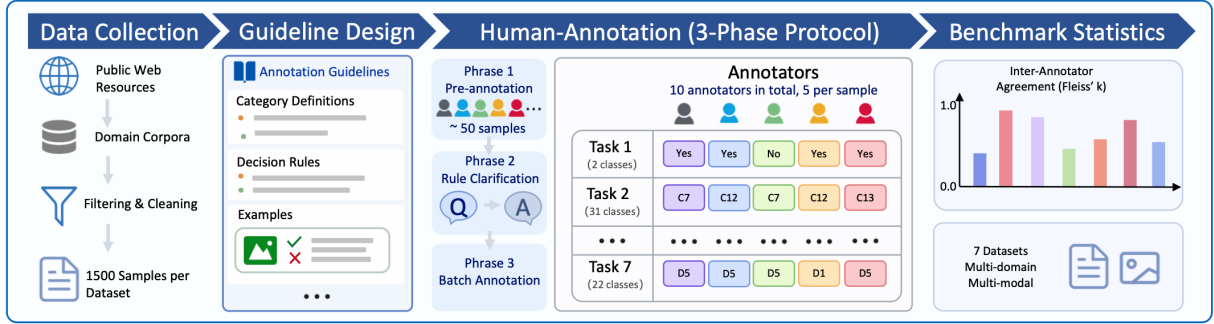


Figure 2: Construction of **ANT-A**. We curate domain-relevant samples, establish annotation guidelines through a pre-annotation pilot, and collect 5 independent annotations per sample to measure inter-annotator agreement.

Dataset	Domain	Modality	#Samples	#Classes	Fleiss' κ
Disease Privacy Assessment	Medical	Text	1,500	2	38.77
Billing Scenario Classification	E-commerce	Text	1,500	31	83.50
Medical Entity Extraction	Medical	Text	1,500	22	82.09
Social Text Classification	General	Text	1,500	15	44.62
Image Attribution Recognition	General	Vision-Language	1,500	3	58.59
Image Classification	E-commerce	Vision-Language	1,500	24	81.27
Image Text Relevance	General	Vision-Language	1,500	3	56.37

Table 1: Summary statistics of **ANT-A**

et al., 2023b), inter-agent dialogue (Du et al., 2024; Liang et al., 2024), or generation-verification pipelines (Hegazy et al., 2025; Ahtisham et al., 2026), but treat prompts as static after initial design. Recent work explores optimizing multi-agent structures via sparse topology learning (Zhang et al., 2025a,b; Wang et al., 2025; Tian et al., 2025) or joint prompt-topology search (Zhou et al., 2025; Wang et al., 2026), yet these target token efficiency rather than task-specific adaptation with cross-iteration transfer. Meanwhile, evolutionary (Guo et al., 2024; Fernando et al., 2024), trajectory-based (Yang et al., 2024; Tang et al., 2025), and feedback-based methods (Pryzant et al., 2023; Xiang et al., 2025; Chen et al., 2025) optimize prompts without fine-grained error analysis or cross-iteration knowledge retention. Memory-based ERM (Yan et al., 2025) accumulates past errors but operates at the instance level, lacking structural abstraction of error patterns. At the model-training level, weakness-case acquisition has also been used to focus iterative preference optimization on model-specific failures (Rao et al., 2025b,a). **ANT-OR** instead mines structured decision rules from iterative error-success contrasts, enabling multi-agent systems to rapidly adapt to complex annotation tasks without training.

3 ANT-A Benchmark

As shown in Figure 2, we introduce **ANT-A** (Aggregated N-annotator Task Annotations), a

multi-source benchmark that jointly provides detailed guidelines and 5 independent human judgments per sample that reflect the inherent ambiguity of tasks and subjectivity of human annotations rendering single labels unreliable (see Appendix B), enabling rigorous comparison between LLM systems and human raters under realistic conditions.

3.1 Dataset Preparation

Data Collection. We collect domain-relevant query samples from public internet data sources for each dataset, anonymize potentially identifiable information, and filter exact duplicates and low-quality instances to obtain 1,500 samples with a mix of clear-cut and boundary cases.

Guideline Document Design. We derive annotation guidelines from real-world requirements and refine them by clarifying ambiguous definitions, adding decision rules for boundary cases, and tailoring to our data distribution. The resulting documents specify fine-grained decision boundaries and edge-case handling rules.

3.2 Human Annotation Process

All annotations are conducted by professional annotators from an internal annotation vendor, and we present the relevant backgrounds of the annotators in Table 10. We use a pool of 10 annotators, and each sample is independently labeled by 5 of them. This setup reflects realistic annotation practice and reduces bias from a fixed small human group. To ensure quality, we adopt a three-phase protocol, with a total annotation cost of USD \$4078.

Phase 1: Pre-annotation. A small subset of each dataset (approximately 50 samples) is first annotated by all 10 annotators. This pilot phase familiarizes annotators with the guideline document and task requirements, and surfaces ambiguities

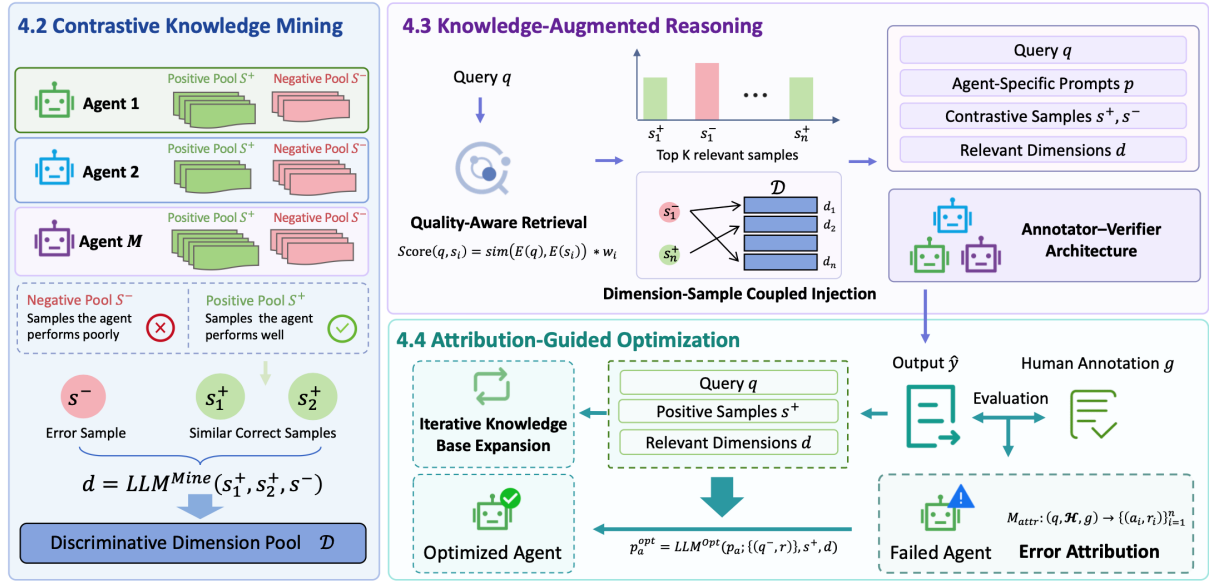


Figure 3: Workflow of **ANT-OR**. Agents construct contrastive knowledge bases (§4.2), perform retrieval-augmented annotation (§4.3), and optimize prompts via failure attribution with iterative knowledge updates (§4.4).

or under-specified rules in the guidelines. Outputs from this phase are reviewed to identify systematic disagreements stemming from guideline ambiguity rather than genuine judgment differences.

Phase 2: Rule Clarification. Based on the pre-annotation results, we collect questions and confusion points raised by annotators regarding category definitions, boundary cases, and decision rules. These are addressed through a structured Q&A process: annotator questions are compiled, unified answers are provided, and the guideline document is revised accordingly to eliminate identified ambiguities. This iterative clarification ensures that remaining disagreements in the formal annotation phase reflect genuine task difficulty rather than inconsistent rule interpretation.

Phase 3: Batch Annotation. After guideline finalization, each of the 1,500 samples in a dataset is independently labeled by 5 annotators selected from the pool of 10. Annotators label only their assigned samples and cannot access others’ labels, enabling meaningful inter-annotator agreement analysis and comparison with LLM-based annotators.

3.3 Benchmark Statistics

Table 1 summarizes the key statistics of our benchmark. Across the 7 datasets, we observe varying levels of inter-annotator agreement as measured by Fleiss’ κ (Fleiss, 1971), reflecting the inherent difficulty differences across tasks and domains.

Across datasets, instances with five completely different annotations are extremely rare (nearly 1%), suggesting that most samples retain partial human consensus. Further details on dataset construction, annotation guidelines, and annotator demographics are provided in Appendix C.

4 Our Proposed ANT-OR Framework

4.1 Overview

As shown in Figure 3, we propose **ANT-OR**, an automated **AN**notation framework **T**hrough multi-agent **O**ptimization with agent-level cont**R**astive knowledge. Designed for complex real-world annotation tasks, ANT-OR optimizes agents using contrastive evidence from negative and similar positive samples, uncovering the decision criteria that distinguish confusable categories in boundary cases.

4.2 Contrastive Knowledge Mining

ANT-OR equips each agent with a structured knowledge base consisting of validated samples and discriminative dimensions, allowing agents to accumulate, in a continual manner, both concrete annotation examples and decision-making knowledge across optimization iterations.

Contrastive Sample Pool. We construct a positive sample $s^+ = (q, \hat{y}^+)$ with correct output $\hat{y}^+$ on a given query q and $s^- = (q, \hat{y}^-)$ as the negative sample. (In our proposed ANT-A settings, a prediction is considered positive if it matches at

least two human labels out of five.) For each agent a , we maintain a positive pool $\mathcal{S}_a^+$ and a negative pool $\mathcal{S}_a^-$ of verified instances. During annotation, instances from both pools are sampled into the agent’s prompt as instance-level references. Each instance s has a quality score w that represents its sampling priority and is initialized to 1.

Discriminative Dimension Pool. Beyond utilizing instance-level references, agents further extract rule-level experiences. When an agent makes a wrong prediction $\hat{y}^-$ on a given query q , we first construct the error sample s^- . Then, we retrieve two semantically similar correct samples $s_1^+, s_2^+ \in \mathcal{S}_a^+$ to form a contrastive tuple (s_1^+, s_2^+, s^-) and an LLM is prompted to summarize *discriminative dimension* d from these confusing cases:

$$d = \text{LLM}^{\text{Mine}}(s_1^+, s_2^+, s^-), \quad (1)$$

The discriminative dimension is formulated as a structured textual unit, containing category-specific decision criteria distilled from contrastive cases and the applicable scenarios in which these criteria are expected to guide annotation decisions.

To avoid redundant dimensions, for each newly extracted discriminative dimension d , we retrieve semantically similar dimensions from the discriminative dimension pool $\mathcal{D}$. When a matching dimension d_{old} is identified, we apply an LLM-based refinement step to fuse d with d_{old} before adding the resulting dimension back to $\mathcal{D}$:

$$d := \text{LLM}^{\text{Refine}}(d_{\text{old}}, d). \quad (2)$$

Each dimension is assigned a support count, which records the number of times it has been involved in fusion. The support count serves as an estimate of how many training samples with inference errors share similar judgment patterns. To suppress spurious patterns, we retain only dimensions whose support count is no less than 2. We maintain a bidirectional many-to-many mapping between dimensions and source samples for traceability.

4.3 Knowledge-Augmented Reasoning

Annotator–Verifier Architecture. Our ANT-OR adopts a hierarchical annotator–verifier architecture (Hegazy et al., 2025; Ahtisham et al., 2026): three annotators independently produce labels and rationales using separately optimized prompts, and a verifier then compares their justifications to determine the final label. This design preserves proposal diversity while centralizing conflict resolution, avoiding the costly multi-round deliberation

of debate-based methods (Du et al., 2024) and the plan-induced error propagation of planner–worker alternatives (Wang et al., 2023a) (see Appendix A).

Quality-Aware Sample Retrieval. Given a query q and a sample s_i from the contrastive sample pool, we compute the retrieval score as

$$\text{Score}(q, s_i) = \text{sim}(E(q), E(s_i)) \cdot w_i, \quad (3)$$

where $E(\cdot)$ denotes the embedding function, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. After each retrieval, quality score w_i corresponding to s_i is updated according to the correctness of the resulting prediction for future retrievals:

$$w_i \leftarrow w_i + \Delta w_i, \quad (4)$$

where

$$\Delta w_i = \alpha \cdot \begin{cases} 1 - \rho, & \text{if correct,} \\ -\rho, & \text{if incorrect,} \end{cases} \quad (5)$$

with ρ denoting the current global accuracy and α the base step size. This update increases the priority of samples that contribute to correct predictions while penalizing those leading to incorrect predictions. As a result, retrieval is guided not only by semantic relevance but also by the historical reliability of stored examples.

Dimension–Sample Coupled Injection. We select the top- K highest-scoring samples, and discriminative dimensions associated with them. ANT-OR injects both instance-level and rule-level into prompts, providing both concrete evidence and abstract decision criteria, while existing methods typically inject either examples or rules alone.

4.4 Attribution-Guided Optimization

ANT-OR performs prompt optimization at the agent level. It first attributes each error to responsible agents and then refine their prompt accordingly.

Error Attribution. After feeding a query into the multi-agent system and obtaining the output, error attribution is triggered when the result fails to match at least two human-annotated references out of five. The current query q , the agent interaction history $\mathcal{H}$, and the majority label g are fed into an error attribution model for detailed error analysis, which identifies the agents responsible for the labeling error along with an analysis of the error causes, such as intrinsic mistakes or misleading information from upstream models:

$$\mathcal{M}_{\text{attr}} : (q, \mathcal{H}, g) \rightarrow \{(a_i, r_i)\}_{i=1}^n \quad (6)$$

where $\mathcal{M}_{\text{attr}}$ is implemented by the same backbone LLM as annotation agent with a dedicated attribution prompt. a_i is the i -th responsible agent, r_i is the corresponding error reason for agent a_i , and n is the total number of responsible agents identified for the current error.

Agent-Level Prompt Optimization. For each flagged agent a , we optimize its prompt p_a using error samples with attributed failure reasons r , most relevant positive sample s^+ , and discriminative dimensions d associated with s^+ :

$$p_a^{\text{opt}} = \text{LLM}^{\text{Opt}}(p_a; \{(q^-, r)\}, s^+, d). \quad (7)$$

This combines concrete failures, contrastive support, and abstract decision rules in a single refinement step, enabling more targeted prompt updates than optimization solely on error samples. After updating the prompts, we re-evaluate the optimized multi-agent system on the current batch and retain the updated prompts only if batch-level performance improves; otherwise, we roll back to the previous prompts to ensure robust updates.

Iterative Knowledge Base Expansion. Optimization proceeds over multiple epochs through an iterative cycle of evaluation, knowledge-base update, and prompt refinement. After each iteration, verified prediction outcomes are added to the contrastive sample pools, $\mathcal{S}_a^+$ and $\mathcal{S}_a^-$, expanding the knowledge base available for subsequent retrieval and contrastive mining. In this way, ANT-OR establishes a self-reinforcing optimization loop, where improved prompts lead to higher-quality annotations, and the enriched knowledge base further supports more informative error attribution and more effective prompt optimization.

5 Experiments

5.1 Experimental Setup

Models and Benchmarks. Considering model performance, cost, and inference efficiency, we conduct experiments using DeepSeek-V3.1-Terminus (DeepSeek-AI et al., 2025) for text-modality tasks and Qwen3-235B-A22B-VL-Instruct (Bai et al., 2025) for vision-language tasks as the backbone models of MAS. Both models also serve as the optimizers themselves. As mentioned in Section 3,

we construct a benchmark covering multiple domains (medical, e-commerce, general) and multiple modalities (including text and vision-language tasks) to evaluate the performance of different optimization methods in data annotation scenarios.

Baselines. We reproduce the relevant baselines in both single-agent and multi-agent system (MAS) scenarios. In single-agent scenarios, we compare with the Chain-of-Thought (CoT) (Wei et al., 2022) and In-Context Learning (ICL) via similarity-based example retrieval (Liu et al., 2022), as well as other single-agent optimization algorithms, including APE (Zhou et al., 2023), ORPO (Yang et al., 2024), ERM (Yan et al., 2025), and SPO (Xiang et al., 2025) (we reproduce the optimization logic of SPO in labeled scenarios). In multi-agent settings, we evaluate vanilla and ICL variants of the Annotator-Verifier architecture. We also adapt SPO, the strongest single-agent optimization baseline, to the same MAS architecture using the same batch schedule for a fair comparison.

Implementation Details We configure ANT-OR with 3 LLM-based annotators and one verifier, all at temperature=0.1. Each annotator is initialized with the same prompt template, while the verifier uses a separate adjudication prompt. Following the optimization procedure of SPO (Xiang et al., 2025), we train on 200 samples per task with batch size 20, 3 iterations per batch over 6 epochs, and select the best epoch based on training set performance. We evaluate on 1000 test samples; the remaining 300 samples are unused. For retrieval, we encode text and images with Qwen3-Embedding-8B (Zhang et al., 2025d) and Qwen3-VL-Embedding-8B (Li et al., 2026). Prompts are detailed in Appendix D.

Metrics. We evaluate annotation performance with two complementary metrics: **agreement@ ≥ 2** measures whether an annotation is supported by at least two human annotators. For model predictions, we count a prediction as correct if it matches a label annotated by at least two of the five human annotators. For human performance, we use a leave-one-annotator setting: each human annotation is treated as the prediction and is counted as correct if at least one of the other four annotators gives the same label. **Fleiss’ κ** measures inter-annotator consistency. We compute the human baseline from the five human annotators, and compute model κ by treating the model as an additional annotator alongside the humans. For Medical Entity Extraction,

Method	MedEE		DisPA		BillSC		SocialTC		ImageAR		ImageCI		ImageTR		Average	
	Acc	F1	Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ	Acc	κ
Human	82.76	82.09	93.86	38.77	94.66	83.50	76.82	44.62	94.06	58.59	83.54	81.27	96.86	56.37	88.94	63.60
Single-Agent Method for Data Annotation																
Vanilla	16.10	73.21	91.80	36.28	79.80	80.12	54.20	41.63	49.10	41.58	40.20	76.19	88.40	48.37	59.94	56.77
CoT (Wei et al., 2022)	24.30	74.07	92.40	38.59	83.30	81.03	57.80	42.92	57.80	45.03	45.10	77.20	88.20	48.22	64.13	58.15
ICL (Liu et al., 2022)	27.60	75.59	91.60	40.13	89.40	82.75	63.30	44.26	59.20	44.91	54.60	78.37	87.20	47.61	67.56	59.09
APE (Zhou et al., 2023)	25.90	73.68	91.50	40.96	90.20	83.14	52.90	41.45	81.30	53.85	60.20	79.12	84.50	46.48	69.50	59.81
ORPO (Yang et al., 2024)	28.10	73.60	93.00	39.82	84.20	81.53	55.70	42.45	81.30	54.59	56.20	78.80	83.60	45.31	68.87	59.44
ERM (Yan et al., 2025)	24.90	74.80	92.80	40.27	91.20	83.46	50.00	40.12	75.60	52.08	55.80	78.57	86.50	48.43	68.11	59.68
SPO (Xiang et al., 2025)	28.20	74.44	87.90	39.29	88.00	82.46	58.30	42.79	79.30	53.56	60.40	79.16	86.20	47.76	69.76	59.92
Multi-Agent Systems for Data Annotation																
Vanilla	25.10	74.80	92.80	40.25	84.60	81.59	60.40	43.48	58.40	44.73	48.00	77.73	88.40	48.34	65.39	58.70
ICL (Liu et al., 2022)	28.70	75.80	92.50	40.83	89.40	82.72	63.30	44.44	59.40	45.04	56.10	78.68	88.40	49.06	68.26	59.51
SPO (Xiang et al., 2025)	27.70	74.78	91.10	40.43	90.50	83.18	59.00	41.89	80.00	54.02	58.40	78.68	85.00	46.56	70.24	59.93
ANT-OR	35.90	76.10	93.50	41.05	92.60	83.82	65.00	44.50	86.40	56.61	60.80	79.22	89.00	50.15	74.74	61.64

Table 2: Performance comparison of different methods on data annotation tasks. We report Accuracy (agreement@ ≥ 2) (%) as the metric for annotation correctness, with Fleiss’ Kappa (κ) measuring inter-annotator consistency. For Medical Entity Extraction, where Fleiss’ κ is not applicable due to span-level entity annotations, we use Entity-Level F1 (Zhuang et al., 2024) instead. The best results are highlighted in **bold**.

where Fleiss’ κ is not applicable, we use **Entity-Level F1** (Zhuang et al., 2024) instead.

5.2 Main Results

Table 2 presents the comprehensive evaluation results comparing ANT-OR against single-agent and multi-agent baselines across all 7 datasets.

Consistent Superiority Across Tasks, Domains, and Modalities. ANT-OR achieves the best overall performance across all 7 datasets, outperforming the strongest baseline by 4.5 points in average agreement@ ≥ 2 accuracy. The gains are consistent across medical, e-commerce, and general-domain tasks, as well as both text and vision-language modalities. These results show that ANT-OR can effectively handle nuanced decision boundaries in diverse guideline-governed annotation settings.

Narrowing the Gap to Human-Level Consistency. ANT-OR not only improves point-wise consensus coverage (agreement@ ≥ 2), but also achieves higher distributional agreement with human judgments, as measured by Fleiss’ κ . Notably, ANT-OR achieves $\kappa = 56.61$ on ImageAR and $\kappa = 83.82$ on BillSC, approaching the inter-annotator agreement among human annotators themselves (58.59 and 83.50 in Table 1). This validates that, compared to static agent systems, optimized multi-agent systems can narrow the gap to human-level consistency on complex annotation tasks.

Method	MedEE		DisPA		BillSC		SocialTC		Avg.	
	Acc	F1	Acc	κ	Acc	κ	Acc	κ	Acc	κ
ANT-OR (Full)	35.90	76.10	93.50	41.05	92.60	83.82	65.00	44.50	71.75	61.37
I. Component Ablation										
w/o Dimension Mining	34.90	75.93	91.20	40.34	87.60	82.01	58.50	42.59	68.05	60.22
w/o Contrastive Triples	32.80	75.70	89.50	40.35	91.90	83.69	63.70	43.18	69.47	60.74
w/o Coupled Injection	27.40	75.18	93.40	40.90	93.50	84.04	63.80	44.07	69.53	61.04
II. Effect of Injected Sample Number (Default=2)										
1 sample	33.60	75.79	93.60	41.50	92.00	83.55	63.50	44.22	70.67	61.27
3 samples	35.40	76.36	93.10	40.95	92.20	83.66	64.20	44.17	71.22	61.29
5 samples	36.70	76.09	93.20	41.03	92.30	83.69	64.30	44.24	71.62	61.26
III. Effect of Retrieval Strategy										
Random	34.80	75.87	92.80	40.85	91.80	83.59	63.80	44.38	70.80	61.17
Similarity only	36.20	76.29	93.10	41.01	92.60	83.81	63.90	44.32	71.45	61.36

Table 3: Ablation and sensitivity analysis of ANT-OR on four text datasets.

Multi-Agent Stability and Effective Optimization. Multi-agent systems provide a more stable foundation for automated annotation: MAS consistently outperforms the basic single-agent CoT setup across all 7 datasets, with gains ranging from 0.2 to 2.9 points. Similarly, applying SPO to multi-agent systems improves average performance despite task-level variance. ANT-OR fully realizes this potential through agent-level structured knowledge base, achieving gains over the vanilla MAS baseline on all seven datasets, establishing multi-agent optimization as a reliable strategy for improving automated annotation performance.

5.3 Ablation Study

Effect of Dimension Mining. Discriminative dimensions are the core knowledge units in ANT-OR, capturing recurring decision boundaries from anno-

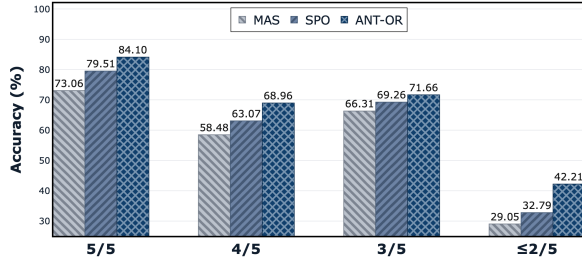


Figure 4: Average accuracy under different levels of human agreement.

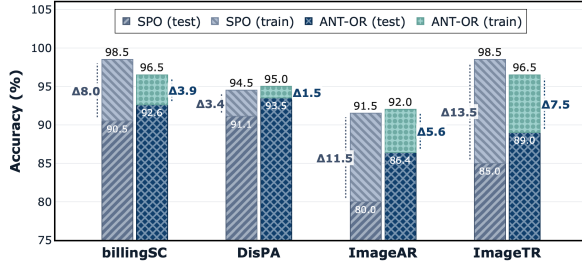


Figure 5: Generalization comparison of ANT-OR and SPO across datasets.

tation errors as reusable guidance. To evaluate their effect, we remove explicit dimension construction while keeping the rest of the pipeline unchanged. As shown in Table 3, performance drops consistently across all four datasets: average accuracy decreases from 71.75 to 68.05, and average κ from 61.37 to 60.22. The larger decline on BillSC and SocialTC suggests that decision-level abstractions are important for modeling stable annotation criteria beyond instance-level corrections.

Effect of Contrastive Triples. We further examine the role of contrastive triples in the mining stage. In the full model, each error case is paired with semantically similar positive cases to expose fine-grained label distinctions. As shown in Table 3, removing this construction consistently hurts performance, lowering average accuracy from 71.75 to 69.47 and average κ from 61.37 to 60.74. These results suggest that contrastive triples provide richer optimization signals than error cases alone.

Effect of Coupled Injection. We further examine coupled injection, which retrieves examples at inference time to contextualize discriminative dimensions. As shown in Table 3, it improves overall accuracy (69.53→71.75) and κ (61.04→61.37), though gains vary across tasks. This reflects differences in boundary complexity: retrieved examples are more helpful when category distinctions are subtle, but contribute less to rule-based tasks and

# Pos Samples	BillSC		MedEE		Average	
	Agree ≥ 2	Fleiss' κ	Agree ≥ 2	Fleiss' κ	Agree ≥ 2	Fleiss' κ
1	91.70	83.51	33.90	75.81	62.80	79.66
2	92.60	83.82	35.90	76.10	64.25	79.96
3	92.70	83.78	36.20	76.21	64.45	79.99

Table 4: Ablation study on the number of positive samples in contrastive mining.

may introduce noise. These results suggest that discriminative dimensions drive the main improvements, while retrieved samples provide complementary inference-time support.

Effect of Injected Sample Number and Retrieval Strategy. Table 3 shows that ANT-OR is relatively insensitive to the number of injected samples: varying from 1 to 5 yields only small performance changes. This suggests that injected examples provide contextual support, but their benefit saturates with few samples. For retrieval strategy, random selection performs worst, and our quality-aware similarity retrieval performs best overall, indicating that effective retrieval should consider both relevance and reliability.

Effect of positive samples. We use a 2:1 ratio of positive to negative samples in contrastive mining. Increasing the number of positive examples is intended to improve the accuracy of attribution. Table 4 presents the ablation study on the number of positive samples used in contrastive mining. The ablation shows that moving from 1 to 2 positive samples yields clear gains on average agreement (≥ 2), while adding a third brings only marginal improvements. This confirms that two positives are sufficient to stabilize decision patterns without unnecessary cost, making the 2:1 ratio the optimal trade-off.

5.4 Analysis

Analysis by Human Agreement Level. To assess performance under varying ambiguity, we group instances by maximum annotator agreement and report agreement@ ≥ 2 averaged across datasets in Figure 4. Performance declines for all methods as agreement decreases. SPO offers limited gains on highly ambiguous cases ($\leq 2/5$ agreement), suggesting that static optimization struggles with subtle decision boundaries. In contrast, ANT-OR improves substantially, outperforming MAS by nearly 8 points on this subset. This suggests that iterative contrastive optimization helps agents learn structured rules for guideline gray areas rather than merely fitting clear-cut cases.

Method	MedEE		DisPA		BillSC		SocialTC		Avg.	
	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost
SPO	27.70	73.87	91.10	54.32	90.50	33.86	59.00	101.71	67.08	65.94
ANT-OR	35.90	116.70	93.50	74.87	92.60	47.15	65.00	166.99	71.75	101.43

Table 5: Optimization Cost Analysis.

Generalization Performance. Figure 5 compares training and test performance of ANT-OR and SPO. Although SPO achieves higher training accuracy (e.g., 98.5% vs. 96.5% on BillSC), ANT-OR performs better on the test sets in Figure 5 by an average of 3.7 percentage points. This suggests that SPO tends to overfit the training data, while ANT-OR generalizes better to unseen instances.

Cost Analysis. Table 5 compares optimization costs between SPO and ANT-OR. The optimization cost is calculated based on DeepSeek’s official token pricing, converted to USD: \$0.26 per million input tokens (assuming 70% of the input tokens hit the cache) and \$1.80 per million output tokens for DeepSeek-V3.1-Terminus. While ANT-OR incurs a higher average cost (\$101.43 vs. \$65.94), it achieves a substantial accuracy improvement of 4.7 points (71.75% vs. 67.08%). The additional cost stems from multi-agent contrastive mining, knowledge base construction and sample injection.

6 Conclusion

We present ANT-A, a multi-annotator benchmark and ANT-OR, a multi-agent framework, to advance automated data annotation toward human-level consistency. Our benchmark spans 7 datasets across multiple domains and modalities, providing 5 independent human annotations per sample alongside detailed guideline documents, thereby enabling rigorous evaluation of inter-annotator agreement. To address the limitations of static agent architectures, ANT-OR introduces iterative prompt optimization with a structured knowledge base and contrastive learning over annotation samples, mining discriminative dimensions from contrastive prediction patterns to transform collective annotation experience into reusable decision rules. Extensive experiments demonstrate that ANT-OR significantly narrows the gap to human-level consistency, consistently outperforming single-agent and multi-agent baselines across diverse tasks.

Limitations

Broader Benchmark Coverage. Although our benchmark spans multiple domains and modali-

ties, it still represents only a subset of real-world annotation settings. Future work can extend this benchmark to cover a wider range of tasks, such as long-document annotation, open-ended judgment, and more specialized expert workflows, in order to further evaluate automated annotation systems under broader and more realistic conditions.

More Efficient Optimization. ANT-OR improves annotation reliability through multi-agent collaboration and iterative optimization, but this design also introduces additional computational overhead. A natural direction for future work is to develop more efficient optimization and inference strategies that preserve the benefits of structured multi-agent refinement while reducing optimization and annotation cost.

Stronger and More General Backbone Models. The current framework is built on the capabilities of the underlying LLMs for retrieval, reasoning, and prompt refinement. As backbone models continue to improve, it will be important to study how ANT-OR scales to stronger models and how well it transfers across different model families, domains, and annotation settings.

Ethics Statement

This work follows the ACL Ethics Policy. Our benchmark is constructed for research on reliable data annotation, and we take care to remove or anonymize personally identifiable or sensitive information before release. Human annotations are collected from professional annotators who are provided with task guidelines and clarification opportunities. Although ANT-OR improves LLM-based annotation reliability, it should not be used as a substitute for human experts in high-stakes domains such as medicine, finance, or legal decision-making. LLM-generated annotations may still contain errors or biases, so users should conduct task-specific privacy, bias, and reliability assessments before deployment.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026B0101100004), and Shenzhen Science and Technology Program (Grant No. KJZD20231023094700001, KQTD2024072910215406).

References

- Mohammad Abdollahi, Khandaker Rifah Tasnia, Soumit Kanti Saha, Jinqui Yang, Song Wang, and Hadi Hemmati. 2025. [Demystifying errors in llm reasoning traces: An empirical study of code execution simulation](#). *Preprint*, arXiv:2512.00215.
- Bakhtawar Ahtisham, Kirk Vanacore, Jinsook Lee, Zhuqian Zhou, Doug Pietrzak, and Rene F. Kizilcec. 2026. [Ai annotation orchestration: Evaluating llm verifiers to improve the quality of llm annotations in learning analytics](#). In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, LAK '26, page 447–456, New York, NY, USA. Association for Computing Machinery.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-vl technical report](#). *arXiv preprint arXiv:2511.21631*.
- Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. 2020. [Products-10k: A large-scale product recognition dataset](#). *arXiv preprint arXiv:2008.10545*.
- Minjeong Ban, Jeonghwan Choi, Hyangsuk Min, Nicole Hee-Yeon Kim, Minseok Kim, Jae-Gil Lee, and Hwanjun Song. 2026. [Completing missing annotation: Multi-agent debate for accurate and scalable relevant assessment for IR benchmarks](#). In *The Fourteenth International Conference on Learning Representations*.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. [We need to consider disagreement in evaluation](#). In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.
- Guanhua Chen, Yutong Yao, Lida A. Chan, Xuebo Liu, and Derek F. Wong. 2025. [SGIC: A self-guided iterative calibration framework for RAG](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *arXiv preprint arXiv:2412.19437*.
- Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. 2025. [Guidebench: Benchmarking domain-oriented guideline following for llm agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11361–11399, Vienna, Austria. Association for Computational Linguistics.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. [Improving factuality and reasoning in language models through multiagent debate](#). In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Chrisantha Fernando, Dylan Sunil Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. [Promptbreeder: Self-referential self-improvement via prompt evolution](#). In *Forty-first International Conference on Machine Learning*.
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023a. [When the majority is wrong: Modeling annotator disagreement for subjective tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023b. [When the majority is wrong: Modeling annotator disagreement for subjective tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Fabrizio Gilardi, Meysam Alizadeh, and Maïl Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. [Connecting large language models with evolutionary algorithms yields powerful prompt optimizers](#). In *The Twelfth International Conference on Learning Representations*.
- Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. [AnnoLLM: Making large language models to be better crowdsourced annotators](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.
- Mahmood Hegazy, Aaron Rodrigues, and Azzam Naeem. 2025. [Mafa: A multi-agent framework for annotation](#). *Preprint*, arXiv:2505.13668.
- Hong Jia, Weibin Li, Jingyao Wu, Xiaofeng Yu, Yan Gao, Jintao Cheng, Xiaoyu Tang, Feng Xia, and Ting Dang. 2026. [Decoding ambiguous emotions with test-time scaling in audio-language models](#). *arXiv preprint arXiv:2602.03873*.

- Xiaopeng Ke, Hexuan Deng, Xuebo Liu, Jun Rao, Zhenxi Song, Jun Yu, and Min Zhang. 2025. [AQUILT: Weaving logic and self-inspection into low-cost, high-relevance data synthesis for specialist LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Doyoung Kim, Zhiwei Ren, Jie Hao, Zhongkai Sun, Lichao Wang, Xiyao Ma, Zack Ye, Xu Han, Jun Yin, Heng Ji, Wei Shen, Xing Fan, Benjamin Yao, and Chenlei Guo. 2026. [Beyond perfect apis: A comprehensive evaluation of llm agents under real-world api complexity](#). *Preprint*, arXiv:2601.00268.
- Fengheng Li, An Liu, Wei Feng, Honghe Zhu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junjie Shen, Zhangang Lin, and Jingping Shao. 2023a. [Relation-aware diffusion model for controllable poster layout generation](#). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*.
- Jianquan Li, Xidong Wang, Xiangbo Wu and Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. [Huatu0-26m, a large-scale chinese medical QA dataset](#). *arXiv preprint arXiv:2305.01526*.
- Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibao Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. [Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking](#). *arXiv preprint arXiv:2601.04720*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.
- Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, and 6 others. 2021. [M6: A chinese multimodal pretrainer](#). *arXiv preprint arXiv:2103.00823*.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024a. [SelectIT: Selective instruction tuning for LLMs via uncertainty-aware self-reflection](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
- Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2024b. [Can LLMs learn uncertainty on their own? expressing uncertainty effectively in a self-training manner](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
- Navya Mehrotra, Adam Visokay, and Kristina Gligorić. 2026. [Multi-perspective llm annotations for valid analyses in subjective tasks](#). *Preprint*, arXiv:2603.21404.
- Benedetta Muscato, Yue Li, Gizem Gezici, Zhixue Zhao, and Fosca Giannotti. 2025. [Bridging the gap: In-context learning for modeling human disagreement](#). *Preprint*, arXiv:2506.06113.
- Melissa Z Pan, Mert Cemri, Lakshya A Agrawal, Shuyi Yang, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Kannan Ramchandran, Dan Klein, Joseph E. Gonzalez, Matei Zaharia, and Ion Stoica. 2025. [Why do multiagent systems fail?](#) In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chengguang Zhu, and Michael Zeng. 2023. [Automatic prompt optimization with “gradient descent” and beam search](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.
- Jun Rao, Yunjie Liao, Xuebo Liu, Zepeng Lin, Lian Lian, Dong Jin, Shengjun Cheng, Jun Yu, and Min Zhang. 2025a. [SeaPO: Strategic error amplification for robust preference optimization of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16540–16557, Suzhou, China. Association for Computational Linguistics.
- Jun Rao, Zepeng Lin, Xuebo Liu, Xiaopeng Ke, Lian Lian, Dong Jin, Shengjun Cheng, Jun Yu, and Min Zhang. 2025b. [APT: Improving specialist LLM performance with weakness case acquisition and iterative preference training](#). In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. [Snorkel: rapid training data creation with weak supervision](#). *Proc. VLDB Endow.*, 11(3):269–282.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.

- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Man-sooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.
- Xinyu Tang, Xiaolei Wang, Wayne Xin Zhao, Siyuan Lu, Yaliang Li, and Ji-Rong Wen. 2025. [Unleashing the Potential of Large Language Models as Prompt Optimizers: Analogical Analysis with Gradient-Based Model Optimizers](#). In *AAAI Conference on Artificial Intelligence*, pages 25264–25272.
- Chunhao Tian, Yutong Wang, Xuebo Liu, Zhexuan Wang, Liang Ding, Miao Zhang, and Min Zhang. 2025. [AgentInit: Initializing LLM-based multi-agent systems via diversity and expertise orchestration for effective and efficient collaboration](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*.
- Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D. Nguyen. 2025. [Multi-agent collaboration mechanisms: A survey of llms](#). *Preprint*, arXiv:2501.06322.
- Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. [Evaluating large language models as expert annotators](#). *arXiv preprint arXiv:2508.07827*.
- Jiayu Wang and Junyoung Lee. 2026. [More human, more efficient: Aligning annotations with quantized slms](#). *Preprint*, arXiv:2604.00586.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*. Association for Computational Linguistics.
- Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. [Human-llm collaborative annotation through effective verification of llm labels](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI ’24*, New York, NY, USA. Association for Computing Machinery.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Zhexuan Wang, Xuebo Liu, Li Wang, Zifei Shan, Yutong Wang, Zhenxi Song, and Min Zhang. 2026. [MASPO: Joint prompt optimization for LLM-based multi-agent systems](#). In *Proceedings of the 43rd International Conference on Machine Learning (ICML 2026)*.
- Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. 2025. [Agent-Dropout: Dynamic agent elimination for token-efficient and high-performance LLM-based multi-agent collaboration](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Jingyao Wu. 2026. [AmbER2: Dual ambiguity-aware emotion recognition applied to speech and text](#). *arXiv preprint arXiv:2601.18010*.
- Jinyu Xiang, Jiayi Zhang, Zhaoyang Yu, Xinbing Liang, Fengwei Teng, Jinhao Tu, Fashen Ren, Xiangru Tang, Sirui Hong, Chenglin Wu, and Yuyu Luo. 2025. [Self-supervised prompt optimization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9017–9041, Suzhou, China. Association for Computational Linguistics.
- Yinuo Xu, Veronica Derricks, Allison Earl, and David Jurgens. 2025. [Modeling annotator disagreement with demographic-aware experts and synthetic perspectives](#). *arXiv preprint arXiv:2508.02853*.
- Cilin Yan, Jingyun Wang, Lin Zhang, Ruihui Zhao, Xiaopu Wu, Kai Xiong, Qingsong Liu, Guoliang Kang, and Yangyang Kang. 2025. [Efficient and accurate prompt optimization: the benefit of memory in exemplar-guided reflection](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 753–779, Vienna, Austria. Association for Computational Linguistics.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. [Large language models as optimizers](#). In *International Conference on Learning Representations*, volume 2024, pages 12028–12068.
- Chenglin Yu, Yang Yu, Songmiao Wang, Yucheng Wang, Yifan Yang, Jinjia Li, Ming Li, and Hongxia Yang. 2025. [Infiagent: Self-evolving pyramid agent framework for infinite scenarios](#). *Preprint*, arXiv:2509.22502.
- Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2025a. [Cut the crap: An economical communication pipeline for LLM-based multi-agent systems](#). In *The Thirteenth International Conference on Learning Representations*.

- Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. 2025b. [G-designer: Architecting multi-agent communication topologies via graph neural networks](#). In *Forty-second International Conference on Machine Learning*.
- Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. [WRENCH: A comprehensive benchmark for weak supervision](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Michael JQ Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2025c. [Diverging preferences: When do annotators disagree and do models know?](#)
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025d. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.
- Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. 2025. [Lifelong learning of large language model based agents: A roadmap](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:5552–5571.
- Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Serkan O Arik. 2025. [Multi-agent design: Optimizing agents with better prompts and topologies](#). In *International Conference on Machine Learning*. PMLR.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. [Large language models are human-level prompt engineers](#). In *The Eleventh International Conference on Learning Representations*.
- Yan Zhuang, Junyan Zhang, Ruogu Lu, Kunlun He, and Xiuxing Li. 2024. [Medner: Enhanced named entity recognition in medical corpus via optimized balanced and deep active learning](#). *ACM Trans. Intell. Syst. Technol.*, 15(5).

A Multi-Agent Architecture Comparison

Table 6 compares different multi-agent system (MAS) architectures for data annotation. We analyze the failure modes of alternative designs:

Debate-based Architecture. Following Du et al. (2024), we implement a multi-agent debate where agents iteratively discuss and refine their predictions. While debate improves reasoning on open-ended tasks, it performs poorly for guideline-governed annotation: (1) debates often diverge from the original task into tangential arguments, amplifying rather than correcting initial errors; (2) the lack of explicit role separation leads to group-think, where agents converge on consensus rather than correct answers; (3) computational costs scale quadratically with debate rounds, making it economically impractical for large-scale annotation.

Planner-Worker Architecture. We implement a Planner-Worker design following standard task-decomposition patterns (Wang et al., 2023a), where a planner agent divides inputs into sub-tasks and delegates to worker agents, then aggregates results. Contrary to expectations, this architecture underperforms single-agent baselines across all tasks (-1.1% on MedEE, -3.7% on DisPA, -0.4% on BillSC). This failure may be attributed to coordination noise in guideline interpretation. Unlike simple task decomposition where sub-tasks are independent, annotation tasks require holistic understanding of interdependent category definitions and boundary rules. The planner’s decomposition forces premature partitioning of decision logic: workers process truncated contexts that lack crucial disambiguating information from the full guidelines, while the aggregation step (simple concatenation or voting) lacks the capability to perform guideline-grounded adjudication. This creates a compounding error effect: the planner introduces interpretation errors at the decomposition stage, workers amplify these errors with incomplete context, and the aggregation fails to reconcile genuine disagreements. While costs are reduced compared to full debates (34-50% lower), the accuracy degradation renders this architecture unsuitable for high-stakes annotation tasks.

Annotator-Verifier Architecture. Consistent with prior findings that annotator-verifier architectures enhance reliability (Hegazy et al., 2025; Ahtisham et al., 2026), ANT-OR uses parallel annotators to diversify perspectives while a dedicated

verifier adjudicates conflicts through guideline-grounded reasoning. This design achieves 1-2% accuracy improvement over baselines with only 3-4 $\times$ cost increase (vs. 10-50 $\times$ for alternatives), demonstrating superior accuracy-cost efficiency for annotation tasks.

B Case Study

In this appendix, we present two representative cases from Billing Scenario Classification Task, to illustrate (1) the realism of our benchmark and (2) the qualitative improvements brought by ANT-OR. To keep the case tables focused and readable, we present the annotation guideline for this task in the table below.

Case 1: A boundary instance with human disagreement. Table 7 shows an example where annotators disagree despite following the same annotation guideline. This case highlights an important property of real-world annotation: some instances lie near genuine decision boundaries, and their difficulty cannot be fully captured by a single gold label. By collecting five independent annotations per sample, our benchmark makes such ambiguity explicit and enables more realistic evaluation of annotation systems.

Case 2: Improvement after optimization. Table 11 compares the predictions of a vanilla multi-agent system (MAS) and ANT-OR on another difficult example from the same task. The baseline prediction is driven mainly by coarse lexical or topical cues, while ANT-OR gives the correct label by following the task guideline more precisely. This example qualitatively shows how iterative optimization improves annotation reliability.

C Benchmark Construction and Annotation Details

This appendix provides comprehensive details on the construction, annotation process, and characteristics of the ANT-A benchmark, supplementing the descriptions in the main paper.

C.1 Dataset Overview and Task Selection

The construction of the seven tasks takes into account different modalities, domains, and scenario requirements. They are obtained by collecting suitable queries according to annotation guidelines used in real application scenarios and then re-annotating them. Most of the queries come from

Method	MedEE		DisPA		BillSC		Average	
	Acc	Cost	Acc	Cost	Acc	Cost	Acc	Cost
CoT	24.3	7.00	92.4	6.83	83.3	8.18	66.67	7.34
Debate	16.7	66.65	86.5	78.34	76.1	77.36	59.77	74.12
Planner-Worker	23.2	37.07	88.3	34.11	82.9	33.31	64.80	34.83
Annotator-Verifier	25.1	29.16	92.8	22.08	84.6	22.80	67.50	24.68

Table 6: Performance comparison on different MAS (w/o Prompt Optimization).

Task	Billing Scenario Classification
Input	Merchant Name: Leisure Tea House Bill Title: Leisure Tea House Counterparty: —
Human Annotations	[Food & Dining, Culture & Leisure, Food & Dining, Food & Dining, Culture & Leisure]
Observation	This case captures the inherent ambiguity of crossover business formats at the semantic boundary. According to the annotation rules, Food & Dining covers “beverages,” while Culture & Leisure encompasses “leisure venue consumption, including cinemas, bars, chess and card rooms, etc.” A tea house simultaneously operates as a beverage provider (Food & Dining) and a leisure venue (Culture & Leisure). Three annotators prioritized the beverage service aspect, whereas two annotators categorized it as a leisure venue analogous to a bar or chess room. This 3:2 split reflects the difficulty of assigning a single category to establishments whose core offering inherently spans multiple semantic boundaries.

Table 7: A representative boundary case with human disagreement in Billing Scenario Classification Task.

existing datasets or web sources, but they are re-annotated according to the guidelines of different tasks.

Our dataset consists of two parts: annotation guidelines and queries. We first created the annotation guidelines, which were adapted from actual annotation requirements used in real-world scenarios, rather than being derived from existing datasets or manually fabricated from scratch. For example, the billing scenario classification task is used for bill-category identification in payment scenarios, and image-text relevance can be used in image search scenarios. We then collected corresponding queries according to the tasks described in the guidelines. After organizing the guidelines and queries, we manually annotated the queries according to the guidelines.

Table 8 presents detailed information about the seven tasks.

C.2 Annotation Guidelines and Query Characteristics

The annotation guidelines are adapted from actual annotation requirements used in real production annotation scenarios (see the Task Usage column in Table 8 for their actual applications), rather than being manually fabricated from scratch. For example, the billing scenario classification task is used for bill-category identification in real payment scenarios, and the image-text relevance task is used in image search scenarios. The queries in different datasets are deduplicated and anonymized. We use a locally deployed DeepSeek-V3.1-Terminus to evaluate and anonymize sensitive information, and further conduct manual spot checks and a pre-annotation stage to ensure that the queries are suitable for annotation under the tasks specified by the guidelines.

It should be emphasized that the complexity of real annotation tasks does not necessarily equate to extremely long guidelines. Many real-world

Dataset	Domain	Modality	Task Usage	Query Source
BillSC	E-commerce	Text	Identifying bill types in payment scenarios	Anonymized billing records
MedEE	Medical	Text	Medical Content Analysis and Structuring	Huatuo-26M (Li et al., 2023b)
DisPA	Medical	Text	Privacy Protection and Content Review	QABasedOnMedicaKnowledgeGraph (GitHub)
SocialTC	General	Text	Knowledge Classification and Content Layering	Zhihu Website
ImageTR	General	Vision-Language	Search images by text	Paraphrased from MUGE (Lin et al., 2021) via LLM; images via Bing Image Search API
ImageCI	E-commerce	Vision-Language	Sorting and managing poster materials	CGL-Dataset (Li et al., 2023a)
ImageAR	General	Vision-Language	Score Image Identifiability	Products-10K (Bai et al., 2020)

Table 8: Overview of the seven tasks in the ANT-A benchmark.

production annotation tasks are based on operational rules centered around specific business goals, such as bill-type identification, image-text relevance judgment, or image recognizability assessment. These guidelines may not be hundreds of pages long, but they require annotators to handle a large number of boundary cases in real samples. Our guidelines are derived from such real annotation requirements. They specify the task objectives, category boundaries, and judgment criteria, but the real difficulty lies in the fact that guidelines can hardly exhaustively cover all possible sample cases: sample information may be incomplete, categories may overlap, and the text-image pair may be supported by only partial evidence. As a result, multiple annotators may provide different yet reasonable judgments. We retain five independent annotation results precisely to characterize this real-world uncertainty. Therefore, annotator disagreement itself is part of the value of the benchmark.

Table 9 presents detailed statistics about the guidelines, queries, and annotation agreement distributions. The lower agreement in SocialTC and ImageAR is expected, as these tasks involve hierarchical semantic classification, subjective content layering on open-domain social text, and visual identifiability assessment, where multiple labels can be defensible due to inherently ambiguous semantic boundaries, a property that single-label benchmarks collapse but our multi-annotator design explicitly preserves.

C.3 Annotation Process

Our three-stage annotation process is designed to differentiate meaningful disagreement from ambiguity caused by unclear guidelines.

Phase 1–2: Pilot Annotation and Guideline Refinement. We use approximately 50 pilot samples to expose ambiguities and missing rules, which are resolved through Q&A and guideline revision before formal annotation begins.

Phase 3: Formal Annotation. After guideline refinement, formal annotation is conducted by five independent annotators per sample.

This process is further supported by the annotation statistics shown in Table 9: the proportion of instances where all 5 annotators give completely different labels is about 1% on average. If disagreements were primarily caused by unclear guidelines, we would expect a much higher rate of random divergence. The fact that full disagreement is rare indicates that most samples reach a consensus under the refined guidelines, and the retained disagreements are concentrated in specific boundary cases that resist unique adjudication. Thus, the disagreements in our experiments reflect differences that persisted after guideline refinement, not systematic ambiguity.

C.4 Annotator Demographics and Quality Control

All annotators are recruited from a professional annotation company. Table 10 presents the background information of the human annotators.

All annotators hold a bachelor’s degree and have at least two years of annotation experience. Their competence is verified through the pilot annotation phase before formal labeling. We note that the two medical-related tasks (DisPA and MedEE) do not require professional medical expertise; their annotation objectives are grounded in explicit textual patterns and predefined schemas, which general an-

Dataset	Avg Guideline Length (chars)	Avg Query Length (chars)	4–5 Agree (%)	2–3 Agree (%)	No Agree (%)
BillSC	9,236	160	86.1	13.9	0.0
MedEE	4,624	261	61.1	33.9	5.0
DisPA	4,499	22	80.1	19.9	0.0
SocialTC	27,529	2,700	44.8	53.3	1.9
ImageTR	3,907	32	89.9	10.1	0.0
ImageCI	7,668	N/A	82.5	17.5	0.0
ImageAR	12,487	N/A	62.3	36.6	1.1

Table 9: Statistics of annotation guidelines, query lengths, and inter-annotator agreement distributions. We define 4–5 agreement as cases where at least four annotators selected the same label, 2–3 agreement as cases where the majority label was selected by two or three annotators, and no agreement as cases without a dominant label.

ID	Age	Degree	Expertise	Years of Experience
1	37	Bachelor	Law	6
2	33	Bachelor	Computer Application Technology	2.5
3	29	Bachelor	Business Administration	2.5
4	25	Bachelor	Electronic Information Engineering Technology	2
5	27	Bachelor	Computer Science and Technology	2
6	26	Bachelor	E-Commerce	2
7	28	Bachelor	Public Administration / Administrative Management	2
8	30	Bachelor	Human Resource Management	2
9	28	Bachelor	Electronic Information Engineering	2
10	28	Bachelor	Land Resource Management	2

Table 10: Demographics of the human annotators.

notators can accurately perform after being trained with the detailed guidelines.

C.5 Qualitative Analysis of Prompt Optimization

To complement the quantitative results, we provide a qualitative analysis of how the prompts evolve through the optimization process. Below is a representative example from the Billing Scenario Classification task, illustrating how the annotation prompt improves across optimization epochs.

Example: Merchant Name Specificity Threshold in Billing Scenario Classification.

Epoch 0 (Simple condition):

“When merchant name contains explicit industry keywords, classify by industry attribute.”

Epoch 2 (Added anti-examples):

“When merchant name is too generic (e.g., only contains ‘company’, ‘industry’, ‘group’ without industry indication) AND transaction description is ambiguous → classify as ‘Other’.”

Epoch 5 (Balanced two-tier rule):

“When merchant name contains explicit industry keywords → prioritize industry over transaction nature; when overly generic + ambiguous content → ‘Other’.”

Why it matters: This evolution prevents both false positives (over-eager industry classification) and false negatives (missing valid industry signals), allowing the model to better align with human understanding.

D Prompts

We present all the prompts used in our paper in the following tables, including:

- **Prompts for Basic Annotation (Single Annotator):** The prompt is used to perform standard annotation when no contrastive samples or dimensions are available. The system provides the base annotation rules and the query to be annotated, then the LLM generates the annotation result with step-by-step reasoning.
- **Prompts for Dimension-Aware Annotation with Contrastive Samples:** The prompt is used to perform annotation when discriminative dimensions and contrastive samples are

retrieved from the example pool. The system provides the base rules, relevant dimensions (rule intent, decision boundary, applicable boundary), and reference samples (both correct and incorrect demonstrations) to guide the LLM in making more accurate annotations by learning from historical error patterns.

- **Prompts for Verification Agent:** The prompt is used to select the best annotation result from multiple annotators by comparing their reasoning logic. When annotators disagree, the LLM evaluates each annotator's reasoning for logical consistency, coherence, and errors, then selects the most sound reasoning. This serves as a meta-judge to resolve disagreements among the annotator ensemble.
- **Prompts for Contrastive Dimension Mining:** The prompt is used to analyze contrastive triples (two correctly judged samples and one incorrectly judged sample that are semantically similar) to uncover the fundamental intent behind rule formulation. The LLM identifies surface differences, decision boundaries, rule intent, meta-rules, and applicable boundaries, producing structured dimension definitions that capture the core logic distinguishing correct from incorrect judgments.
- **Prompts for Dimension Field Refinement:** The prompt is used to integrate newly analyzed contrastive cases with existing discriminative dimensions. The LLM determines the relationship between new cases and historical rules (support, supplement, correction, or extension), extracts generalizable rule intent, resolves conflicts, and tests counterexamples to produce refined rule_intent, decision_boundary, and applicable_boundary definitions.
- **Prompts for Verifier Prompt Optimization:** The prompt is used to optimize the verification agent's selection criteria based on error analysis. The system provides current verification errors, learned discriminative dimensions, and contrastive samples as auxiliary reference. The LLM analyzes root causes of verification failures and produces an improved verification prompt that can more accurately select the best annotator result by checking reasoning logic.
- **Prompts for Annotator Prompt Optimization:** The prompt is used to optimize the annotation instructions based on error analysis. The system provides the existing annotation prompt, current error samples (core optimization basis), learned discriminative dimensions, and contrastive samples as auxiliary reference. The LLM analyzes root causes of annotation errors and produces an improved annotation prompt that directly addresses the identified problems while following optimization principles.

Task	Billing Scenario Classification
Input	Merchant Name: Convenience Retail Store Bill Title: Personal Transfer Receipt Counterparty: Individual Customer
Human Annotations	[Daily Necessities, Daily Necessities, Daily Necessities, Daily Necessities, Daily Necessities]
MAS	Result: Transfer & Red Envelope Reasoning: Focused on the surface-level keyword “transfer” in the bill title and treated the transaction as a peer-to-peer fund transfer between individuals, completely overlooking the merchant’s industry attribute implied by “Retail Store.”
ANT-OR	Result: Daily Necessities Reasoning: Applied the dimension-based rule that merchant names containing explicit industry keywords (e.g., “retail store”) should take priority over payment-behavior descriptions in the title. Referenced analogous cases where merchants with clear industry labels were correctly classified by business type despite transfer/receipt descriptors.
Observation	This case demonstrates a classic failure mode of literal keyword matching. MAS is misled by the word “transfer” in the bill title and ignores the structural priority between merchant industry and transaction behavior. In contrast, ANT-OR’s multi-dimensional reasoning framework correctly elevates the merchant name’s explicit industry signal (“retail store” → <i>Daily Necessities</i>) over the payment-operation description. The unanimous human consensus (5/5) on <i>Daily Necessities</i> confirms that ANT-OR’s reasoning aligns with human judgment while MAS falls into the trap of superficial pattern matching.

Table 11: Qualitative comparison between MAS and ANT-OR on Billing Scenario Classification Task.

Billing Scenario Classification Guidelines (Part 1)

Bill Scenario Classification Task - Annotation Rules Document

I. Rules and Examples

1.1 Annotation Basics

Categorization should be determined based on a comprehensive assessment of the transaction content. The transaction content must clearly correspond to the definition of a specific category; otherwise, it should be classified as "Others."

1.2 Data Description

Transaction Content: Describes the specifics of the current transaction, indicating its exact nature.

1.3 Annotation Instructions

Category Names and Definitions

Level-1 Category Name	Level-1 Category Semantics
-----------------------	----------------------------

_____	_____
-------	-------

Food & Dining	Encompasses meals, fast food, beverages, fresh food, snacks, and tobacco-related consumption and services.
---------------	--

Apparel & Accessories	Retail consumption of men's and women's clothing, shoes, bags, and jewelry, covering specialty stores, e-commerce platforms, and accessory brands.
-----------------------	--

Daily Necessities	Retail services for daily necessities, including department stores, convenience stores, flowers and plants, and community group-buying consumption.
-------------------	---

Home Furnishings & Renovation	Covers furniture, building materials, and renovation services; provides home textiles, lighting, and other household products to optimize living spaces.
-------------------------------	--

Electronics & Appliances	Covers major appliances, small appliances, and 3C digital product purchases and related consumption.
--------------------------	--

Sports & Outdoors	Covers gym memberships, yoga and dance, sports venues, and outdoor sports equipment-related consumption.
-------------------	--

Beauty & Hair Care	Covers beauty and hair care-related consumption, including beauty and personal care, skincare, facial masks, nail art, as well as haircuts, perms, and coloring services and products.
--------------------	--

Mother & Baby	Consumption of infant and child products, children's clothing, toys, and related maternal and infant goods and services.
---------------	--

Billing Scenario Classification Guidelines (Part 2)

| Pets | Covers pet-related service consumption, including pet supplies, pet toys, pet hospitals, pet clothing, and grooming services. |

| Transportation| Covers transportation and travel consumption, including intercity transport such as air tickets, ferry tickets, train tickets, and bus tickets, as well as urban travel such as subway, bus, and toll fees. |

| Vehicle Maintenance | Fuel and charging, maintenance and repair, and parking services; covers auto repair shops, gas stations, charging stations, and sales of car/bicycle accessories. |

| Housing & Property| Rent, property fees, and agency services; covers property management, real estate leasing, and agency services. |

| Hotels & Travel| Covers hotel accommodation and travel-related consumption, including scenic spot tickets, hotels and homestays, travel packages, tour services, and annual passes. |

| Culture & Leisure| Covers film and entertainment, performances and events, and leisure venue consumption, including cinemas, bars, chess and card rooms, etc., for both offline venues and online platforms. |

| Education & Training | Covers education and training-related consumption, including tuition payment, online course purchases, public and private educational institutions, and training services. |

| Medical & Health | Covers hospitals and clinics, physical examination and health care, medical beauty, medical equipment, and nutritional supplements-related services and consumption. |

| Lifestyle Services | Addresses transactional life needs, such as rentals, logistics, housekeeping and repairs, same-city delivery, legal and wedding services, and other specialized service fees. |

| Public Services | Fees for administrative, judicial, social security, housing provident fund, and other public affairs management provided by government agencies, as well as non-profit organization services. |

| Business Services | Fees arising from corporate operational support such as finance, transportation, marketing, and tax consulting, as well as e-commerce transaction guarantee services. |

| Charity & Donations | Transactions for charitable donations, such as foundations (e.g., One Foundation), charity platforms (e.g., Shuidichou), or temples (e.g., incense donations), or charitable fundraising (e.g., rural primary schools). |

| Mutual Aid & Protection | Payments for mutual aid protection plans, such as online mutual aid (e.g., Xianghubao sharing) or community insurance (e.g., neighborhood medical fund) consumption. |

| Investment & Wealth Management | Covers financial product purchases, investment services, and financial institution-related transactions and expense payments. |

| Insurance | Covers various insurance products and services, including policy purchases, claims, compensation, and related protection expenses. |

Billing Scenario Classification Guidelines (Part 3)

Credit & Rental	Covers shared rentals, lending platforms, and financial leasing-related services and expense payments.
Recharge & Utilities	Covers utilities (water, electricity, gas), communication, and network recharge services, including public utility payments, phone credit and data top-ups, and telecommunications operator services.
Income	Various types of interest income.
Transfer & Red Envelope	Content related to transfers and red envelopes.
Payment on Behalf of Relatives/Friends	Content related to payments made on behalf of relatives or friends.
Account Deposit & Withdrawal	Content related to account deposits and withdrawals.
Refund	Transactions explicitly containing the word “refund” in the title, or other refund bills.
Others	Content that does not belong to any of the above level-1 categories.

Prompts for Basic Annotation (Single Annotator)

System: You are a helpful assistant.

User:

{prompts/guidelines}

Query to be Annotated:

{query}

Please output in the following format:

Reasoning Process: [Step-by-step reasoning]

Annotation Result: {output_format}

Prompts for Dimension-Aware Annotation with Contrastive Samples (Part 1)

System: You are a helpful assistant.

User:

{prompts/guidelines}

Key Discriminative Dimensions (Summarized from Historical Errors)

The following dimensions are associated with the provided reference samples. Please annotate by combining dimension understanding with sample demonstrations:

[Dimension 1: {dim_name}]

- Rule intent: {rule_intent}
- Decision boundary: {decision_boundary}
- Applicable scenario: {applicable_boundary}

[Dimension 2: ...]

Reference Samples (Associated with the Above Dimensions)

[Reference Sample 1 (Correct Demonstration)] (Related dimensions: {dim_names})

[Quality Tag] (similarity: {sim_score})

Query: {query}

Annotation result: {annotation}

Correct answer: {ground_truth}

[Reference Sample 2 (Incorrect Demonstration)] (Related dimensions: {dim_names}) [Quality Tag] (similarity: {sim_score})

Query: {query}

Incorrect annotation: {annotation}

Correct answer: {ground_truth}

Error reason: {error_reason}

[Contrastive Analysis Guidance]

Before making your judgment, carefully compare:

1. Core similarities between the current query and the correct sample(s)
2. Key differences between the current query and the incorrect sample(s)
3. How to avoid the mistakes made in the incorrect samples
4. Look for consistent patterns across multiple samples to improve judgment confidence

Usage Guidance

1. First understand the rule intent and decision boundary of key discriminative dimensions
2. Refer to correct samples to understand how to judge, refer to incorrect samples to learn common pitfalls
3. Combine with the rule document to make an independent judgment on the current Query

Prompts for Dimension-Aware Annotation with Contrastive Samples (Part 2)

Query to Annotate:
{query}

Please output in the following format:

Reasoning process: [Step-by-step analysis of Query features, relevant dimensions, and comparison with reference samples]

Annotation result: {output_format}

Prompts for Final Judgment Aggregation (Part 1)

System: You are a helpful assistant.

User:

You are a data annotation expert. Your task is to carefully analyze the query to be annotated and multiple candidate annotation results, and provide the final annotation result for the query.

Original Rule Document:
{doc_content}

Query to be Annotated:
{query}

Candidate Annotation Results:

Annotation result 1:

{annotation_result}

Annotation result 2:

{annotation_result}

Annotation result 3:

{annotation_result}

Please note that the rule document serves only as a simple reference. Please carefully compare multiple candidate annotations and provide the final annotation result based on these candidate results.

Judgment Suggestions:

Key Discriminative Dimensions (Summarized from Historical Errors)

[Dimension 1: {dim_name}]

- Rule intent: {rule_intent}
- Decision boundary: {decision_boundary}
- Applicable scenario: {applicable_boundary}

Prompts for Final Judgment Aggregation (Part 2)

Similar Sample Reference

[Reference Sample 1 (Correct Demonstration)]

Query: {query}

Annotation result: {annotation}

Correct answer: {ground_truth}

[Reference Sample 2 (Incorrect Demonstration)]

Query: {query}

Incorrect annotation: {annotation}

Correct answer: {ground_truth}

Error reason: {error_reason}

Please output in the following format:

Analysis Process: [Your analysis of the query to be annotated and candidate annotation results]

Annotation Result: {output_format}

Prompts for Verification Agent

System: You are a helpful assistant.

User:

Check the reasoning logic of multiple annotators and select the best result.

Focus on:

1. Compare the reasoning logic of each annotator
2. Check for logical consistency and coherence
3. Identify any reasoning errors or inconsistencies
4. Select the annotator with the most sound reasoning

Note: When all three annotators disagree (three-way tie), you must select the best one based on reasoning quality.

Query: {query_str}

[Key Discriminative Dimensions]

[Dimension 1: {dim_name}]

- Rule intent: {rule_intent}
- Decision boundary: {decision_boundary}
- Applicable scenario: {applicable_boundary}

Annotations:

=== annotator1 (label: {label}) ===
{annotation_result}

=== annotator2 (label: {label}) ===
{annotation_result}

=== annotator3 (label: {label}) ===
{annotation_result}

Vote Info: Majority vote: {winning_label} ({winning_count}/{total_votes})

OR Vote: Three-way tie (1:1:1), no majority

[Similar Sample Reference]

[Reference samples with correct/incorrect demonstrations]

Output format:

```
““json
{
  "reason": ["logic error description for each annotator"],
  "result": ["annotator1/annotator2/annotator3"]
} ““
```

Prompts for Contrastive Dimension Mining (Part 1)

System: You are a helpful assistant.

User:

Analyze this group of cases that are 'similar but judged differently' to deeply uncover the fundamental intent behind the rule formulation.

These samples are very similar in semantic input, but some are correctly judged while others are incorrectly judged.

Original Rule Document:
{doc_content}

Historical Discriminative Dimensions:
Before inducing new dimensions, please check the following existing discriminative dimensions. If you find that the discriminative dimension reflected in the current case is substantially the same as a historical dimension (i.e., describing the same type of meta-rule), please directly use that historical dimension's meta-rule as the name instead of creating a new name.

{existing_dims_text}

Case group to analyze:
[Contrastive Group]
Sample A (Correctly judged):
Query: {query}
Model output: {annotation}
Correct answer: {ground_truth}

Sample B (Correctly judged):
Query: {query}
Model output: {annotation}
Correct answer: {ground_truth}

Sample C (Incorrectly judged):
Query: {query}
Model output: {annotation}
Correct answer: {ground_truth}

Please deeply analyze this contrastive case group to uncover the fundamental intent behind rule formulation:

Prompts for Contrastive Dimension Mining (Part 2)

1. [Surface Differences] What are the literal differences between samples? Especially between correct and incorrect samples
2. [Decision Boundary] What factors cause samples to be judged as correct or incorrect? (How should similar samples be annotated?)
3. [Rule Intent] What is the core reason why rule makers require this annotation? (i.e., what is the general core annotation logic?)
4. [Meta-Rule] If this distinguishing principle is abstracted into a general rule, how should it be expressed?
5. [Applicable Boundary] Under what circumstances does this rule apply? When might it not apply?

Please note: Please analyze carefully based on the samples themselves. Since some samples may have deviations, please avoid using overly absolute judgment logic.

Please output your complete reasoning process first, then output the final JSON result:

Reasoning process: [Your reasoning process]

Output JSON format:

```
“json
{
  "surface_difference": "Description of surface differences",
  "decision_boundary": "Decision boundary (How should similar samples be annotated?)",
  "rule_intent": "Core logic of rule formulation (i.e., general task annotation logic)",
  "meta_rule": "Reusable meta-rule expression (used as dimension name, prioritize using historical dimension names)",
  "applicable_boundary": "Applicable boundary conditions (what types of samples it applies to)"
}
“
```

Note:

- meta_rule will be used as the unique identifier name for the discriminative dimension; if it is substantially the same as a historical dimension, please use the historical dimension name

Prompts for Dimension Field Refinement (Part 1)

System: You are a helpful assistant.

User:

You are an annotation rule expert. Based on the main rule document, integrate historical rules with new cases to extract generalizable rule intent (core annotation logic).

Definitions

1. [Rule Intent] What is the core reason why rule makers require this annotation? (i.e., the general core annotation logic)
2. [Decision Boundary] What factors cause one case to be judged as X and another as Y?
3. [Applicable Boundary] Under what circumstances does this rule apply? When might it not apply?

Input Information

[Original Rule Document]
{doc_content}

[Historical Discriminative Dimensions]
- Rule Intent: {existing_intent}
- Decision Boundary: {existing_boundary}
- Applicable Boundary: {existing_applicable}

[New Case Analysis]
[Contrastive Group]
Sample A (Correctly judged):
Query: {query}
Model output: {annotation}
Correct answer: {ground_truth}

Sample B (Correctly judged):
Query: {query}
Model output: {annotation}
Correct answer: {ground_truth}

Sample C (Incorrectly judged):
Query: {query}
Model output: {annotation}
Correct answer: {ground_truth}

Prompts for Dimension Field Refinement (Part 2)

- Preliminary intent: {new_intent}
- Preliminary boundary: {new_boundary}
- Preliminary applicability: {new_applicable}

Fusion Analysis Steps (Please follow this reasoning)

1. Relationship Determination: Determine the relationship between the new case and historical rules
 - Support: New case validates and reinforces historical rules
 - Supplement: Fills gaps in scenarios not covered by historical rules
 - Correction: Discovers misjudgments or boundaries that are too wide/narrow in historical rules
 - Extension: Discovers that historical rules can be generalized to new categories/scenarios
2. Essence Extraction: Go beyond surface differences of specific cases to induce general task annotation logic (but don't over-speculate; the extracted logic should serve the task itself)
3. Conflict Resolution: If conflicts exist, analyze whether it's a "special handling of boundary cases" or a "fundamental flaw in historical rules", and provide a resolution plan
4. Counterexample Testing: Consider what scenarios would produce incorrect judgments if this rule were over-generalized, and tighten the applicable boundary accordingly

Output Requirements

Note: Please analyze carefully based on the samples themselves. Since some samples may have deviations, please avoid using overly absolute judgment logic. Please avoid academic or abstract vocabulary (such as "mechanism", "paradigm", "dimension", "attribute", etc.), and use plain language to induce general core annotation logic.

Please output your complete reasoning process first, then output the final JSON result:

Reasoning process: [Your reasoning process]

Final JSON format:

```
“json
{
  "rule_intent": "Fused rule intent (reflecting core reasons and generalization principles, but don't deviate from the annotation task itself)",
  "decision_boundary": "Fused decision boundary (clearly define annotation logic)",
  "applicable_boundary": "Fused applicable boundary (define clear scope and exceptions based on sample types, don't over-associate)"
} “
```

Prompts for Verifier Prompt Optimization (Part 1)

System: You are a helpful assistant.

User:

You are a verification prompt optimization expert. Your task is to analyze current error samples, identify the root causes in the verification prompt that lead to wrong selections, and perform targeted optimization.

Current Verification Prompt (to optimize)
{base_rules}

Current Verification Error Samples (core optimization basis)
Please carefully analyze the following error samples, understand in which scenarios current verification prompt leads to wrong annotator selections:

[Error Sample 1:]
Query: {query}
Ground Truth: {ground_truth}
Annotator Results:
- annotator1: {annotation} (label: {label})
- annotator2: {annotation} (label: {label})
- annotator3: {annotation} (label: {label})
Vote Info: {vote_info}
Verification Selected: {selected_agent}
Error Analysis: {error_reason}

[Error Sample 2: ...]

Error Reason Summary
{failure_reasons}

[Optional Auxiliary Reference:]
Auxiliary Reference Information
The following dimensions and contrastive samples can help understand error patterns, but optimization should be primarily based on the error samples above:

Learned Discriminative Dimensions Reference
[Dimension: {dim_name}] (Evidence count: {evidence_count})
- Rule intent: {rule_intent}
- Decision boundary: {decision_boundary}
- Applicable boundary: {applicable_boundary}

Prompts for Verifier Prompt Optimization (Part 2)

Contrastive Sample Reference (Different judgments for similar Queries)

[Contrastive Sample 1]

Sample A (System judged correct):

Query: {query}

System output: {annotation}

Correct answer: {ground_truth}

Sample B (System judged incorrect):

{full_error_entry}

Reference Optimization Principles

1. Error samples as core: Carefully analyze each error sample, understand why current prompt leads to these errors, optimization should directly solve these problems.
2. Reference dimensions and contrastive samples: Dimensions and contrastive samples provide summaries of historical error patterns, can serve as optimization reference, but should not overshadow the main focus.
3. Fully integrate related suggestions, don't pile up at the end: Avoid appending "additional notes", "additional clarifications", "supplementary rules" patch paragraphs at the end of the prompt.
4. Prioritize reusing existing content: Most optimization should be achieved by extending or rewriting existing prompt content when referencing dimensions and error samples.
5. Eliminate contradictions: When modifying or adding content based on dimensions and error samples, must check and synchronously correct other parts that may conflict.

Please strictly output in the following format:

Analysis process: [First analyze the root cause of each error sample one by one, then combine with reference information to propose verification prompt optimization plan]

Optimized verification prompt: [Fine-tuned complete verification prompt, can be directly used by the verification agent]

Prompts for Annotator Prompt Optimization (Part 1)

System: You are a helpful assistant.

User:

You are an annotation Prompt optimization expert. Your task is to analyze current error samples, identify the root causes in the Prompt that lead to annotation errors, and perform targeted optimization.

Existing Annotation Prompt (to optimize):
{prompts/guidelines}

Current Annotation Error Samples (Core Optimization Basis)
Please carefully analyze the following error samples, understand in which scenarios the current Prompt leads to annotation errors:

[Error Sample 1:]
Query: {query}
Ground Truth: {ground_truth}
Model Output: {annotation}
Error Analysis: {error_reason}

[Error Sample 2: ...]

Error Reason Summary:
{failure_reasons}

[Optional Auxiliary Reference:]

Auxiliary Reference Information

The following dimensions and contrastive samples can help you understand error patterns, but optimization should be primarily based on the error samples above:

Contrastive Sample Reference (Different judgments for similar Queries)

[Contrastive Sample 1]

Sample A (Current Annotator judged correct):

Query: {query}
Current Annotator output: {annotation}
Correct answer: {ground_truth}

Sample B (Current Annotator judged incorrect):
{full_error_entry}

Learned Discriminative Dimensions Reference

The following dimensions are general discriminative logic induced from historical contrastive samples, for reference only:

Prompts for Annotator Prompt Optimization (Part 2)

[Dimension: {dim_name}] (Evidence count: {evidence_count})

- Rule intent: {rule_intent}
- Decision boundary: {decision_boundary}
- Applicable boundary: {applicable_boundary}

Reference Optimization Principles

1. Error samples as core: Carefully analyze each error sample, understand why current prompt leads to these errors, optimization should directly solve these problems.
2. Reference dimensions and contrastive samples: Dimensions and contrastive samples provide summaries of historical error patterns, can serve as optimization reference, but should not overshadow the main focus.
3. Fully integrate related suggestions, don't pile up at the end: Avoid appending "additional notes", "additional clarifications", "supplementary rules" patch paragraphs at the end of the prompt.
4. Prioritize reusing existing content: Most optimization should be achieved by extending or rewriting existing prompt content when referencing dimensions and error samples.
5. Eliminate contradictions: When modifying or adding content based on dimensions and error samples, must check and synchronously correct other parts that may conflict.

Please strictly output in the following format:

Analysis process: [First analyze the root cause of each error sample one by one, then combine with reference information to propose Prompt optimization suggestions]

Optimized Prompt: [Fine-tuned complete Prompt, can be directly used for annotation]