Refining and Reusing Annotation Guidelines for LLM Annotation

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) demonstrates remarkable zero-shot annotation tasks, they often struggle with the specialized conventions of gold-standard benchmarks. We propose the systematic reuse and refinement of annotation guidelines as an alignment mechanism, introducing an iterative moderation framework that simulates the early phases of annotation projects. We evaluate three hypotheses: (1) the efficacy of guideline integration, (2) the advantage of reasoning-optimized models, and (3) the viability of moderation under minimal supervision. Testing across biomedical NER tasks (NCBI Disease, BC5CDR, BioRED) with three LLM families (GPT, Gemini, DeepSeek), our results empirically confirm all three hypotheses. While the iterative moderation framework shows a good potential in effectively refining guidelines, our analysis also reveals a significant room for improvement.

1 Introduction

Text annotation serves as the foundation for various text processing tasks, including semantic search and text mining. Recently, zero-shot or few-shot prompting of Large Language Models (LLMs) have demonstrated impressive performance in general labeling tasks. However, when annotations must serve specific downstream requirements, regulating LLM outputs to adhere to these concrete constraints becomes non-trivial.

Traditionally, in the corpus annotation community, annotation guidelines have been the primary mechanism for regulating the behavior of human annotators. Even when domain experts possess sufficient background knowledge, the realization of that knowledge into specific annotations can vary significantly when they encounter actual textual expressions. Discrepancies often arise regarding the scope of target entities, the resolution of "gray-zone" cases, and the precise determination of span boundaries. Annotation guidelines exist to standardize these behaviors, preventing what would otherwise be a stochastic or inconsistent process. Consequently, these guidelines encode not only high-level conceptual goals but also the granular, task-specific requirements necessary to optimize annotations for practical applications. This is precisely where the value of annotation guidelines lies: they are the result of rigorous optimization efforts designed to translate downstream application requirements into concrete labeling rules. As such, they constitute an invaluable resource for any text annotation effort.

For LLM-based annotation, prompting has become the primary mechanism for generating annotations; however, the challenge lies in effectively encoding these specialized requirements within the prompt. This work addresses this challenge by testing the following hypothesis:

> H1: The incorporation of annotation guidelines into the prompt will improve the LLM-based annotations.

Guideline-driven annotation requires sophisticated cognitive processing, as the application of complex rules to diverse textual contexts can be viewed as a deductive reasoning process. Therefore, it is expected that models with superior reasoning capabilities will better perform on this task. This leads to our second hypothesis:

> H2: Reasoning models will outperform their non-reasoning counterparts in guideline-driven annotation tasks.

Guidelines are often incomplete. This is evident in traditional corpus projects where new annotators require a "training" phase. During this phase, annotators annotate a small set of documents to identify discrepancies against gold annotations. This process, known as moderation, results in either the

refinement of the annotator's understanding or the clarification of the guidelines themselves. Notably, this process is typically conducted with minimal supervision. Because guidelines are inherently iterative and subject to initial ambiguity, our final hypothesis is:

> H3: Annotation guidelines can be effectively improved through a moderation process with minimal supervision.

Note that the task of guideline refinement itself can be framed as a process of inductive reasoning. This dual requirement for both deductive and inductive reasoning further motivates our second hypothesis.

The contributions of this work are three-folds:

1. We provide an empirical validation of the three hypotheses concerning the role of guidelines, reasoning capabilities, and moderation in LLM-based annotation.

2. We introduce an iterative moderation framework for guideline refinement and demonstrate its efficacy in aligning LLM outputs with gold-standard conventions under minimal supervision.

3. We provide a detailed qualitative analysis of the refinement process, offering insights into the evolution of guidelines and the specific linguistic discrepancies addressed during iteration to facilitate future research.

2 Related Work

Biomedical corpora and document-level annotation. The NCBI Disease Corpus (Islamaj Doğan et al., 2014) provides manually curated disease mentions and concept normalization on PubMed abstracts, while BC5CDR (Li et al., 2016) supports chemical and disease entities with relation annotations for the BioCreative CDR task. BioRED (Luo et al., 2022) extends document-level relation extraction with multiple entity types and relation pairs. These datasets exemplify annotation projects in which detailed human-written guidelines and moderation are essential for handling boundary cases and ambiguity (Islamaj et al., 2020). We focus on *NER-only* (exact span+type), as guideline refinement in our setting primarily targets mention boundaries and type assignments.

LLMs as annotators and guideline followers. Until recently, many annotation projects have been implemented through crowd-sourcing due to faster and cheaper execution (Callison-Burch, 2009). Beyond traditional annotation pipelines, prior work has explored LLMs as substitutes or complements to human annotators (Wang et al., 2021; Ding et al., 2023). Some studies show that ChatGPT can match or exceed crowd-worker quality on several text-annotation tasks, and motivates LLM-based labeling pipelines (Gilardi et al., 2023; Zhu et al., 2023b). However, reliable deployment depends on whether LLMs can adhere to detailed labeling rules rather than generic task descriptions. Another study investigates guideline-following behavior across domains, finding that performance can be brittle and sensitive to guideline structure, ambiguity, and edge cases (Fonseca and Cohen, 2024).

Evaluation with LLMs as judges and evaluator. Several studies show the potential of LLMs as judges for open-ended outputs (Zhu et al., 2023a) and discuss key biases, such as position and verbosity effects (Zheng et al., 2023). Subsequent work explores LLMs as structured evaluator rather than just scorers (Liu et al., 2023). Our setting differs in that the judge is constrained by explicit domain guidelines and disagreement evidence, and is tasked with producing actionable guideline refinements rather than a scalar preference score.

Utilization of Guidelines for LLMs. Some studies reveal that using guidelines helps LLMs with consistency by providing clear definitions (Huang et al., 2025) and instructions (Srivastava et al., 2025). A case study illustrates that LLMs help identify guideline issues and propose refinements and suggests that disagreement patterns can be exploited to iteratively strengthen guidelines (Bibal et al., 2025). Another recent work also investigates whether existing human annotation guidelines can be directly repurposed for LLM-based annotation by highlighting both the promise of guideline-driven prompting and the challenges of faithfully translating complex, domain-specific rules into model behavior (Kim et al., 2025). This research builds upon prior work by introducing an iterative moderation framework designed for minimal supervision and validating three core hypotheses across diverse experimental settings.

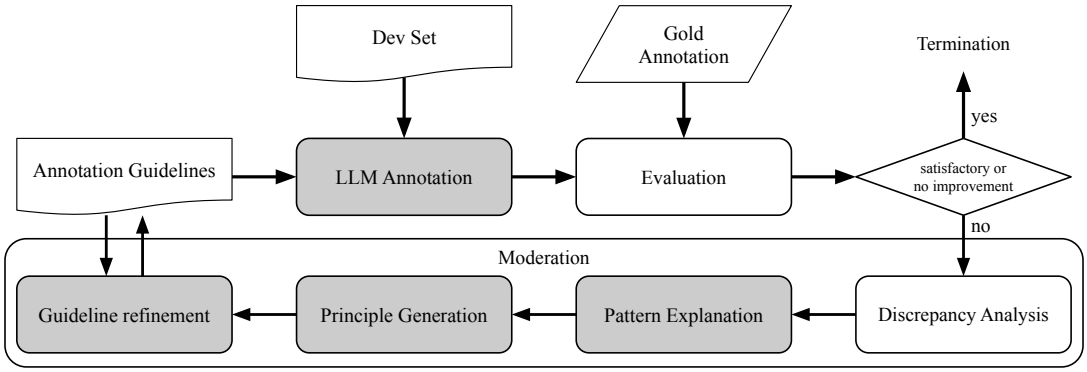

Figure 1: An overview of the iterative moderation framework

3 Methodology

3.1 Task formulation

The objective of the task is to adapt existing annotation guidelines - originally developed for human annotators - to regulate the performance of LLMs.

In traditional annotation projects, guidelines are established during a pilot phase where multiple human annotators label the same document set. During this *moderation* stage, *Inter-Annotator Agreement (IAA) rate* is measured, and guidelines are iteratively refined to resolve ambiguities and improve consistency. If new annotators join mid-project, they undergo a *training* phase: they annotate a small set of documents with gold annotations, and their performance is evaluated against the gold annotations. If the IAA rate is not high enough, the annotators are either retrained or, preferably, the guidelines are refined to ensure better reproducibility and objective clarity.

In this work, we explore the potential of using existing annotation guidelines to enable LLMs to reproduce annotations based on the guidelines. We frame this process as being equivalent to training new human annotators through iterative guideline refinement, however with LLM annotators instead of human annotators. As outlined in Figure 1, the proposed framework operates as an iterative cycle encompassing (1) annotation, (2) evaluation, and (3) moderation.

3.2 LLM-based Annotation

At the iteration k, a LLM annotator is prompted to annotate the documents in the development set D using the current set of annotation guidelines G_k. This process yields a set of predicted annotations A_k.

3.3 Evaluation

To assess the LLM's performance, the predicted annotations A_k are compared against the gold annotations A_g. As a measure of IAA rate, the F1-score is calculated based on a *strict matching criterion*, which requires exact match of both the entity boundary and the entity type.

After evaluation, the termination criteria is assessed. The iteration concludes if either:

1. The alignment reaches a predefined quality threshold ($\text{IAA}_k \geq \tau$).

2. For refined versions ($k > 0$), the most recent refinement fails to yield a performance gain ($\text{IAA}_k \leq \text{IAA}_{k-1}$)

3.4 Moderation

When the evaluation results fall below the required threshold, a moderation process is triggered to refine the guidelines. The LLM annotator then re-annotates the development set using the updated guidelines, forming an iterative loop. This process continues until the desired performance is achieved or a termination condition is met. Specifically, the loop terminates if a moderation cycle fails to yield measurable improvement, in which case the most recent refinement is discarded to prevent performance degradation.

At each iteration k, it takes two inputs: (1) the current set of guidelines G_k, and (2) a prioritized group of discrepancy patterns identified during the analysis phase. A refined version of the guidelines G_{k+1} is then generated, through a process we call *moderation*. The moderation process is designed to take three steps as shown in Figure 2: (1) pattern explanation, (2) principle generation, and (3) guideline refinement.

(A) Discrepancy Pattern Inference (Dominant Error Pattern Group)

Pattern Name: Phenotypic feature-list contexts suppress `DiseaseClass` tagging.
Confusion Trigger: Condition terms used as items in dependent "feature lists" (e.g., `with/of` complements, or coordinated arguments of report/causal verbs) are treated as mere symptoms/processes and dropped.
Contrastive Evidence: True positives concentrate in definitional disease-head frames (e.g., "X is an autosomal recessive disorder"), where an explicit disease-head noun cues `DiseaseClass`.
Rule (Proposed): If an NP denotes a clinical condition and appears as an item in a dependent feature-list (e.g., within a `with/of` complement or a coordinated report/causal frame), annotate each conjunct as `DiseaseClass`; do not suppress solely due to list presentation.

(B) Moderation Principle Generation (Single General Principle)

IF a noun phrase denoting a clinical condition appears as an item in a dependent "feature-list" construction (e.g., as a conjunct within a `with/of` complement of a patient/syndrome noun, or as a coordinated argument governed by observation/reporting/causation verbs),
THEN annotate that phrase (and propagate to all coordinated conjuncts) as `DiseaseClass` using the minimum necessary span;
EXCEPT when the condition phrase functions only as a modifier of a non-disease head (e.g., gene/locus/region/protein/variant) or is a bare/general process term that should not be annotated as a disease.

(C) Guideline Refinement (Inserted Rule; Truncated)

#4. Annotate coordinated headless phenotype/pathology complements as `DiseaseClass`

When a coordinated sequence of *headless* phenotype/pathology noun phrases (e.g., `anemia`, `lymphadenopathy`, `rash`) functions as the complement of a clinical-description or diagnostic construction, annotate *each coordinated item* as `DiseaseClass` using the minimum span.

Applicable constructions include:

- Patient/disease-centered descriptions (e.g., `patients with ...`, `characterized by ...`, `presenting with ...`).
- Syndrome-level frames introducing a phenotypic list (e.g., `a syndrome characterized by ...`).

Do not apply when:

- Items are bare general or process terms (e.g., `symptoms`, `carcinogenesis`).
- The coordination is governed by an overt disease head (e.g., `breast and ovarian cancer`; follow `CompositeMention`).

Note. This rule refines `DiseaseClass` only and does not affect `SpecificDisease` or disease `Modifier` handling (e.g., `ALPS` vs. `ALPS phenotype`).
(*Truncated due to space limit. Full text is provided in Figure 6.*)

Figure 2: Example of LLM-simulated moderation for DiseaseClass guideline refinement. The moderator identifies a recurring failure where DiseaseClass mentions embedded in dependent feature-list contexts (e.g., `with/of` complements and coordinated report/causal frames) are suppressed, generates an actionable principle, and inserts a targeted guideline rule in order to reduce DiseaseClass false negatives. The following guideline from (C) is inserted between Section 3 (Modifiers) and Section 4 (Duplicate mentions) in the "What to Annotate" section of the original human annotation guidelines, as shown in Figure 5b.

3.4.1 Discrepancy Analysis

To facilitate guideline refinement, we collect and analyze all discrepancies between A_k and A_g. While the primary evaluation is strict, we apply a *soft matching criterion* (requiring at least one character of overlap) to categorize errors for finer-grained diagnostic analysis. Each discrepancy case is assigned to one of four mutually exclusive categories, prioritized in the following order:

1. **Label mismatch**: A gold entity and a predicted entity overlap in text (at least one character) but have different types.
2. **Boundary mismatch**: a gold entity and a predicted entity overlap in text and have the same type, but their start/end offsets differ.
3. **False Negative (FN)**: a gold entity with zero character overlap with any predicted entity.
4. **False Positive (FP)**: a predicted entity with zero character overlap with any gold entity.

Discrepancy cases are grouped based on their predicted and gold label pairs, and the most frequent group is selected as the target of moderation.

Each discrepant case is represented by (1) the mention string(s), (2) the surrounding context (window size = 60), (3) the types of predicted and gold entities, and (4) the discrepancy category.

3.4.2 Pattern Explanation

The goal of this step is to generate an explanations of the cases of the targeted discrepancy pattern group. We provide (a) discrepancy examples from the selected group and (b) verified true positives handled correctly under G_k. Verified true positives are automatically sampled at each iteration from instances that are strict true positives under G_k. We enforce a contrastive analysis procedure: the model explicitly compares discrepancy cases against true positives to isolate a single discriminating feature (syntactic, positional, or semantic) that separates the two sets, and outputs exactly one pattern insight in a fixed format.

3.4.3 Principle Generation

This step is to generate a generalized principle from the pattern explanation. In this step, we convert the linguistic insight into one actionable rule suitable for annotators and future LLM annotators. We cast the model as an *AI moderator* and provide entity definitions, the linguistic analysis, and the current guideline. The moderator must synthesize a single general principle in a strict IF/THEN form and include an explicit negative constraint to specify when the rule *does not* apply to avoid overcorrection and false positives. To improve generalization, we require abstract phrasing rather than instantiations tied to specific tokens. Finally, the model checks consistency against existing entity definitions and guideline rules.

3.4.4 Guideline Refinement

Finally, this step integrates the principle generated in the previous step into the current set of guidelines G_k, and generates a revised version of guidelines G_{k+1}. The primary role of this step is to maintain the integrity during the revision. We provide the current guideline, the newly generated principle, and the discrepancy context. We include verified examples (true positives) as in-prompt checks: the model is instructed not to introduce changes that would flip these known-correct cases. To preserve document integrity, the model outputs the full updated guideline text without omission or summarization, while maintaining the original formatting and structure.

4 Experimental Setup

4.1 Implementation

We explored biomedical named entity recognition (NER) tasks as a testbed of the methodology. For NER tasks, an individual annotation is represented by triplets of the character offsets of the begin and end of the entity, and the type of the entity.

We used the PubAnnotation framework for management, evaluation, visualization of annotations (Kim et al., 2019). Accordingly, the PubAnnotation JSON format is used to represent annotations. All annotations and evaluation results are stored in PubAnnotation for reproducibility and are also available at **anonymized URLs**, which will be updated upon acceptance.

4.2 Datasets

We evaluate on biomedical datasets with established human annotation guidelines:

- **NCBI Disease**: PubMed abstracts annotated for disease mentions.
- **BC5CDR**: PubMed abstracts annotated for Chemical and Disease mentions (NER-only).
- **BioRED**: PubMed abstracts annotated with biomedical entity types (and relations); we focus on NER-only entity denotations.

Since our primary objective is to evaluate the feasibility of inducing high-level annotation principles, rather than achieving state-of-the-art performance, we departed from the standard train/dev/test splits provided by the original dataset authors. Instead, we adopted a "low-resource" approach to test the hypothesis that high-level guidelines can be induced from minimal supervision.

Specifically, we randomly sampled a minimal set of 10 documents from the original training split of each benchmark to serve as our *development dataset* for iterative moderation. However, to ensure a statistical reliability during evaluation, we also prepared a larger set of 100 documents from the original development split as our *evaluation dataset*.

For NCBI Disease and BioRED, we used their respective development splits in full, as both contain

exactly 100 documents. For BC5CDR, which provides a larger development split of 500 documents, we randomly sampled 100. Finally, we intentionally withheld the original test sets to preserves their integrity for future benchmarking and to ensures our refinement process remains strictly separated from the final unseen data.

4.3 Selection of Large Language Models

We selected three representative, state-of-the-art Large Language Model (LLM) families for our experiments: *GPT*, *Gemini*, and *DeepSeek*. Within each family, we distinguish between non-reasoning and reasoning models.

Our primary hypothesis is that the iterative process of guideline refinement and application requires high-order reasoning capabilities. Specifically, synthesizing new guidelines from discrepancies requires inductive reasoning, while applying those guidelines to specific instances requires deductive reasoning. To test this hypothesis, our first experiment compares the performance of non-reasoning and reasoning models from the same families to evaluate their relative effectiveness in this iterative loop. Following is the specific models and their configurations used in the experiments:

- **GPT**: `gpt-5-2025-08-07` with `reasoning_effort` $\in \{\texttt{low}, \texttt{high}\}$.
- **Gemini**: `gemini-2.5-pro` with `thinking_budget` $\in \{\texttt{min}, \texttt{max}\}$.
- **DeepSeek**: `deepseek-chat` (non-reasoning) vs. `deepseek-reasoner` (reasoning).

We run all models with default hyperparameter settings, as reasoning models expose non-comparable configuration options (e.g., GPT-5 does not support "temperature", whereas Gemini does).

4.4 Design of Experiments

We designed our experiment to explore the three hypothesis presented in the section 1. Accordingly, we compare three approaches:

1. **Prompt-only** (baseline): minimal task instruction, annotation entities, and required output schema (see Figure 5a).
2. **Original-guidelines**: original human guideline text with light formatting (see Figure 5b).
3. **Guideline-refinement**: discrepancy-driven guideline refinement.

For the second approach, we experimented with both reasoning and non-reasoning models of each LLM family, and as it is reported in 5.2, across all datasets, reasoning models consistently outperform their non-reasoning counterparts (Table 2). Based on this result, we use only the reasoning models for the third approach.

5 Results

5.1 With vs. Without Annotation Guidelines

As reported in the *S* and *G* column groups of Table 1, providing annotation guidelines substantially improved the performance across all datasets and models. This observation confirms the hypothesis that the provision of guidelines enhances the performance is LLM annotators.

5.2 Reasoning vs. Non-reasoning models

As reported in Table 2, reasoning models consistently achieved higher performance than non-reasoning models across all datasets, with the approach of guideline provision. This observation confirms the hypothesis that reasoning capability is essential for guideline-driven annotation.

5.3 With vs. Without Guideline Refinement

As reported in the *M* column group of Table 1, the guideline refinement via the moderation approach yielded further performance gains. While these improvements are marginal (typically +0.01 – 0.03 in F1), they are observed constantly across all datasets and models. These results demonstrate the potential of the moderation-based guideline refinement approach to enhance LLM annotation performance. However, the marginal nature of the gains suggests significant opportunities for further methodological optimization.

5.4 Analysis of Iterative Moderation

Figure 3 illustrates the evolution of confusion matrices through moderation iterations performed by GPT on the NCBI Disease dataset. In this setup, the iterative process concluded after four iterations. For each step, the LLM annotator's predictions $(A_0, \ldots, A_3)$ are compared against the gold annotations (A_g) to generate the confusion matrices. This analysis demonstrates how the refinement process progressively resolves targeted label discrepancies while also revealing the secondary effects of these updates on other annotation cases.

Dataset (#Entity)	Model	#Iters	S				G (Δ)				M (Δ)			
			P	R	F1	TP	P	R	F1	TP	P	R	F1	TP
NCBI (791)	GPT-5	3	0.45	0.48	0.46	378	0.78	0.68	0.73 (+0.27)	540	0.82	0.71	**0.76 (+0.03)**	565
	Gemini	5	0.36	0.47	0.40	369	0.69	0.57	0.63 (+0.23)	453	0.72	0.61	**0.66 (+0.03)**	479
	DeepSeek	2	0.32	0.30	0.31	236	0.72	0.45	0.55 (+0.24)	356	0.71	0.47	**0.56 (+0.01)**	369
BC5CDR (2,146)	GPT	1	0.84	0.78	0.80	1,664	0.89	0.81	0.85 (+0.05)	1,735	0.92	0.81	**0.86 (+0.01)**	1,737
	Gemini	1	0.74	0.63	0.68	1,359	0.84	0.68	0.76 (+0.08)	1,469	0.86	0.70	**0.77 (+0.01)**	1,503
	DeepSeek	1	0.80	0.45	0.58	968	0.89	0.50	0.64 (+0.06)	1,072	0.86	0.52	**0.65 (+0.01)**	1,119
BioRED (3,531)	GPT-5	2	0.75	0.74	0.74	2,598	0.81	0.72	0.76 (+0.02)	2,548	0.82	0.81	**0.82 (+0.06)**	2,871
	Gemini	1	0.62	0.60	0.61	2,111	0.74	0.61	0.67 (+0.06)	2,137	0.71	0.69	**0.69 (+0.02)**	2,371
	DeepSeek	1	0.71	0.33	0.45	1,179	0.77	0.41	0.53 (+0.08)	1,442	0.76	0.42	**0.54 (+0.01)**	1,480

Table 1: Performance of LLM annotation in three approaches: simple prompting (S), guideline (G), and moderation (M). Precision (P), Recall (R), F1-score (F1), and the number of true positives (TP) are reported. #Iters indicates the number of moderation iterations.

Dataset	Model	Non-Reason	Reason	Δ
NCBI	GPT	0.69	**0.73**	+0.04
	Gemini	0.48	**0.63**	+0.15
	DeepSeek	0.29	**0.55**	+0.26
BC5CDR	GPT	0.78	**0.85**	+0.07
	Gemini	0.70	**0.76**	+0.06
	DeepSeek	0.57	**0.64**	+0.07
BioRED	GPT	0.72	**0.76**	+0.04
	Gemini	0.66	**0.67**	+0.01
	DeepSeek	0.43	**0.53**	+0.10

Table 2: Comparison of (Non)-reasoning models.

In the initial iteration, the prioritized discrepancy pattern involved falsely predicted DISEASE-CLASS entities (Predicted: DISEASECLASS, Gold: No Entity), which occurred with a frequency of 7. Figure 2 illustrates how the annotation guidelines are refined in the iteration to address the specific pattern. The confusion matrix for the iteration 1 shows that the specific type of discrepancy was largely resolved from 7 to 1.

However, secondary effect were also observed, where new false cases were inadvertently introduced as a trade-off. This suggests that moderation does not monotonically reduce all discrepancies, but instead rebalances precision–recall trade-offs throughout the iterative process. Figure 4a shows some positive changes (true positives) caused by the guideline refinement, while Figure 4b) highlights a newly introduced false case (false positive) as a trade-off. Detailed before and after examples with gold and predicted annotations are available in Appendix, Figure 7.

Notably, these secondary effects also include positive changes. For instance, the five label mismatch cases, where entities were incorrectly labeled as SPECIFICDISEASE instead of the gold label MODIFIER, were entirely resolved in Iteration 1. This suggests that guideline refinements targeted at one specific error group can successfully propagate to, and resolve, other related discrepancy patterns.

6 Discussions

6.1 Performance Improvement Interpretation

The simple prompt approach serves as a zero-shot annotation method that leverages the innate linguistic and world knowledge of LLMs. While effective in domains where the model possesses sufficient background knowledge, performance on benchmark datasets often remains suboptimal. This gap typically stems not from a deficiency in the model's underlying knowledge, but from a misalignment between the specific annotation conventions of the benchmark and the LLM's independent interpretation of the text. Consequently, the F1-score improvements reported in Section 5 suggest that the use of guidelines successfully aligns the LLM's output with the rigorous requirements and boundary conventions encoded in the gold annotations.

6.2 Limited Amount of Development Data

The experiments are designed to simulate the moderation process typically conducted during the early stages of annotation projects, where gold-standard supervision is highly limited. This methodology represents a fundamental departure from traditional machine learning paradigms that rely on large-scale statistical learning. While those empirical approaches require extensive datasets to capture representative statistics, our work adopts a rationalistic approach that leverages the latent linguistic and world knowledge of LLMs. Our primary hypothesis is that the advanced reasoning capabilities of modern LLMs can bridge the gap between their

(a) Iter 0 (G_0)

Gold ↓ / Pred →	C	D	M	S	O
C	0	0	0	0	0
D	3	0	1	0	7
M	1	0	0	0	1
S	0	5	5	0	3
O	0	1	0	1	0

(b) Iter 1 (G_1)

Gold ↓ / Pred →	C	D	M	S	O
C	0	0	0	0	0
D	0	1	0	0	1
M	1	0	0	0	2
S	0	6	0	0	2
O	1	6	0	1	0

(c) Iter 2 (G_2)

Gold ↓ / Pred →	C	D	M	S	O
C	0	0	0	0	0
D	0	1	0	0	1
M	1	0	0	0	1
S	0	6	1	0	2
O	0	3	0	1	0

(d) Iter 3 (G_3)

Gold ↓ / Pred →	C	D	M	S	O
C	0	0	0	0	0
D	0	1	0	0	1
M	1	0	0	0	1
S	0	5	0	0	2
O	0	3	0	1	0

Figure 3: Confusion matrices over moderation iterations ($k = 0, \ldots, 3$): Rows correspond to the gold labels and columns to the model predictions. The cells of the same gold and prediction labels show the number of span boundary mismatches. C = CompositeMention, D = DiseaseClass, M = Modifier, S = SpecificDisease, O = No Entity.

> Homozygosity mapping in a family with microcephaly, mental retardation, and short stature to a Cohen syndrome region on 8q21.3–8q.

(a) Example of correct changes (resolved FNs).

> (WD) is an autosomal recessive disorder characterized by copper accumulation in the liver, brain, kidneys, and corneas, and ...

(b) Example of false changes (newly introduced FP).

Figure 4: Example of correctly and falsely changed annotations after refinement. (a) The three underlined entities are correctly revolved originally FNs, and (b) The underlined entity is a newly introduced FP.

Dataset	Model	i	c_i ($)	t_i (min)	$\hat{C}_{proc}$	$\hat{T}_{proc}$
NCBI	GPT-5	3	1.186	5.2	3.557	15.6
	Gemini	5	0.092	3.0	0.460	14.8
	DeepSeek	2	0.054	15.8	0.109	31.6
BC5CDR	GPT-5	1	1.729	9.9	1.729	9.9
	Gemini	1	0.099	8.8	0.099	8.8
	DeepSeek	1	0.055	15.4	0.055	15.4
BioRED	GPT-5	2	1.991	14.0	3.982	28.0
	Gemini	1	0.251	5.9	0.251	5.9
	DeepSeek	1	0.048	29.8	0.048	29.8

Table 3: The cost of each iteration and the whole iterations of each experimental setup. c_i and t_i denote the cost and time of the *final* iteration i. The end-to-end overhead is estimated as $\hat{C}_{proc} = i \cdot c_i$ and $\hat{T}_{proc} = i \cdot t_i$.

internal knowledge and the specific conventions observed in concrete annotations.

However, our experimental results suggest that relying on only 10 documents as a basis for guideline refinement may be overly ambitious. First, such a small sample is highly susceptible to selection bias; there is a significant risk that substantial discrepancy patterns present in the wider dataset may not manifest within the sample. Future research should examine the learning curve by incrementally increasing the volume of development data. Second, certain components of the proposed method – specifically the loop-termination criteria – rely on statistical observations derived from this limited development set. Because these metrics are calculated from a small sample size, they are inherently prone to instability. Consequently, future work should focus on devising more robust or noise-tolerant termination conditions.

6.3 Cost and Time Estimation

Table 3 details the economic and temporal overhead of the moderation process. While GPT-5 demonstrates high performance, its operational cost is significantly higher, often by an order of magnitude compared to other models. Conversely, while DeepSeek offers the lowest financial cost per iteration, it suffers from significantly higher computational latency and lower overall performance. Despite these variances, it demonstrates that LLM-based moderation has a good potential to be a highly cost-effective alternative to human labor, although specialized human experts still retain an edge in absolute annotation quality.

7 Conclusion

This work investigated the systematic reuse and refinement of annotation guidelines within an LLM-based annotation paradigm. We empirically validated three core hypotheses regarding (1) the efficacy of explicit guideline integration, (2) the advantage of reasoning-optimized architectures, and (3) the viability of iterative moderation under minimum supervision. While our results across multiple biomedical benchmarks support all three hypotheses, a granular analysis of discrepancy evolution against gold-standard annotations suggests significant opportunities for further improvement.

Limitations

Our approach assumes the availability of an initial human-written guideline document that is sufficiently detailed to support discrepancy-driven refinement. In practice, many benchmarks and domains do not provide publicly available annotation guidelines, or only release high-level task descriptions, which limits the direct applicability of our framework. Addressing this limitation will require methods for *guideline generation*, for example by LLM-consumable guidelines from annotation examples, label ontologies, or limited curator input.

In addition, while we evaluate refined guidelines extrinsically through strict span-and-type performance on held-out data, intrinsic assessment of guideline quality remains challenging. Guideline refinements may introduce unintended side effects or regressions, yet there is currently no standardized quality assurance procedure for evaluating guideline documents themselves. Developing principled *guideline evaluation* and QA mechanisms, such as consistency checks or automated detection of overly broad rules, is an important direction for future work.

Ethical Considerations

This work studies guideline-driven annotation and moderation using large language models on existing, publicly available biomedical named entity recognition benchmarks (NCBI Disease, BC5CDR, and BioRED), which are released under the Creative Commons Attribution 4.0 (CC BY 4.0) license . No new data collection or human subject involvement is required, and all experiments are conducted on previously released datasets that are widely used in prior research.

The proposed framework does not introduce new predictive models or deploy systems in real-world or clinical settings. Instead, it focuses on analyzing systematic annotation discrepancies and refining annotation guidelines in an offline, research-only context. As such, the work does not pose additional ethical risks beyond those commonly associated with benchmark-based NLP research.

References

Adrien Bibal, Nathaniel Gerlek, Goran Muric, Elizabeth Boschee, Steven C. Fincke, Mike Ross, and Steven N. Minton. 2025. Automating annotation guideline improvements using LLMs: A case study. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 129–144, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 286–295.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is gpt-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.

Marcio Fonseca and Shay Cohen. 2024. Can large language models follow concept annotation guidelines? a case study on scientific and financial domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8027–8042, Bangkok, Thailand. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120(30):e2305016120.

Shizhou Huang, Bo Xu, Yang Yu, Changqun Li, and Xin Alex Lin. 2025. Guidener: Annotation guidelines are better than examples for in-context named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24159–24166.

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. Teamtat: a collaborative text annotation tool. *Nucleic acids research*, 48(W1):W5–W11.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Jin-Dong Kim, Yue Wang, Toyofumi Fujiwara, Shujiro Okuda, Tiffany J Callahan, and K Bretonnel Cohen. 2019. Open agile text mining for bioinformatics: the pubannotation ecosystem. *Bioinformatics*, 35(21):4372–4380.

Kon Woo Kim, Rezarta Islamaj, Jin-Dong Kim, Florian Boudin, and Akiko Aizawa. 2025. Repurposing annotation guidelines to instruct llm annotators: A case study. In *International Conference on Applications of Natural Language to Information Systems*, pages 140–151. Springer.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

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

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Saurabh Srivastava, Sweta Pati, and Ziyu Yao. 2025. Instruction-tuning llms for event extraction with annotation guidelines. *arXiv preprint arXiv:2502.16377*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023a. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023b. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Appendix

Inter-Annotator Agreement (IAA) Threshold To contextualize the termination criterion used in our moderation experiments, we report previously published inter-annotator agreement (IAA) statistics for the benchmark datasets considered in this work. These values reflect the level of consistency typically achieved by trained human annotators under the corresponding annotation guidelines.

For the BioCreative V Chemical–Disease Relation (BC5CDR) dataset, the reported IAA scores on the test set are 87.49% for disease entities and 96.05% for chemical entities. The BioRED dataset reports an entity-level IAA of 97.01%, indicating high agreement among expert annotators. The NCBI Disease corpus reports an inter-annotator agreement of approximately 87.5% for disease mention annotation (Leaman et al., 2013).

Considering these established human agreement levels, we set the moderation termination criterion to a strict-match F1 score of 0.9. This threshold is intentionally conservative relative to reported IAA values and is used solely as a stopping heuristic to prevent over-iteration during guideline refinement, rather than as a claim of exceeding or matching human annotation quality.

PROMPT TEMPLATE

SYSTEM INSTRUCTION:
You are an expert AI for text annotation. Your task is to annotate all entities from the provided text based on a strict schema and guidelines.

ENTITY SCHEMA:
{{entitySchema}}

ANNOTATION GUIDELINES:
{{guidelines}}

ANNOTATION RULES:
- Output MUST be a valid JSON object with a single key "annotations".
- "annotations" MUST be an array of objects following this JSON schema:
{{jsonSchema}}
- If no entities are found, return {"annotations": []}.
- Spans must match the original text exactly. Do not alter spacing, casing, or punctuation.

TEXT TO ANNOTATE:
—
{{inputText}}
—

Provide your response as a single JSON object.

(a) Prompt template.

Biomedical Annotation Guidelines (Excerpt)

What to Annotate

1. Annotate all Specific Disease mentions
A disease mention may refer to a Specific Disease or a Disease Class.
- Disease Class: A family of multiple specific diseases.
- Specific Disease: A single, well-defined disease entity.

Example:
Diastrophic dysplasia is an autosomal recessive disease.
Annotate "Diastrophic dysplasia" as Specific Disease and
"autosomal recessive disease" as Disease Class.

2. Annotate contiguous text strings
Composite mentions referring to multiple diseases are annotated as a single span.

3. Annotate disease mentions used as modifiers
Disease names modifying other noun phrases are annotated as Modifier.

4. Annotate duplicate mentions
All disease mentions within a sentence are annotated, including duplicates.

5. Annotate the minimum necessary span
Prefer the smallest span expressing the most specific disease form.

6. Annotate all synonymous mentions
Long forms and abbreviations are annotated separately.

What NOT to Annotate
- Organism names unless clearly referring to diseases.
- Gender terms unless defining a distinct disease subtype.
- Overlapping mentions.
- General terms (e.g., disease, syndrome), except cancer and tumor.
- Biological processes (e.g., tumorigenesis).
- Mentions interrupted by nested spans.

Examples include composite mentions, disease classes, and exclusion cases.

(b) Human guideline excerpt.

Figure 5: Prompt template and Human guideline. (Left) Prompt template used for LLM-based annotation. (Right) Excerpt from the original human annotation guidelines - NCBI Disease Corpus.

#4. Annotate coordinated headless phenotype/pathology complements as Disease Class

When a coordinated sequence of *headless* phenotype/pathology noun phrases (i.e., each item's head is a pathological/phenotypic noun such as `anemia`, `lymphadenopathy`, `splenomegaly`, `thrombocytopenia`, `rash`) functions as the complement of a clinical-description or diagnostic construction, annotate *each coordinated item* as `DiseaseClass` using the minimum contiguous span.

Clinical-description / diagnostic constructions include:

- Phrases headed by or modifying a patient/disease/syndrome NP: `patients with ...`, `the disorder is characterized by ...`, `a syndrome of ...`, `manifesting/presenting with ...`, `features include ...`, `diagnosed with ...`.
- Constructions governed by a superordinate disease/syndrome head that introduces the list (e.g., `a syndrome characterized by ...`).

Annotate:

- Each coordinated phenotype/pathology item as `DiseaseClass`.
- Use the smallest span that conveys the most specific pathological content (retain subtype-defining adjectives; avoid non-diagnostic intensifiers unless clinically defining).

Do not apply this rule when:

- Items are bare general terms without specific pathological content (e.g., `complications`, `abnormalities`, `symptoms`).
- Items denote biological processes rather than disorders (e.g., `tumorigenesis`, `carcinogenesis`).
- The coordination depends on an overt disease head outside the items (e.g., `breast and ovarian cancer`); in such cases follow `CompositeMention` rules instead.

Important. This rule adds `DiseaseClass` annotations for diagnostic/phenotypic lists and does not change how `SpecificDisease` or disease `Modifiers` are annotated elsewhere (e.g., `ALPS` remains `SpecificDisease`; `ALPS phenotype` remains a `Modifier` and the generic head `phenotype` is not annotated).

Figure 6: Inserted guideline rule after moderation. Full text of the newly added rule (§4) that instructs annotating coordinated headless phenotype/pathology complements as `DiseaseClass`, including scope conditions and exceptions.

(a) **DiseaseClass FN (Iteration 0)**

Homozygosity mapping in a family with **microcephaly**, **mental retardation**, and **short stature** to a Cohen syndrome region on 8q21.3–8q.

Gold: microcephaly, mental retardation, short stature (DISEASECLASS)
Prediction: none (all missed) *[FN]*

(b) **DiseaseClass fixed (Iteration 1)**

Homozygosity mapping in a family with microcephaly, mental retardation, and short stature to a Cohen syndrome region on 8q21.3–8q.

Gold: microcephaly, mental retardation, short stature (DISEASECLASS)
Prediction: all correctly recognized as DISEASECLASS

(c) **Correct non-annotation (Iteration 0)**

(WD) is an autosomal recessive disorder characterized by **copper accumulation** in the liver, brain, kidneys, and corneas, and ...

Gold: none
Prediction: none (correct)

(d) **FP introduced (Iteration 1)**

(WD) is an autosomal recessive disorder characterized by copper accumulation in the liver, brain, kidneys, and corneas, and ...

Gold: none
Prediction: DISEASECLASS (copper accumulation) *[FP]*

Figure 7: Detailed before/after examples across moderation iterations. Top row: a coordinated DISEASECLASS list is missed at iteration 0 but recovered after refinement (FN reduction). Bottom row: "copper accumulation" is correctly unannotated at iteration 0 but becomes a false positive after refinement (precision–recall trade-off).