Aggregating Crowd of LLMs for Cost-Effective Data Annotation

Jiacheng Liu
Shanghai Jiao Tong University
liujiacheng@ieee.org

Xiaofeng Hou
Shanghai Jiao Tong University
hou-xf@cs.sjtu.edu.cn

Abstract

Recent advancements in Large Language Models (LLMs) have shown promise for automated data annotation, yet reliance on expensive commercial models like GPT-4 limits accessibility. This paper rigorously evaluates the potential of open-source smaller LLMs (sLLMs) as a cost-effective alternative. We introduce a new benchmark dataset, Multidisciplinary Open Research Data (MORD), comprising 12,277 annotated sentence segments from 1,500 scholarly articles across five research domains, to systematically assess sLLM performance. We further propose to build the Crowd of LLMs, which aggregates annotations from multiple sLLMs using label aggregation algorithms. This approach not only outperforms individual sLLMs but also reveals that combining sLLM annotations with human crowd labels yields superior results compared to either method alone. Our findings highlight the viability of sLLMs for democratizing high-quality data annotation while underscoring the need for tailored aggregation methods to fully realize their potential.

1 Introduction

The exponential growth of digital data has transformed the landscape of knowledge discovery and data mining, creating unprecedented opportunities and challenges in extracting meaningful insights from vast information repositories (Cambria and White, 2014). The development of increasingly sophisticated machine learning models has enabled groundbreaking advances in data analytics, pattern recognition, and predictive modeling across diverse domains (Oppenheim et al., 2000; Ko et al., 2022; Adamopoulou and Moussiades, 2020). At the heart of this progress lies a critical yet often overlooked component: high-quality annotated data for training machine learning models. Data annotation, the process of labeling raw data to create training sets, plays a pivotal role in developing robust models for tasks ranging from text classification and entity recognition to complex pattern mining and knowledge extraction (Handschuh and Staab, 2003). The quality and quantity of annotated data directly impact the performance, reliability, and generalizability of machine learning models that power modern data mining applications (Cheng et al., 2018; Neves and Leser, 2014).

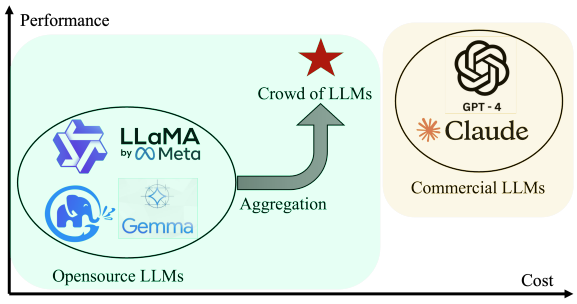

Figure 1: Comparison of Crowd of LLMs with existing LLM annotation. Low-cost open-source LLMs like LLaMA and Gemma can be aggregated to form a "Crowd of LLMs" that potentially rivals the performance of closed source and expensive commercial LLMs such as GPT-4 and Claude.

Traditionally, data annotation has relied heavily on human annotators, often recruited through crowdsourcing platforms like Amazon Mechanical Turk (Howe et al., 2006; Amazon.com, 2024). This approach, while widely adopted, faces significant challenges. Human annotation is inherently time-consuming and expensive, especially for large-scale datasets or specialized domains requiring expert knowledge (Chittilappilly et al., 2016). Moreover, human annotators are susceptible to fatigue, bias, and inconsistency, potentially introducing errors that propagate through the machine learning pipeline (Eickhoff, 2018). As data mining applications become more complex and diverse, addressing these limitations has become increasingly crucial for advancing the field and ensuring the reliability of intelligent systems in real-world scenarios.

Findings of the Association for Computational Linguistics: EACL 2026, pages 2407–2419
March 24-29, 2026

The emergence of Large Language Models (LLMs) like GPT-3 and GPT-4 has opened new possibilities for automating the data annotation process (Tan et al., 2024; Li, 2024a; He et al., 2024a; Schoenegger et al., 2024; He et al., 2024b). These models, trained on vast amounts of text data, demonstrate remarkable capabilities in understanding context, generating human-like text, and performing complex reasoning tasks. Recent studies have shown promising results using LLMs for various annotation tasks, from sentiment analysis to complex reasoning tasks, offering the prospect of faster, more consistent, and potentially more accurate labeling (Wang et al., 2021). However, current LLM-based annotation approaches face several critical limitations. First, reliance on commercial LLMs incurs prohibitive costs and raises serious privacy concerns when dealing with sensitive or confidential data(Yao et al., 2024). The high computational requirements and associated costs make these models inaccessible to many researchers and organizations, particularly those with limited resources (Yuan et al., 2024). Second, current evaluation methodologies suffer from a dual crisis of accessibility and methodological rigor. Most studies rely on legacy datasets like RTE or QUIZ (Li, 2024a,b; Snow et al., 2008), despite growing evidence that these benchmarks were partially ingested during LLM pre-training (He et al., 2024b). This methodological flaw particularly impacts domains requiring frequent annotation updates, such as emerging scientific literature, where post-training data distributions are crucial for accurate evaluation. Neither the commercial LLM-dependent approach nor the potentially contaminated evaluation protocols adequately address the needs of resource-constrained researchers working with novel data distributions, creating a significant barrier to advancement in the field.

We address these challenges through a comprehensive investigation of open-source sLLMs (Figure 1). Our study reveals that while individual sLLMs may not match GPT-4's performance, careful ensemble strategies can unlock remarkable capabilities. Through extensive experimentation with twelve state-of-the-art sLLMs, including recent models like Llama-3, Mistral, and Gemma, we demonstrate that ensembles can approach GPT-4's accuracy at merely 3% of the cost. This dramatic cost reduction, coupled with the transparency and adaptability of open-source architectures, represents a significant advancement for democratizing high-quality data annotation. To facilitate reproducible research and enable rigorous evaluation, we introduce Multidisciplinary Open Research Data (MORD), a carefully curated benchmark comprising 12,277 annotations from 1,500 research papers across five distinct domains. Crucially, all papers in MORD were published after August 2024, ensuring zero overlap with the training data of evaluated models and providing a truly uncontaminated test bed for assessing annotation quality.

Through extensive empirical evaluation on both MORD and existing benchmarks, we uncover several significant insights. First, sLLM ensembles demonstrate consistent superiority over individual human annotators across all evaluated domains, with particularly strong performance in technical and domain-specific content where crowdworkers often struggle. Second, our cost analysis reveals that the Crowd of LLMs approach reduces annotation costs by two orders of magnitude compared to GPT-4 while maintaining comparable quality. This dramatic cost reduction enables annotation at scales previously impossible for many research groups. Third, we find that integrating sLLM annotations with even a small number of human labels yields synergistic improvements, suggesting a promising direction for hybrid annotation pipelines that combine the efficiency of automated systems with human domain expertise.

2 Related Work

2.1 LLMs in Data Annotation

LLMs have recently shown promise for data annotation tasks (Yao et al., 2024). Gao et al. (2020) demonstrated GPT-3's few-shot learning potential across NLP tasks with minimal fine-tuning, while Brown (2020) explored zero-shot and few-shot capabilities that could reduce the need for large task-specific datasets.

He et al. (2024a) introduced AnnoLLM, improving annotation by having LLMs explain their reasoning before providing labels. Their subsequent work showed GPT-4 can outperform standard crowdsourcing pipelines in annotation quality (He et al., 2024b). Li (2024a) compared LLMs and crowdsourcing across different NLP tasks, providing insights into their relative strengths.

Recent research has evaluated LLMs on specific annotation tasks. Törnberg (2023) assessed ChatGPT-4's ability to classify political affiliations of Twitter users, while Huang et al. (2023) exam-

ined its use for generating natural language explanations in hate speech detection. Ding et al. (2022) evaluated GPT-3 as a general-purpose data annotator, and Cegin et al. (2023) found ChatGPT-created paraphrases were more diverse than human annotations. Wu et al. (2023) explored LLMs' ability to replicate crowdsourcing pipelines.

While existing research often relies on large commercial LLMs, our work investigates how sLLMs can achieve comparable annotation results.

2.2 Crowdsourcing and Label Aggregation

Crowdsourcing has been widely used for data annotation (Howe et al., 2006). Snow et al. (2008) showed that aggregating non-expert annotations could rival expert quality. Kittur et al. (2008) emphasized the importance of task design, while Gadiraju et al. (2015) developed a taxonomy of microtask types to optimize crowdsourcing workflows.

Label aggregation is essential when combining annotations from multiple sources (Zheng et al., 2017; Liu et al., 2023; Ustalov et al., 2024; Li et al., 2019b; Whitehill et al., 2009; Li et al., 2019a; Welinder et al., 2010; Zhou et al., 2012; Liu et al., 2025, 2026). Dawid and Skene (1979) proposed an influential model using expectation-maximization to estimate true labels and annotator reliabilities. Whitehill et al. (2009) extended this with their GLAD algorithm, accounting for annotator expertise and item difficulty. Ratner et al. (2016) introduced data programming, combining multiple noisy labeling functions using a generative model.

While label aggregation techniques are widely used in crowdsourcing, our work is the first to investigate how these techniques can help smaller language models in data annotation tasks.

3 Background & Motivation

3.1 LLMs in Data Annotation and Key Requirements

LLMs represent a significant advancement over traditional crowdsourcing annotation approaches, which face limitations in scalability, consistency, and cost-effectiveness—especially in specialized domains requiring expert knowledge. Models like GPT-3 and GPT-4 have demonstrated capabilities that position them as potential automated annotators for diverse data mining tasks.

Research in LLM-based annotation has established several critical requirements: annotation accuracy comparable to expert human annotators

Algorithm 1 Statistical-based Aggregation

Require: Set of tasks T, Set of sLLMs M, Aggregation method A
Ensure: Aggregated labels L

$A_M \leftarrow$ CollectAnnotations(T, M)
$Q_M \leftarrow$ EstimateQuality(A_M)
$L \leftarrow \{\}$
for each task t in T **do**
    $l_t \leftarrow$ AggregateLables($A_M[t]$, Q_M, A)
    $L \leftarrow L \cup \{l_t\}$
end for
return L

across diverse contexts, consistency in producing stable labels, and addressing privacy and data sovereignty concerns—particularly in sensitive domains. While commercial LLMs demonstrate impressive capabilities, their API-based deployment and privacy implications have motivated exploration of smaller, open-source models and ensemble methods that can be deployed locally while maintaining quality.

3.2 Motivation

The democratization of high-quality data annotation represents a critical challenge in advancing modern NLP applications, yet current methodologies face fundamental barriers that significantly impede progress in the field. The first major obstacle is the prohibitive cost structure associated with commercial LLM. The significant annotation cost creating an insurmountable financial barrier for many academic researchers, non-profit organizations, and smaller research groups. This cost structure effectively creates a two-tiered research ecosystem, where only resource-rich institutions can fully leverage state-of-the-art annotation capabilities.

A second, equally concerning challenge lies in the methodological foundations of current evaluation frameworks. Many widely-used benchmarks, including RTE and QUIZ, were likely incorporated into the training data of modern LLMs (Li, 2024a,b; Snow et al., 2008). This methodological flaw becomes particularly problematic in dynamic domains requiring frequent annotation updates, such as emerging scientific literature or rapidly evolving technical fields, where the ability to generalize to post-training data distributions is crucial for practical applications.

The potential of smaller open-source LLMs

(sLLMs) offers a promising direction, with 100-1000x lower inference costs and inherent privacy advantages through local deployment. However, their adoption faces several unresolved challenges in the context of data annotation. The field has been dominated by an implicit assumption that annotation quality scales strictly with model size, discouraging systematic exploration of sLLM capabilities, particularly in ensemble settings. This assumption has limited research into potentially more efficient and accessible annotation solutions.

Furthermore, the lack of contemporary, contamination-free benchmarks has masked the true generalization capabilities of LLMs. Previous research has predominantly treated LLMs as singular annotators, overlooking the potential complementary strengths that could emerge from diverse sLLM architectures working in concert. This oversight has left unexplored a rich space of potential solutions that could combine the strengths of multiple smaller models to achieve robust annotation performance.

These challenges present a crucial opportunity for innovation in the NLP community. There is a pressing need for new approaches that can make high-quality annotation more accessible while ensuring reliable performance metrics that reflect real-world applicability. This is particularly crucial for sensitive domains like clinical text analysis and legal document processing, where both cost constraints and privacy requirements must be carefully balanced. Addressing these fundamental barriers could significantly accelerate progress across the broader field of data mining.

4 LABEL AGGREGATION

Label aggregation, inspired by the wisdom of the crowds' phenomenon (Alon et al., 2015) and traditional crowdsourcing techniques, can be adapted to the context of the crowd LLMs. This approach combines the outputs of various sLLMs to mitigate individual model biases and leverage the diverse strengths of different sLLMs in different areas.

The general framework for label aggregation in our sLLM-based annotation system is formalized in Algorithm 1. The process begins with the collection of annotations from all sLLMs for each task. Subsequently, the quality or reliability of each sLLM is estimated based on its performance across different benchmarks. This can also be obtained through comparison with a small set of gold

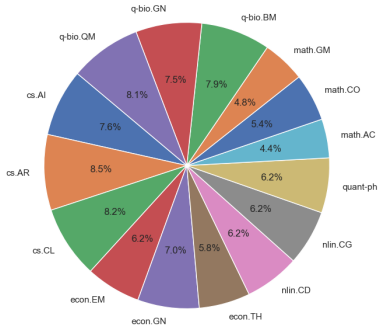

Figure 2: MORD dataset number of segment distribution.

standard labels. This estimation can be applied to a wide range of aggregation methods as the initial abilities. The core of the process is the aggregation step, where annotations from all sLLMs are combined using a method that considers the estimated quality of each model. This produces a final aggregated label for each task. Some methods may iterate this process, refining quality estimates and aggregated labels until convergence.

Within this framework, we implement and evaluate several SoTA label aggregation methods. These include simple approaches such as Majority Voting, which selects the most common label among all sLLMs for each task. We also explore probabilistic methods such as the DawidSkene Algorithm (Dawid and Skene, 1979), which employs an expectation-maximization approach to jointly estimate sLLM qualities and true labels. We also include methods like MMSR, Wawa, ZeroBasedSkill, GLAD, and MACE from crowdkit toolkit (Ustalov et al., 2024). Additionally, we also include the representative BWA, IBCC and EBCC algorithms based on Bayesian model (Li et al., 2019a). For each sentence segment in our dataset, we collect annotations from all 12 sLLMs and apply each aggregation method to determine the final aggregated annotation.

5 DATA COLLECTION AND ANNOTATION

Most of the existing works use existing datasets to evaluate the performance of LLMs which is shown to be problematic (He et al., 2024b). It is crucial for an unbiased assessment in which new datasets are curated and used. The process of data collection and annotation is crucial for evaluating the

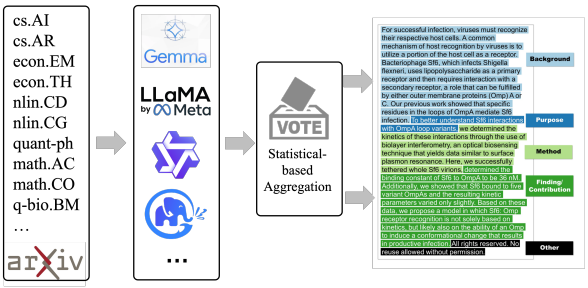

Figure 3: Annotation with the crowd of LLMs.

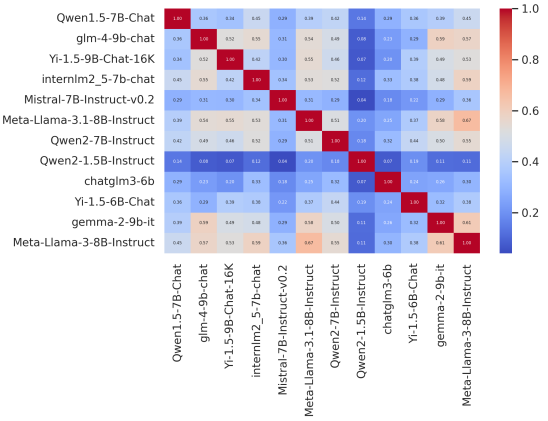

Figure 4: Label correlation of different sLLMs.

performance of language models in understanding and categorizing scientific text. This section details our methodology for selecting a diverse range of scientific papers, extracting relevant segments, and annotating them using both sLLMs and GPT-4 as a benchmark. Our approach aims to provide a comprehensive and fair assessment of sLLMs' capabilities in scientific text annotation across various disciplines.

5.1 Annotation Scheme and Data

We adopted the CODA-19 annotation scheme (Huang et al., 2020) for our study, which categorizes sentence segments in research papers into five aspects: Background, Purpose, Method, Finding/Contribution, and Other. This scheme was chosen for its proven effectiveness in capturing the essential components of scientific discourse across various disciplines.

We select scholarly articles spanning 5 distinct research fields with each have 3 subcategories. We selected papers from the arXiv repository in August 2024, focusing on the following categories:

1. Computer Science: cs.AI (Artificial Intelligence), cs.AR (Hardware Architecture), cs.CL (Computation and Language)

2. Economics: econ.EM (Econometrics), econ.GN (General Economics), econ.TH (Theoretical Economics)

3. Physics: nlin.CD (Chaotic Dynamics), nlin.CG (Cellular Automata and Lattice Gases), quant-ph (Quantum Physics)

4. Mathematics: math.AC (Commutative Algebra), math.CO (Combinatorics), math.GM (General Mathematics)

5. Quantitative Biology: q-bio.BM (Biomolecules), q-bio.GN (Genomics), q-bio.QM (Quantitative Methods)

We specifically chose recent arXiv papers to ensure a fair test for the language models. By selecting papers published after the training cutoff dates of the sLLMs, we mitigate the risk of data contamination and provide a more accurate assessment of the models' generalization capabilities on truly unseen data.

For each field, we randomly sampled 100 papers, resulting in a total of 1,500 papers, and then the abstract is split into segments. The distribution of the segments are shown in Figure 2. This diverse selection allows us to evaluate the sLLMs' performance across a wide range of scientific disciplines.

5.2 Annotating with sLLMs and GPT-4

We annotated the extracted sentence segments using 12 different sLLMs (detailed in Appendix Table 2) and GPT-4 as shown in Figure 3. The sLLMs were carefully selected from various organizations, each with fewer than 10 billion parameters, to ensure a diverse and manageable set of models. To visualize the relationships between the predictions of these models, we created a correlation plot with Cohen's kappa coefficient (Figure 4). Interestingly, this plot reveals that these models do not exhibit strong correlations with each other, suggesting a diversity in their approaches and outputs that could be beneficial for aggregation. We also included GPT-4 in our study as a representative of state-of-the-art performance, providing a high-quality benchmark for comparison with the sLLMs. To ensure consistency and fairness in our evaluation, all models, including GPT-4, were given identical annotation prompts. This standardized approach allows for a direct and meaningful comparison of performance across all models in the task of scien-

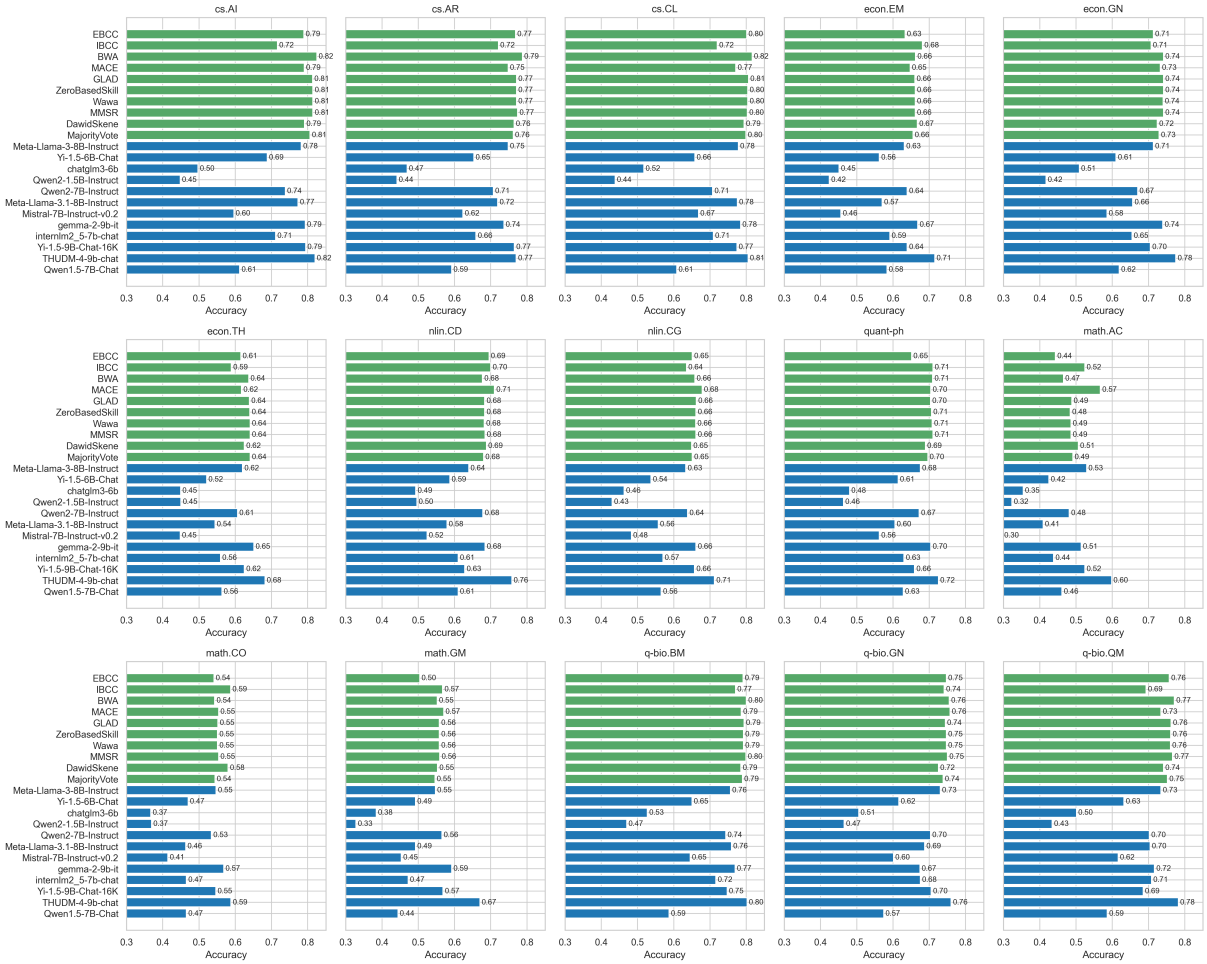

Figure 5: Comparison of annotation accuracy across research fields in MORD dataset: single sLLM model (blue) vs aggregated annotations (green).

tific text annotation.

The annotation process was fully automated, with each sLLM and GPT-4 independently processing all 12,277 sentence segments. This approach enables a direct comparison of different models' performance on the same task, offering valuable insights into their relative strengths and weaknesses in understanding and categorizing scientific text across various disciplines. In this process, we encountered a common challenge in working with language models: inconsistent adherence to output format instructions. To address this, we implemented a re-execution strategy for annotations that did not conform to the specified format. Rather than relying on human inspection, which can be time-consuming and potentially biased, we found that re-executing the annotation task for non-compliant outputs was more effective and efficient. This approach not only ensured consistency in our dataset but also maintained the integrity of our automated process. Importantly, we have included the additional computational costs from these re-executions in our overall cost estimations, providing a comprehensive and realistic view of the resources required for this annotation method. We also explored structure generation approaches (Willard and Louf, 2023), but their incomplete support across API providers could introduce bias in our comparisons. Therefore, we chose to implement the more straightforward re-execution strategy. Following previous work use GPT-4 to evaluate the LLM's performance (Zheng et al., 2023). We use GPT-4's annotations as the gold standard. By comparing the sLLMs' label against GPT-4, we can estimate the performance of the proposed method. This dataset is called the MORD dataset.

In addition to this, we also include the dataset in He et al. (2024b)'s work that includes crowdsourcing annotations. We also annotate this with 12 sLLMs as described above, we call this dataset the CORD dataset. This helps us to assess how

well smaller, more efficient models can approximate the capabilities of their larger counterparts in this specialized task.

6 RESULTS ANALYSIS

6.1 Performance in Different Research Fields

Figure 5 illustrates the performance of different annotation methods across various research fields, revealing important nuances in model and aggregation method effectiveness. The performance of both individual sLLMs and aggregation methods varies notably across different research domains, emphasizing the importance of considering field-specific characteristics when selecting models or designing annotation strategies.

However, our analysis also reveals that certain research fields present greater challenges for all methods (e.g., math). In these more difficult domains, we observe a general downward shift in performance across all models and aggregation techniques. This trend points to areas where further advancements in both model design and aggregation strategies could yield substantial benefits. It also underscores the need for domain-specific fine-tuning or the development of specialized models for particularly challenging scientific disciplines.

These demonstrate the significant potential of sLLMs as a cost-effective alternative to large commercial models for scientific text annotation. However, the variations in performance across different research fields and annotation categories also reveal areas for future improvement and specialization in both model development and aggregation techniques.

6.2 Compare LLMs with Crowdsourcing

Table 1 presents a detailed comparison of the performance of individual sLLMs, GPT-4, traditional crowdsourcing, and various label aggregation methods on the CORD dataset. The results are broken down by the five categories of the CODA-19 annotation scheme: Background, Purpose, Method, Finding, and Other.

Our analysis reveals that most sLLMs significantly outperform traditional crowdsourcing in terms of overall accuracy. The best-performing sLLM, *glm-4-9b-chat*, achieves an impressive accuracy of 0.816, which is substantially higher than the 0.464 accuracy obtained through crowd workers. This stark difference underscores the potential of sLLMs as a viable alternative to human annotators for scientific text categorization tasks. However, it's important to note that GPT-4 still demonstrates superior performance with an accuracy of 0.836, highlighting the current gap between large commercial models and open-source alternatives.

Examining the category-specific performance, we observe that sLLMs exhibit varying strengths across different annotation categories. Many models excel in identifying "Background" and "Finding" sections, with F1 scores often exceeding 0.8. For instance, gemma-2-9b-it achieves an F1 score of 0.811 for the "Background" category and 0.802 for "Finding". However, most models struggle with the "Purpose" and "Other" categories, indicating areas for potential improvement in model training or prompt engineering.

Interestingly, traditional crowdsourcing shows particularly low performance in identifying "Purpose" (F1 score 0.209) and "Other" (F1 score 0.320) categories. This suggests that human annotators may face difficulties in consistently recognizing these elements in scientific texts, possibly due to the nuanced nature of these categories or variations in annotator interpretation. The superior performance of sLLMs in these categories highlights their potential to capture subtle textual patterns that human annotators might overlook.

6.3 LLM-Human Cooperation Performance

To explore potential synergies between human annotators and language models, we investigated the performance of combining crowdsourced annotations with LLM-generated labels. As we can see from the last part of Table 1, the top-performing aggregation methods (Wawa, ZeroBasedSkill, and MACE), all achieving an accuracy of 0.840, seem to leverage strengths from different sources. These methods appear to combine the high precision of LLMs in categories like "Background" and "Method" with the high recall of crowdsourcing in categories like "Other".

For instance, in the "Other" category, Wawa achieves an F1 score of 0.842, significantly higher than most individual LLMs and approaching crowdsourcing performance in terms of recall. This suggests the successful incorporation of human annotators' strengths in identifying less common or more ambiguous categories.

6.4 Annotation Cost

To compare the annotation cost of different LLMs, we visualize the token usage and total cost associ-

Table 1: Performance in CORD dataset. The bold numbers indicate better than the best single LLM, and the underlined numbers indicate better than GPT-4.

Annotation	Background			Purpose			Method			Finding			Other			Acc
Methods	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
Crowdsourcing	.338	.692	.454	.571	.128	.209	.678	.408	.509	.415	.883	.565	.190	1.000	.320	.464
GPT-4	.913	.860	.885	.843	.499	.627	.871	.775	.820	.784	.982	.872	.905	.322	.475	.836
Qwen1.5-7B-Chat	.689	.737	.712	.318	.457	.375	.407	.505	.451	.689	.758	.722	.143	.088	.109	.600
Qwen2-7B-Instruct	.838	.703	.765	.406	.454	.428	.646	.670	.658	.781	.860	.819	.190	.667	.296	.735
Qwen2-1.5B-Instruct	.590	.574	.582	.023	.167	.040	.110	.556	.184	.805	.680	.737	.000	.000	.000	.551
glm-4-9b-chat	.913	.725	.808	.700	.543	.612	.850	.772	.809	.776	.973	.863	.762	.762	.762	.816
chatglm3-6b	.682	.524	.593	.313	.247	.276	.407	.483	.442	.575	.733	.644	.333	.044	.077	.543
Yi-1.5-9B-Chat-16K	.950	.572	.714	.585	.516	.549	.762	.730	.745	.619	.972	.756	.476	.909	.625	.719
Yi-1.5-6B-Chat	.821	.551	.660	.249	.342	.288	.453	.474	.463	.626	.778	.694	.429	.220	.290	.605
internlm2_5-7b-chat	.830	.721	.771	.544	.674	.602	.751	.692	.721	.786	.912	.844	.429	.113	.178	.769
gemma-2-9b-it	.881	.752	.811	.839	.426	.565	.769	.801	.785	.681	.976	.802	.476	.071	.124	.753
Mistral-7B-Instruct-v.2	.798	.487	.605	.544	.522	.533	.415	.786	.543	.545	.916	.684	.286	.038	.067	.571
Llama-3.1-8B-Instruct	.930	.635	.755	.608	.520	.561	.729	.765	.747	.715	.954	.817	.762	.390	.516	.758
Llama-3-8B-Instruct	.917	.647	.759	.677	.531	.595	.740	.747	.744	.693	.962	.805	.762	.176	.286	.751
MajorityVote	.778	.778	.778	.700	.400	.509	.815	.689	.747	.783	.951	.859	.429	1.000	.600	.781
DawidSkene	.881	.798	.837	.724	.447	.553	.826	.785	.805	.732	.972	.835	.571	.160	.250	.783
MMSR	.791	.802	.797	.700	.427	.531	.825	.695	.755	.800	.948	.868	.429	1.000	.600	.794
Wawa	.900	.778	.835	.705	.582	.637	.832	.780	.805	.837	.957	.893	.762	.941	.842	**.840**
ZeroBasedSkill	.903	.775	.834	.700	.587	.639	.829	.782	.805	.837	.957	.893	.762	.889	.821	**.840**
GLAD	.895	.783	.836	.710	.568	.631	.838	.777	.806	.832	.957	.890	.762	.941	.842	**.839**
MACE	.908	.776	.837	.696	.588	.637	.822	.812	.817	.839	.953	.892	.762	.444	.561	**.840**
BWA	.838	.797	.817	.710	.531	.607	.849	.718	.778	.814	.946	.875	.238	1.000	.385	.816
IBCC	.875	.810	.842	.760	.437	.555	.832	.794	.813	.811	.964	.881	.762	.842	.800	**.826**
EBCC	.891	.802	.844	.733	.447	.555	.835	.798	.816	.819	.967	.887	.476	1.000	.645	**.830**

ated with different LLMs for the annotation task in Figure 6. The cost analysis reveals a significant advantage for open-source sLLMs in terms of economic efficiency. While GPT-4 offers the highest accuracy at 0.836, its substantially higher cost may not justify its use in all scenarios, especially for projects with budget constraints or large-scale annotation requirements.

Among the sLLMs, models like glm-4-9b-chat and Qwen2-7B-Instruct offer an attractive balance between performance and cost. For instance, glm-4-9b-chat achieves an accuracy of 0.816, very close to that of GPT-4 (0.836), at a fraction of the cost. Specifically, our analysis shows that glm-4-9b-chat uses approximately 70% fewer tokens than GPT-4 for the same annotation task, translating to a cost reduction of over 80%. This cost-effectiveness becomes particularly crucial when considering large-scale annotation projects, where the cumulative cost savings could be substantial without significantly compromising annotation quality.

Furthermore, our analysis reveals that models from the same family, such as Qwen1.5-7B-Chat and Qwen2-7B-Instruct, can have notably different cost-performance ratios. Qwen2-7B-Instruct, for example, shows improved accuracy (0.735 vs 0.600) with only a marginal increase in token usage compared to its predecessor. This highlights the importance of considering not just the model size, but also the specific architecture and training approach when selecting an sLLM for annotation tasks. Additionally, the analysis also verify the effectiveness of the re-execution strategy, models such as Mistral-7B-Instruct-v2 occasionally consume up to $\approx 3\times$ tokens due to weaker instruction following, yet the low per-token cost of sLLMs makes automated re-execution far cheaper than human inspection by orders of magnitude.

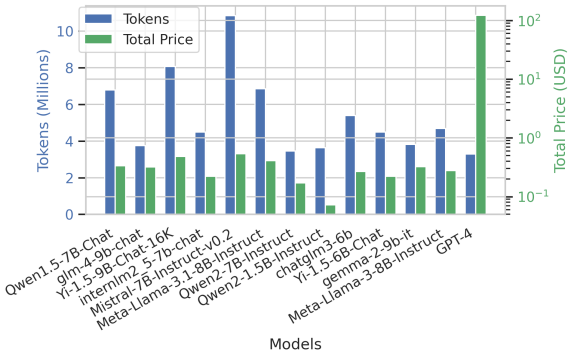

Figure 6: Used tokens and total cost of different LLMs.

7 CONCLUSION

This research demonstrates the significant potential of open-source sLLMs for cost-effective data annotation. Our findings show that sLLMs can achieve annotation quality surpassing individual human annotators and approaching state-of-the-art models like GPT-4, at a fraction of the computational and financial cost. The aggregation techniques can significantly enhance overall annotation quality. Our cost analysis reveals that sLLM-based annotation can be orders of magnitude more cost-effective than

commercial LLM solutions while maintaining high quality, a crucial finding for resource-constrained environments.

Acknowledgments

We sincerely thank all the anonymous reviewers for their valuable comments and feedback. This work was supported by the National Natural Science Foundation of China (No. 62441225). Xiaofeng Hou is the corresponding author.

Limitations

Despite the comprehensive nature of our experiments, several limitations should be acknowledged. Firstly, our study focused on a specific set of scientific disciplines and annotation tasks in English. The generalizability of our findings to other domains, language or more complex annotation tasks remains to be established. Additionally, the rapid pace of development in language model technology means that the relative performance of different sLLMs may evolve quickly, potentially affecting the stability of our results over time.

Another limitation lies in the potential biases inherent in the training data of the sLLMs and GPT-4. These biases could influence the annotation results in ways that are difficult to quantify or control for, especially in scientific domains where ground truth can be ambiguous or contested. Furthermore, our study did not extensively explore the prompts used in the experiments, which could be a significant factor in performance.

The reliance on the GPT-4 model to evaluate the performance of CrowdLLM in different fields may introduce bias, since the GPT-4 cannot get perfect annotation performance. Future research may need more annotation effort to evaluate against real labels.

References

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with applications*, 2:100006.

Noga Alon, Michal Feldman, Omer Lev, and Moshe Tennenholtz. 2015. How robust is the wisdom of the crowds? In *Twenty-Fourth International Joint Conference on Artificial Intelligence*. Citeseer.

Inc. Amazon.com. 2024. Amazon mechanical turk. Accessed: 2025-01-01.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

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

Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. 2018. A survey and analysis on automatic image annotation. *Pattern Recognition*, 79:242–259.

Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. 2016. A survey of general-purpose crowdsourcing techniques. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

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

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1631–1640.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Siegfried Handschuh and Steffen Staab. 2003. *Annotation for the semantic web*, volume 96. IOS press.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024a. 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.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024b. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25.

Jeff Howe and 1 others. 2006. The rise of crowdsourcing. *Wired magazine*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Ting-Hao'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C Lee Giles. 2020. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. *arXiv preprint arXiv:2005.02367*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Jiyi Li. 2024a. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529. IEEE.

Jiyi Li. 2024b. Human-llm hybrid text answer aggregation for crowd annotations. *arXiv preprint arXiv:2410.17099*.

Yuan Li, Benjamin IP Rubinstein, and Trevor Cohn. 2019a. Truth inference at scale: A bayesian model for adjudicating highly redundant crowd annotations. In *The World Wide Web Conference (WWW)*.

Yuan Li, Benjamin Rubinstein, and Trevor Cohn. 2019b. Exploiting worker correlation for label aggregation in crowdsourcing. In *International conference on machine learning (ICML)*.

Jiacheng Liu, Feilong Tang, and Xiaofeng Hou. 2023. Label aggregation with self-supervision enhanced graph transformer. *European Conference on Artificial Intelligence (ECAI)*.

Jiacheng Liu, Feilong Tang, Hao Liu, Long Chen, Yichuan Yu, Yanmin Zhu, Jiadi Yu, Xiaofeng Hou, and Pheng-Ann Heng. 2025. Bat: A versatile bipartite attention-based approach for comprehensive truth inference in mobile crowdsourcing. *IEEE Transactions on Mobile Computing*.

Jiacheng Liu, Feilong Tang, Hao Liu, Long Chen, Yanmin Zhu, Jiadi Yu, Yichuan Yu, and Xiaofeng Hou. 2026. Noisy multi-label aggregation with self-supervised graph transformer in mobile crowdsourcing. *IEEE Transactions on Mobile Computing*.

Mariana Neves and Ulf Leser. 2014. A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics*, 15(2):327–340.

Charles Oppenheim, Anne Morris, Cliff McKnight, and S Lowley. 2000. The evaluation of www search engines. *Journal of documentation*, 56(2):190–211.

Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for llm-as-a-judge. *arXiv preprint arXiv:2407.03479*.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, and Philip E Tetlock. 2024. Wisdom of the silicon crowd: Llm ensemble prediction capabilities match human crowd accuracy. *arXiv preprint arXiv:2402.19379*.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good?

evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Dmitry Ustalov, Nikita Pavlichenko, and Boris Tseitlin. 2024. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96):6227.

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*.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems (NIPS)*.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems (NIPS)*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, and 1 others. 2023. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, and 1 others. 2024. Llm inference unveiled: Survey and roofline model insights. *arXiv preprint arXiv:2402.16363*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment (PVLDB)*.

Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. 2012. Learning from the wisdom of crowds by minimax entropy. *Advances in neural information processing systems (NIPS)*.

A Additional Related Works

A.1 LLM-as-judge

A complementary line of work studies LLMs as evaluators ("LLM-as-judge"), where a powerful model adjudicates the quality of system outputs, either by providing direct scores or by ranking alternatives (Li et al., 2024; Pan et al., 2024). Such evaluators can reduce human adjudication cost and offer flexible, task-agnostic criteria, but they also introduce risks of bias, self-consistency artifacts when the judged and judging models share training data, and sensitivity to prompt phrasing and context length.

In contrast, our approach treats sLLMs as noisy "annotators" and applies label aggregation to combine multiple, diverse signals. Rather than relying on a single adjudicator, we leverage classical truth-inference methods (e.g., Dawid–Skene, GLAD, MACE) to estimate model reliabilities and item difficulty, which can improve robustness to individual model biases and yield calibrated uncertainty. We consider LLM-as-judge complementary but orthogonal to our goal of multi-annotator aggregation. Because it entails a different objective and evaluation pipeline (single adjudicator with rubric/prompt design), we do not include LLM-as-judge baselines here and leave that empirical comparison to future work.

B Used sLLMs

The list of sLLMs used in this paper is shown in Table 2.

C Data Visualization

To visualize the content distribution within each research field, we generated word clouds for each of the 15 sub-fields. Figure 7 presents these word clouds, offering an intuitive glimpse into the most frequently occurring terms and concepts within each domain. This visualization helps illustrate the thematic focus of the papers in our dataset and highlights the diversity of topics across the selected arXiv categories.

D Implications of Findings

Our study reveals several important insights into the use of sLLMs for data annotation and the effectiveness of various aggregation methods. Firstly, we observe that while individual sLLMs may not consistently outperform traditional crowdsourcing methods, their collective intelligence, when properly harnessed through sophisticated aggregation techniques, can yield superior results. This finding underscores the potential of sLLMs as a viable alternative or complement to human annotators.

The performance variations across different aggregation methods highlight the critical role of choosing an appropriate aggregation strategy. Methods that account for model-specific strengths and weaknesses, such as MACE, generally outperform simpler approaches like majority voting. This suggests that future annotation systems should incorporate adaptive aggregation strategies that can dynamically adjust to the characteristics of the input data and the performance patterns of individual sLLMs. It also opens new research opportunities to propose novel label aggregation methods specifically designed for aggregating crowds of LLMs.

Our experiments with prompt-based aggregation using LLMs demonstrate the potential of leveraging LLMs to reason over LLM outputs. Compared to traditional label aggregation algorithms, this method has the potential to reason with rich features. However, this also requires LLMs to have better reasoning abilities and may require more tokens to include answers from multiple LLMs. Despite the current performance of this method not meeting our expectations, we believe it also has great potential in the future.

Table 2: List of sLLMs used for annotation.

Model	Parameters	Context Length	Release Date
Qwen1.5-7B-Chat (Bai et al., 2023)	7B	32K	August 2023
chatglm3-6b (GLM et al., 2024)	6B	32K	October 2023
glm-4-9b-chat (GLM et al., 2024)	9B	128K	January 2024
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	7B	32K	March 2024
Meta-Llama-3-8B-Instruct (Dubey et al., 2024)	8B	8K	April 2024
Yi-1.5-6B-Chat (Young et al., 2024)	6B	4K	May 2024
Yi-1.5-9B-Chat-16K (Young et al., 2024)	9B	16K	May 2024
Qwen2-1.5B-Instruct (Yang et al., 2024)	1.5B	32K	June 2024
Qwen2-7B-Instruct (Yang et al., 2024)	7B	32K	June 2024
gemma-2-9b-it (Team et al., 2024)	9B	8K	June 2024
internlm2_5-7b-chat (Cai et al., 2024)	7B	32K	July 2024
Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024)	8B	128K	July 2024

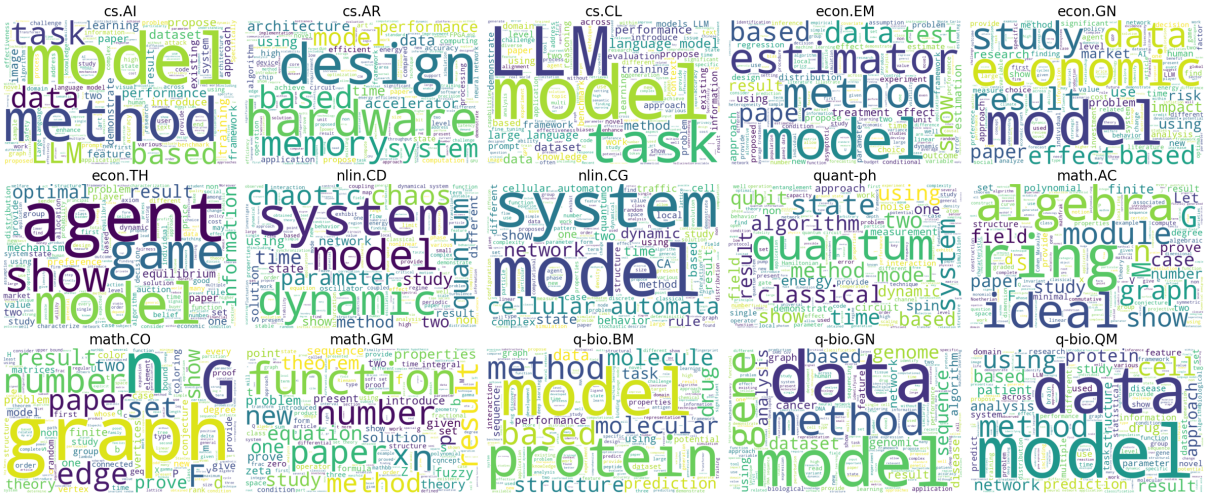

Figure 7: Word clouds representing the most frequent terms in each of the 15 arXiv sub-categories used in our study. The size of each word is proportional to its frequency in the respective sub-category.