Are Expert-Level Language Models Expert-Level Annotators?

Anonymous ACL submission

Abstract

Data annotation refers to the labeling or tagging of textual data with relevant information. A large body of works have reported positive results on leveraging LLMs as an alternative to human annotators. However, existing studies focus on classic NLP tasks, and the extent to which LLMs as data annotators perform in domains requiring expert knowledge remains underexplored. In this work, we investigate comprehensive approaches across three highly specialized domains and discuss practical suggestions from a cost-effectiveness perspective. To the best of our knowledge, we present the first systematic evaluation of LLMs as expert-level data annotators.

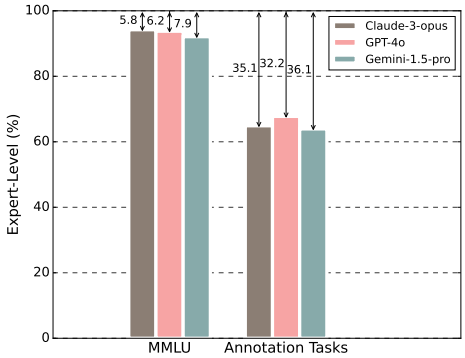

Figure 1: The degree of expert-level performance reached by state-of-the-art (SOTA) LLMs. For MMLU, we report model scores from the HELM (Liang et al., 2023a) website divided by human-expert score (89.8) from Hendrycks et al. (2020).

1 Introduction

Data annotation refers to the task of labeling or tagging textual data with relevant information (Tan et al., 2024). For example, adding topic keywords to social media contents. Typically, data annotation is carried out by crowd-sourced workers (*e.g.*, MTurkers) or specialized annotators (*e.g.*, researchers), depending on the tasks, to ensure high-quality annotations. However, the annotating procedures are often costly, time-consuming, and labor-intensive, particularly for tasks that require domain expertise.

With the rise of large language models (LLMs), a series of works have explored using them as an attractive alternative to human annotators (Ding et al., 2023; Zhang et al., 2023; Choi et al., 2024; He et al., 2023). Empirical results show that, in certain scenarios, LLMs such as ChatGPT and GPT-3.5 even outperform master-level MTurk workers, with substantially lower per-annotation cost (Gilardi et al., 2023; Alizadeh et al., 2023; Bansal and Sharma, 2023; Zhu et al., 2023). However, existing studies mainly focus on classic NLP tasks (*e.g.*, sentiment classification, word-sense disambiguation) on general domain datasets. The extent to which LLMs as data annotators perform in domains requiring expert knowledge remains unexplored.

On the other hand, LLMs have exhibited striking performance in a variety of benchmarks, both professional and academic (Jin et al., 2019; Hendrycks et al., 2020; Chen et al., 2021; Rein et al., 2023; Achiam et al., 2023). Leveraging the abundant domain-specific knowledge encoded in the parameters, LLMs could pass exams that require expert-level abilities (Choi et al., 2021; Singhal et al., 2023a; Callanan et al., 2023; Singhal et al., 2023b; Katz et al., 2024). These findings prompt the question – Can LLMs apply their parametric knowledge to perform expert-level annotation tasks?

To address this, we investigate three specialized domains: finance, biomedicine, and law. Specifically, we adopt six existing datasets that *(i)* provide fully-detailed annotation guidelines and *(ii)* are manually labelled by domain experts. We format the annotation task, the guideline, and unlabelled data instances as instructional inputs to the most performant, publicly-available LLMs, and evaluate their annotation results against ground-truth labelled by human experts. Experimental results

Model / Method	Finance		Biomedicine		Law		Avg.
	REFinD	FOMC	AP-Rel	CODA-19	CUAD	FoDS	
GPT-3.5-Turbo	47.4	60.4	58.9	64.4	71.8	37.1	56.7
GPT-4o	**67.2**	**67.6**	65.8	**79.3**	**82.2**	44.4	**67.8**
Gemini-1.5-Pro	64.6	**67.6**	54.8	73.2	80.6	42.8	63.9
Claude-3-Opus	61.2	63.6	**71.2**	65.6	80.8	**46.9**	64.9
GPT-4o	67.2	67.6	65.8	79.3	82.2	44.4	67.8
CoT	71.0 (↑3.8)	*68.2 (↑0.6)	*68.5 (↑2.7)	*81.1 (↑1.8)	79.8 (↓2.4)	43.9 (↓0.5)	68.7
Self-Consistency	*72.4 (↑5.2)	*70.4 (↑2.8)	*68.5 (↑2.7)	78.9 (↓0.4)	*†82.4 (↑0.2)	*†45.0 (↑0.6)	**69.6**
Self-Refine	70.0 (↑2.8)	*69.2 (↑1.6)	*69.9 (↑4.1)	*81.5 (↑2.2)	78.0 (↓4.2)	*45.5 (↑1.1)	69.0

Table 1: Performance of SOTA LLMs as annotators (accuracy) and comparison of GPT-4o with advanced techniques for expert-level annotation tasks. An asterisk (*) indicates statistical significance over the vanilla method ($p < 0.05$), and a dagger (†) indicates statistical significance of self-consistency over CoT ($p < 0.05$).

in our vanilla setting suggest that LLMs show substantial rooms for improvements, with an average of around 35% behind human expert annotators.

Towards a more comprehensive evaluation, we employ a variety of approaches tailored to elicit the capabilities in LLMs, including chain-of-thought (CoT), self-consistency, and self-refine promptings. Additionally, drawing inspiration from how human annotators reach consensus, we introduce a multi-agent annotation framework which incorporates a peer-discussion process for producing annotations. Lastly, we discuss practical suggestions on leveraging LLMs for expert annotation tasks, from a cost-effectiveness perspective. We summarized our main contributions as follows:

- We present, to the best of our knowledge, the first systematic evaluation of LLMs as expert-level data annotators.
- We explore comprehensive approaches, including prompt-based methods and multi-agent frameworks, across three highly specialized domains.
- We provide a cost-effectiveness analysis and practical suggestions on leveraging LLMs for expert annotation tasks.

2 Datasets

We adopt six datasets across three specialized domains: REFinD (Kaur et al., 2023) and FOMC (Shah et al., 2023) for finance, AP-Relation (Gao et al., 2022) and COVID-19 Research Aspect (CODA-19; Huang et al., 2020) for biomedicine, and Contract Understanding Atticus Dataset (CUAD; Hendrycks et al., 2021) and Function of Decision Section (FoDS; Guha et al., 2024) for law. To ensure data quality and fair comparison, we have checked that annotation labels in each dataset are annotated by domain expert. We provide descriptions, statistics, and annotation guidelines of each dataset in Appendix A and B.1.

3 LLMs as Expert Annotators

3.1 Methods

Vanilla The vanilla method refers to standard direct-answer prompting, where instructional input consists of the annotation task, guideline, and the instance are given to the LLMs. LLMs are tasked to conduct annotation as a domain expert of relevant fields. We utilized a uniform prompt template that is easily generalizable across domains and datasets. The vanilla prompt also serves as the base of other sophisticated approaches (described below). We provide all prompt templates in Appendix B.2.

CoT Prompting with chain-of-thought (CoT) improves LLMs' complex reasoning ability significantly (Wei et al., 2022). Specifically, we employ zero-shot CoT (Kojima et al., 2022), where a trigger phrase "*Let's think step by step*" augments the prompt to elicit reasoning chain from LLMs and leads to a more accurate answer.

Self-Consistency Self-consistency (SC; Wang et al., 2022) improves upon CoT via a sample-and-marginalize decoding procedure, which selects the most consistent answer rather than the greedily decoded one. Concretely, we sample 5 diverse reasoning paths with temperature 0.7, and take the majority vote to determine the final answer.

Self-Refine Self-refine (SR; Madaan et al., 2024) method includes three steps: generate, review, and refine. An LLM first generates an initial answer (*i.e.*, draft). Then, the model review its draft and provide feedback. Lastly, the LLM refine the draft

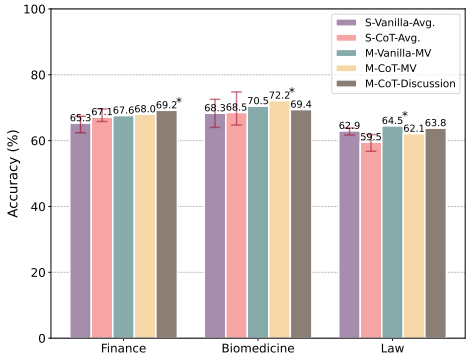

Figure 2: Comparison of performance between single LLM settings (S) and multi-agent frameworks (M) across three domains. For S, numbers represent average performance of GPT-4o, Gemini-1.5-Pro, and Claude-3-Opus, with red bars showing performance range. An asterisk (*) indicates statistical significance ($p < 0.05$).

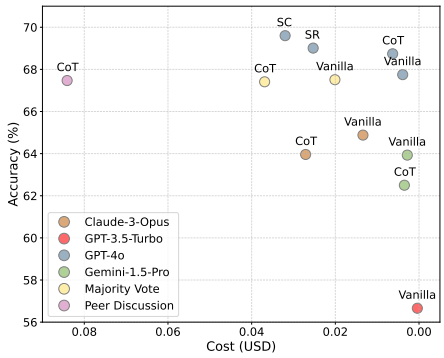

Figure 3: Illustration of the cost-effectiveness relationship of various setups. The x-axis shows cost per instance (USD), and the y-axis shows accuracy. Note that higher costs are on the left, making the orientation counter-intuitive. The upper right corner indicates better performance, with lower cost and higher accuracy.

by incorporating its feedback, and outputs an improved answer. The same LLM is used in all steps.

3.2 Results

We report our main results in Table 1. We compare four models, including GPT-3.5-Turbo (OpenAI, 2023), GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Reid et al., 2024), and Claude-3-Opus (Anthropic, 2024), and report their annotation accuracy. We use labels annotated by human experts from the corresponding dataset as ground-truth answers.

As observed, under the vanilla method (upper block), GPT-4o records the best overall performance. Claude-3-Opus and Gemini-1.5-Pro achieve similar scores, while GPT-3.5-Turbo performs notably worse. However, all LLMs show substantial rooms for improvements, with an average of 32.2% $\sim$ 43.3% behind human expert annotations. The best single score (GPT-4o on CUAD dataset) still lacks around 20%. The results suggest that naive standard prompting is *not* feasible to obtain satisfactory annotation quality from LLMs in tasks involving domain expertise.

To further probe the capabilities of LLMs, we experiment GPT-4o with three methods: CoT, SC, and SR, proposed to improve LLMs factual knowledge and reasoning. As shown in the lower block, all methods generally exhibit improvements, with an average of 1% $\sim$ 2% accuracy gain. However, comparing with the huge performance boosts of how these methods benefit general domain datasets, their efficacy on expert-level annotation tasks is relatively low. This might imply that the models inherently lack necessary knowledge and reasoning capability to perform as expert annotators.

4 Multi-Agent Annotation

The multi-agent framework, where multiple LLMs communicate with each other to solve tasks in a collaborative manner, has become a prevalent research direction (Liang et al., 2023b; Du et al., 2023; Chen et al., 2023; Tseng et al., 2024). A common scenario in annotation is the disagreement among multiple annotators. A typical way for resolving such discrepancy is by discussing with others to reach a consent. Motivated by this, we design a multi-agent annotation framework, which incorporates a peer-discussion process mimicking human annotators for better annotations. Our multi-agent annotation framework consists of three LLMs: GPT-4o, Gemini-1.5-Pro, and Claude-3-Opus.

4.1 Methods

Majority Vote Majority vote (MV) represents a minimal form of discussion, reducing the process to selecting the majority output as the final annotation. We apply two settings for MV: vanilla and CoT.

Peer-Discussion Peer-Discussion involves three steps: (*1*) Generate initial annotations, (*2*) Check consensus, (*3*) Discuss and re-annotate. Initially, each agent generates its own annotation through CoT prompting given the same annotation task, guideline, and instance. Next, we check for consensus (*i.e.*, if all annotations are the same labels). If consensus is achieved, the instance is successfully annotated and the process completes. If not, we compile all agents' reasoning and labels into a "*Discussion History*". Agents then re-annotate using the same input and discussion history. This check-discuss-re-annotate cycle continues until consensus

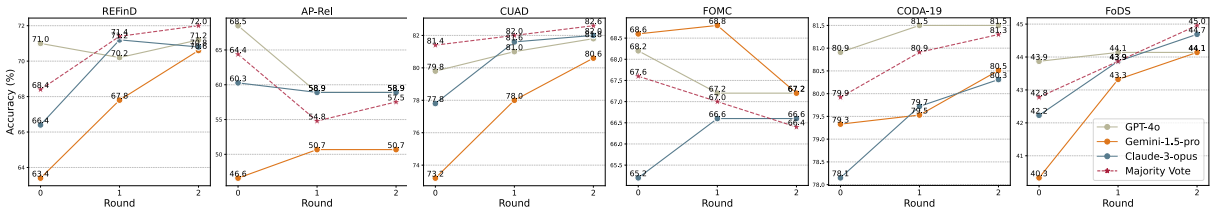

Figure 4: Marginal performance of each LLM during multi-agent peer-discussion process. The performance in Round 0 indicates the LLMs' initial annotation performance, while the performance in Round 1 and Round 2 indicates the LLMs' annotation performance after one or two rounds discussion, respectively.

is achieved or the maximum discussion round is reached, which we set to 2 in our experiments. We provide prompt templates in Appendix B.2.

4.2 Results

As shown in Figure 2, multi-agent frameworks slightly outperform the average of single-LLM settings. However, the most performant singe LLM (primarily GPT-4o) still exhibits superior results in vanilla and CoT setting, similar to the observation by Wang et al. (2024). Though multi-agent methods seem inferior to the best single-LLM setting, it may be a more suitable approach when we are unable to infer which model to adopt in advance.

To further investigate the discussion process in the multi-agent peer-discussion setting, we present the marginal performance of each LLM over two rounds in Figure 4. As observed, the peer-discussion process benefits most datasets, leading to improved annotation performance (i.e., increasing *Majority Vote* lines in the figure). For the AP-Rel dataset, however, peer discussion alternately hurt the performance. We hypothesize a potential reason is that AP-Rel contains many highly specialized, biomedical-domain specific entities, which are especially challenging for the models.

Experimental results suggest that GPT-4o tends to provide good insight to others, in the mean time, maintaining its own correct answer, especially shown in the REFinD, CUAD, and FoDS datasets. Under the circumstances, not only does the multi-agent framework performance increase, the individual performance of the three LLMs also benefits from the discussion process. In contrast, Gemini-1.5-Pro exhibit a tendency to follow and agree with others points of views, instead of defending its own correct annotation, notably demonstrated within the FOMC and CODA-19 datasets. Therefore, affecting the final discussion performance and its own performance.

5 Cost-Effectiveness Discussion

We aggregate our empirical results and compile a cost-effectiveness illustration in Figure 3. In sum, GPT-4o with vanilla or CoT method presents as the best cost-effective options within our experiments. GPT-4o with SC achieves the best overall performance at the expense of tripling the cost. Considering to the real-world annotation scenarios, there are no oracles to determine which LLM could perform the most accurately against an annotation task. Thus, an intermediate option would be multi-agent frameworks, which demonstrate competitive performance and could be a more robust option when access to different LLMs are available. Despite LLMs do not present as a direct alternative for annotation tasks requiring domain expertise, their collective performance of over 50% and profoundly lower cost present a promising human-LLM hybrid annotation schema in the future.

6 Conclusion

In this work, we present a pilot focused contribution, as a first attempt, towards answering the research question – "Can top-performing LLMs, which might be perceived as having expert-level proficiency in academic and professional benchmarks, be an alternative of human-expert annotators?". The core finding, from the empirical results on both single-LLM and multi-agent methods across three domains with six datasets, indicates that the current generation of LLMs do *not* present as a direct alternative for annotation tasks requiring domain expertise. Solely relying on the parametric knowledge in LLMs to perform domain-specific, expert-level annotation tasks is non-trivial. Considering that these specialized domains are often relevant to high-risk sectors (*e.g.*, medical application), it is crucial to ensure the annotated data has a higher precision and accuracy.

Limitations

As we aim to provide direct insight and observation on whether top-performing LLMs can perform as expert annotators *out-of-the-box*, we minimize efforts in prompt engineering. Some works have demonstrated that, for specific scenarios, one can achieve sizable improvement through carefully-crafted prompts. Consequently, our results may further benefit from a more exhaustive prompt optimization.

Another potential limitation is that we primarily focus on natural language understanding (NLU) tasks with fixed label space. Towards a more comprehensive evaluation, natural language generation (NLG) tasks could be further incorporated.

References

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

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *arXiv preprint arXiv:2310.08678*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. Chatgpt goes to law school. *J. Legal Educ.*, 71:387.

Juhwan Choi, Eunju Lee, Kyohoon Jin, and Young-Bin Kim. 2024. Gpts are multilingual annotators for sequence generation tasks. *arXiv preprint arXiv:2402.05512*.

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.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M Churpek, and Majid Afshar. 2022. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources & Evaluation*, volume 2022, page 5484. NIH Public Access.

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

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

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. In *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3054–3063.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023a. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023b. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-3.5 turbo.

OpenAI. 2024. Hello gpt4-o.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

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*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

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

A Datasets

Finance We adopt the REFinD (Kaur et al., 2023) and FOMC datasets (Shah et al., 2023) for financial domain. REFinD is the largest relation extraction dataset over financial documents, comprising 8 entity pairs and 29 relations, with labels reviewed by financial experts. In this task, annotators are tasked to extract relations between finance-specific entity pairs, such as [*person*] is an employee of [*organization*]. FOMC is constructed for identifying sentiments about the future monetary policy stances, annotated by experts with a correlated financial knowledge. The labels of this annotation task are Dovish, Hawkish, and Neutral, where a Dovish sentence indicates easing and a Hawkish sentence indicates tightening.

Biomedicine For the biomedical domain, we utilize AP-Relation dataset (Gao et al., 2022) and COVID-19 Research Aspect Dataset (CODA-19; Huang et al., 2020). AP-Relation is designed for extracting the relationship between Assessment and Plan Subsections in daily progress notes. The Assessment describes the patient and establishes the main symptoms or problems for their encounter, while the Plan Subsection addresses each differential diagnosis or problem with a daily action or treatment plan. The annotation label schemes for different relations are categorized as *direct*, *indirect*, *neither*, or *not relevant*. CODA-19 codes each segment aspect of English abstracts in the COVID-19 Open Research Dataset (Wang et al., 2020). In this task, annotators are tasked to label each segment as *background*, *purpose*, *method*, *finding/contribution*, or *other* sections. To ensure the quality of the labels, we only adopt instances annotated by biomedical experts.

Law In the legal domain, we adopt Contract Understanding Atticus Dataset (CUAD; Hendrycks et al., 2021) and Function of Decision Section dataset (FoDS; Guha et al., 2024). CUAD consists of legal contracts with extensive annotations from legal experts, created with a year-long effort by dozens of law student annotators, lawyers, and machine learning researchers. The annotation task is to label 41 types out of legal clauses, classified into 5 answer categories, that are considered important in contract review related to corporate transactions. We manually use "Yes/No" answer category to construct our annotation task as the identification of 32 types of clauses. FoDS comprises one-paragraph excerpts from legal decisions, annotated by legal professionals who are included as authors. In this task, annotators are tasked to review a legal decision and identify one out of seven function categories that each section (*i.e.*, excerpt) of the decision serves. We provide data statistics of the six existing specialized datasets in Table 2.

B Implementation Details

B.1 Annotation Guidelines

We provide annotation guidelines of each dataset from Figure 5 to Figure 11.

B.2 Prompt Templates

We provide prompt templates of each methods from Figure 12 to Figure 18.

Domain	Dataset	Instance Type	#Instances	#Labels
Finance	REFinD (Kaur et al., 2023)	Sentence	500	29
	FOMC (Shah et al., 2023)	Sentence	500	3
Biomedicine	AP-Rel (Gao et al., 2022)	Pair	73	4
	CODA-19 (Huang et al., 2020)	Paper Abstract	508	5
Law	CUAD (Hendrycks et al., 2021)	Clause	500	32
	FoDS (Guha et al., 2024)	Excerpt	367	7

Table 2: The statistics of existing domain-specific datasets.

Relation Extraction (RE) is the task of extracting relationships between entities in a sentence.

You will be given a sentence that contains two entities: entity 1 and entity 2.
Entity 1 is enclosed in double asterisks (i.e., **entity**) and entity 2 is enclosed in double underscores (i.e., __entity__).
Each entity has its own entity type specified in square brackets before the entity (e.g., [PERSON]**entity 1**).
The definition of the entity types are as follows:
- PERSON: People, including fictional.
- ORG: Companies, agencies, institutions, etc.
- UNIV: Universities, colleges, etc.
- GOV_AGY: Government agencies and departments.
- DATE: Absolute or relative dates or periods.
- GPE: Countries, cities, states.
- MONEY: Monetary values, including unit.
- TITLE: Positions or titles, including military.

Please annotate the relation between entity 1 and entity 2 described in the given sentence according to the following label descriptions.
Note that the relation is directional, meaning that the order of entity 1 and entity 2 matters.
Note that you can only select the most appropriate label that is consist of the given type of entities.
If you think there is no relation or other relation between entity 1 and entity 2, please select the label 0.

- 0: **entity 1** has no relation or other relation to __entity 2__
- 1: [PERSON]**entity 1** has/had the job title of [TITLE]__entity 2__
- 2: [PERSON]**entity 1** is/was an employee of [ORG]__entity 2__
- 3: [PERSON]**entity 1** is/was a member of [ORG]__entity 2__
- 4: [PERSON]**entity 1** is/was a founder of [ORG]__entity 2__
- 5: [PERSON]**entity 1** is/was a employee of [UNIV]__entity 2__
- 6: [PERSON]**entity 1** is/was a member of [UNIV]__entity 2__
- 7: [PERSON]**entity 1** has/had attended [UNIV]__entity 2__
- 8: [PERSON]**entity 1** is/was a member of [GOV_AGY]__entity 2__
- 9: [ORG]**entity 1** is/was formed on [DATE]__entity 2__
- 10: [ORG]**entity 1** is/was acquired on [DATE]__entity 2__
- 11: [ORG]**entity 1** is/was headquartered in [GPE]__entity 2__
- 12: [ORG]**entity 1** has/had operations in [GPE]__entity 2__
- 13: [ORG]**entity 1** is/was formed in [GPE]__entity 2__
- 14: [ORG]**entity 1** has/had shares of [ORG]__entity 2__
- 15: [ORG]**entity 1** is/was a subsidiary of [ORG]__entity 2__
- 16: [ORG]**entity 1** is/was acquired by [ORG]__entity 2__
- 17: [ORG]**entity 1** has/had a agreement with [ORG]__entity 2__
- 18: [ORG]**entity 1** has/had a revenue of [MONEY]__entity 2__
- 19: [ORG]**entity 1** has/had a profit of [MONEY]__entity 2__
- 20: [ORG]**entity 1** has/had a loss of [MONEY]__entity 2__
- 21: [ORG]**entity 1** has/had a cost of [MONEY]__entity 2__

Figure 5: The annotation guideline of REFinD dataset.

Hawkish-Dovish classification is to classify the sentiment about the future monetary policy stance into Dovish, Hawkish, or Neutral.
In general:
- 0: Dovish sentences were any sentence that indicates future monetary policy easing.
- 1: Hawkish sentences were any sentence that would indicate a future monetary policy tightening.
- 2: Neutral sentences were those with mixed sentiment, indicating no change in the monetary policy, or those that were not directly related to monetary policy stance.

You will be given a sentence that falls into one of the following eight categories enclosed in square brackets. Please annotate the sentiment of the sentence according to the following detailed label descriptions.
Note that you can only select one label that is most appropriate.

Detailed label descriptions:
[Economic Status: A sentence pertaining to the state of the economy, relating to unemployment and inflation.]
- 0: when inflation decreases, when unemployment increases, when economic growth is projected as low.
- 1: when inflation increases, when unemployment decreases when economic growth is projected high when economic output is higher than potential supply/actual output when economic slack falls.
- 2: when unemployment rate or growth is unchanged, maintained, or sustained.
[Dollar Value Change: A sentence pertaining to changes such as appreciation or depreciation of value of the United States Dollar on the Foreign Exchange Market.]
- 0: when the dollar appreciates.
- 1: when the dollar depreciates.
- 2: N/A
[Energy/House Prices: A sentence pertaining to changes in prices of real estate, energy commodities, or energy sector as a whole.]
- 0: when oil/energy prices decrease, when house prices decrease.
- 1: when oil/energy prices increase, when house prices increase.
- 2: N/A
[Foreign Nations: A sentence pertaining to trade relations between the United States and a foreign country. If not discussing United States we label neutral.]
- 0: when the US trade deficit decreases.
- 1: when the US trade deficit increases.
- 2: when relating to a foreign nation's economic or trade policy.
[Fed Expectations/Actions/Assets: A sentence that discusses changes in the Fed yields, bond value, reserves, or any other financial asset value.]
- 0: Fed expects subpar inflation, Fed expecting disinflation, narrowing spreads of treasury bonds, decreases in treasury security yields, and reduction of bank reserves.
- 1: Fed expects high inflation, widening spreads of treasury bonds, increase in treasury security yields, increase in TIPS value, increase bank reserves.
- 2: N/A
[Money Supply: A sentence that overtly discusses impact to the money supply or changes in demand.]
- 0: money supply is low, M2 increases, increased demand for loans.
- 1: money supply is high, increased demand for goods, low demand for loans.
- 2: N/A
[Key Words/Phrases: A sentence that contains key word or phrase that would classify it squarely into one of the three label classes, based upon its frequent usage and meaning among particular label classes.]
- 0: when the stance is "accommodative", indicating a focus on "maximum employment" and "price stability".
- 1: indicating a focus on "price stability" and "sustained growth".
- 2: use of phrases "mixed", "moderate", "reaffirmed".
[Labor: A sentence that relates to changes in labor productivity.]
- 0: when productivity increases.
- 1: when productivity decreases.
- 2: N/A

Figure 6: The annotation guideline of FOMC dataset.

A/P Relation classification is to classify the relation between Assessment and Plan Subsection in daily progress notes into DIRECT, INDIRECT, NEITHER, or NOT RELEVANT.

You will be given a pair of passages, Assessment and Plan Subsection, from daily progress notes.
Assessment describes the patient and establishes the main symptoms or problems for their encounter.
Plan Subsection addresses each differential diagnosis/problem with an action plan or treatment plan for the day.

Please annotate the relation between Assessment and Plan Subsection in the given pair according to the following label descriptions.
Note that you can only select one label that is most appropriate.

Label descriptions:
- 0: DIRECT. Assessment section includes a primary diagnosis/problem and it is mentioned in the Plan subsection, or Progress note includes a primary diagnosis/problem for hospitalization and it is mentioned in the Plan subsection, or Plan subsection contains a problem/diagnosis related to the primary signs/symptoms in the Assessment section.
- 1: INDIRECT. Plan subsection contains complications/subsequent events or organ failure related to the primary diagnosis/problem from the Assessment section, or Plan subsection contains other listed diagnoses/problems from the overall Progress Note or in the Assessment section that are not part of the primary diagnosis/problem, or Plan subsection contains a diagnosis/problem that is not previously mentioned but closely related (i.e., same organ system) to the primary diagnoses/problems mentioned in the overall Progress Note or Assessment section.
- 2: NEITHER. None of the criteria for Directly Related or Indirectly Related are met but a diagnosis/problem or other signs/symptoms are mentioned.
- 3: NOT RELEVANT. Plan subsection does not include a diagnosis/problems OR signs/symptoms.

Figure 7: The annotation guideline of AP-Relation dataset.

You will be given one paper abstract comprising several segments.
Each segment is a short text describing a specific aspect of the paper, including background, purpose, method, finding/contribution, or other.

Please annotate the aspects of each segment according to the following label descriptions.
Note that you can only select one label that is most appropriate for each segment. The total number of labels must be equal to the number of segments in the abstract.

Label descriptions:
- 0: Background. “Background” text segments answer one or more of these questions: Why is this problem important?, What relevant works have been created before?, What is still missing in the previous works?, What are the high-level research questions?, How might this help other research or researchers?
- 1: Purpose. “Purpose” text segments answer one or more of these questions: What specific things do the researchers want to do?, What specific knowledge do the researchers want to gain?, What specific hypothesis do the researchers want to test?
- 2: Method. “Method” text segments answer one or more of these questions: How did the researchers do the work or find what they sought?, What are the procedures and steps of the research?
- 3: Finding/Contribution. “Finding/Contribution” text segments answer one or more of these questions: What did the researchers find out?, Did the proposed methods work? Did the thing behave as the researchers expected?
- 4: Other. Text segments that do not fit into any of the four categories above. Text segments that are not part of the article. Text segments that are not in English. Text segments that contain only reference marks (e.g., “[1,2,3,4,5]”) or dates (e.g., “April 20, 2008”). Captions for figures and tables (e.g. “Figure 1: Experimental Result of ...”). Formatting errors. Text segments the annotator does not know or is not sure about.

Figure 8: The annotation guideline of CODA-19 dataset.

You will be given a clause from a legal contract. Please annotate the category of the given clause according to the following label descriptions.
Note that you can only select one label for each segment that is most appropriate.

Label descriptions:
- 0: Most Favored Nation. This clause provides that if a third party gets better terms on the licensing or sale of technology/goods/services described in the contract, the buyer of such technology/goods/services under the contract shall be entitled to those better terms.
- 1: Non-Compete. This clause imposes a restriction on the ability of a Party to compete with the other party or operate in a certain geography or business or technology sector.
- 2: Exclusivity. This clause provides for an exclusive dealing commitment between the parties of a contract. This clause also includes: a commitment by a party to procure all "requirements" from the other party of certain technology, goods, or services; or a prohibition against licensing or selling technology, goods or services to third parties, or a prohibition on collaborating or working with other parties.
- 3: No-Solicit of Customers. This clause restricts a party from soliciting, contacting or doing business with the other party's customers, vendors or partners.
- 4: Competitive Restriction Exception. This clause states the exception(s) to one of the following three labels: Exclusivity, Non-Compete, or No-Solicit of Customers.
- 5: No-Solicit of Employees. A No-Solicit of Employee clause prohibits a party from soliciting or hiring the other party's employees or consultants for itself or for a third party, during the contract or after the contract ends (or both).
- 6: Non-Disparagement. This clause requires a party not to disparage or defame the other party's goodwill, reputation or image.
- 7: Termination for Convenience. This clause allows a party to terminate a contract without cause or penalty. It allows a party to unilaterally terminate a contract by giving notice and oftentimes after a waiting period expires.
- 8: Right of First Refusal, Offer or Negotiation (Rofr/Rofo/Rofn). This clause grants one party a right of first refusal, right of first offer or right of first negotiation to purchase, license, market, or distribute equity interest, technology, assets, products or services.
- 9: Change of Control. This clause requires consent or notice of the other party if a party undergoes a change of control, such as a merger, stock sale, transfer of all or substantially all of its assets or business (collectively, "CIC").
- 10: Anti-Assignment. This clause requires a party to seek consent or notice if the contract is assigned, transferred or sublicensed to a third party, in whole or in part.
- 11: Revenue/Profit Sharing. This clause requires one party to share revenue or profit with the other party for any technology, goods, or services.
- 12: Price Restriction. This clause restricts the ability of a party to raise or reduce prices of technology, goods, or services provided.
- 13: Minimum Commitment. This clause requires a minimum order size or minimum amount or units per-time period that one party must buy from the counterparty under the contract.
- 14: Volume Restriction. This clause charges a fee or requires consent if one party's use of the product/services exceeds a certain threshold.
- 15: IP Ownership Assignment. This clause provides that intellectual property created by one party becomes the property of the other party, either per the terms of the contract or upon the occurrence of certain events.

Figure 9: The annotation guideline of CUAD dataset (1-1).

- 16: Joint IP Ownership. This clause provides for joint or shared ownership of intellectual property between the parties to the contract.
- 17: License Grant. This clause authorizes a party to use intellectual property or intangibles of the other party. It can be an authorization to use or to reproduce, distribute, manufacture, etc. certain content, technology, or other items that are protected by intellectual property rights. This clause is very common, and is considered one of the "factual" clauses. The purpose of this label is to help human reviewers to understand what IP is licensed under a contract and what restrictions are imposed on the license, including restrictions on duration, territory and purpose of use.
- 18: Non-Transferable License. This clause prohibits one party to transfer, assign or sublicense IP in the contract.
- 19: Affiliate IP License-Licensor. This clause contains a license grant by affiliates of the licensor or that includes intellectual property of affiliates of the licensor.
- 20: Affiliate IP License-Licensee. This clause contains a license grant to a licensee (incl. sublicensor) and the affiliates of such licensee/sublicensor.
- 21: Unlimited/All-You-Can-Eat License. This clause contains a provision granting one party an "enterprise," "all you can eat" or unlimited usage license.
- 22: Irrevocable or Perpetual License. This clause contains an irrevocable and/or perpetual license of IP. An irrevocable license is a perpetual license that cannot be cut short or terminated. A perpetual license, on the other hand, may not be irrevocable. Namely, a perpetual license can be terminated upon specified events such as material breach. Many license grant clauses use "irrevocable" and "perpetual" in the same sentence. The intent of some contracts may be to use the two terms interchangeably. As a result, for the purpose of CUAD, you should label the two types of licenses under the same label.
- 23: Source Code Escrow. This clause requires one party to deposit its source code into escrow with a third party or into a deposit account with the other party, which can be released to the other party upon the occurrence of certain events (bankruptcy, insolvency, etc.).
- 24: Post-Termination Services. This clause imposes obligations on a party after the termination or expiration of a contract, including any post-termination transition, payment, transfer of IP, wind-down, last-buy, or similar commitments.
- 25: Audit Rights. This clause grants one party the right to audit the books, records, or physical locations of the other party to ensure compliance with the terms of a contract.
- 26: Uncapped Liability. This clause leaves a party's liability uncapped upon the breach of its obligation in the contract. This also includes uncap liability for a particular type of breach such as IP infringement or breach of confidentiality obligation.
- 27: Cap On Liability. This clause includes a cap on liability upon the breach of a party's obligation. This includes time limitation for the counterparty to bring claims or maximum amount for recovery.
- 28: Liquidated Damages. This clause is an agreement to pay a party a pre-determined amount of damages if the other party breaches the contract. For the purpose of CUAD, this clause also includes an early termination fee.
- 29: Insurance. This clause requires a party to maintain insurance for the benefit of the other party.
- 30: Covenant not to Sue. This clause restricts a party from contesting the validity of the other party's ownership of intellectual property or otherwise bringing a claim against the other party that goes beyond the scope of standard Limitation on Liability clauses.
- 31: Third Party Beneficiary. This clause provides that a non-contracting party is a beneficiary to some or all of the clauses in the contract and therefore can enforce its rights against a contracting party.

Figure 10: The annotation guideline of CUAD dataset (1-2).

You will be given a one-paragraph excerpt of a legal decision. Please annotate the category of the given excerpt according to the following label descriptions.
Note that you can only select one label that is most appropriate for the excerpt.

Label descriptions:
- 0: Facts. A section of the decision that recounts the historical events and interactions between the parties that gave rise to the dispute.
- 1: Procedural History. A section of the decision that describes the parties' prior legal filings and prior court decisions that led up to the issue to be resolved by the decision.
- 2: Issue. A section of the decision that describes a legal or factual issue to be considered by the court.
- 3: Rule. A section of the decision that states a legal rule relevant to resolution of the case.
- 4: Analysis. A section of the decision that evaluates an issue before the court by applying governing legal principles to the facts of the case
- 5: Conclusion. A section of the decision that articulates the court's conclusion regarding a question presented to it.
- 6: Decree. A section of the decision that announces and effectuates the court's resolution of the parties' dispute, for example, granting or denying a party's motion or affirming, vacating, reversing, or remanding a lower court's decision.

Figure 11: The annotation guideline of FoDS dataset.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the label in the format of: 'The label is ...'. Do not include any reasoning or explanation.

Figure 12: Vanilla prompt template.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: '**Let's think step by step.** ... The label is ...'.

Figure 13: Chain-of-Thought prompt template.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: '**Let's think step by step.** ... The label is ...'.

Figure 14: Self-Refine prompt template. Step 1: Generate.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

{**[model response from step 1.]**}

Review your previous reasoning and annotation and find potential problems. For example, whether the annotation guideline is violated, whether the reasoning is not conclusive.

Figure 15: Self-Refine prompt template. Step 2: Review.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

{**[model response from step 1.]**}

Review your previous reasoning and annotation and find potential problems. For example, whether the annotation guideline is violated, whether the reasoning is not conclusive.

Review:
{**[model response from step 2.]**}

Based on the problems you found in the above review, improve your annotation quality and reasoning and output in the format of: 'Let's think step by step ..... The label is ...'.

Figure 16: Self-Refine prompt template. Step 3: Refine.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: '**Let's think step by step.** ...
The label is ...'.

Figure 17: Multi-agent peer-discussion prompt template. Step 1: Generate initial annotation.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ...
The label is ...'.

Discussion History:
{**[discussion_history]**}

You need to consider the above discussion history carefully. You can maintain your point of view and annotation if others' reasons are not concrete or cannot convince you.
Please strictly follow the guideline and output the reasoning and the label in the format of: '**Let's think step by step.** ...
The label is ...'.

Figure 18: Multi-agent peer-discussion prompt template. Step 2: Discuss and re-annotate.