EVALUATING LLMS WITHOUT ORACLE FEEDBACK: AGENTIC ANNOTATION EVALUATION THROUGH UNSUPERVISED CONSISTENCY SIGNALS

Cheng Chen[1,3], Haiyan Yin[3], Ivor W. Tsang[2,3]

[1] Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia
[2] College of Computing and Data Science, Nanyang Technological University, Singapore
[3] Centre for Frontier AI Research, Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
Cheng.chen-16@student.uts.edu.au, {ivor_tsang, yin_haiyan}@cfar.a-star.edu.sg

ABSTRACT

Large Language Models (LLMs), when paired with prompt-based tasks, have significantly reduced data annotation costs and reliance on human annotators. However, evaluating the quality of their annotations remains challenging in dynamic, unsupervised environments where oracle feedback is scarce and conventional methods fail. To address this challenge, we propose a novel agentic annotation paradigm, where a student model collaborates with a noisy teacher (the LLM) to assess and refine annotation quality without relying on oracle feedback. The student model, acting as an unsupervised feedback mechanism, employs a user preference-based majority voting strategy to evaluate the consistency of the LLM's outputs. To systematically measure the reliability of LLM-generated annotations, we introduce the **Consistent and Inconsistent (CAI)** Ratio, a novel unsupervised evaluation metric. The CAI Ratio not only quantifies the annotation quality of the noisy teacher under limited user preferences but also plays a critical role in model selection, enabling the identification of robust LLMs in dynamic, unsupervised environments. Applied to ten open-domain NLP datasets across four LLMs, the CAI Ratio demonstrates a strong positive correlation with LLM accuracy, establishing it as an essential tool for unsupervised evaluation and model selection in real-world settings.

1 INTRODUCTION

Large Language Models (LLMs), when combined with prompt optimisation (Brown et al., 2020; Chen & Tsang, 2024; Kojima et al., 2022; Wei et al., 2021; 2022; Huang et al., 2022; Yao et al., 2022; Diao et al., 2023; Liu et al., 2023; Wang et al., 2023; Yao et al., 2024; Long, 2023; Huang et al., 2023; Madaan et al., 2024; Shinn et al., 2024), have demonstrated remarkable capabilities in text and data annotation across diverse open-domain tasks (Meng et al., 2022; Ye et al., 2022; Wang et al., 2024; Liu et al., 2024; Wu et al., 2024), including spoken language understanding (Chen et al., 2024). Often outperforming traditional crowdsourcing and manual annotation methods (Gilardi et al., 2023), LLM-generated annotations have become pivotal for supervised fine-tuning, alignment training, and real-time inference (Tan et al., 2024). However, evaluating the quality of LLM-generated annotations remains challenging in unsupervised environments where oracle feedback is unavailable. Traditional evaluation methods (Table 2 in Appendix) fall short in these settings, and LLMs often exhibit overconfidence and inconsistent behavior without external supervision (Xiong et al., 2023; Zhou et al., 2024), underscoring the need for robust unsupervised evaluation strategies.

To address this, we propose a novel agentic annotation evaluation paradigm, where a student model collaborates with a noisy teacher (the LLM) to assess annotation quality through model agreement. This paradigm embodies the essence of agentic reasoning: in the absence of oracle feedback, *reliability emerges from the consistency of interactions between models,* with agreement serving as an implicit signal of annotation quality. When external supervision is missing, the alignment or misalignment between the student and the LLM offers a self-regulating mechanism to gauge annotation reliability. Building on this, we introduce the **Consistent and Inconsistent (CAI)** Ratio, a novel metric that effectively evaluates the LLM reliability by exploiting the unsupervised structural patterns within the

data through the student model. By harnessing the intrinsic consistency patterns between both models, the CAI Ratio provides a powerful unsupervised signal for assessing LLM annotations. Beyond evaluation, it also serves as decisive criterion for model selection, enabling the identification of most appropriate LLMs without relying on oracle feedback. We demonstrate in Figure 4 and Table 1 that the CAI Ratio exhibits a strong positive correlation with LLM annotation accuracy and effectively distinguishes the best-performing LLM models in unsupervised settings.

2 METHODOLOGY

We propose an agentic annotation evaluation paradigm to assess LLM reliability through the collaboration between a noisy LLM teacher and a student model. To enable annotation in an unsupervised setting, we leverage the student model's ability to capture the structural relationships in data, assigning annotations through a majority voting mechanism in its embedding space (Equation 1& 2). Meanwhile, the LLM generates outputs via an autoregressive process. Consistent samples emerge when both models agree, indicating reliable annotations, while inconsistent samples reflect the LLM's overconfident outputs that diverge from the student's predictions. By systematically analyzing these agreement patterns, we capture both confidence signals and overconfidence biases, enabling robust unsupervised evaluation through the proposed Consistent and Inconsistent (CAI) Ratio.

2.1 PROBLEM DEFINITION

Given unsupervised text corpus distributions for training and testing, denoted as $\mathcal{D}_U = \{x_1, \ldots, x_N\}, \mathcal{D}_{U_t} = \{x'_1, \ldots, x'_L\}, x, x' \in \mathcal{X} \subseteq \mathbb{R}^d$. Additionally, since annotations are generated according to user preferences, a small-size user-preference distribution is provided, denoted as: $H = \{(x_i, \bar{y}_i)\}_{i=1}^{s}, s = 5\% \times |\mathcal{D}_U|$. These samples are clustered into k non-overlapping clusters $C_1, C_2, ..., C_k$. A set of preference annotations is denoted as $\mathcal{A} = \{\bar{y}_1, \bar{y}_2, ..., \bar{y}_k\}$. Each cluster C_j is defined as $C_j = \{x_i \mid x_i \in H_j\}, H_j \subseteq H, \forall x_i \in C_j, \bar{y}_i = \bar{y}_j$. The clusters are disjoint, satisfying $\bigcup_{j=1}^{k} C_j = H, C_i \cap C_j = \emptyset, \forall i \neq j$. The goal is to evaluate quality of LLMs assigned annotation by estimating the latent consistent and inconsistent sets $\mathcal{C}^*$ and $\mathcal{I}^*$.

2.2 AGENTIC ANNOTATION EVALUATION THROUGH A NOISY LLM TEACHER AND STUDENT MODEL COLLABORATION

An **agentic annotation evaluation process** is constructed upon an LLM teacher with noisy annotation labels and a student model with limited user preferences, incorporating a *preference-based majority voting* mechanism. We utilize MINILM (Wang et al., 2020) as the student model, denoted as $\mathcal{S}$, which is a sentence-transformer designed for efficient sentence encoding. The student model encodes each input instance x_i into its corresponding *sentence embedding* as $\mathcal{S}(x_i) = e_i$. To assign annotations, the student model employs a *user preference-based majority voting* strategy, leveraging our proposed **Average Similarity (AS)** function, which is defined as:

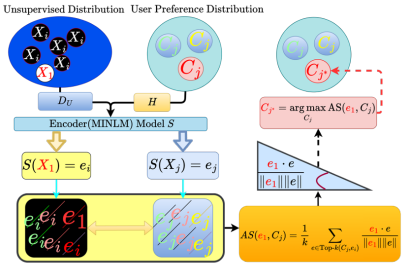

Figure 1: **Agentic Annotation Evaluation Process**: the teacher LLM generates noisy annotations in zero-/single-shot settings, which are compared against student model's **AS** output to measure the consistency score.

$$AS(e_i, C_j) = \frac{1}{k} \sum_{e \in \text{Top-}k(C_j, e_i)} \frac{e_i \cdot e}{\|e_i\| \|e\|}, \tag{1}$$

where e_i denotes the embedding for x_i, and e represents the embedding of each sample in cluster C_j. The term Top-$k(C_j, e_i)$ refers to the subset of samples in C_j with the top k cosine similarity scores with e_i. Formally, Top-$k(C_j, e_i) = \{e \in C_j \mid \text{AS}(e_i, e)$ ranks among the top k in $C_j\}$. Based on the computed similarity scores, the most similar examples to e_i are identified, and the average cosine similarity is computed for the top-selected samples in each cluster. In our experiments, we set k to five. Lastly, for the annotation assignment, we assign the annotation of the cluster C_j with the highest average cosine similarity score to the unlabelled sample $x_i \in D_u$. The cluster C_{j^*}, which has the highest average cosine similarity with the embedding e_i of a sample x_i, is defined as:

$$C_{j^*} = \arg\max_{C_j} \text{AS}(e_i, C_j), \tag{2}$$

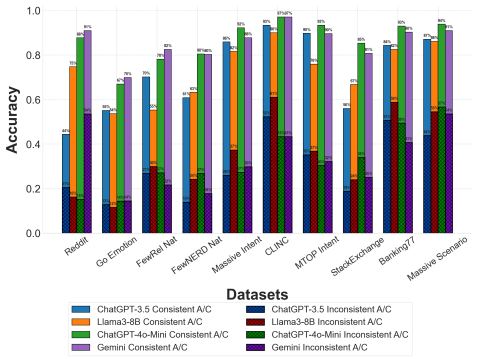

Figure 2: **An illustrative figure highlighting the importance of *consistent-and-inconsistent* sample identification in evaluating LLM performance.** LLM annotations on inconsistent samples (**dark-colored bars**) exhibit significantly lower accuracy compared to those on consistent samples (**light-colored bars**).

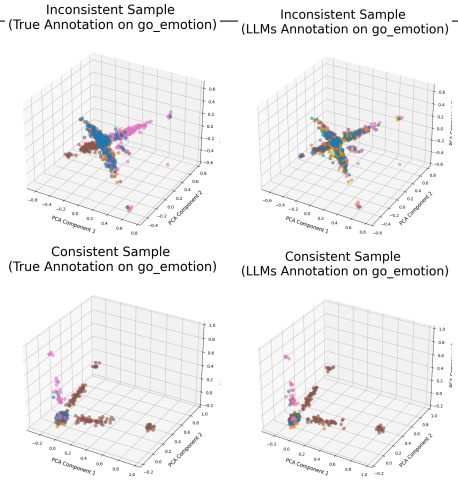

Figure 3: **Visualization of t-SNE Clustering** (better viewed in color, enlarged) comparing LLM vs Ground-Truth Annotations on *Go_Emotion* Dataset. LLM outputs exhibit ***high similarity*** with ground-truth labels on **consistent** samples, while showing ***significant divergence*** on **inconsistent** samples.

where $\text{AS}(e_i, C_j)$ is the average cosine similarity of e_i with the embeddings in C_j. The annotation $\bar{y}_{j^*}$ associated with C_{j^*} is then assigned to x_i, i.e., $\bar{y}_i = \bar{y}_{j^*}$. This process is represented by the annotation assignment function $h(x_i)$. Subsequently, the annotation associated with C_{j^*} is as defined by the user, is assigned to x_i. Finally, the student agent-annotated dataset is constructed as $D_s = \{(x_i, \bar{y}_i)\}_i^N$, where each $\bar{y}_i$ represents the student annotation obtained using the user preference-based majority voting approach. Given the acquired dataset $D_s = \{(x_i, \bar{y}_i)\}_{i=1}^N$ generated by the SA, we further leverage a noisy teacher LLM to generate annotations through a *group prompting* mechanism, applying both *zero-shot* and *single-shot* strategies. Specifically, in the zero-shot setting, the noisy teacher LLM generates annotations independently, defined as $\bar{y}_i^t = T(x_i)$. $\bar{y}_i^t = T(x_i) \text{with} P\left(\bar{y}_i^t \mid x_i\right) = \prod_{t=1}^{T_i} P\left(\bar{y}_{i,t}^t \mid x_i, \bar{y}_{i,1}^t, \ldots, \bar{y}_{i,t-1}^t\right)$. In contrast, the *single-shot* setting incorporates student-generated annotations as additional context, yielding $\hat{y}_i^t = T(x_i, \bar{y}_i)$, where $(x_i, \bar{y}_i) \in D_s$. $\hat{y}_i^t = T(x_i, \bar{y}_i) \text{with} P\left(\hat{y}_i^t \mid x_i, \bar{y}_i\right) = \prod_{t=1}^{T_i} P\left(\hat{y}_{i,t}^t \mid x_i, \bar{y}_i, \hat{y}_{i,1}^t, \ldots, \hat{y}_{i,t-1}^t\right)$. Since the LLM follows an autoregressive generation framework, we query the noisy teacher LLM to provide the annotation for each instance x_i without including the student labels $\bar{y}_i$ for zero-shot prompting, producing the noisy teacher distribution $D_t = \{(x_i, \bar{y}_i^t)\}_{i=1}^N$. For the single-shot setting, the SA annotations are incorporated, resulting in the student-noisy teacher distribution $\hat{D}_t = \{(x_i, \hat{y}_i^t)\}_{i=1}^N$.

2.3 Evaluation of LLMs without Oracle Feedback

After acquiring the SA-generated dataset D_s, the Noisy Teacher-generated dataset D_t, and the SA-Noisy Teacher dataset $\hat{D}_t$, we introduce the Consistent-and-Inconsistent (CAI) Identification and Ratio framework. Specifically, CAI Identification determines consistent and inconsistent samples across D_s, D_t, and $\hat{D}_t$ by comparing annotation agreement between the Student Agent (SA) $\mathcal{S}$ and the Noisy Teacher (NT) $\mathcal{T}$. Samples with identical predictions from both the SA and NT models are classified as consistent samples; otherwise, they are considered inconsistent samples. For each $x \in \mathcal{D}_u$, the annotation assignment process is represented by the function h. The annotation label from the SA is denoted as $\bar{y}_{\mathcal{S}}$, while the NT's annotation labels are represented as $\bar{y}_{\mathcal{T}}$ (zero-shot) and $\hat{y}_{\mathcal{T}}$ (single-shot). A sample is classified as consistent if $\bar{y}_{\mathcal{S}} = \bar{y}_{\mathcal{T}} = \hat{y}_{\mathcal{T}}, x \in \mathcal{C}$ where $\mathcal{C}$ denotes the set of consistent samples. Conversely, a sample is classified as inconsistent if at least one of the assigned annotations differs, represented as $\exists\,(y, y') \in \{\bar{y}_{\mathcal{S}}, \bar{y}_{\mathcal{T}}, \hat{y}_{\mathcal{T}}\}, y \neq y', x \in \mathcal{I}$, where $\mathcal{I}$ represents the set of inconsistent samples. Identifying annotation inconsistencies is crucial, as is rigorously assessing the teacher model's annotation quality, especially in the absence of ground truth.

Definition 1 (Consistent-and-Inconsistent (CAI) Ratio)**.** *Let N_C and N_{IC} denote the number of consistent samples (LLM and student model agree) and the number of inconsistent samples (LLM and student model disagree), respectively. The CAI Ratio is defined as CAI Ratio $= \frac{N_C}{N_{IC}}$.*

3 Experiments

Experimental Setup We collected the CAI Ratio for LLMs—GPT-3.5 Turbo, GPT-4o Mini, Google Gemini 1.5 Flash, and Llama-8B Instruct and evaluated these across ten textual datasets.

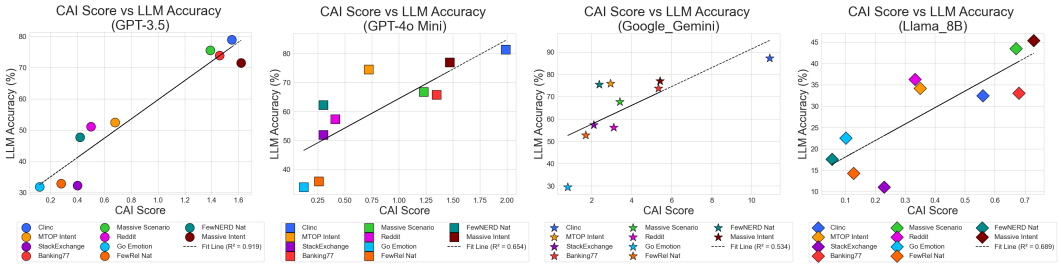

Figure 4: **Correlation analysis between LLM annotation accuracy and the CAI ratio**, evaluated across 4 principled LLMs (also see statistical test results in Sec 3). The Pearson correlation coefficients and corresponding p-values confirm the statistical significance of the positive correlation between CAI ratio and LLMs accuracy.

Dataset	Best CAI Model		Best Accuracy Model		Match	Accuracy Difference (%)
	Model	Accuracy (%)	Model	Accuracy (%)		
CLINC	Google Gemini	87.24	Google Gemini	87.24	✓	0.00
MTOP Intent	Google Gemini	75.85	Google Gemini	75.85	✓	0.00
StackExchange	Google Gemini	57.31	Google Gemini	57.31	✓	0.00
Banking77	Google Gemini	73.76	GPT-3.5	73.93	✗	**-0.17**
Massive Scenario	Google Gemini	67.72	GPT-3.5	75.55	✗	**-7.83**
Reddit	Google Gemini	56.23	ChatGPT-4o Mini	57.39	✗	**-1.16**
Go Emotion	Google Gemini	29.44	ChatGPT-4o Mini	33.82	✗	**-4.38**
FewRel Nat	Google Gemini	52.74	Google Gemini	52.74	✓	0.00
FewNERD Nat	Google Gemini	75.48	Google Gemini	75.48	✓	0.00
Massive Intent	Google Gemini	77.03	Google Gemini	77.03	✓	0.00

Table 1: **Model Selection Using CAI Ratio as a Metric**: The model selected based on CAI ratio exhibits a strong correlation with the model achieving the highest accuracy.

These datasets include Bank77 (Casanueva et al., 2020), CLINC, Go Emotion, MTOP, Massive (Intent) (Larson et al., 2019; FitzGerald et al., 2022; Li et al., 2020), StackExchange, Reddit (Geigle et al., 2021), FewRel Nat, and FewNerd Nat (Han et al., 2018). Covering domains such as intent classification, topic modeling, and unsupervised intent discovery (Zhang et al., 2021; 2022), their annotation practices follow (Zhang et al., 2023).

Proof-of-Concept Experiments on *Consistency and Inconsistency Identification* We first investigate the impact of identifying consistent and inconsistent samples in our framework. Figure 2 shows that LLMs achieve significantly higher accuracy on *consistent* samples, reflecting greater confidence in their predictions, whether observed with a student model or within the LLM's own outputs. The t-SNE visualization in Figure 3 further confirms that LLM annotations align closely with ground-truth labels for *consistent samples*, while diverging significantly for *inconsistent* samples. This contrast highlights the importance of our identification process for evaluating LLM annotations.

Correlation Results between CAI Ratio and LLM Accuracy We performed a Pearson correlation analysis to investigate the relationship between CAI Ratio and LLM accuracy. The correlation analysis between the Consistent-over-Inconsistent (CAI) ratio and accuracy across different LLMs demonstrates a strong relationship between these two metrics. GPT-3.5 shows the highest correlation ($\rho = 0.93$, $p = 8.22 \times 10^{-5}$), indicating a very strong positive relationship between CAI and accuracy, with high statistical significance. GPT-4o Mini shows a strong correlation ($\rho = 0.86$, $p = 1.61 \times 10^{-3}$), suggesting that CAI is a reliable predictor of accuracy for this model. Llama-8B-Instruct ($\rho = 0.81$, $p = 1.44 \times 10^{-2}$) and Google Gemini ($\rho = 0.72$, $p = 1.80 \times 10^{-2}$) exhibit moderate-to-strong correlations with significant statistical confidence.

Model Selection with CAI Ratio Model selection based solely on the CAI Ratio correctly identifies the best-performing LLMs in 60% of cases. Among the mismatched cases, the accuracy differences are not significant. Although CAI Ratio alone is not a perfect indicator of LLMs accuracy, it serves as a reliable heuristic for selecting well-performing LLMs in unsupervised settings. We have chosen the Best CAI Model and the Best Accuracy Model from the candidate LLM set, which includes GPT-3.5 Turbo, GPT-4o Mini, Google Gemini 1.5 Flash, and Llama-8B Instruct.

4 Conclusion

In this work, we propose a novel and effective metric, the **CAI Ratio**, based on a Agentic Annotation Evaluation Paradigm for unsupervised dataset annotation aligned with user preferences. The CAI Ratio has demonstrated its effectiveness in both **LLM annotation evaluation** and **model selection**. Evaluated on ten domain-specific NLP datasets, the CAI metric exhibited a strong positive correlation with LLM performance, confirming its efficacy as a model selection and evaluation tool for unsupervised dataset annotation tailored to user preferences.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.

Cheng Chen and Ivor Tsang. Self-teaching prompting for multi-intent learning with limited supervision. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=DeoamI1BFh.

Cheng Chen, Bowen Xing, and Ivor W Tsang. Low-hanging fruit: Knowledge distillation from noisy teachers for open domain spoken language understanding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 107–125, 2024.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. Active prompting with chain-of-thought for large language models, 2023.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. Tweac: transformer with extendable qa agent classifiers. *arXiv preprint arXiv:2104.07081*, 2021.

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

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*, 2018.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2023.

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

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*, 2020.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024.

Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference 2023*, 2023.

Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023.

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

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and Quoc V Le. H. chi, sharan narang, aakanksha chowdhery, and denny zhou. self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, volume 1, 2023.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. Codeclm: Aligning language models with tailored synthetic data. *arXiv preprint arXiv:2404.05875*, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*, 2022.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14365–14373, 2021.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. New intent discovery with pre-training and contrastive learning. *arXiv preprint arXiv:2205.12914*, 2022.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. Clusterllm: Large language models as a guide for text clustering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*, 2024.

A APPENDIX

This supplementary material is organized as follows. In Sec A, we provide a detailed interpretation of the CAI Ratio, highlighting its advantages and distinctions compared to traditional evaluation metrics. In Sec B, we present comprehensive experimental results, including model accuracy and CAI ratios across various datasets evaluated with different LLMs. Finally, in Sec C, we showcase t-SNE visualization results, illustrating clustering patterns for consistent and inconsistent samples on additional datasets.

A.1 CONSISTENT AND INCONSISTENT RATIO INTERPRETATION

The CAI ratio provides a principled means to assess the reliability of LLM-generated annotations in the absence of labelled supervision. A significantly higher CAI ratio (CAI Ratio $\gg 1$) may indicate consistency or higher annotation accuracy, while a lower CAI ratio (CAI Ratio $\ll 1$) suggests greater lower annotation accuracy and inconsistency in the LLM's outputs. In these cases, the ratio suggests that the LLM's outputs are unreliable, necessitating refinements with external human annotations or additional prior knowledge to improve annotation accuracy.

Furthermore, the relationship between the CAI ratio and LLM annotation accuracy can be formalized as the *Law of Consistency*. This principle states that if both the LLM and student model are optimal hypotheses, denoted as T^* and S^* for a given dataset D_u, the number of consistent samples should asymptotically exceed the number of inconsistent samples as the dataset size approaches infinity. [Law of Consistency] Let T^* and S^* be the optimal teacher (LLM) and student model hypotheses for an unsupervised dataset D_u. Define N_C and N_{IC} as the number of consistent and inconsistent samples, respectively, identified by the CAI ratio. As the dataset size $|D_u| \to \infty$, the probability that that of consistent samples surpasses the number of inconsistent samples approaches one: $\lim_{|D_u|\to\infty} P(N_C > N_{IC}) = 1$.

A.2 COMPARISON WITH TRADITIONAL EVALUATION METRICS

Metric	Ground-Truth Labels?	Data Drift?	Tracks Annotation Quality Over Time?
Accuracy	✓	✗	✗
Precision/Recall	✓	✗	✗
F1-score	✓	✗	✗
CAI Ratio	✗	✓	✓

Table 2: Comparison of Traditional Metrics and CAI Ratio

A.3 PEARSON CORRELATION TEST FOR CONSISTENT AND INCONSISTENT RATIO

We have performed a Pearson correlation, the correlation coefficient r is calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i symbolises the CAI Ratio. y_i denotes the LLM annotation accuracies. $\bar{x}$ and $\bar{y}$ are the average mean of x_i and y_i, accordingly. n is the number of samples we have used for evaluation. To assess the statistical significance, we use a hypothesis test for the correlation coefficient, calculating a t-statistic (Schober et al., 2018):

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

The P-value is then calculated from the t-distribution with $n - 2$ degrees of freedom.

A.4 IMPORTANCE OF STUDENT MODEL IN AGENTIC ANNOTATION EVALUATION PARADIGM

The inclusion of the student model is essential as it provides a safeguard against underperformance by the LLM. Additionally, the student model serves as a reference point for "course tracking,"

meaning that it allows us to monitor and guide the annotation process by comparing the student model's output with the teacher model's output. This is particularly evident in our experiments where the Meta-8B Instruct model, serving as a low-competency "noisy teacher," exhibited suboptimal performance on most of the eight datasets, as reflected in its low CAI scores. The student model addresses this challenge by collaborating with the teacher model to iteratively refine annotations, ensuring robustness even when the teacher model lacks competency in specific tasks. We justify the necessity of the student model through experimental analysis.Another key motivation for our approach, including the use of a student model, is to enhance efficiency. This efficiency is evident in two significant aspects:

- Computational Efficiency: Our method requires access to the teacher model only twice per dataset, with minimal or no reliance on demonstrations for prompting. This substantially reduces computational overhead.
- Cost-Effectiveness: For closed-source models with API service fees, our approach offers a cost-efficient solution. By utilizing the student model alongside our proposed clustering operation and a limited number of teacher model predictions, our method achieves superior performance compared to both models individually. Importantly, it does so at a lower cost, particularly when compared to methods that rely on iterative self-correction.

B EXPERIMENTAL RESULTS

Implementation Details The top-k selection and proportions of consistent and user-preference samples are as follows. For CLINC and Massive Scenario, 'top-k' is set to 5, with 'proportion' at 0.2. For MTOP Intent, 'proportion' is set to 1, and 'top-k' is updated to 15 after printing the current value. In StackExchange, 'top-k' is set to 5 and 'proportion' to 1, while in Banking77, 'top-k' is set to 3 and 'proportion' is 0.2. In massive intent, 'top-k' is 20 and 'proportion' is 0.5), proportion=0.2, and few real nat has top-k=30, and proportion is 1. In 'reddit', 'top-k' is set to 7, and the proportion is 0.2. All tests are done with two random seeds with temperature parameters (0.5 and 1) for user preference samples, student model-assigned annotation, and LLMs with and without student annotations.

Dataset	GPT-3.5		ChatGPT-4o Mini		Google Gemini		Llama-8B	
	Accuracy (%) ± Std	CAI Ratio	Accuracy (%) ± Std	CAI Ratio	Accuracy (%) ± Std	CAI Ratio	Accuracy (%) ± Std	CAI Ratio
Banking77	73.93 ± 0.81	1.46	65.78 ± 0.24	1.35	73.73 ± 0.03	5.34	33.06 ± 1.92	0.68
Clinc	79.01 ± 1.08	1.55	81.46 ± 0.36	1.99	87.50 ± 0.26	10.90	32.49 ± 6.73	0.56
Massive Scenario	75.55 ± 1.76	1.39	66.83 ± 1.31	1.23	67.95 ± 0.23	3.41	43.52 ± 1.85	0.67
MTOP Intent	52.49 ± 2.52	0.68	74.54 ± 0.32	0.72	75.61 ± 0.23	2.94	34.17 ± 6.70	0.35
Stack Exchange	32.27 ± 0.65	0.40	51.90 ± 0.18	0.30	57.48 ± 0.17	2.11	11.02 ± 2.78	0.23
Reddit	51.12 ± 1.27	0.50	57.39 ± 0.40	0.41	56.73 ± 0.50	3.10	36.31 ± 0.97	0.333
Go Emotion	31.84 ± 0.87	0.12	33.82 ± 0.25	0.12	29.72 ± 0.28	0.81	22.53 ± 0.21	0.102
Few Rel Nat	32.87 ± 1.72	0.28	35.87 ± 0.22	0.26	52.96 ± 0.21	1.70	14.25 ± 0.36	0.128
Few Nerd Nat	47.70 ± 1.36	0.42	62.20 ± 0.19	0.30	75.35 ± 0.13	2.37	17.60 ± 2.02	0.055
Massive Intent	71.52 ± 0.95	1.62	76.93 ± 0.16	1.47	76.90 ± 0.13	5.41	45.41 ± 0.06	0.730

Table 3: Model Selection Results Using CAI as a Metric

Dataset	GPT-3.5 CAI	ChatGPT-4o Mini CAI	Google Gemini CAI	Llama 8B CAI	GPT-3.5 Accuracy	ChatGPT-4o Mini Accuracy	Google Gemini Accuracy	Llama 8B Accuracy	Best CAI Model	Best Accuracy Model
CLINC	1.55	1.9974	10.900	0.560	79.01	81.46	87.24	32.49	Google Gemini	Google Gemini
MTOP Intent	0.68	0.7236	2.940	0.670	52.49	74.54	75.85	43.52	Google Gemini	Google Gemini
StackExchange	0.40	0.3014	2.110	0.350	32.27	51.90	57.31	34.17	Google Gemini	Google Gemini
Banking77	1.46	1.3494	3.545	0.680	73.93	65.78	73.76	33.06	Google Gemini	GPT-3.5
Massive Scenario	1.39	1.2269	4.375	0.230	75.55	66.83	67.72	11.02	Google Gemini	GPT-3.5
Reddit	0.50	0.4151	3.100	0.333	51.12	57.39	56.23	36.31	Google Gemini	ChatGPT-4o Mini
Go Emotion	0.12	0.1238	0.810	0.102	31.84	33.82	29.44	22.53	Google Gemini	ChatGPT-4o Mini
FewRel Nat	0.28	0.2613	1.700	0.128	32.87	35.87	52.74	14.25	Google Gemini	Google Gemini
FewNERD Nat	0.42	0.3064	2.370	0.055	47.70	62.20	75.48	17.60	Google Gemini	Google Gemini
Massive Intent	1.62	1.4701	5.410	0.730	71.52	76.93	77.03	45.41	Google Gemini	Google Gemini

Table 4: Accuracy with Consistent Samples and Inconsistent Samples Across Four LLMs

Dataset	ChatGPT-3.5		Llama-8B		ChatGPT-4o Mini		Google Gemini	
	Consistent (%)	Inconsistent (%)	Consistent (%)	Inconsistent (%)	Consistent (%)	Inconsistent (%)	Consistent (%)	Inconsistent (%)
Reddit	44.37	20.53	74.68	16.21	87.70	15.15	86.37	13.66
Go Emotion	55.10	12.87	53.61	11.57	66.93	14.38	69.88	14.48
FewRel Nat	70.16	26.86	55.32	29.97	78.11	27.11	82.47	21.67
FewNERD Nat	60.82	13.94	63.21	24.17	80.48	26.83	80.27	17.82
Massive Intent	85.86	25.92	81.59	37.34	92.26	27.30	87.79	29.81
CLINC	93.37	52.27	90.02	61.04	97.09	43.33	90.98	53.53
MTOP Intent	89.75	35.21	75.88	36.78	93.42	30.38	89.63	32.19
StackExchange	55.99	18.76	66.71	24.01	85.32	34.05	80.77	25.07
Banking77	84.30	50.64	82.48	58.77	93.01	49.43	90.31	40.77
Massive Scenario	87.04	43.86	86.09	54.55	93.83	56.65	90.98	53.54

C t-SNE VISUALIZATION FOR CLUSTERING ON MORE DATASETS FOR CHATGPT-4O MINI

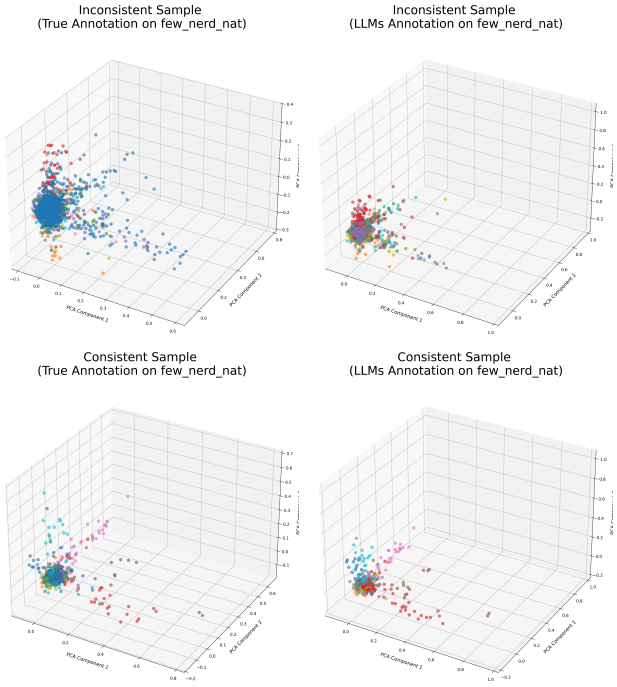

Figure 5: **Visualization of t-SNE Clustering** for LLM vs True Annotations on *Few_Nerd_Nat* Dataset. LLM outputs exhibit ***high similarity*** with ground-truth labels on **consistent** samples, while showing ***significant divergence*** on **inconsistent** samples.

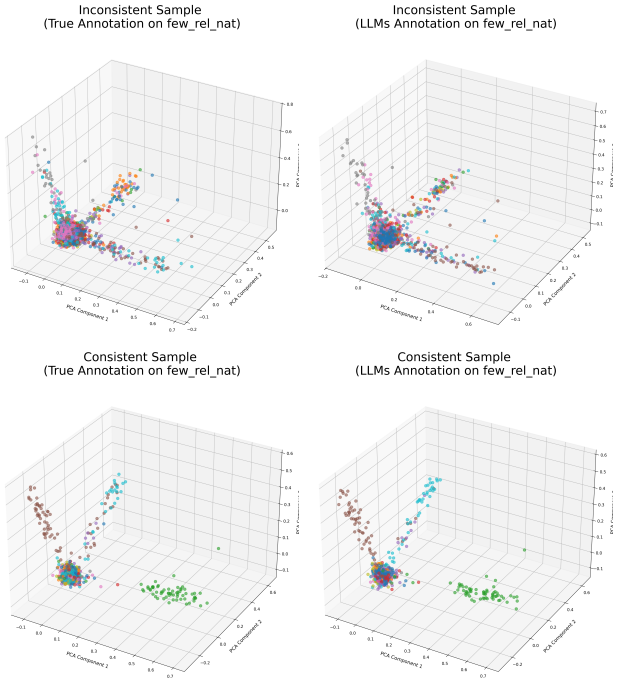

Figure 6: **Visualization of t-SNE Clustering** for LLM vs True Annotations on *Few_Rel_Nat* Dataset. LLM outputs exhibit ***high similarity*** with ground-truth labels on **consistent** samples, while showing ***significant divergence*** on **inconsistent** samples.

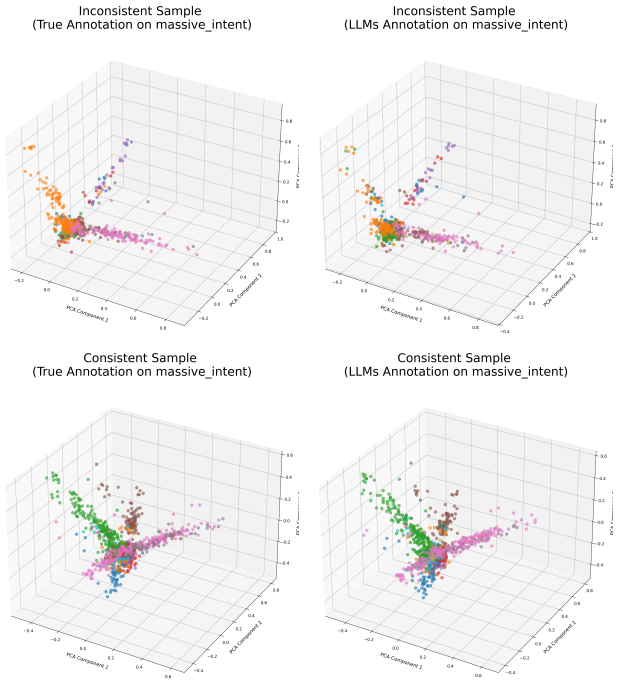

Figure 7: **Visualization of t-SNE Clustering** for LLM vs True Annotations on *Massive_Intent* Dataset. LLM outputs exhibit ***high similarity*** with ground-truth labels on **consistent** samples, while showing ***significant divergence*** on **inconsistent** samples.

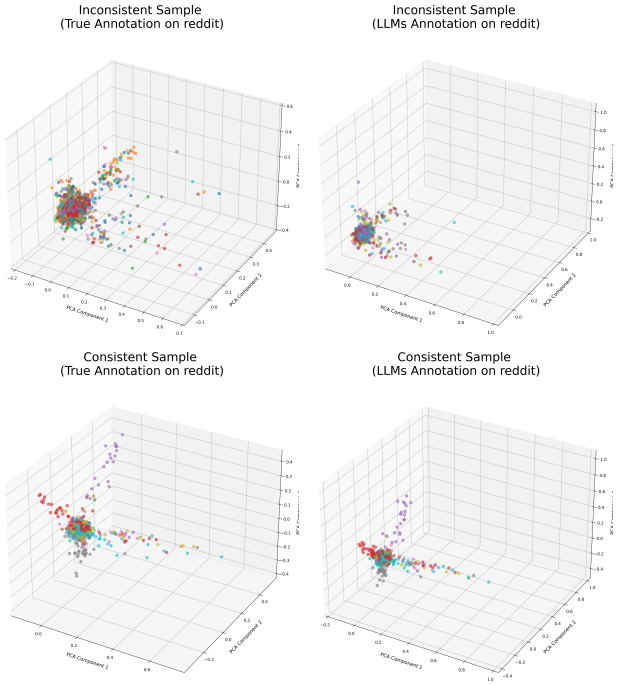

Figure 8: **Visualization of t-SNE Clustering** for LLM vs True Annotations on *reddit* Dataset. LLM outputs exhibit ***high similarity*** with ground-truth labels on **consistent** samples, while showing ***significant divergence*** on **inconsistent** samples.