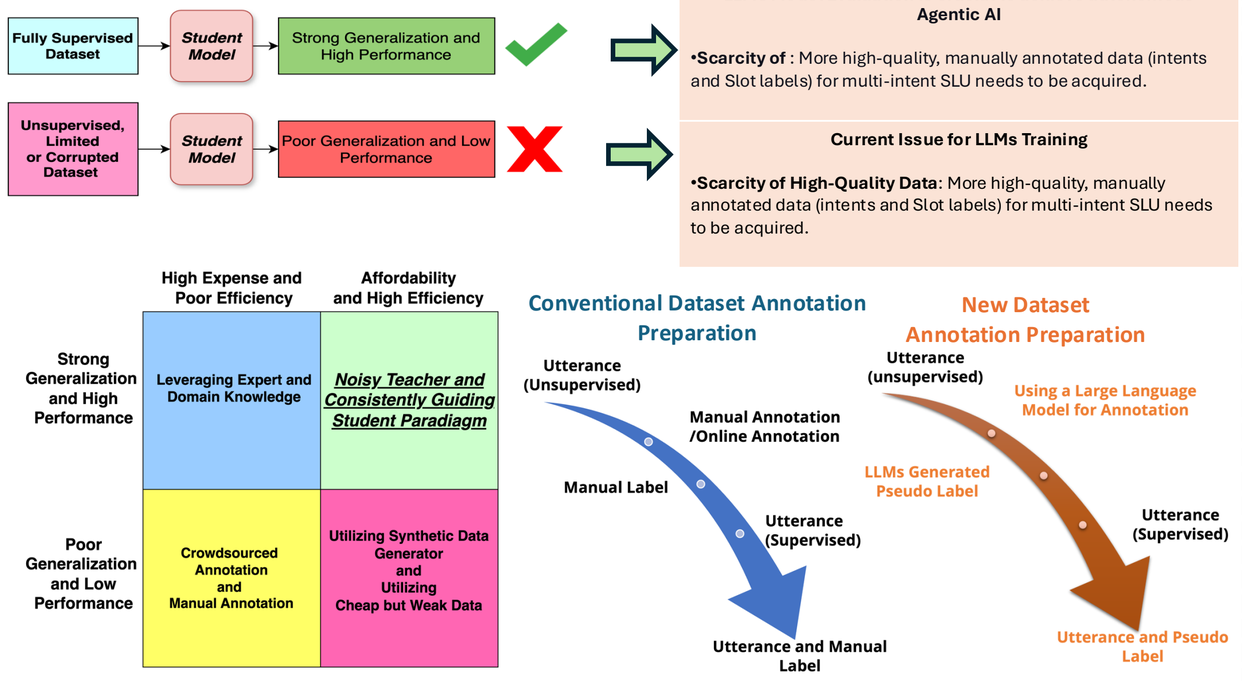
Evaluating LLMs Without Oracle Feedback: Agentic Annotation Evaluation Through Unsupervised Consistency Signals

Cheng Chen [1,2,3] Haiyan Yin[2,3], Ivor W Tsang [2,3,4]

[1]University of Technology Sydney & [2]CFAR, Agency for Science, Technology and Research, Singapore
& [3]IHPC, Agency for Science, Technology and Research, Singapore
& [4]College of Computing and Data Science, Nanyang Technological University

Introduction

- **Background**



- **Evaluation is Crucial for Unsupervised Annotation Task**

No ground-truth safety net – without oracle labels, errors in pseudo-annotations can propagate unchecked.

Bias & mode-collapse detection – evaluation surfaces class imbalance, shortcut learning, or repetitive mistakes early.

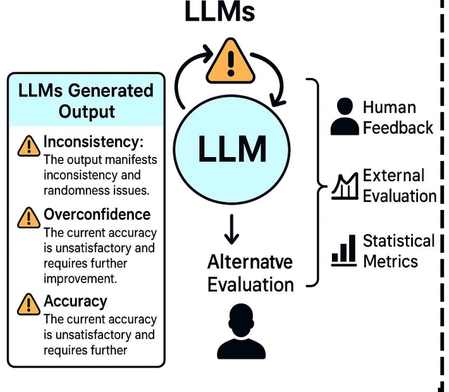
Progress tracking & stop-criteria – metrics like CAI or disagreement rate tell you when the iterative loop has really improved (or plateaued), saving compute.

Downstream quality guard – small sanity-checks on a held-out slice predict whether noisy labels will hurt final task performance.

❖ **We can not just rely on LLMs itself for evaluation ?**

LLMs Generated Output

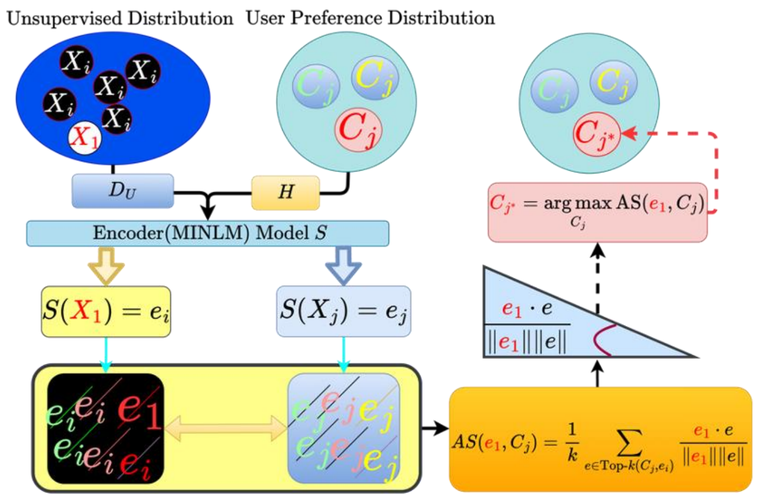
1. **Inconsistency**: The output manifests inconsistency and randomness issues.
2. **Overconfidence**: The current accuracy is unsatisfactory and requires further improvement.
3. **Accuracy**: The current accuracy is unsatisfactory and requires further improvement.



A **novel** agentic paradigm Framework

Student Model

- **A user preference-based majority voting strategy**



Teacher Model

- **Zero-Shot and One-Shot Prompting**

Zero-Shot Prompting

1.Q: "Can you find out about the ground transportation available in Atlanta and then what is restriction AP57"

Q: For this sentence "what's the airport at Orlando and what time zone is Denver in.", please identify the intentions based on the provided intention set {'abbreviation': 0, 'airport': 1, 'city': 2, 'capacity': 3, 'aircraft': 4, 'ground_service': 5, 'meal': 6, 'quantity': 7, 'flight_no': 8, 'flight_time': 9, 'ground_fare': 10, 'cheapest': 11, 'flight': 12, 'distance': 13,'restriction': 14, 'airfare': 15, 'airline': 16, 'day_name': 17} using the above the chain_of_thoughts strategy:

(Output) The answer is
['Airport', 'City','Time Zone'] - Partially Correct

One-Shot Prompting

1.Q: "What airport is at Tampa, and where is General Mitchell International located?"

A: ***Student Model Assigned Response (expressed)***: The sentence asks for the airport at Tampa (airport) and wants to know the location of General Mitchell International (city).

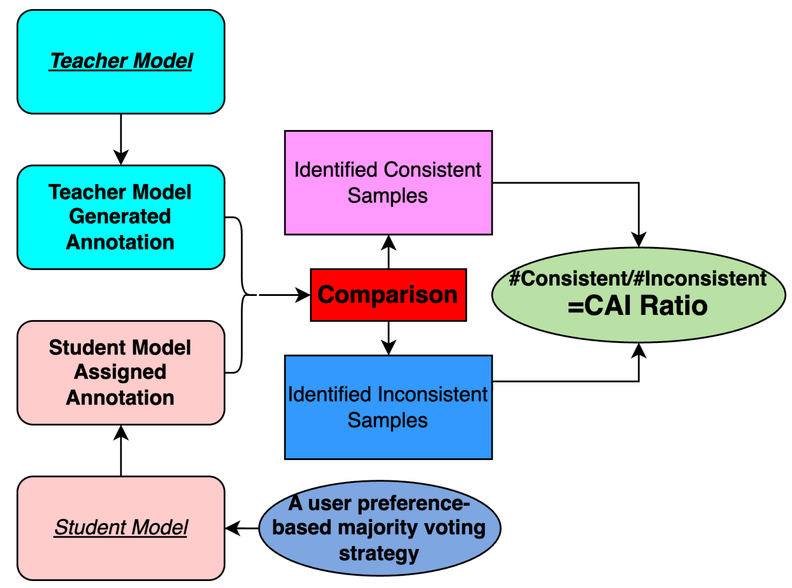
Q: For this sentence "what's the airport at Orlando and what time zone is Denver in.", please identify the intentions based on the provided intention set {'abbreviation': 0, 'airport': 1, 'city': 2, 'capacity': 3, 'aircraft': 4, 'ground_service': 5, 'meal': 6, 'quantity': 7, 'flight_no': 8, 'flight_time': 9, 'ground_fare': 10, 'cheapest': 11, 'flight': 12, 'distance': 13, 'restriction': 14, 'airfare': 15, 'airline': 16, 'day_name': 17} using the above the chain_of_thoughts strategy:

(Output) The answer is
['Airport','City'] - Correct

- **Consistent and Inconsistent Ratio for LLM evaluation and Model Selection**

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 $= \frac{N_C}{N_{IC}}$.*

A **novel Consistent and Inconsistent Ratio (CAI),** where a student model collaborates with a noisy teacher (the LLM) to **assess** and refine annotation quality without relying on oracle feedback.



Experiments

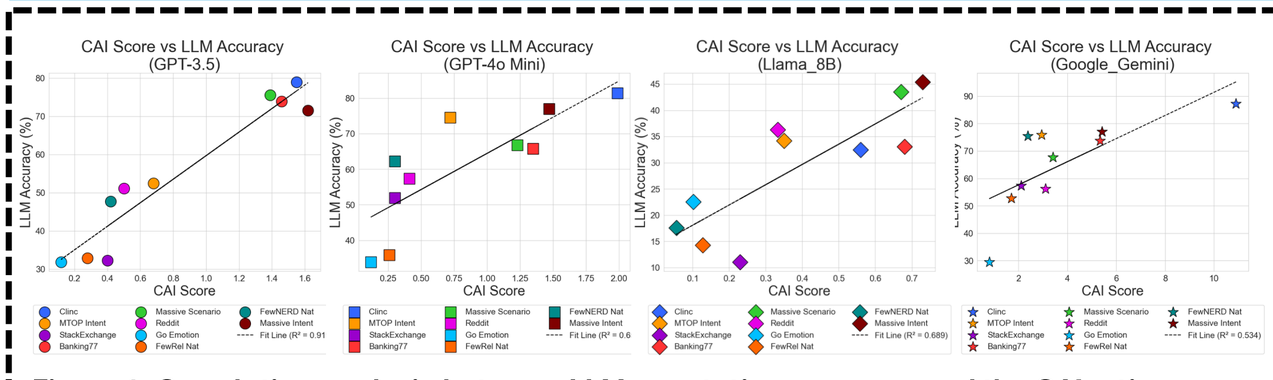


Figure 1: Correlation analysis between LLM annotation accuracy and the CAI ratio, evaluated across 4 principled LLMs (also see statistical test results in Sec 3). The Pearson correlation coefficients and corresponding p-values confirm the statistical significance of the positive correlation between CAI ratio and LLMs accuracy

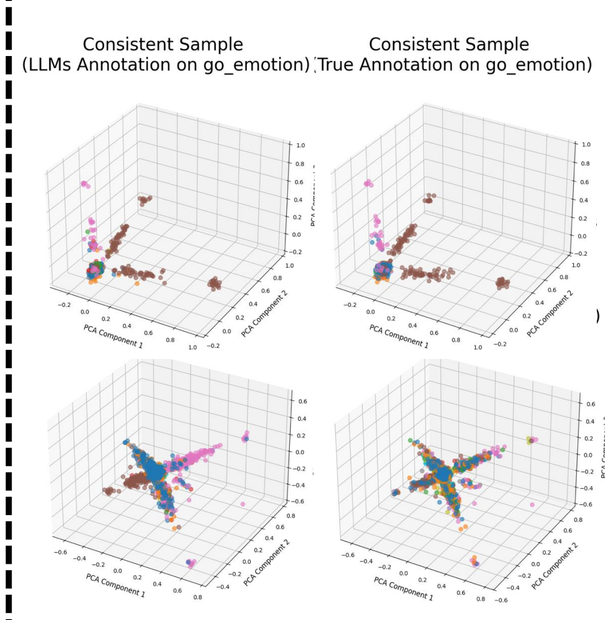


Figure 2: Visualization of t-SNE Clustering (better viewed in color, enlarged) comparing LLM vs Ground-Truth Annotations on Go Emotion Dataset. LLM outputs exhibit high similarity with ground truth labels on consistent samples, while showing significant divergence on inconsistent samples.

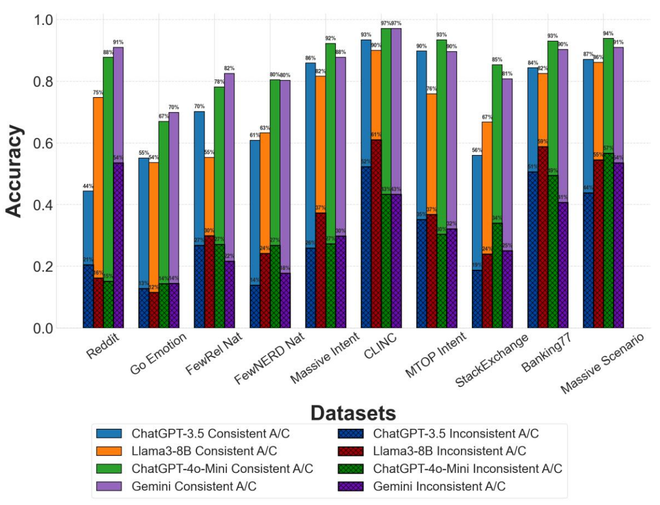


Figure 3: An illustrative figure highlighting the importance of consistent-and-inconsistent sample identification in evaluating LLM performance. LLM annotations on inconsistent samples (dark-colored bars) exhibit significantly lower accuracy compared to those on consistent samples (light-colored bars).

LLM	Pearson Correlation (ρ)	p-value (p)
GPT-3.5	0.93	8.22×10^{-5}
GPT-4o Mini	0.86	1.61×10^{-3}
Llama-8B-Instruct	0.81	1.44×10^{-2}
Google Gemini	0.72	1.80×10^{-2}

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.