Data-Driven Prompt Refinement for Consistent LLM-assisted Corpus Pre-annotation

Mujun Xu[1]; Ethan Wang[2]; Jingqi Wang[3], PhD; and Tavleen Singh[3], PhD

[1]Round Rock High School, Austin, Texas; [2]St. Thomas Episcopal School, Houston, Texas; [3]IMO Health, Houston, Texas

OVERVIEW

Human annotation is a critical but time-consuming and labor-intensive process as high-quality training data is required to support supervised machine learning and fine-tuned large language models (LLMs).

While using LLMs as a pre-annotation step can significantly reduce manual effort, limited context windows prevent LLMs from seeing all documents, which leads to inconsistencies and discrepancies in annotation quality.

In this study, we propose a data-driven feedback method that iteratively refines prompts and selects the most informative examples to improve pre-annotation performance.

RESULTS

For SpecificDisease entities, F1-scores improved from 0.693 → 0.726 (4.8% improvement), showing stable performance gains across iterations. This indicates that the data-driven feedback loop effectively resolved consistency issues for well-defined medical concepts.

For Modifier entities, F1-scores rose from 0.333 → 0.613 (84% improvement) in the first iteration, with recall increasing from 0.208 → 0.733. Although the second iteration saw a slight dip (to 0.600), the overall improvement remained substantial..

For DiseaseClass entities and CompositeMention entities, the performance declined across iterations, likely due to complex entity boundaries and limited domain understanding of the foundation model in differentiating DiseaseClass vs SpecificDisease concepts.

	Initial Round			1st Iteration			2nd Iteration		
Entity type	P	R	F1	P	R	F1	P	R	F1
CompositeMention	0.591	0.448	0.510	0.333	0.483	0.394	0.200	0.414	0.270
DiseaseClass	0.559	0.443	0.494	0.557	0.430	0.486	0.537	0.430	0.478
Modifier	0.835	0.208	0.333	0.647	0.582	0.613	0.508	0.733	0.600
SpecificDisease	0.632	0.766	0.693	0.694	0.735	0.714	0.757	0.698	0.726

METHOD

Our proposed method consists of an iterative refinement process that includes the following steps:

1) Initial Setup: We begin by creating baseline prompts and one manually curated example representing typical annotation scenarios for the target task.

2) Pre-annotation Phase: The LLM (Gemini 2.5 Flash) is applied to perform pre-annotation on a subset of the corpus (200 random notes from NCBI Disease corpus).

3) Consistency Check: Automated consistency checking identifies three primary inconsistency types: overlapped entities, potentially missing entities, and ambiguous entities across contexts.

4) Feedback Analysis and Prompt Refinement: Using feedback generated by Claude Sonnet 4, prompts are automatically updated to resolve identified issues, and the example set is augmented with challenging cases.

The entire pipeline is implemented within LangExtract, which enables automated tracking, customization, and versioning of prompts and examples.

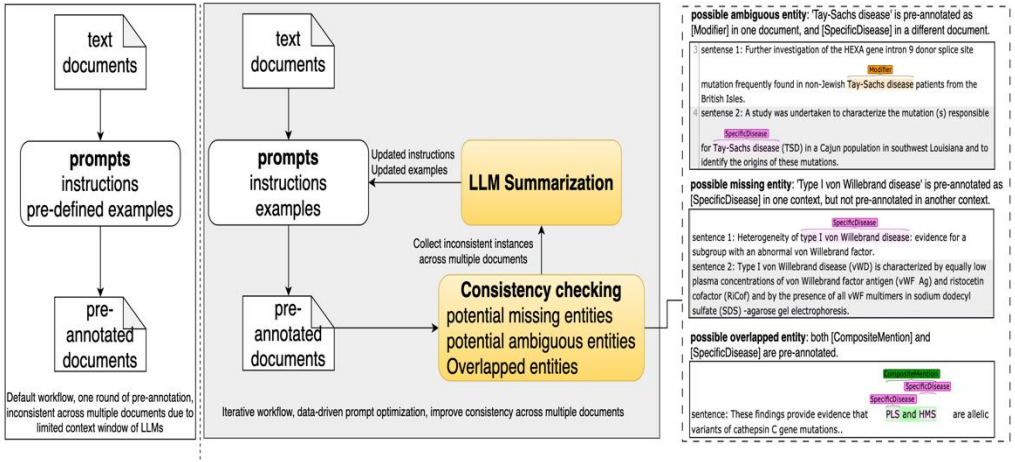

Figure 1: Overall Methodology

CONCLUSION & DISCUSSION

Our data-driven iterative method enhances annotation consistency while reducing human intervention, effectively leveraging LLM strengths and mitigating their variability across large corpora.

This approach minimizes reliance on prompt-engineering expertise, scales efficiently across domains, and enables continuous quality improvement, making it well-suited for large-scale annotation projects.

Performance depends on the robustness of consistency checks and LLM feedback quality. Future work will extend the framework into an autonomous agent system and explore integration with active learning for adaptive example selection.

CONTACT

Mujun Xu: mujun.xu@gmail.com