UNIVERSITY OF TORONTO

# Towards Efficient And Reliable Data Curation for Machine Learning

Presenter: Naiqing Guan
Supervisor: Nick Koudas

*"AI is akin to building a rocket ship. You need a huge engine and a lot of fuel. The rocket engine is the learning algorithms, but the fuel is the huge amounts of data we can feed to these algorithms."*

— Andrew Ng

# Data Curation for ML Pipelines

## Data Creation

Collecting and annotating datasets for training ML models

## Data Organization

Structuring and indexing the data for efficient queries

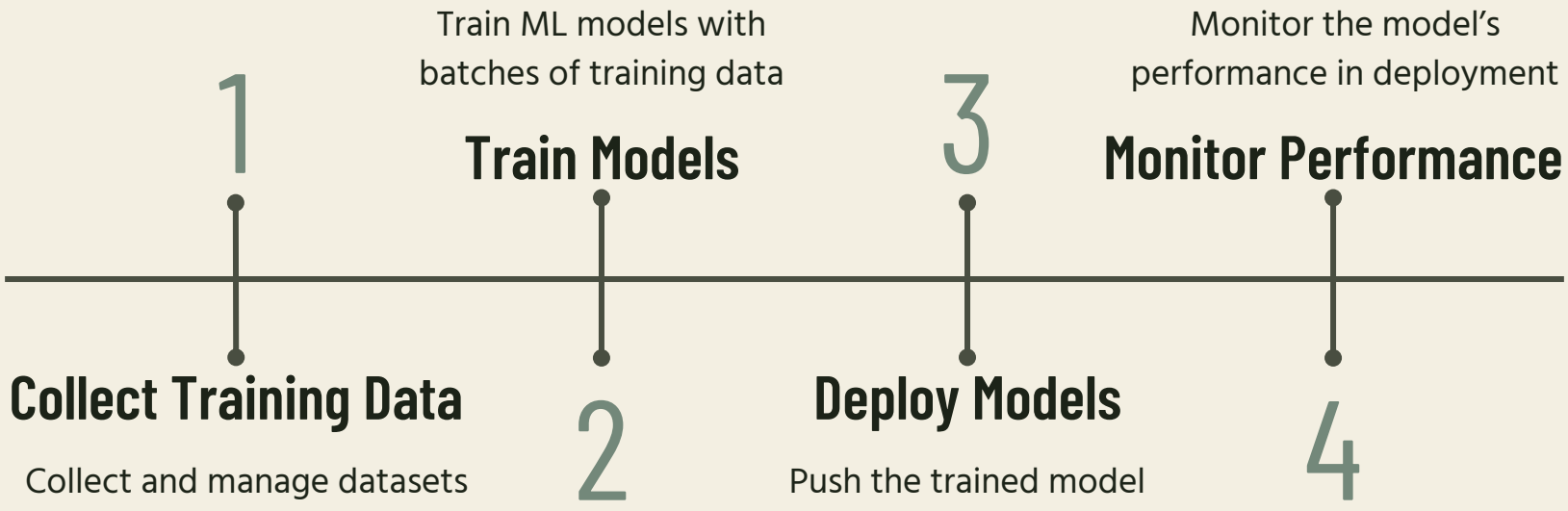
## Data Maintenance

Integrate, update and clean datasets to maintain their value

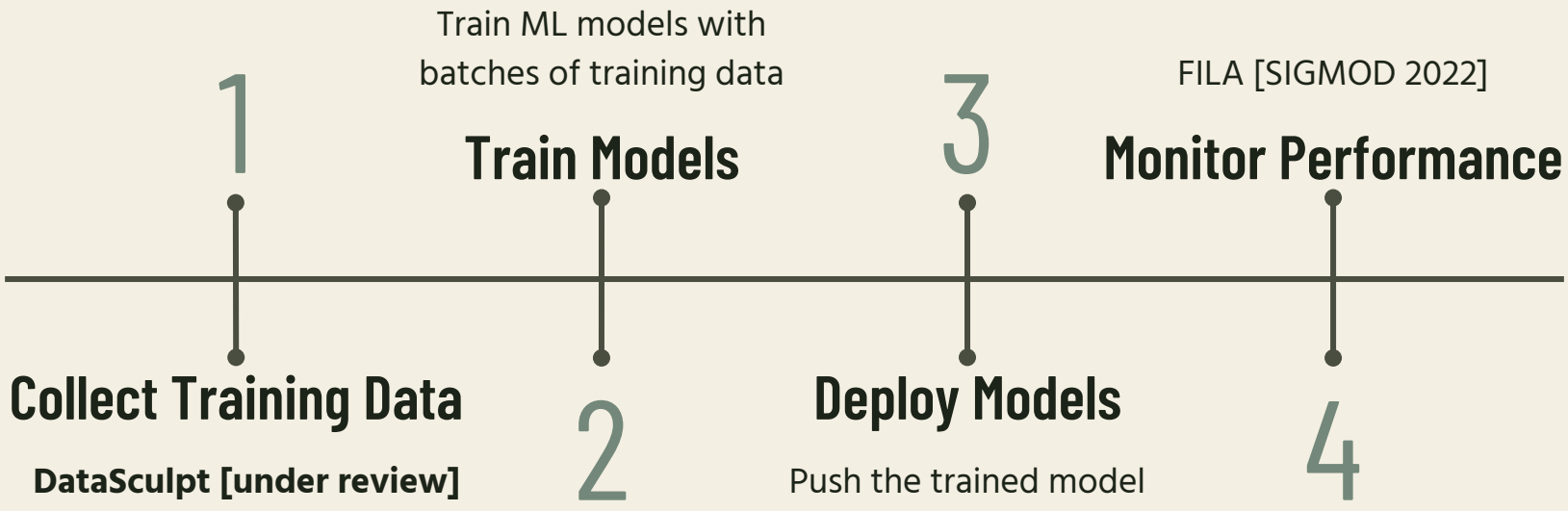
## Data Evaluation

Connect data to business values or task-specific values

# Data Curation for ML Pipelines



# Data Curation for ML Pipelines

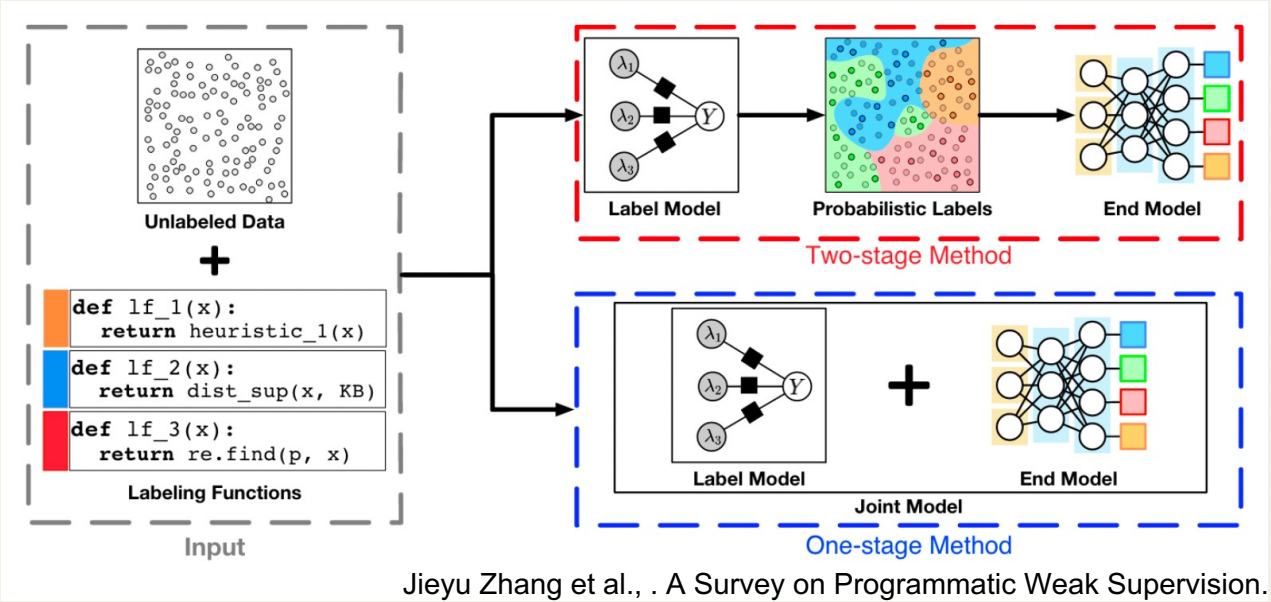


02

# DataSculpt

Automatically design label functions by prompting large language models

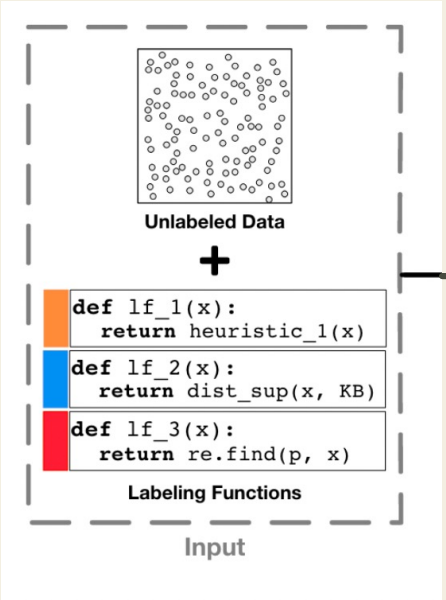
# Programmatic Weak Supervision




Jieyu Zhang et al., . A Survey on Programmatic Weak Supervision.

Design LFs → Train Label Model → Generate Labels → Train End Model

# Programmatic Weak Supervision




Ask human experts to design LFs

- Require nontrivial efforts and costs

DataSculpt: Ask LLMs to design LFs

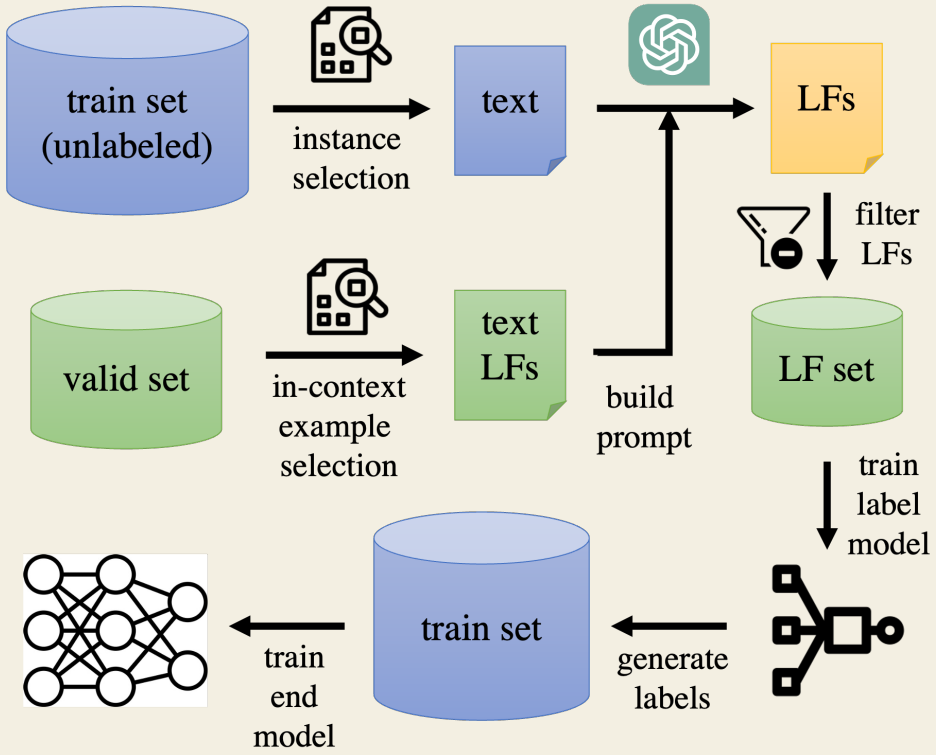
- Will the generated LFs be accurate?

Design LFs → Train Label Model → Generate Labels → Train End Model

# Research Questions

- RQ1: In which cases can large language models design accurate label functions?
- RQ2: How will the current prompting methods, such as chain-of-thought and self-consistency, affect the performance of label function design?
- RQ3: How do different LLMs (GPT-3.5, GPT-4, Llama-2) perform in designing label functions?

# DataSculpt Overview



# DataSculpt Prompts

task description

SYSTEM PROMPT:

You are a helpful assistant who helps users in a sentiment analysis task. In each iteration, the user will provide a movie review. Please decide whether the review is positive or negative. (0 for negative, 1 for positive)
After the user provides input, *first explain your reason process step by step.* Then identify a list of keywords that helps making prediction. Finally, provide the class label for the input.

in-context examples

USER PROMPT:

Query: dead husbands is a somewhat silly comedy about a bunch of wives conspiring to bump off each others husbands…
*Explanation: the review is negative as it thinks the movie is silly.*
Keywords: silly
Label: 0
Query: this movie is an extremely funny and heartwarming story about an orphanage…
*Explanation: the review is positive as it describes the movie as funny and heartwarming.*
Keywords: funny, heartwarming
Label: 1

user query

Query: first the cgi in this movie was horrible I watched it during a marathon of bad movies on the scifi channel…

SYSTEM PROMPT:

You are a helpful assistant who helps users in a chemical disease relation extraction task. In each iteration, the user will provide a biomedical passage, followed by a question asking whether a chemical causes a disease, Please decide whether the chemical causes the disease based on the passage. (0 for the chemical does not cause the disease, 1 for the chemical causes the disease.)
After the user provides input, *first explain your reason process step by step.* Then provide a list of regular expression such that if a passage matches the regex, it is likely to have the same label with the current input. Use {{A}} to represent the first entity and {{B}} to represent the second entity occur in the user's query. Use [SEP] to separate multiple regular expressions. Finally, provide the class label for the input.

USER PROMPT:

Query: During dipyridamole-induced hyperemia, 12 of the 16 dogs with a partial coronary stenosis had a visible area of hypoperfusion… Does dipyridamole cause hyperemia?
*Explanation: The claim states that dipyridamole induced hyperemia, indicating a causal relationship between them.*
Regex: {{A}}-induced {{B}}
Label: 1
…

Query: In the present study we aimed to investigate plasma levels of CGRP during headache induced by the NO donor glyceryl trinitrate (GTN) … Does GTN cause headache?

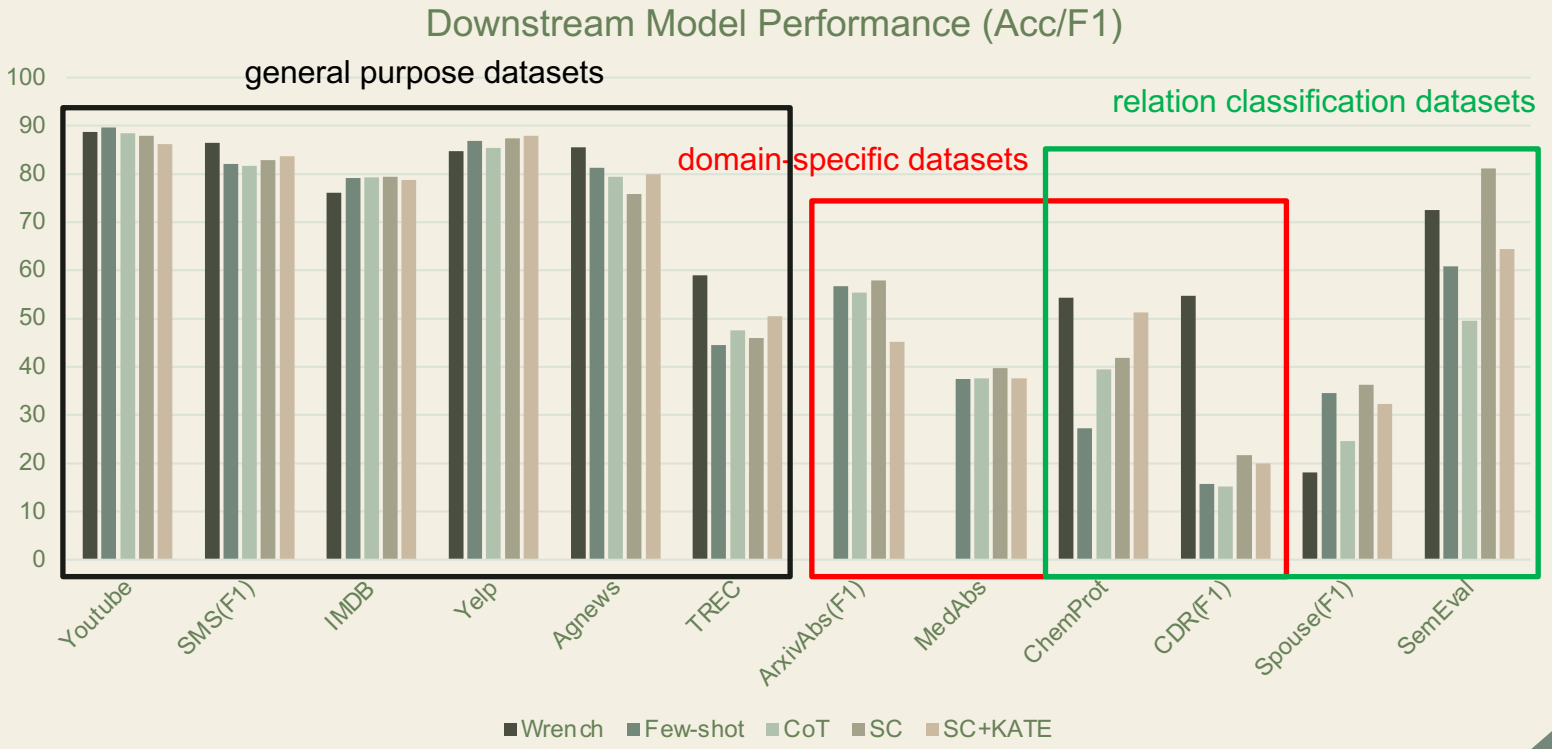
# Experiment Setup

- 12 real-world datasets, 8 for text classification and 4 for relation classification
- Iteratively prompts 50 query instances to the LLM to design LFs

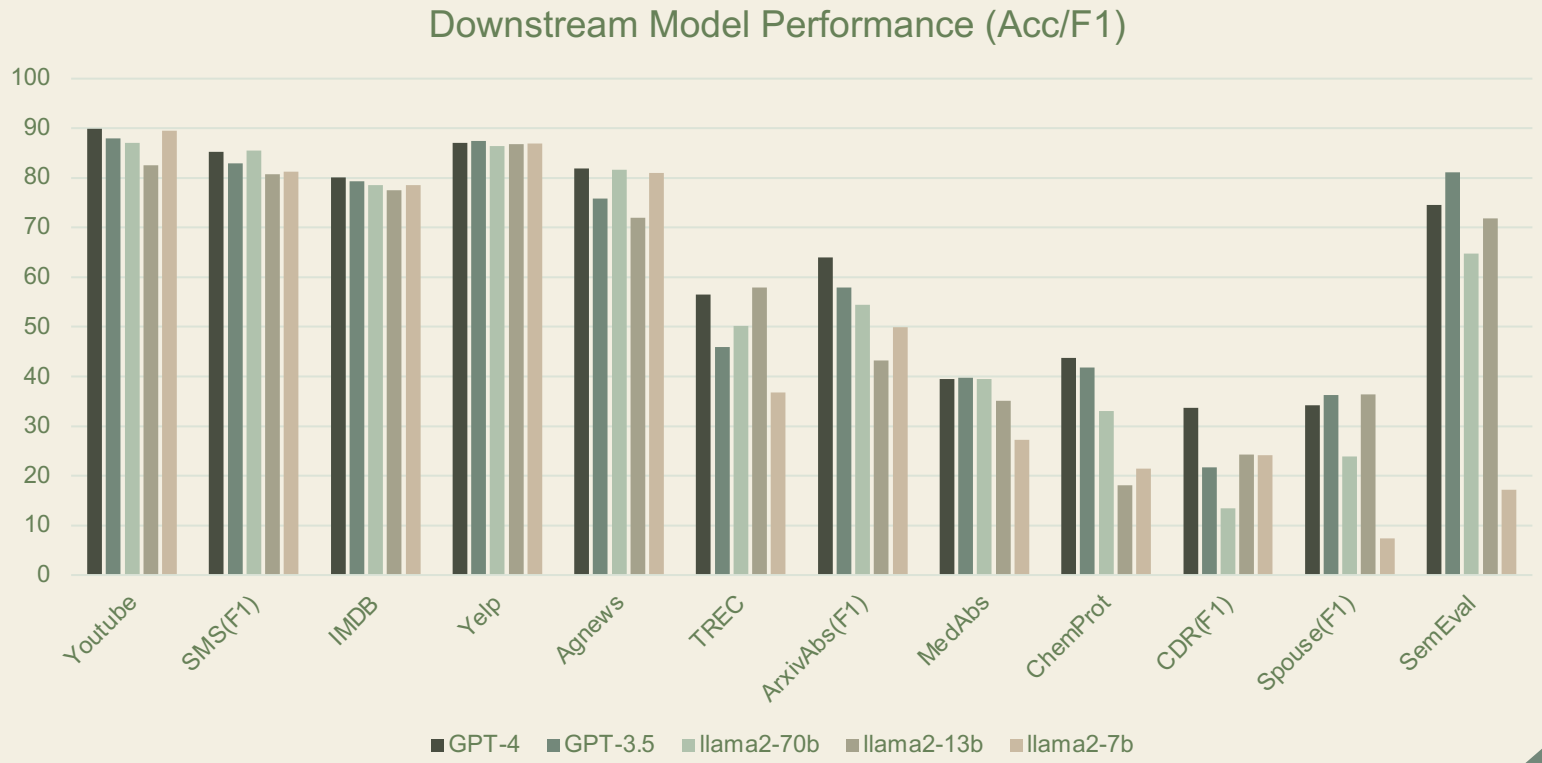
**Table 1: Datasets used in Evaluation.**

| Task | Domain | Dataset | #Class | #Train | #Valid | #Test |
|---|---|---|---|---|---|---|
| Spam Cls. | Review | Youtube [1] | 2 | 1586 | 120 | 250 |
| | Text Message | SMS [2, 4] | 2 | 4571 | 500 | 500 |
| Sentiment Cls. | Movie | IMDB [27, 35] | 2 | 20000 | 2500 | 2500 |
| | Review | Yelp [35, 47] | 2 | 30400 | 3800 | 3800 |
| Topic Cls. | News | Agnews [35, 47] | 4 | 96000 | 12000 | 12000 |
| | Paper Abstract | ArxivAbs [36] | 2 | 21367 | 4579 | 4579 |
| | Biomedical | MedAbs [37] | 5 | 8085 | 3465 | 2888 |
| Question cls. | Web Query | TREC [4, 22] | 6 | 4965 | 500 | 500 |
| Relation Cls. | News | Spouse [9, 31] | 2 | 22254 | 2811 | 2701 |
| | Biomedical | CDR [10, 31] | 2 | 8430 | 920 | 4673 |
| | Web Text | SemEval [16, 49] | 9 | 1749 | 200 | 692 |
| | Chemical | ChemProt [20, 44] | 10 | 12861 | 1607 | 1607 |

# Prompting Methods



# Pre-trained LLMs



# Key Takeaways

- RQ1: In which cases can large language models design accurate label functions?

  *The evaluated LLMs can design accurate LFs for tasks requiring general knowledge, but falls short in tasks requiring specific domain expertise, or developing pattern-based LFs for relation classification tasks.*

- RQ2: How will the current prompting methods, such as chain-of-thought and self-consistency, affect the performance of label function design?

  *While the prompting methods help the LLM makes more accurate predictions, they do not help improve LF accuracy in general. However, combining multiple responses to create a larger candidate LF set helps improve the end-to-end performance.*

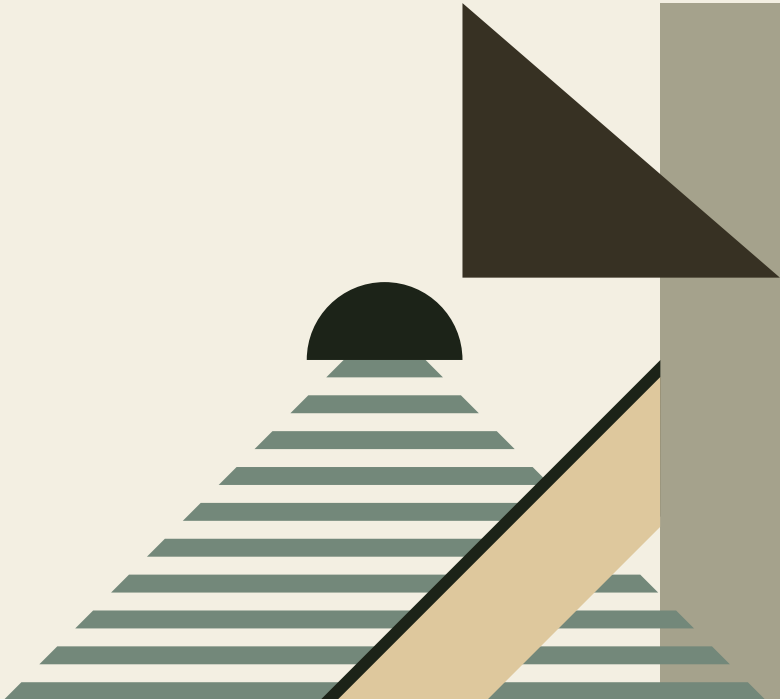
- RQ3: How do different LLMs (GPT-3.5, GPT-4, Llama-2) performs in designing label functions?

  *In general, GPT-4 has the best performance, and Llama-2-70b model has similar end-to-end performance with GPT-3.5. Smaller Llama-2 models (7b and 13b) have problems following the response format.*

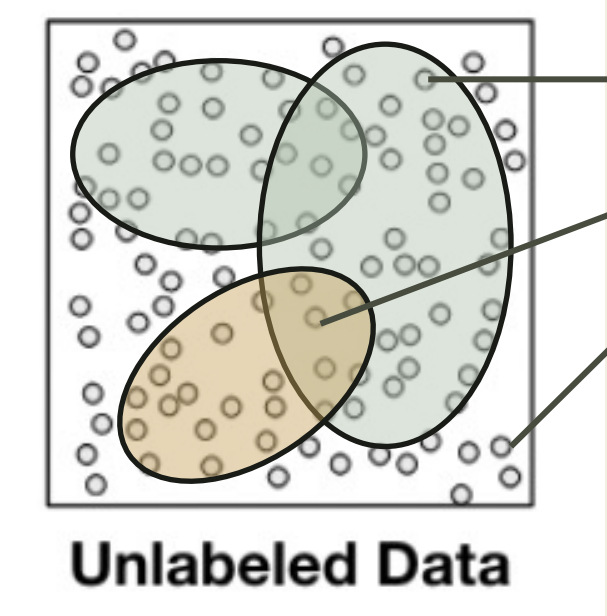
03

# ActiveDP

Combine active learning with PWS to improve label quality



# Motivation



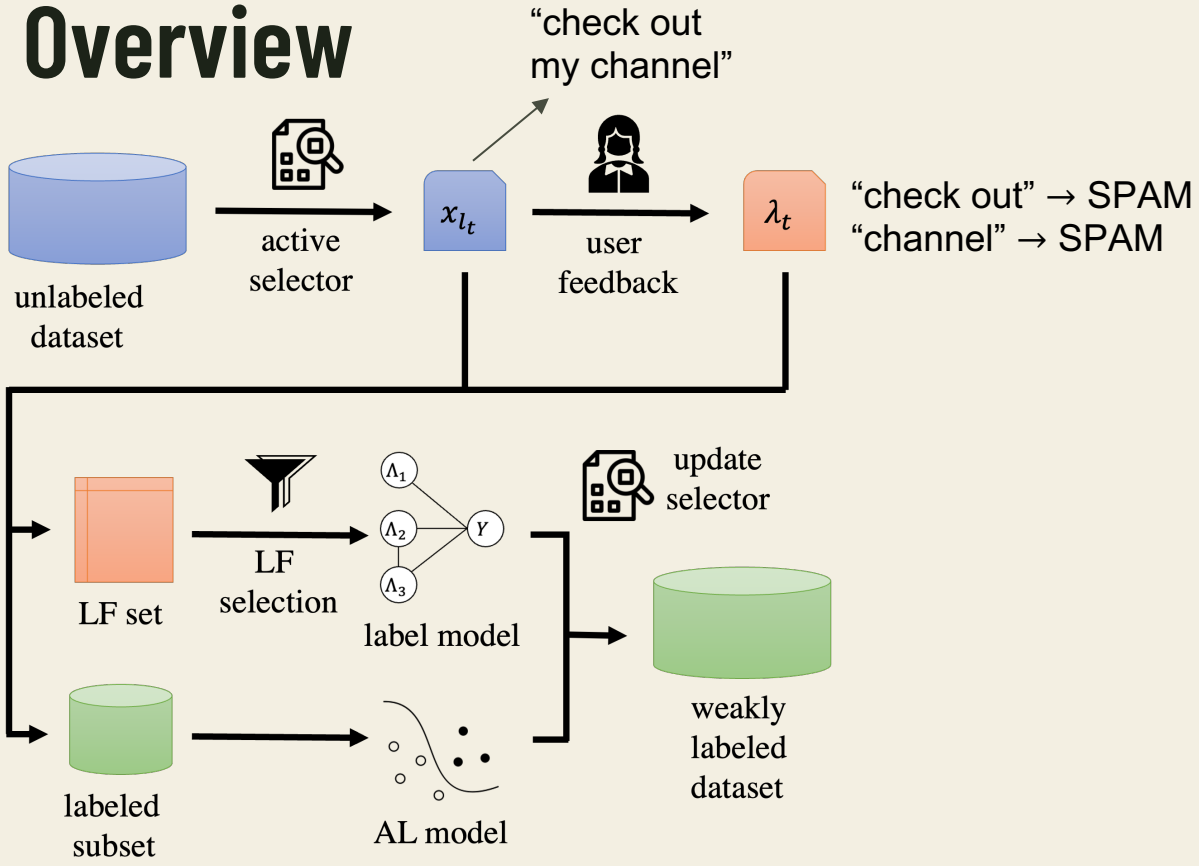

instances with only a few weak labels

conflicting weak labels

uncovered instances

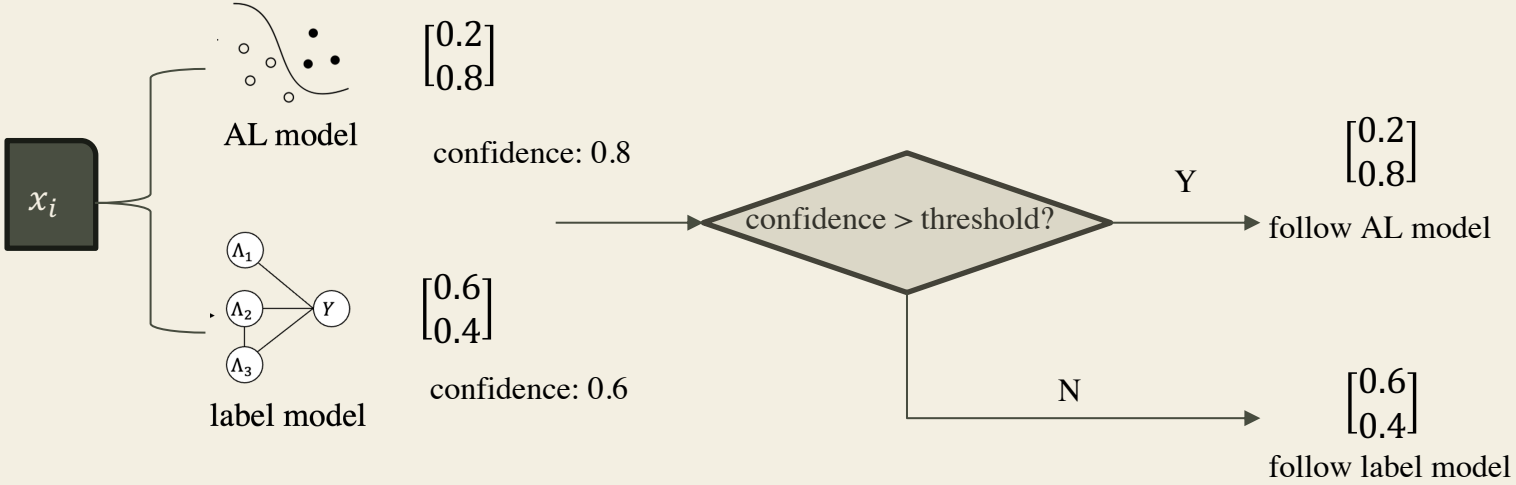
Can we combine weak supervision with strong supervision to improve label quality?

# ActiveDP Overview



# Label Aggregation

We design a confidence-based method for label aggregation. The threshold parameter is tuned on validation dataset to maximize predicted label accuracy.



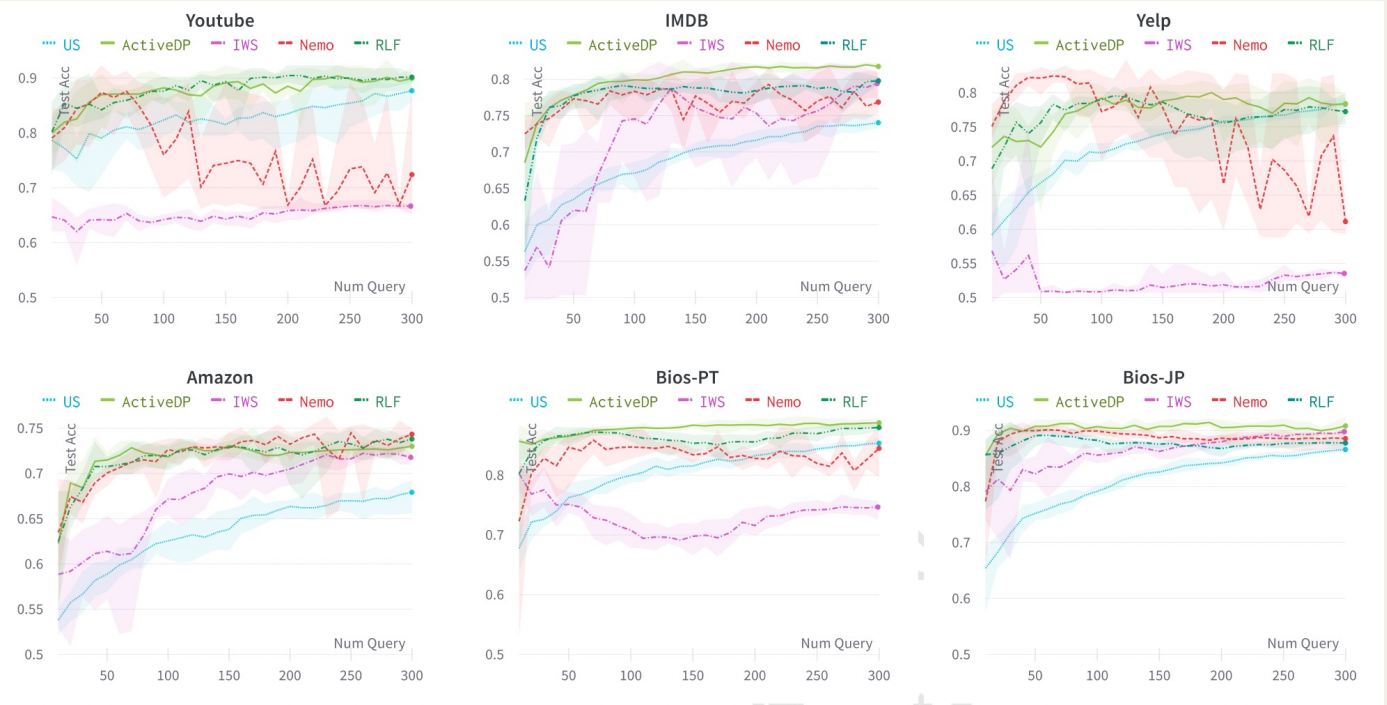
# Active Sampler

The active sampler should select samples that are helpful for both the label model and the AL model, we thus propose a hybrid sampler to balance between these two goals

$$x^* = argmax_x[Entr\big(f_a(x)\big)^{\alpha} * Entr\big(f_l(x,\Lambda)\big)^{1-\alpha}]]$$

Where $f_a(x)$ and $f_l(x,\Lambda)$ are the soft labels predicted by the AL model and the label model respectively, and $Entr(p) = -\sum_j p_j \log(p_j)$ is the entropy of soft labels.

# Experiments




Downstream model's accuracy on 6 evaluated datasets

# Future Directions

## Specialized LLMs for annotation

Domain specific pre-training and finetuning

## Active learning for LLMs

Efficient query instance selection methods for imperfect models

## Synergize multiple paradigms

Combining weak supervision with instance annotations

# Q&A