What Is Wrong with My Model?
Identifying Systematic Problems with **Semantic Data Slicing**

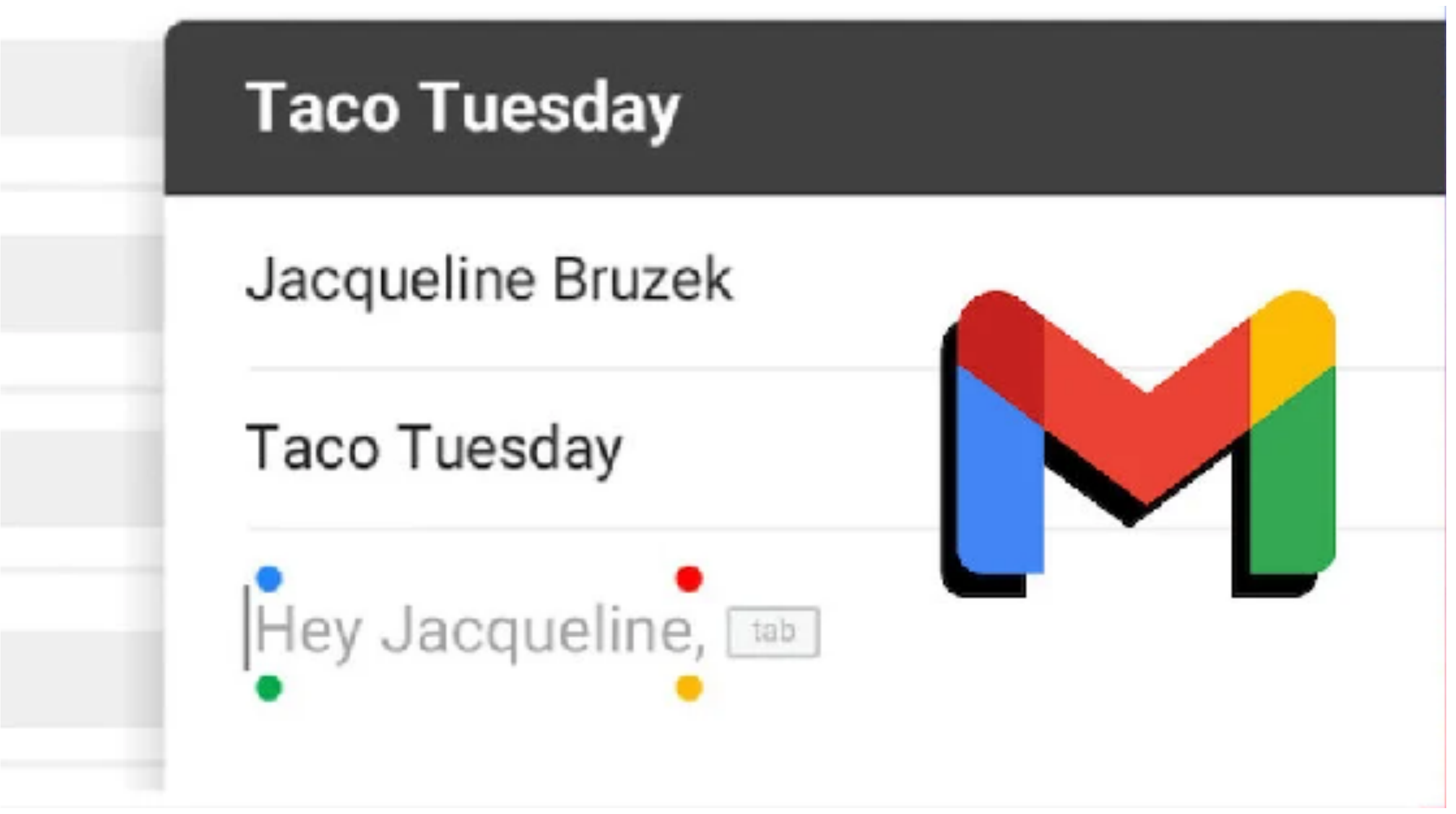


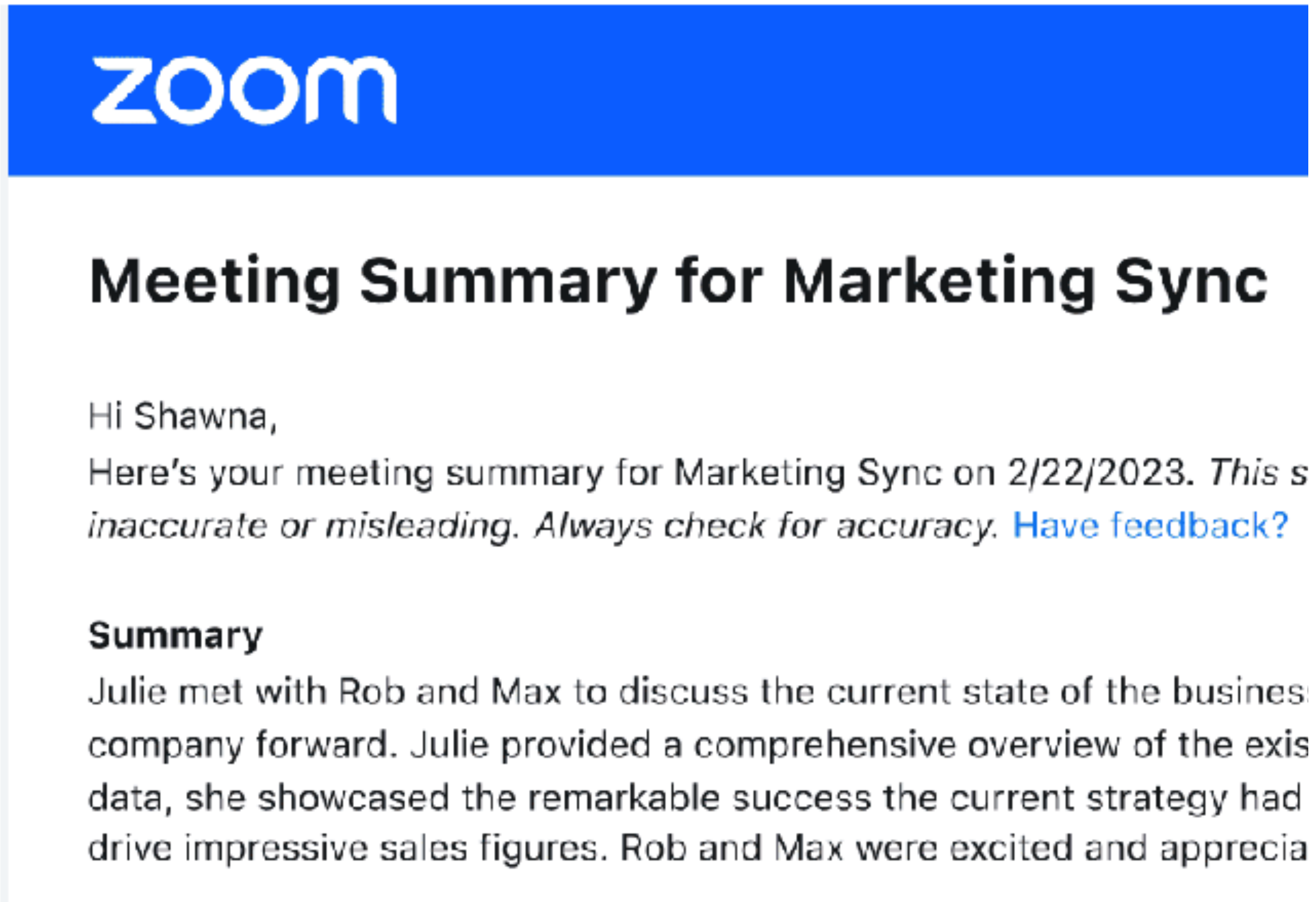


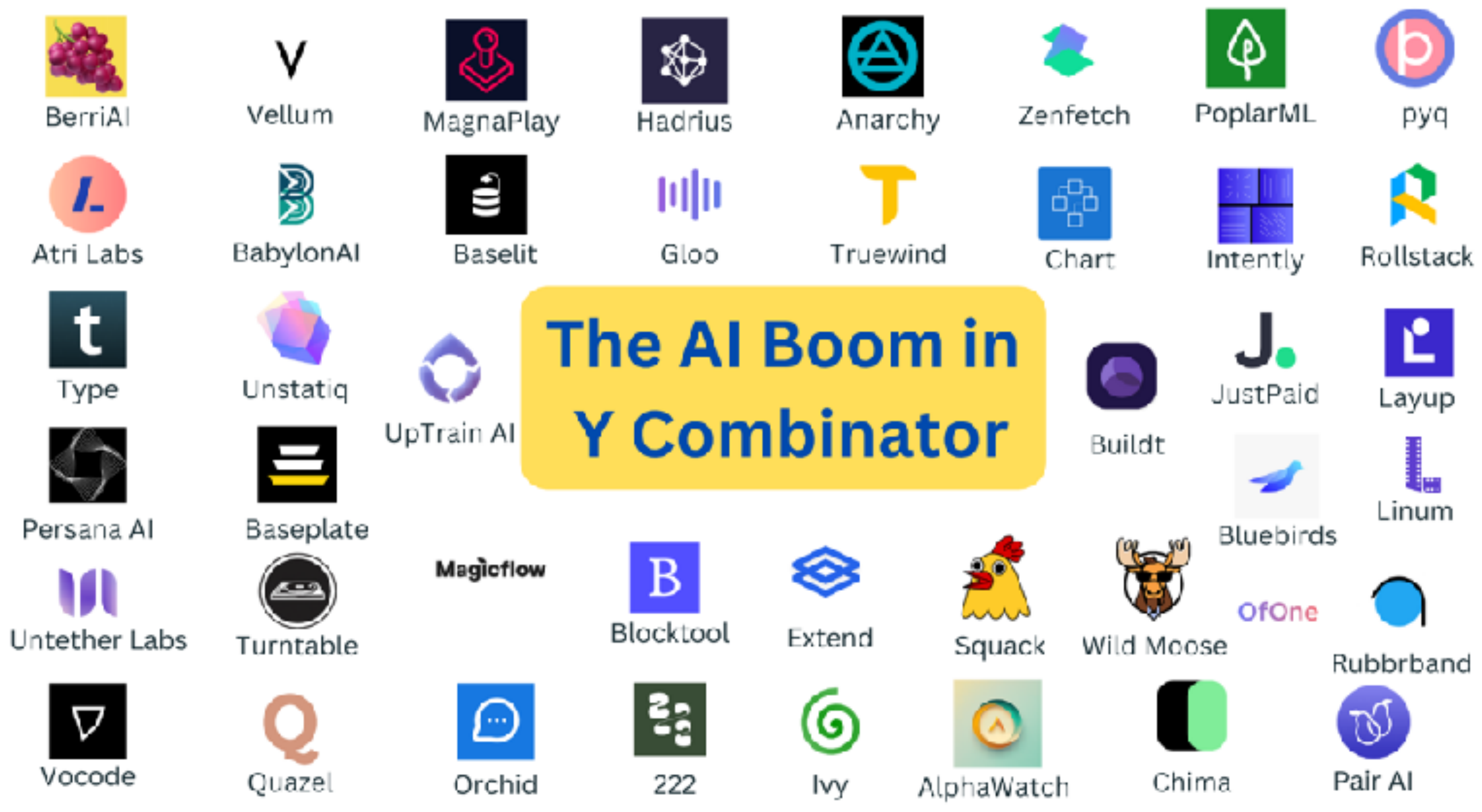
Chenyang Yang, Yining Hong, Grace A. Lewis, Tongshuang Wu, Christian Kästner

Carnegie Mellon University

ML Models are Increasingly Integrated into Software







ML Models are Increasingly Integrated into Software … and Make Mistakes

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said

Whisper is a popular transcription tool powered by artificial intelligence, but it has a major flaw

Melbourne lawyer referred to complaints body after AI generated made-up case citations in family court

Legal professional used software to generate a case citation list, but did not use documents that had undergone human verification

AI Detectors Falsely Accuse Students of Cheating—With Big Consequences

About two-thirds of teachers report regularly using tools for detecting AI-generated content. At that scale, even tiny error rates can add up quickly.

ML Models are Increasingly Integrated into Software ... and Make Mistakes

Air Canada ordered to pay customer who was misled by airline's chatbot

Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said

Whisper is a popular transcription tool powered by artificial intelligence, but it has a major flaw

Compa
when gi

Needs for **ML model quality assurance**,
just like traditional software analysis / testing!

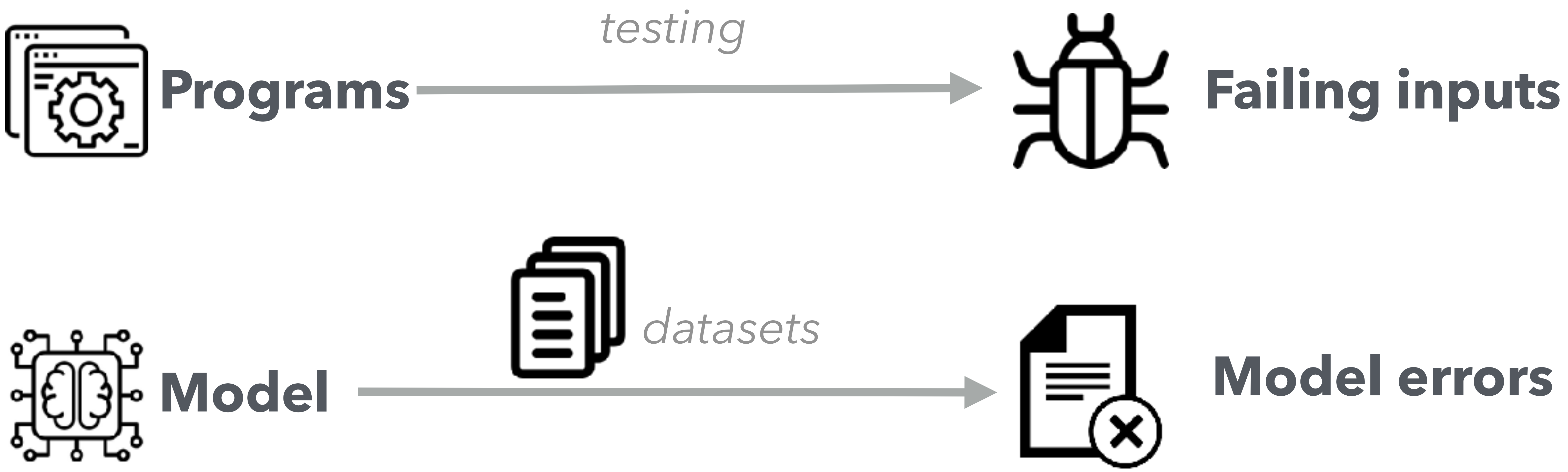
Melb
com
made-up case citations in family court

Legal professional used software to generate a case citation list, but did not use documents that had undergone human verification

Consequences

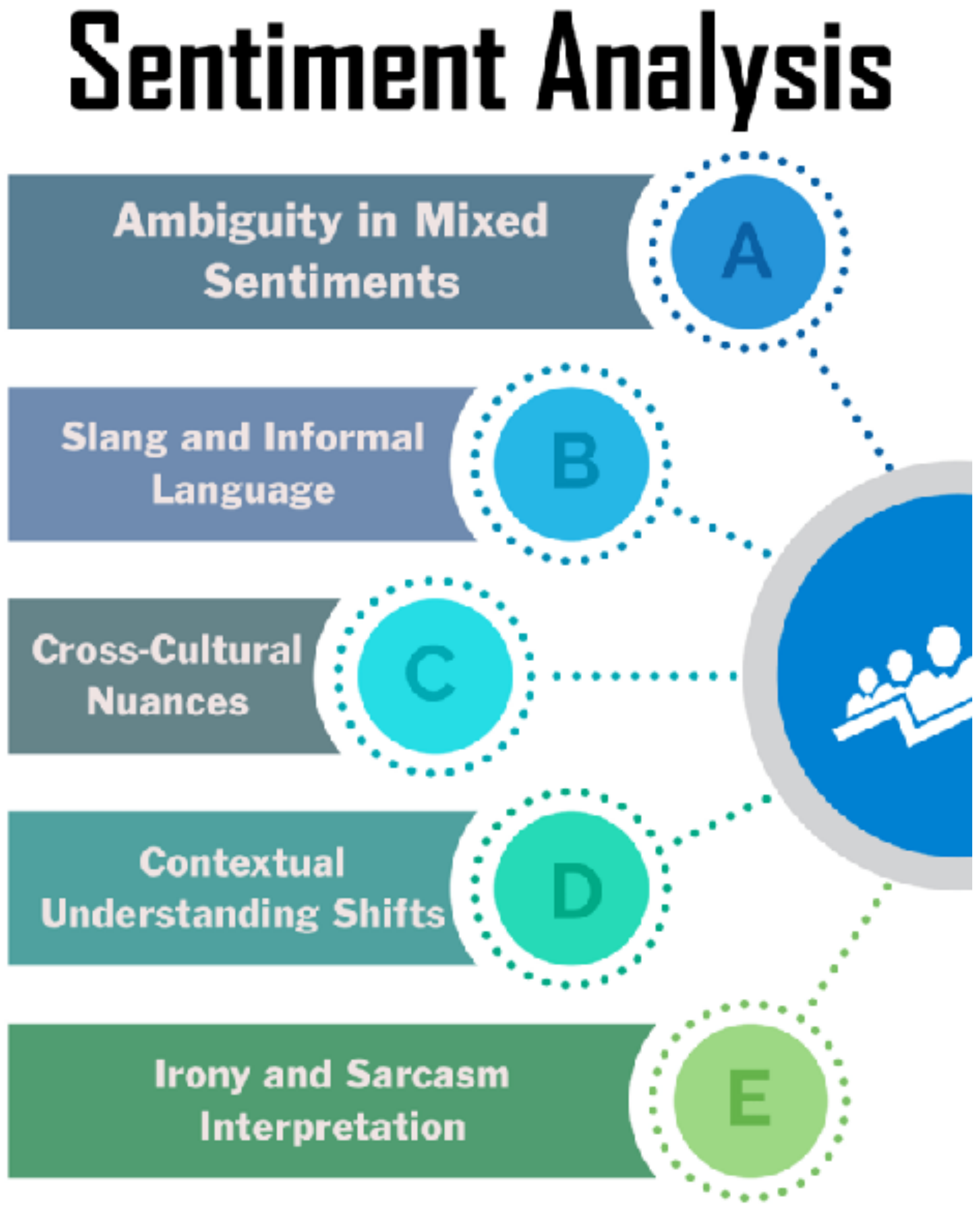
About two-thirds of teachers report regularly using tools for detecting AI-generated content. At that scale, even tiny error rates can add up quickly.

Model Quality Assurance is Different from Programs



ML models always make mistakes – we can not fix every single of them like fixing software bugs.
The question is, **what is the systematic problem behind individual model errors?**

Step 1: Error Analysis to Hypothesize Problems



Error analysis: Go through model errors and hypothesize *high-level patterns*

Error hypothesis: The model is inaccurate *when classifying reviews on locations*

But is this hypothesis true? If so, how prevalent is the problem in production?

Given an error hypothesis and a dataset,
how can we automatically identify all relevant examples?

Step 2: Data Slicing to Validate Problems

Data slicing can identify a subset of examples sharing common characteristics from existing data

Sentiment Analysis

Usually rule-based with significant manual efforts 😩

Hypothesis The model is inaccurate when classifying reviews on locations

Slicing conditions re.search(f"location|walk|far from|close to|neighborhood|near", x)

Data slices

- Close to one major train station
- No walkability for people with mobility issues

…

accuracy gap

Overall Dataset

Traditional Data Slicing is Rule-based & Struggle with Semantic Criteria

Data slicing can identify a subset of examples sharing common characteristics from existing data

Sentiment Analysis

Usually rule-based with significant manual efforts 😩

Hypothesis The model is inaccurate when classifying texts using slangs

The model is inaccurate when classifying texts using sarcasm

Can not slice on arbitrary semantic criteria 😢

Our Work: Semantic Data Slicing

We propose the concept of **semantic data slicing** that can identify a semantically coherent data subset, from arbitrary slicing criteria and datasets

Sentiment Analysis

Little manual efforts ✅
Slice on arbitrary semantic criteria ✅

Hypothesis: The model is inaccurate when classifying reviews on locations

Slicing Criterion: Location "location|walk|far from|close to|neighborhood|near", x)

Data slices:
- Slightly isolated from the area's attractions
- No walkability for people with mobility issues

…

Our Work: Semantic Data Slicing

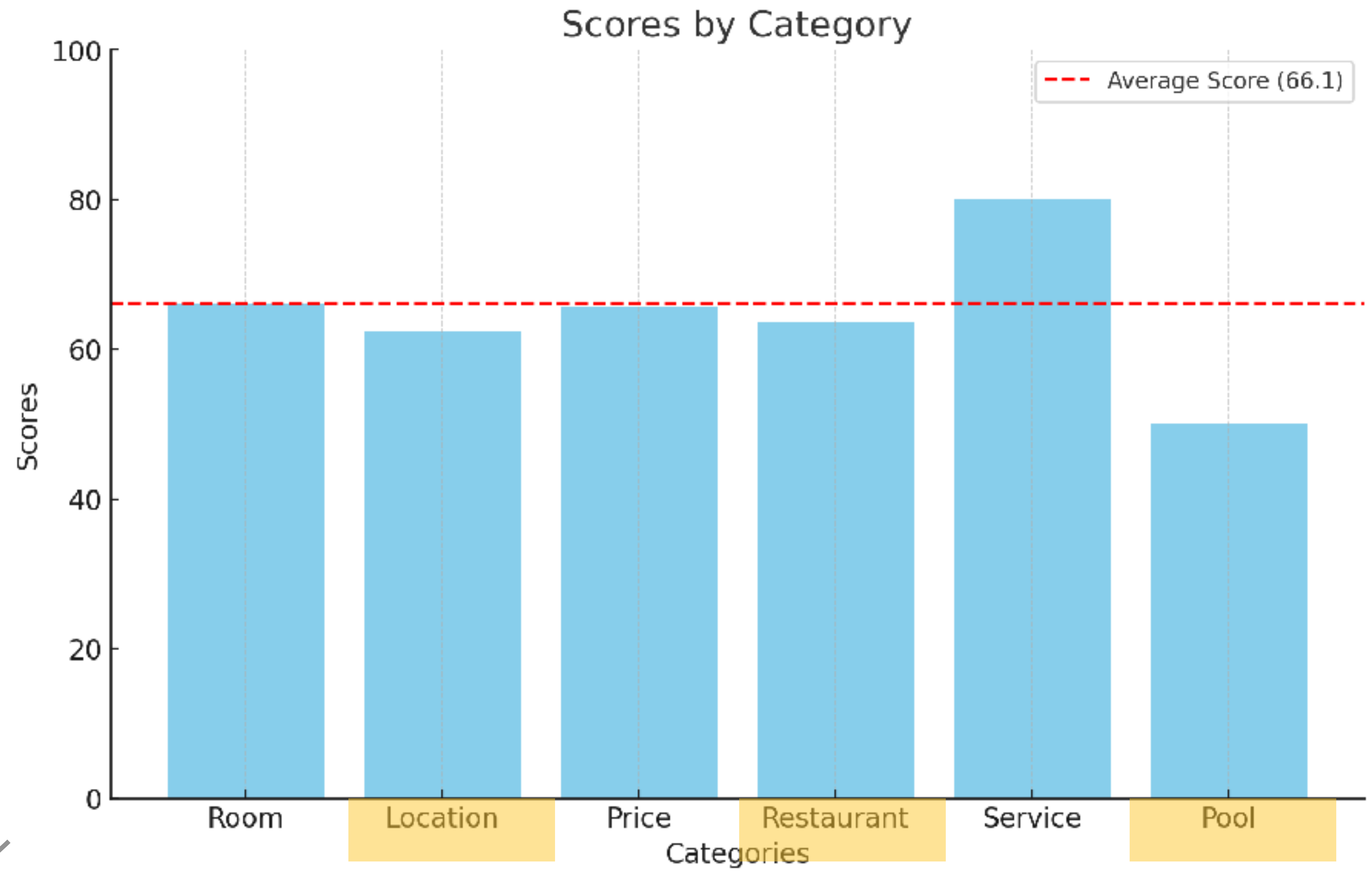
We make semantic data slicing **fast, cheap, and accurate enough** for large-scale model evaluation and analysis

re.search(f"location|walk|far from|close to|neighborhood|near", x)

↓

Location

The model under-performs on location/restaurant/pool related examples



Applications of Semantic Data Slicing

Model debugging: *Can I generalize this model mistake?*

Model evaluation: *Where does my model under-perform?*

Semantic data slicing

Model monitoring: *Does my model regress on the slices?*

Model fixing: *Can I re-train the model to fix the problem?*

Data curation: *Can I curate more data for under-performing slices?*

Designing Semantic Data Slicing

Designing Semantic Data Slicing

Goal: Identify a semantically coherent subset, from arbitrary slicing criteria and datasets
Intuition: LLMs can accurately classify texts given a properly designed prompt

Sentiment Analysis

Hypothesis The model is inaccurate when classifying reviews on locations

Slicing Criterion Location

Simple, fast, but less accurate

Prompt

Is the text related to locations?

Accurate, but high human labor

Does the text discuss aspects of the hotel's location in reviews, including proximity to local attractions, accessibility, [text omitted] ...?

Designing Semantic Data Slicing: Trade-offs

Challenge: How to construct a good prompt for arbitrary semantic slicing criteria, with no training data available, while considering different trade-offs?

How do we produce slicing instructions?

Simple templates vs. complex human-written prompts vs. LLM generated + refined

- Slicing accuracy needed
- Slicing latency expected
- Human effort available
- Computational resources available

Designing Semantic Data Slicing: Trade-offs

Challenge: How to construct a good prompt for arbitrary semantic slicing criteria, with no training data available, while considering different trade-offs?

How do we produce slicing instructions?

Simple templates vs. complex human-written prompts vs. LLM generated + refined

How many few-shot examples do we provide?

Zero-shot vs. few-shot

Which model do we use for data slicing?

Smaller model vs. larger model

…

Slicing accuracy needed

Slicing latency expected

Human effort available

Computational resources available

Designing Semantic Data Slicing

Stage 1: Prompt Construction

Slicing Criterion Location

Data

Instructions

① Instruction generation

② Instruction refinement

Does the text discuss aspects of the hotel's location in reviews, including proximity to local attractions, accessibility, [text omitted] …?

Few-shot Examples

③ Example sampling

④ Example labeling

⑤ Example synthesis

Input: Slightly isolated from the area's attractions
Output: Yes

Input: Crazy expensive for what you get
Output: No
…

Stage 2: Data Slicing

Slicing Prompt

Input: No walkability for people with mobility issues
Output: Yes

Evaluating Semantic Data Slicing

Evaluation

Comparing **accuracy, cost, and latency** of
9 configurations of our semantic slicing framework across 4 datasets

75.9% average F1-score with full automated workflow + human intervention

Most important steps: Few-shot examples & instruction refinement from humans

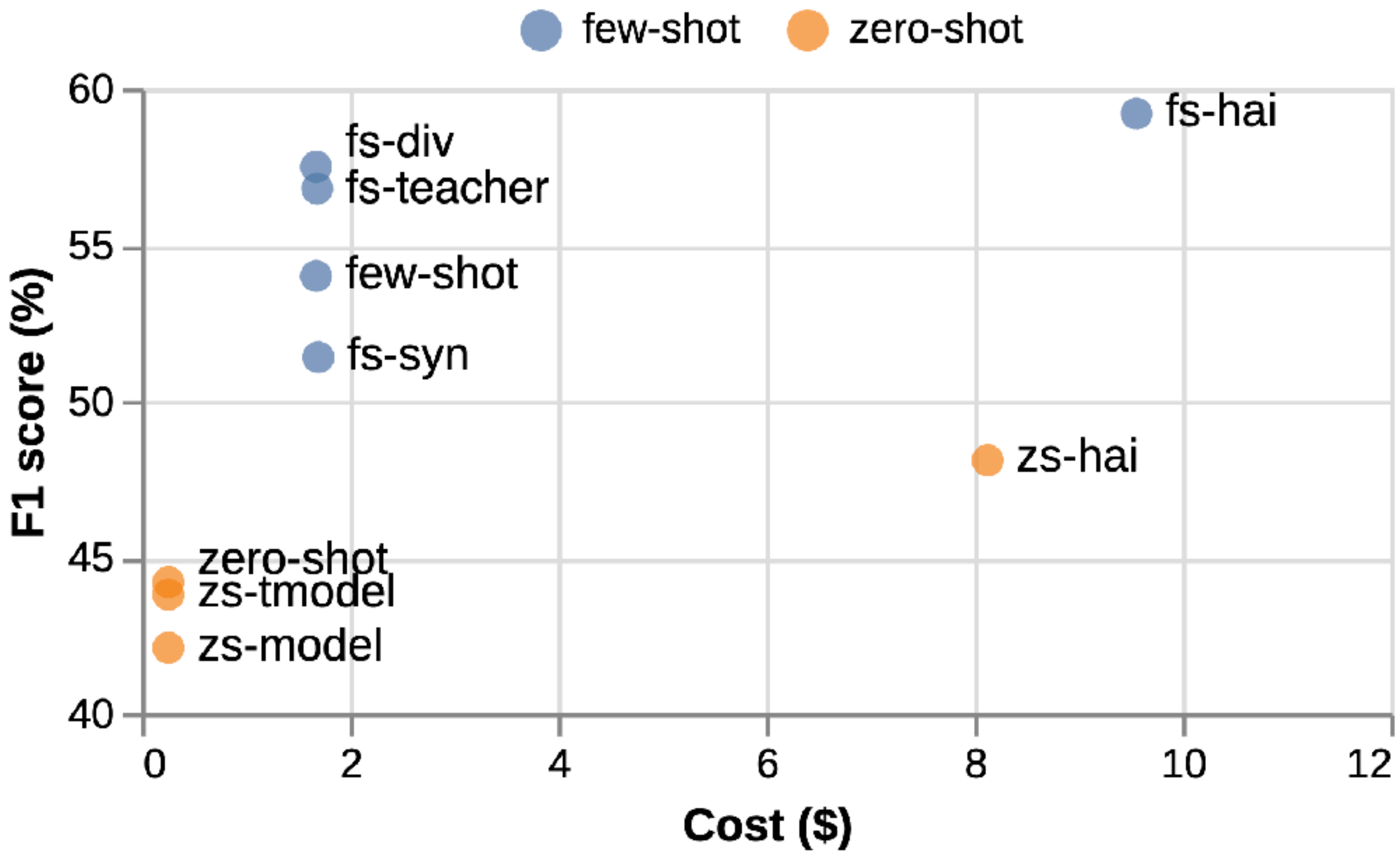
13.5 minutes to generate a slice for 6000 examples **costing $1.7**

*We use 2 A6000 GPUs for local model inference and estimate the cost from cloud providers

Evaluation

Comparing **accuracy, cost, and latency** of
9 configurations of our semantic slicing framework across 4 datasets

Flexible trade-offs with different configurations



Evaluation: Usefulness

Use our semantic slicing framework to **identify under-performing slices** in existing datasets +
Invite practitioners to use our framework to conduct model evaluation

7 out of 7 known under-performing slices can be successfully identified

Practitioners generate additional insights for model evaluation

- Task: Understand model alignment with different demographics

- Insight: slice on "age-related power imbalance" aligns well with millennials but poorly with people older than 40

Takeaways

Check out our paper!

Carnegie Mellon University

ML Models are Increasingly Integrated into Software ... and Make Mistakes

Air Canada ordered to pay customer who was misled by airline's chatbot

Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said

Needs for **ML model quality assurance**,
just like traditional software analysis / testing!

Error Analysis + Data Slicing to Identify Systematic Errors

Traditional data slicing is rule-based & struggle with semantic criteria

Semantic Data Slicing

```
from semslicer.slicer import InteractiveSlicer
data = load_training_data()
criterion = "Muslim"

slicer = InteractiveSlicer(criterion, data, config={
        'few-shot': True,
        'few-shot-size': 8,
        'instruction-source': 'template',
        'student-model': 'flan-t5-xxl',
        'teacher-model': 'gpt-4-turbo-preview'})
```

https://github.com/malusamayo/SemSlicer

Semantic data slicing is accurate, fast, and of low cost

75.9% F1-score with full automated workflow + human intervention

Most important steps: Few-shot examples & instruction refinement from humans

13.5 minutes to generate a slice for 6000 examples **using $1.7**

7 out of 7 known under-performing slices can be successfully identified

Practitioners generate additional insights for model evaluation

- *"age-related power imbalance" aligns well with millennials but poorly with people older than 40*