Towards A Production-Ready Customer Feedback LLM:
Leveraging LLM Evals for Advanced Feedback Analysis

Ilya Boytsov
NLP Lead, Wayfair

Jason Lopatecki
Co-Founder and CEO, Arize AI

Agenda

- Intro about me and my team at Wayfair
- Customer feedback LLM discovery and a role of LLM evaluation pipelines
- Case study: LLM evals with Arize & Phoenix for feedback annotation
- How Evals insights help to design production LLM system
- General Evals findings, best practices and challenges
- Conclusions

About Me

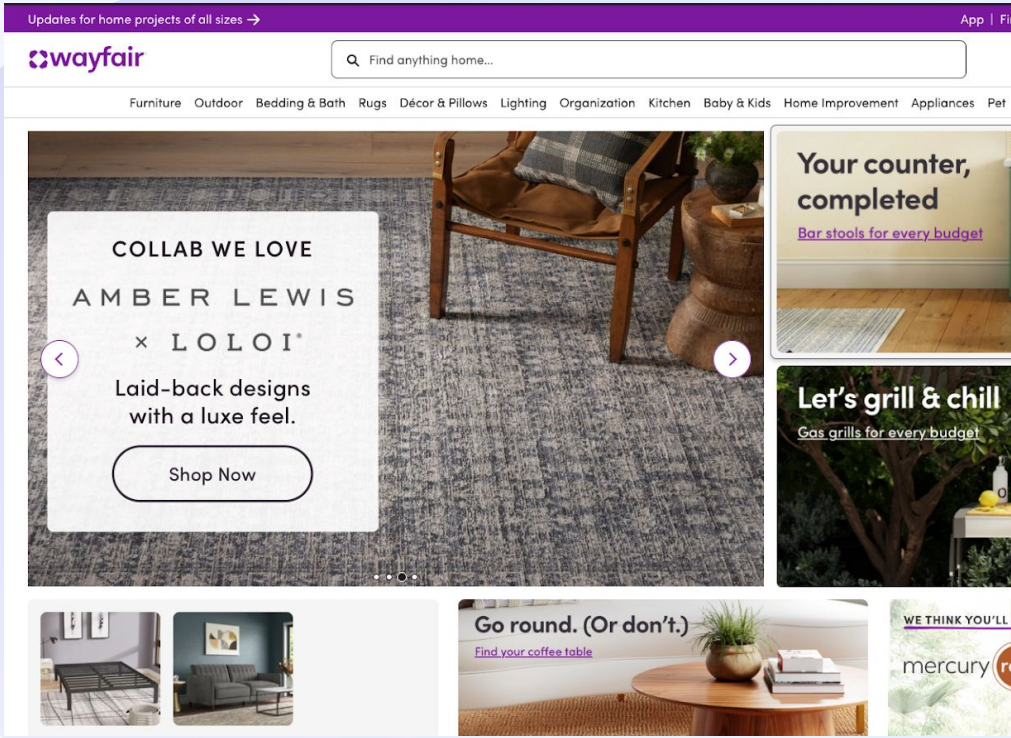
Applied NLP at Wayfair, working and living in Berlin

Co-creator of SSAI - a series of meetups for industry practitioners

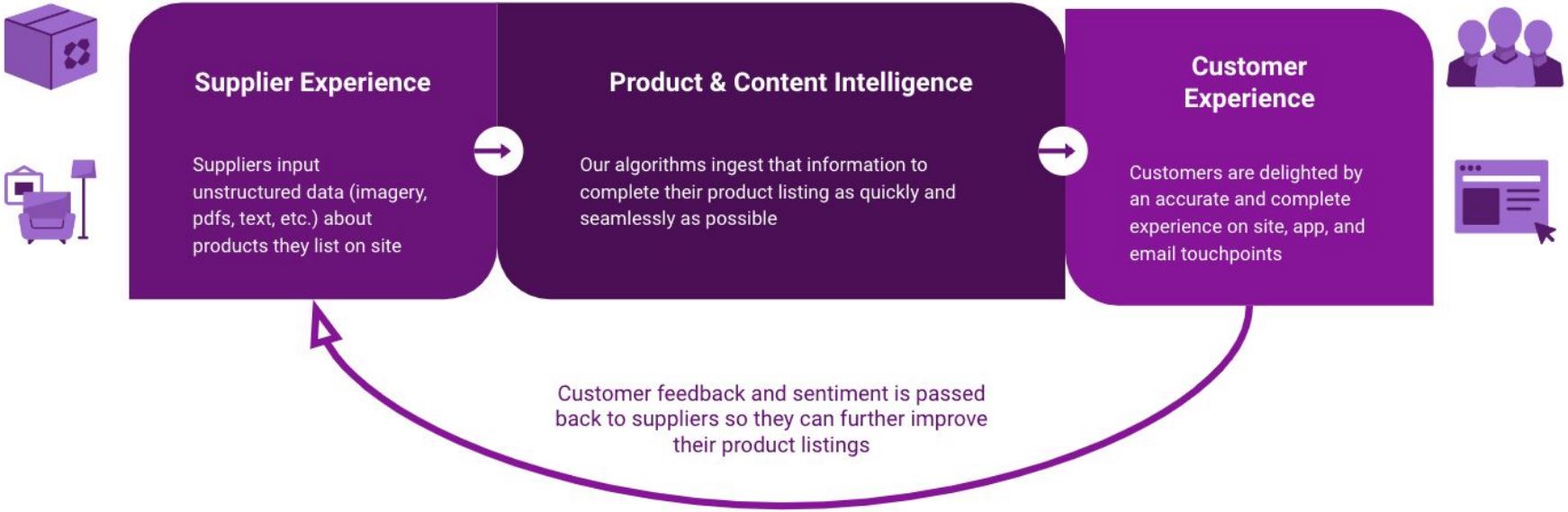
Guest Lecturer, University of Oxford (LLM summer school)

About Wayfair

- > 22 million customers
- > 40 million products
- > 20,000 suppliers
- > 100 million product reviews



Product & Content Intelligence Team (PCI)

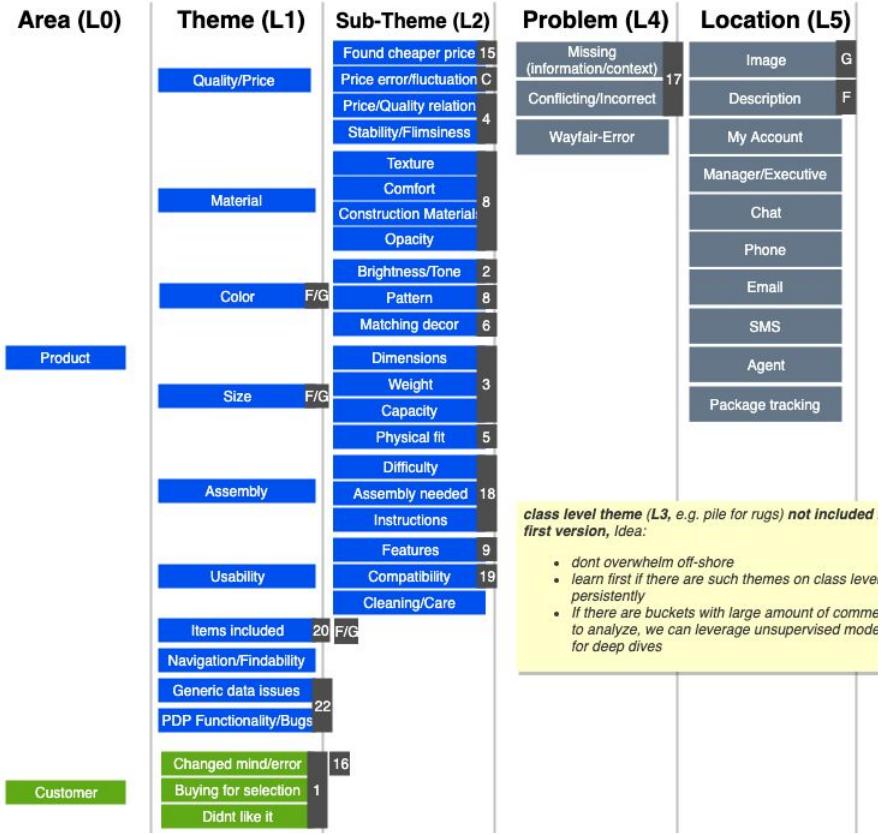


PCI Team Projects

Customer Feedback Annotation

Goal: annotate customer feedback.
[Blog post](#) with initial model.

Not gloss white as shown, shelf too small and wiring up too complicated



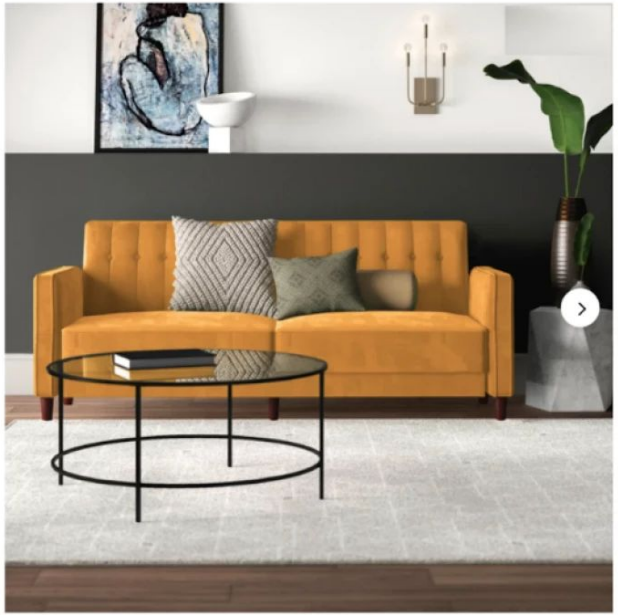
Bubble Filters

Goal: provide customers with the ability to see (and filter) reviews by “common topics” that are mined from review comments.

In past we open sourced our implementation of a topic model ExtRA motivated by the paper.

Product Tags Extraction

Goal: improve customer experience by integrating extracted tags to filters and Search.



ML

Product Class:	Sofa
Arm style:	Square arm
Color:	Yellow
Style:	Vintage, Modern
Design:	Tufted
#of seats:	2
Legs:	Wood

Product Class:	Table
Top Shape:	Round
Frame color:	Black
#of Legs:	4

Product Class:	Accent Pillows
Color:	Gray
Pattern:	Geometric

Customer Feedback LLM Discovery

Status before discovery:

Separate models for customer feedback annotation, aspect-based sentiment analysis, bubble filters extraction, review moderation, etc.

Problem:

Hard to maintain and keep developing a large number of models

Proposed solution:

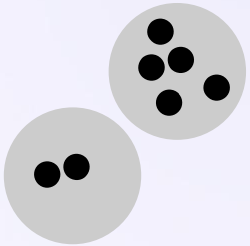
Replace existing production models with a single LLM-based consolidated model

Customer Feedback LLM Design

Challenges to Build a Production-Ready LLM App

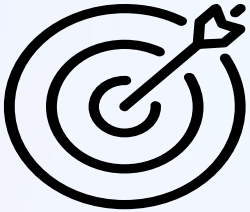
- LLMs output format is not deterministic
- LLMs tend to hallucinate
- Not enough empirical observations what else might go wrong
- Engineering challenges
- ...

Key Components



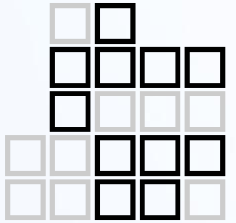
LLM fine-tuning

SFT, RLHF, DPO



Experiment runner

langchain, mlflow



Eval pipelines

Arize, Phoenix

Evaluation of LLM-Based Applications

There are no industry standards yet to evaluate LLM-based solutions, especially within specific domains

Case Study

Customer Feedback Annotation

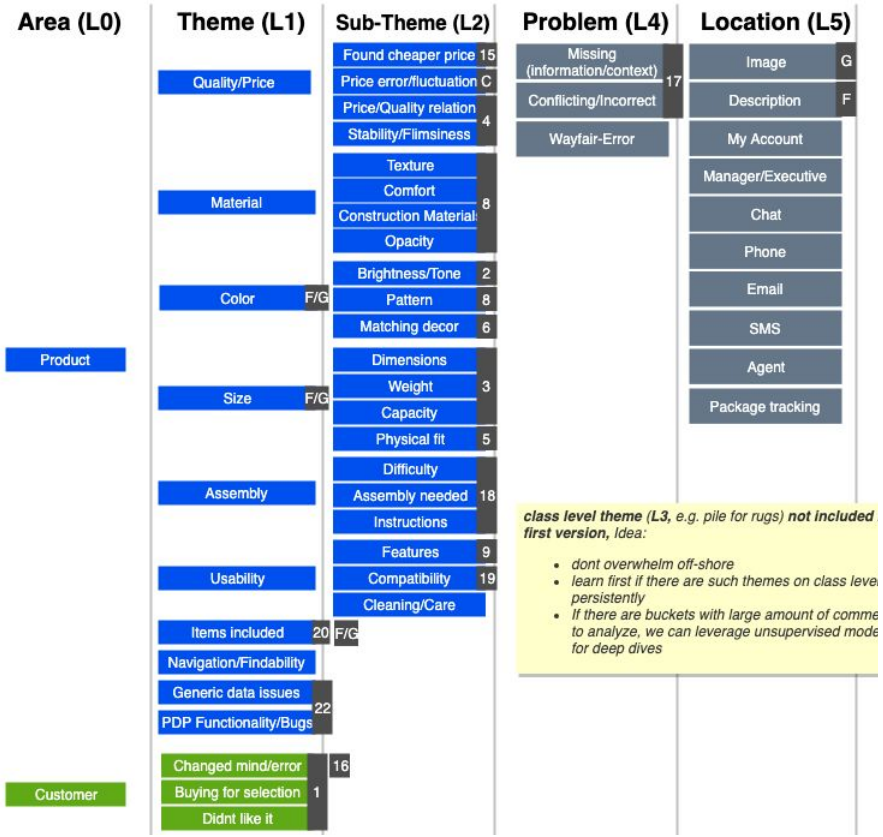
Existing Customer Feedback Annotation Model

- Is a **hierarchical classification** transformer model used to predict customer feedback taxonomy (85 topics)
- It **tags each comment** with one or more topic from a predefined feedback taxonomy
- Is used for **various sources of data**: customer reviews, return comments, etc.
- is **very smart** but also a legacy!

Customer Feedback Annotation

Goal: annotate customer feedback.

Not gloss white as shown, shelf too small and wiring up too complicated



LLM-Based Approach

Model: Zephyr-7b-SFT (outperforms Gemini-pro and GPT-4 on our data)

Prompt:

```
You are given a text from a customer and the goal is to classify customer feedback into the list of predefined topics.

Predefined topics: {topics}

Choose all topics that are mentioned in the text. Do not add topics that are not from the list of predefined topics.

For example:

Text is: Service was as expected. I am planning another order from Wayfair. I liked the quality of what I bought.

Topics: ["service", "product"]

Text is: {text}

Topics:
```

Initial Observations

- LLMs don't always respond with expected output format (list, json)
- LLMs sometimes make up topics that are not present in the taxonomy:
  - **Style** is predicted but is **absent** in the taxonomy
  - **Delivery** is predicted, but the taxonomy has only **Logistics**
  - **Difficult** is predicted but the taxonomy has only **Difficulty**

Note: presence of new topics is not exactly bad, it **may be a useful signal to expand the existing taxonomy**, but it is important to validate new topics before we make such a decision!

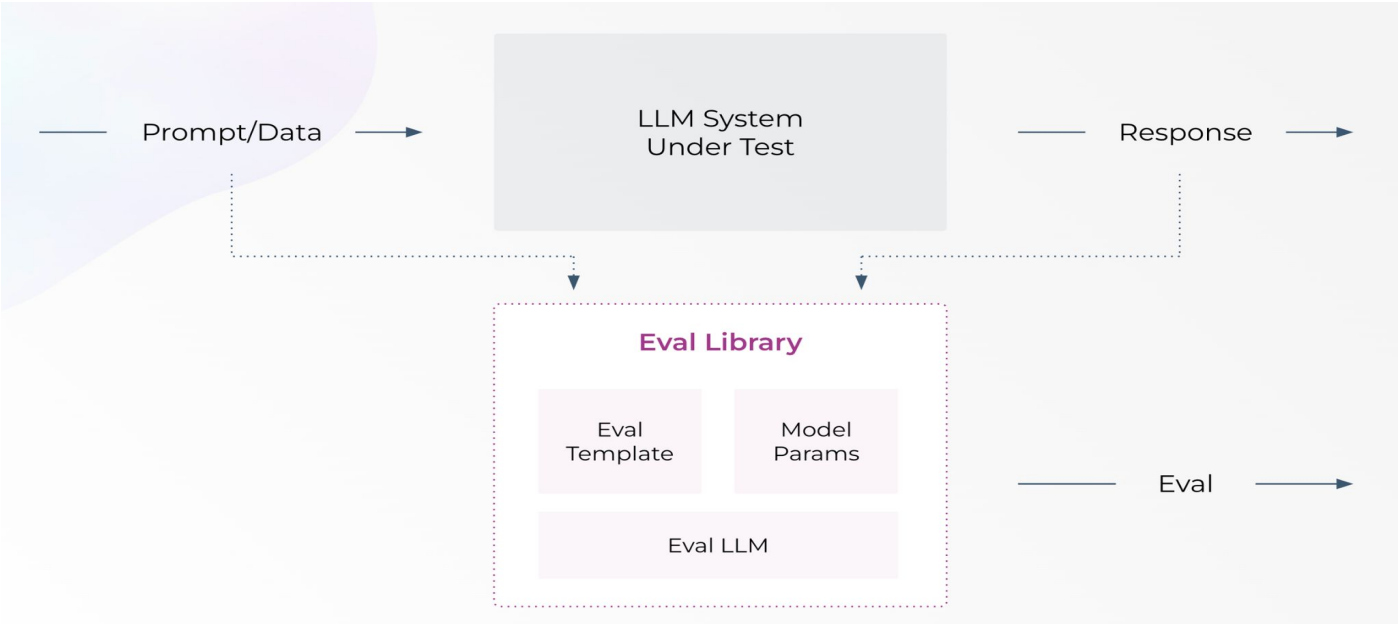
Evaluation Tools

ML observability platform

Tools to evaluate LLM applications

LLM as a Judge Grading Approach

Key idea: ask an LLM to do the grading for you. The method was proposed in [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena]



LLM as a Judge: Correctness Evaluation (using GPT-4)

```
TOPIC_CORRECTNESS_TEMPLATE = """
In this task, you will be presented with feedback from a customer and an extracted topic from an AI system.
Your goal is to determine if the extracted topic is accurate given the customer feedback. Here is the data:
    [BEGIN DATA]
    [CUSTOMER FEEDBACK]: {Content}
    ************
    [EXTRACTED TOPIC]: {topics}
    [END DATA]
Look at the extracted topic and the customer feedback to determine if the extracted topic is accurate based on the feedback text.
Focus on the content of the customer feedback when determining if the extracted topic is accurate.
First, write out in a step by step manner an EXPLANATION to show how to arrive at the correct answer.
Avoid simply stating the correct answer at the outset. Your response LABEL must be a single word, either "accurate" or "unrelated",
and should not contain any other text or characters aside from that word. "accurate" means that the extracted topic is relevant based
on the customer feedback. "unrelated" means that the topic doesn't capture a theme/category of the feedback.
Example response:
************
EXPLANATION: An explanation of your reasoning for why the label is "accurate" or "unrelated"
LABEL: "accurate" or "unrelated"
************
EXPLANATION:
"""
```

Example 1: LLM Decisions Well Explained

Customer review:

“These are a wash for me. The color was more like a pea green, not quite the color I was going for. Also the one we put together wasn’t easy. The screws are quite difficult to thread. I tried to have my kids do it to learn how to put something together but I had to keep jumping in to help because they couldn’t thread them.

I bought six for my dining table but they leaned back so far that no one thought they were comfortable enough to sit at the table. They are a little more reclined than I would have liked and they rockedbounced back even more when sitting. Sent all three sets back despite the 40box return shipping fee. I figured I still get back more than if I tried to sell them. I took an expensive risk and paid for it.”

LLM-predicted new topics: ['use', 'smell', 'taste']

LLM Correctness labels: ['accurate', 'unrelated', 'unrelated'], correctness score: 0.33

LLM explanation: The customer feedback primarily discusses issues related to the assembly and use of the chairs, including the difficulty of threading screws during assembly, the discomfort of the chairs when used at a dining table due to their reclined position, and the decision to return them because of these issues. The topic 'use' accurately captures these concerns as they all relate to the customer's experience with using the product, from assembly to actual sitting. Therefore, the extracted topic of 'use' is relevant and accurately reflects a significant theme of the feedback…”

Example 2: Very Unexpected Response by the LLM

Customer review:

“Being the mom to 6 kids, I use a lot of different types of containers. The Rubbermaid container is amazing! Not to mention, you can use it in the microwave plus wash it in your dishwasher! Very family friendly and this mom gives a huge thumbs up!”

LLM-predicted new topics: ['间隔']

LLM Correctness labels: ['unrelated'], correctness score: 0

LLM explanation: The extracted topic given is "间隔" which, when translated from Chinese, means "interval" or "space between".
The customer feedback primarily discusses the effectiveness of a Rubbermaid container. It does not discuss intervals, spaces between objects, or anything related to the concept of "间隔".
Therefore, the extracted topic does not accurately capture a theme or category of the feedback provided by the customer.”

Topics Similarity Analysis: Merge Similar Topics Together

Some of the new topics may be very *similar* to the ones from the original taxonomy

Predicted topic: **Delivery** (not present in the taxonomy)

Topic in the taxonomy: **Logistics** (very similar by meaning)

Compute **Cosine similarity** between **Delivery** **and** **Logistics** embeddings

If cosine similarity is greater than the upper threshold then **merge** Delivery into Logistics.

Topics Similarity Analysis: Taxonomy Expansion

Some of the new topics may be very *dissimilar* from the existing ones, but can be good candidates for the taxonomy expansion!

Predicted topic: **Portability** (not present in the taxonomy)

Compute **Cosine similarity** between **Portability** **and** **all other existing topics**

If max cosine similarity is lower than the lower threshold then **mark** **Portability** as a potential candidate for the taxonomy expansion.

Topic Similarity Analysis

Embedding model: DistilBERT fine tuned on customer reviews

Visualisation: Arize Phoenix

Semantic Similarity Between New (Accurate) Topics and Existing Taxonomy

Dataset	Raw Data
new_topics	text message
predefined_topics	email
predefined_topics	phone
predefined_topics	chat
predefined_topics	sms

Dataset	Raw Data
new_topics	different price
predefined_topics	found cheaper price

Semantic Similarity Between New (Accurate) Topics and Existing Taxonomy

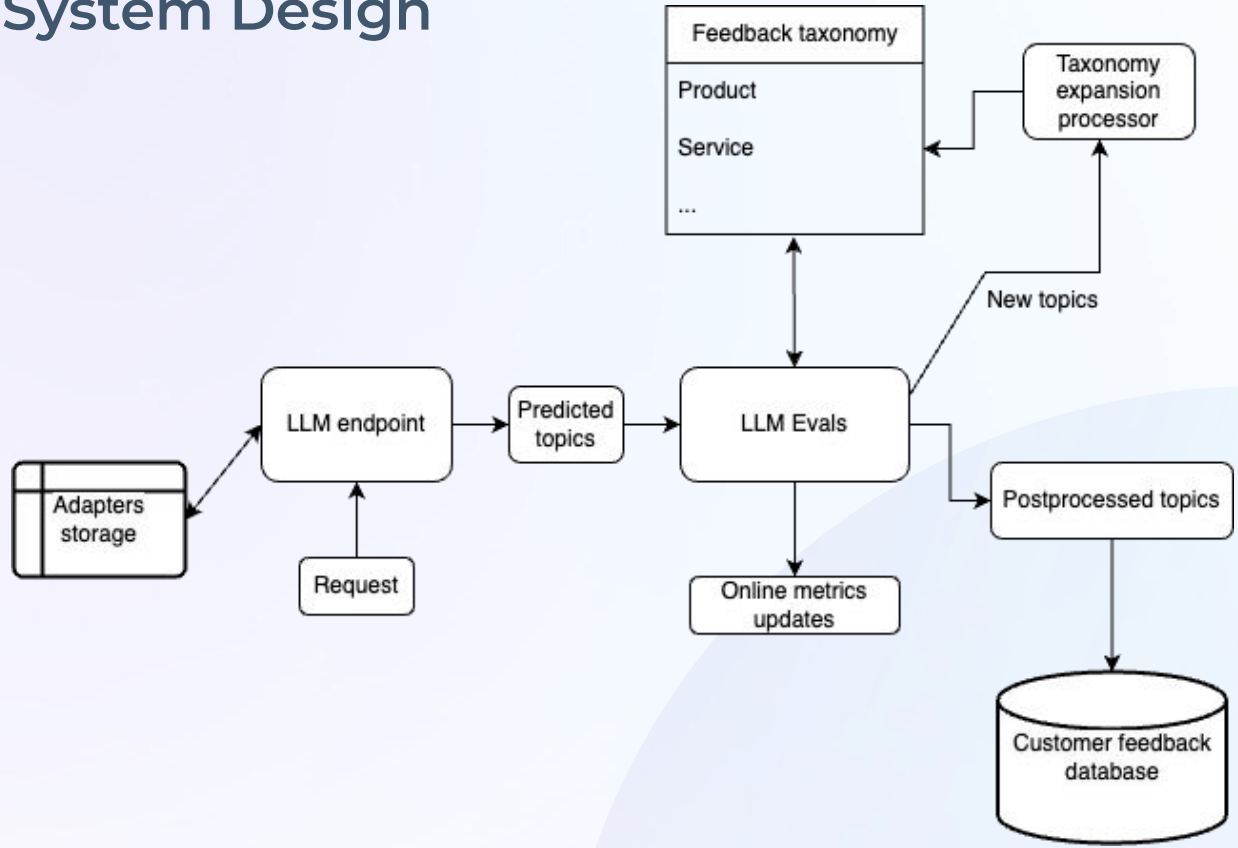
Dataset	Raw Data
new_topics	difficult
predefined_topics	generic data issues
predefined_topics	difficulty

Dataset	Raw Data
new_topics	damaged
new_topics	defect/
new_topics	missing
predefined_topics	conflicting/incorrect
predefined_topics	missing (information/context)
predefined_topics	wayfair-error
predefined_topics	changed mind/error
predefined_topics	didnt like it
predefined_topics	defect/damage
predefined_topics	missing parts
predefined_topics	mis-ship

Key Observations From LLM Evals

- Topics from the predefined taxonomy are predicted with a required precision
- LLMs are able to generate new topics (that are not present in the predefined taxonomy)
- Some of new topics may be unrelated, while others are accurate
- Most of new topics are semantically similar to the ones from the taxonomy
- LLM judges are capable of providing useful explanations of responses
- It makes sense to monitor topics tagged as unrelated
- Don’t ask the LLM judge to evaluate too many things at once

Expected System Design



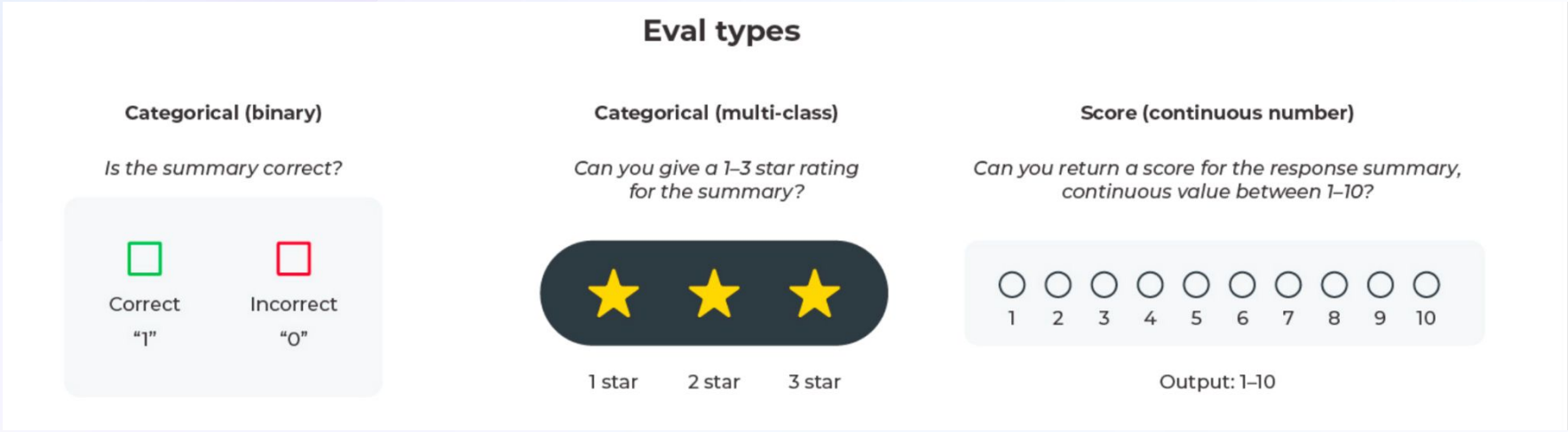
LLM Evals: findings, best practices, challenges

Model Evals vs Task Evals

Category	Foundation of Truth	Nature of Questions	Frequency and Purpose	Value of Explanations	Persona
Model Evals	Relies on benchmark datasets.	Involves a standardized set of questions, ensuring a broad evaluation of capabilities.	Conducted as a one-off test to grade general abilities, using established benchmarks.	Explanations don't typically add actionable value; focus is more on outcomes.	LLM Researcher
Task Evals	Relies on the golden dataset curated by internal experts and augmented with LLMs.	Utilizes unique, task-specific prompts, adaptable to various data scenarios, to mimic real-world scenarios.	An iterative process, applied repeatedly for system refinement and tuning, reflecting ongoing real-world applications.	Explanations provide actionable insights for improvements, focusing on understanding performance in specific contexts.	ML Practitioner

Source

Different Eval Options

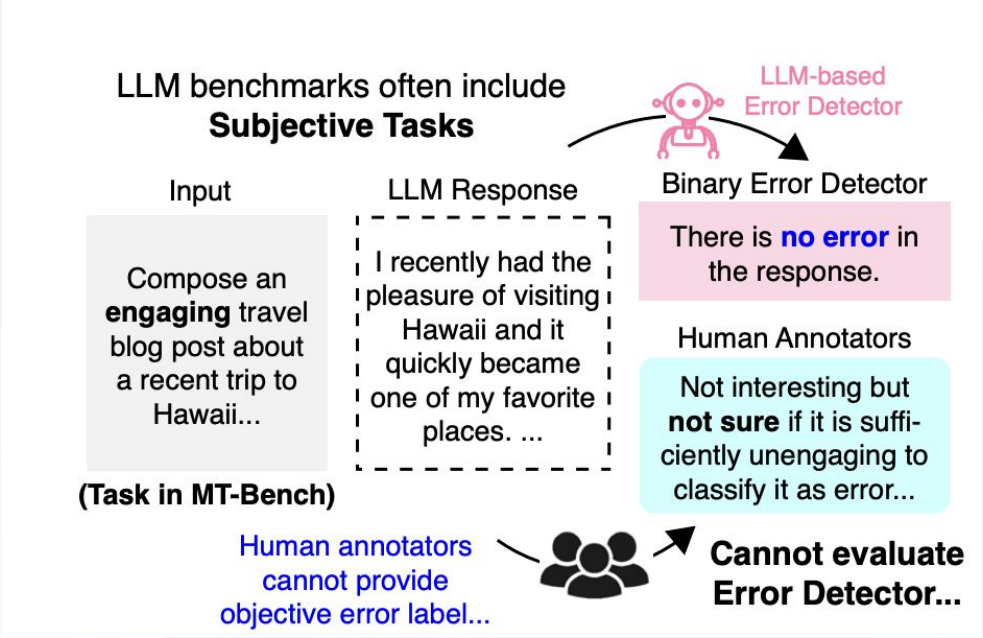


Research shows it is better to use labels over scores to evaluate your LLM.

Its ok to use a binary or multi-class label that is a number such as "1" or "0"

Why is it a Challenge to Detect an Error in LLMs in General?

Collecting error annotations on LLM responses is challenging due to the subjective nature of many NLP tasks



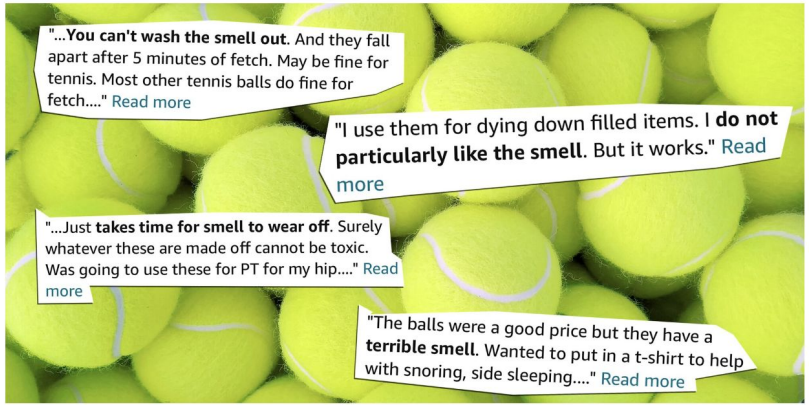
[Evaluating LLMs at Detecting Errors in LLM Responses]

Summarisation as an Example of a Subjective Task

Generated summary may look good but still may be incorrect

Amazon's AI Product Reviews Seen Exaggerating Negative Feedback

Review summaries created by generative artificial intelligence also sometimes mischaracterize products.



A tiny percentage of reviews saying Penn tennis balls smelled were represented in an AI-generated summary. *Photo illustration: 731; Source: Amazon, Getty Images*

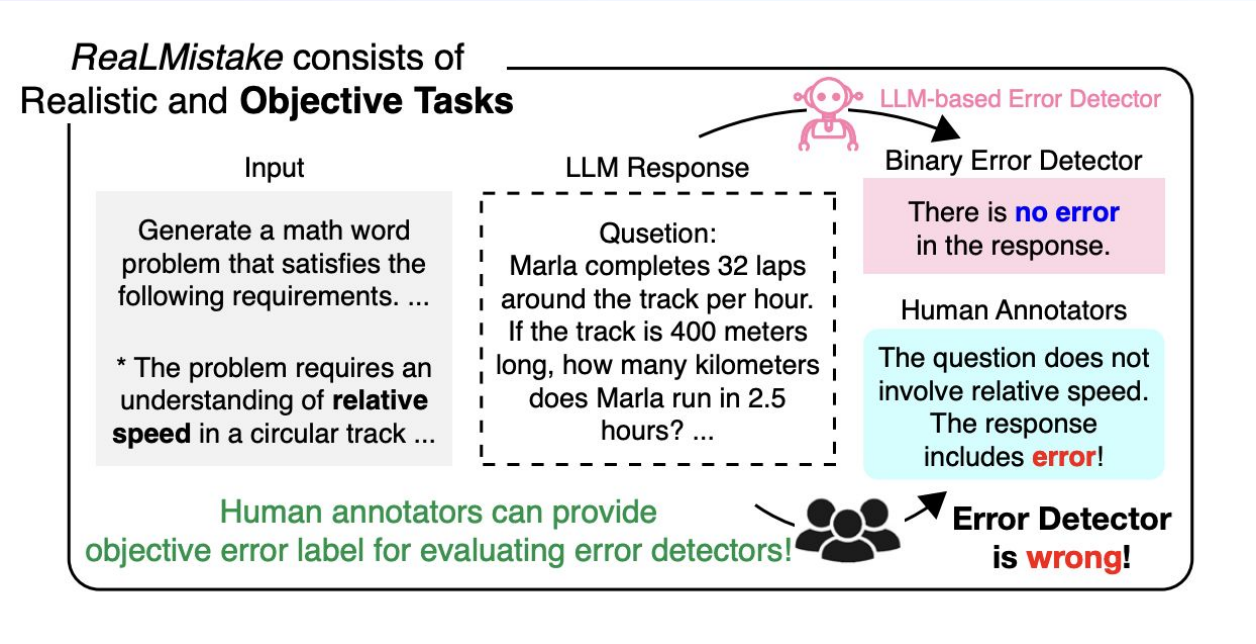
Amazon's AI Product Reviews Seen Exaggerating Negative Feedback

ReaLMistake benchmark

Paper focus: to create an evaluation benchmark for error detection from LLM responses

Main insights: top LLMs still detect errors with a low recall

.



Conclusions

- Domain-specific task evals can help to build a healthy LLM system
- Designing a set of useful evals for your task is a work of art :)
- Both research and industry are moving incredibly fast: we should expect even more from LLM evals
- There is no industry standards in LLM evals: be the one who makes it!

arize

Thank you!

Feel free to contact me:

ilyaboytsov1805@gmail.com
https://www.linkedin.com/in/ieboytsov