Working with LLMs

Stanford CS Talk
Feb 2024
Introductions

Shashaank Shekhar Singhal
Group Product Manager
Customer Obsession, Uber

Arindam Bhattacharya
TLM, Staff MLE
Customer Obsession, Uber
What Customer Obsession (CO) does

CO builds the **technology and products** to power the best in class customer care for gig economy

- Power millions of customer care interaction
- Deliver **efficient** care with **magical** customer experience
FAQs are the most common form of support self service

Usual Customer Challenges

1. Too long to read
2. Too many to scan to get to the exact answer
3. Not always easy to find the relevant ones
One of the most common use of LLMs in customer care is to answer customers questions using company FAQs

You
I am an Uber driver, the customer prefers a different route than the one suggested in the app, which one should I take?

ChatGPT
If a customer prefers a different route than the one suggested by the Uber app, it's generally best to honor their preference, as long as the following conditions are met:

1. **Safety**: The alternative route must be safe for both the driver and the passengers.
2. **Legality**: The route must comply with all local traffic laws and regulations.
3. **Feasibility**: The requested route should not significantly increase the trip duration or cost without the passenger's consent. If the alternate route will take considerably longer or could result in a higher fare, it's a good practice to inform the customer about these potential changes.

Customers may request a different route for various reasons, such as personal comfort, avoiding traffic jams not yet reflected in the app, or simply preferring a more scenic path. Accommodating such requests can lead to a more pleasant experience for the passenger, potentially resulting in higher ratings or tips for you.

Remember to communicate clearly with the customer about any concerns you might have regarding their preferred route, including potential delays or additional costs. This open communication ensures that both parties are aligned expectations about the trip.

Ultimately, the goal is to provide a service that meets or exceeds the customer's expectations.
But what the company might say could actually be different…

An LLM based AI Assistant using company’s internal knowledge

In private ‘Beta’ for some customers
But answers sometimes differ and can be challenging to know what is the right ans from the model

Assume the following questions* were asked

**Customer:** “The rider has a preferred route that is different than the route suggested on app, what should I do”

**AI Assistant:** “The app suggested route might not always be the most efficient or preferred one. If a rider suggests a route, you can take that to ensure a pleasant journey. However you should report the feedback for the route from the Help Section”

**Customer:** “I’d like to take a route that is different from the one suggested on the app, can I do this”

**AI Assistant:** “Unfortunately you are not allowed to take a route different from the preferred route. To ensure safety and satisfaction of our riders, it’s important to take the route suggested by the app. If you have any concerns about the route, please contact the rider via the app to discuss alternatives”

*Hypothetical scenario
LLM model outputs are reviewed by a team of human labellers to capture accuracy and other specific factors.

Team of **human labellers** who sample and evaluate on different parameters like:

1. Accuracy based on internal Knowledge / policy
2. Completeness
3. Brand adherence

and other factors....
Human labelling is helpful has limitations at scale to remain cost effective

- From 1,000s to .... 1,000,000+ interaction
- Extremely costly to maintain high samples rate

- Types of use cases start changing
  - Countries
  - Languages
  - Lines of Business

- Frequent updation to prompt to improve
- Impacts product Launch Velocity
Can LLMs/ AI help mitigate these challenges when operating at a very large scale
What not to say!

1. Evaluations.
   a. Arbitrator LLM.
   b. Manual-eval
2. LogProbs
Auto-Evaluations.

Need for auto-eval

External/internal factors Factors (chatbot/summarization)

- Distribution of conversations changes over time.
- 3P model capability changes (often without any information).
- Prompt iterations.
- Language/regions expansions.

Type of Evals

- Manual Evals
  - Expensive but is Gold Standard, we can get specific labelers with expertise.
- Auto Evals
  - Inexpensive in-terms of time and effort.
Metrics for evals

Google's eval criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follows instructions</td>
<td>To what extent does the model's response demonstrate an understanding of the instruction from the prompt?</td>
</tr>
<tr>
<td>Grounded</td>
<td>Does the response include only information from the inference context and inference instruction?</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>To what extent does the model capture key details in the summarization?</td>
</tr>
<tr>
<td>Brief</td>
<td>Is the summarization verbose? Does it include flowery language? Is it overly terse?</td>
</tr>
</tbody>
</table>

Summarization eval Metrics

We define 4 auto-evaluation metrics:

1. **Accuracy**: scores the summary based on whether it contains correct factual information like delivery details, cancellation fees, etc.
2. **Completeness**: scores the summary based on how well each segment of the conversation is represented in the summary.
3. **Conciseness**: scores the summary based on brevity.
4. **Consistency**: scores a summary based on its agreement with facts in the source conversation. This metric is intended to measure hallucinations, i.e. details generated by the model in the summary which are not present in the conversation log.
Eg Metric: Accuracy

a) Are any facts about the chat history wrong?
b) Was anything made up? Are there any false claims? For example, products that don’t exist, URLs that don’t work, or Uber policy info that is wrong.
c) Are there any reasoning mistakes?

Demo

CLASSIFICATION_PROMPT = """"You will be given a headline of a news article. Classify the article into one of the following categories: Technology, Politics, Sports, and Art. Return only the name of the category, and nothing else. MAKE SURE your output is one of the four categories stated.

Article headline: "Tennis Champion Showcases Hidden Talents in Symphony Orchestra Debut"
Google Ultra
Google Pro 1.5
Google Pro 1.0
OpenAI GPT4
Eg Metric: Accuracy

a) Are any facts about the chat history wrong?
b) Was anything made up? Are there any false claims? For example, products that don't exist, URLs that don't work, or Uber policy info that is wrong.
c) Are there any reasoning mistakes?

Demo

Given a classification task and the output, if you think the classification is correct return True else return False. Return a reason for your judgement.

```{  
You will be given a headline of a news article. Classify the article into one of the following categories: Technology, Politics, Sports, and Art. Return only the name of the category, and nothing else. MAKE SURE your output is one of the four categories stated.

Article headline: "Tennis Champion Showcases Hidden Talents in Symphony Orchestra Debut"
output : Arts
}
```
AutoSxS

Perform automatic side-by-side evaluation | Vertex AI | Google Cloud

At a high level, the diagram shows how AutoSxS compares the predictions of models A and B with a third model, the autorater.
Auto_eval
Log Probs - reduce hallucinations

Log probabilities of output tokens indicate the likelihood of each token occurring in the sequence given the context.

To simplify, a logprob is $\log(p)$, where $p = \text{probability of a token occurring at a specific position based on the previous tokens in the context}$. Some key points about logprobs:

Higher log probabilities suggest a higher likelihood of the token in that context. This allows users to gauge the model's confidence in its output or explore alternative responses the model considered.

Logprob can be any negative number or 0.0. 0.0 corresponds to 100% probability.
"logprobs": {
    "tokenLogProbs": [
        -0.002076329430565238, (98.0%)
        -8.3449995145201683e-06
    ],
    "tokens": [
        "▁Sports",
        "\u003ceos\u003e"
    ]
},
User
hi

Uber Order Taking Longer than expected support bot
Hello! How can I assist you today?

User
my order is an hour late give me back my money

Uber Order Taking Longer than expected support bot
I have checked the details, and your order is still on track to be delivered on time. It's currently being prepared at {tea corner} and is expected to be delivered by 5:00pm. A courier has already been assigned to your order. Please keep an eye on the app for updates, and thank you for your patience.

User
you are wrong it is an hour late, give me back my money

Uber Order Taking Longer than expected support bot
I understand this must be frustrating, but according to our records, your order is not late. It is expected to be delivered by {tea corner} between 5:00pm and 5:30pm, and we are currently before that window. A courier has been assigned to your order, indicating it is on its way. Unfortunately, if you choose to cancel now, the order will be fully charged as per our policy. Please let me know if there's anything else I can do to assist you.