

Unlocking User Insights with LLMs to Design Smarter UI and Flows

Amber Hoak

Microsoft Research





Goals

- Introduce Semantic Telemetry
- Application for AI/User Chat workflows
- Benefits and Costs
- Creating an LLM-based classifier with your data



Traditional Telemetry



Click through rate



Geography



Dwell time

Semantic Telemetry



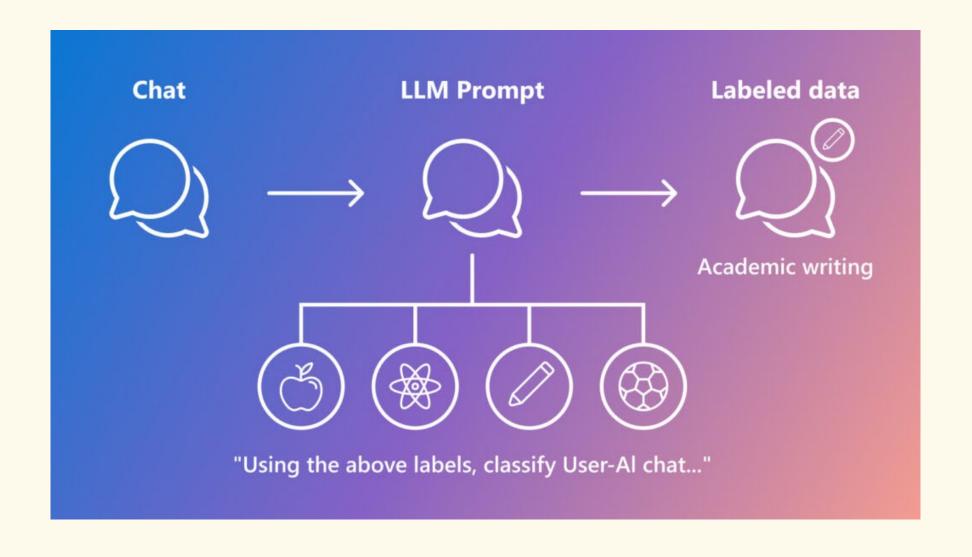


Topic





LLM-based Classifier





How can we apply this to the UI?

Travel Assistant

Ask about destinations, tour styles, seasons or planning phases — I'll tailor context & offers.

Ask anything travel...

No offers yet. Ask about destinations or trip styles to surface suggestions.

Travel Assistant

Ask about destinations, tour styles, seasons or planning phases - I'll tailor context & offers.

Labels:

- User Intent
- Booking Phase
- Type of Tour

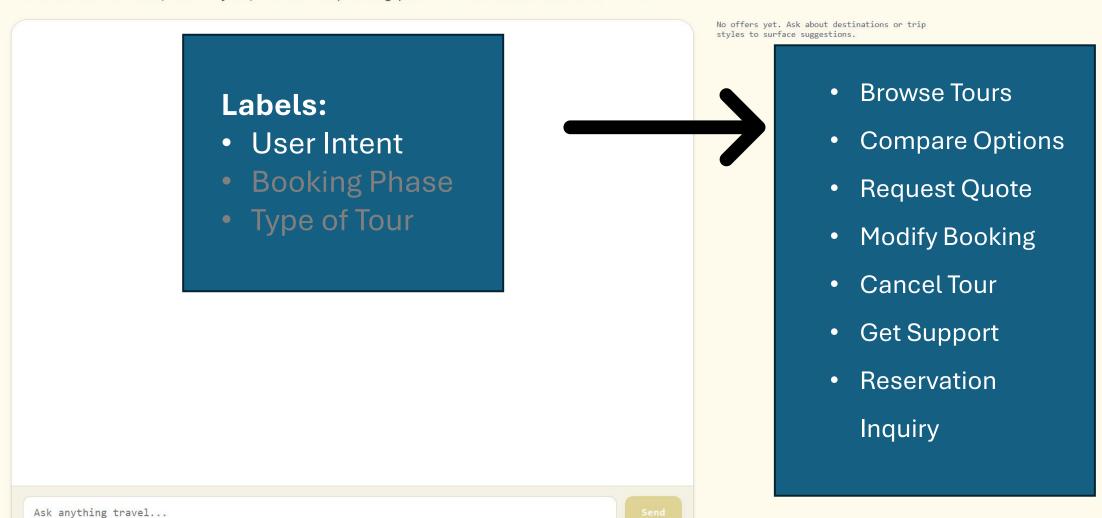
No offers yet. Ask about destinations or trip styles to surface suggestions.

Ask anything travel...

Send

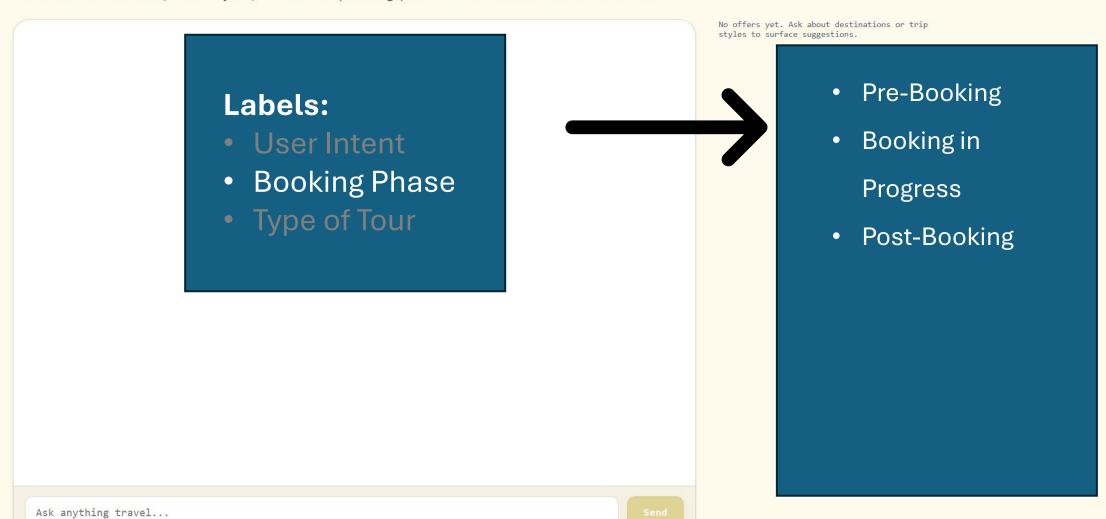


Travel Assistant



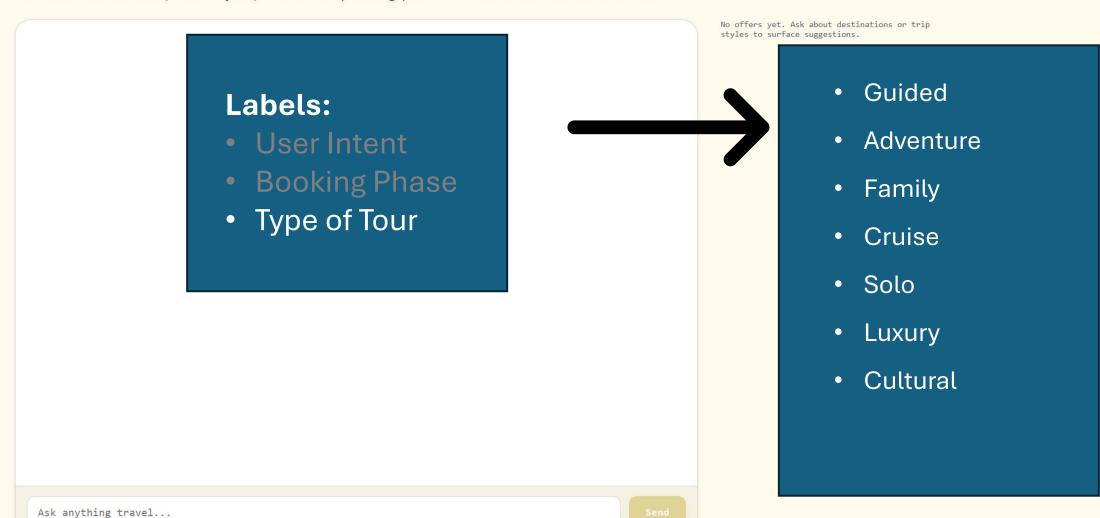


Travel Assistant



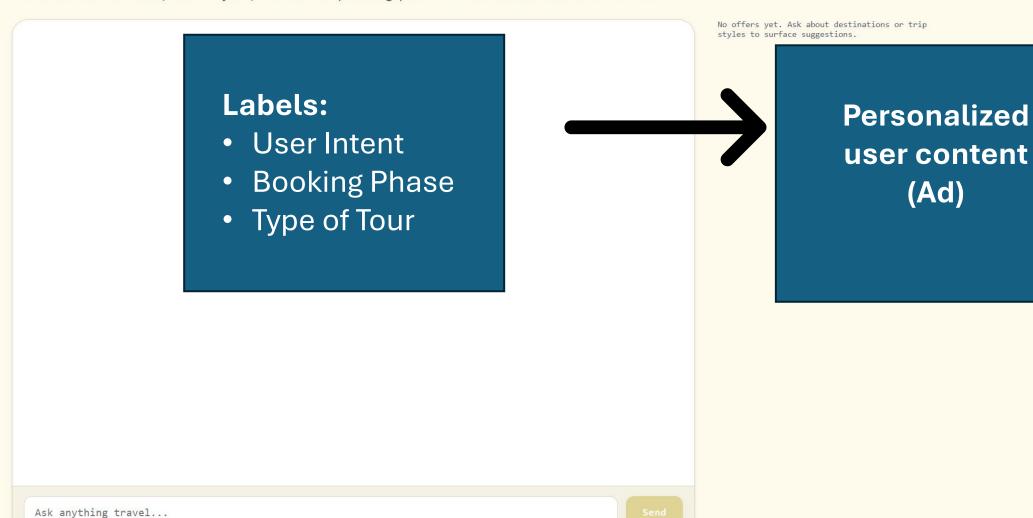


Travel Assistant

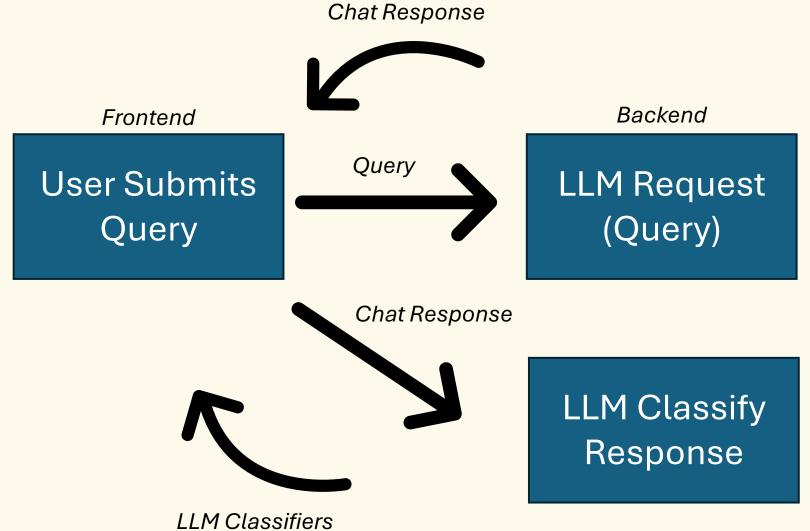




Travel Assistant



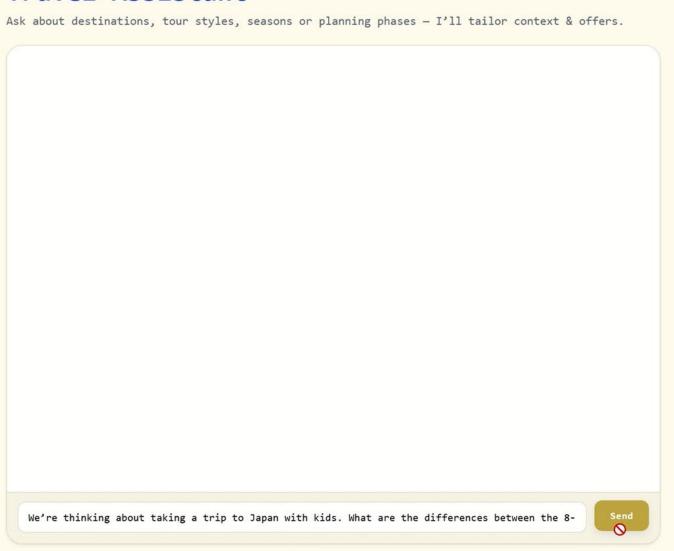




(Intent, Booking Phase, Tour Type)



Travel Assistant



No offers yet. Ask about destinations or trip styles to surface suggestions.



Travel Assistant

Ask about destinations, tour styles, seasons or planning phases - I'll tailor context & offers.

permeen rue o- and in-day roun obrions:

AI: Both tours to Japan are great options for families, but here are the key differences between the two:

- 1. **Duration:**
 - The "8-Day Cherry Blossom Season Tour" lasts 8 days.
- The "10-Day Southern Japan Explorer Tour" is actually an 8-day tour despite the name suggesting 10 days.
- 2. **Season:**
- The Cherry Blossom Season Tour is offered in **Spring**, ideal for experiencing Japan's famous cherry blossoms.
- The Southern Japan Explorer Tour is available in **Winter**, showcasing a different seasonal experience.
- 3. **Tour Types and Focus:**
- The 8-Day Cherry Blossom Tour is a **Guided Tour, Family Tour, and Cultural Tour**, focusing heavily on cultural experiences and family-friendly activities centered around historic sites during cherry blossom season.
- The 10-Day Southern Japan Explorer Tour is a **Solo Traveler Tour, Luxury Tour, and Family Tour**, combining guided tours with opportunities for solo exploration, suitable for those wanting a mix of structure and freedom with a luxury touch.
- 4. **Price:**
 - Both tours are priced at around **\$2800 USD**.

If traveling with kids and wanting to experience Japan's iconic cherry blossom season with lots of cultural activities suited for families, the 8-Day Cherry Blossom Season Tour would be perfect. If you prefer a winter trip with some luxury and a balance of guided and solo experiences, the Southern Japan Explorer Tour could be a better fit.

started making a reservation for this trip, but the payment never went through. Tan you help?

Send

Recommended Offers



Cultural Destination kids will love in Japan

Get digital guidebooks for top cultural destinations.



Benefits & Costs

- Understand user interaction & modify user flows
 - Individual case-Personalized content
 - Aggregate case-Identify unsuccessful user flows
- Cost per LLM call
 - Bundle classifiers into single call
 - LLM classification on client





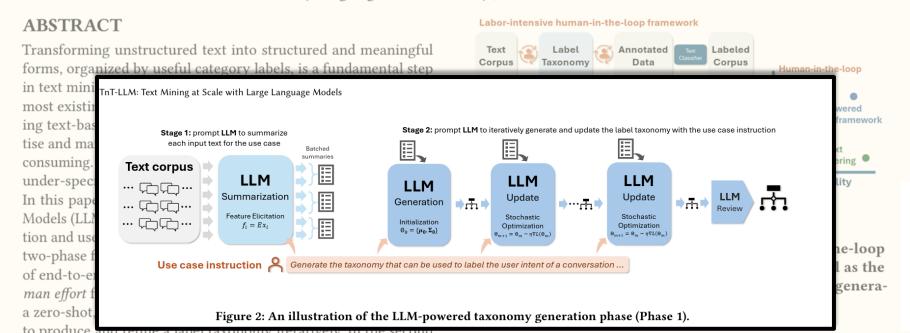
TnT-LLM Approach

TnT-LLM: Text Mining at Scale with Large Language Models

Mengting Wan*[‡], Tara Safavi*[‡], Sujay Kumar Jauhar*, Yujin Kim*, Scott Counts*, Jennifer Neville*, Siddharth Suri*, Chirag Shah[†], Ryen W. White*, Longqi Yang*, Reid Andersen*, Georg Buscher*, Dhruv Joshi*, Nagu Rangan*

*Microsoft Corporation, †University of Washington

{mengting.wan,tarasafavi}@microsoft.com



phase, LLMs are used as data labelers that yield training samples

Link to
Publication









Text Corpus

Summaries



"Substitution for oil in recipe"



"Insert a text box in PDF"



"Comparing noise-canceling headphones"



"Getting started with unicycling"

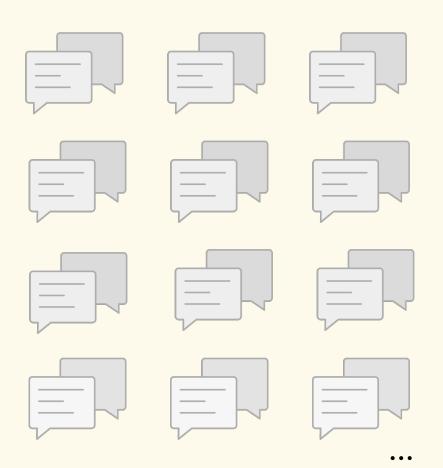
```
export function generateSummarizationPrompt(
          data: string,
          useCase: string,
          summaryLength: number,
 94
          customOutputFormat: string = '',
        ): string {
 96
          return `
                                                                                                      Summarize based
                # GOAL:
 98
                Summarize the input text for the given use case.
 99
                                                                                                      on use case
                You input is a conversation history between a User and an AI agent.
100
                use case is: ${useCase}
101
                 # Data:
102
                ${data}
103
                                                Chat Data
104
                # Requirements:
                -Provide a summary of input text **in ${summaryLength} words or less** that captures the use case.
106
                -The summary will represent the input data for clustering in the next step.
107
                # Tips
108
                -The summary will represent the input data for clustering in the next step.
109
                -Be concise and clear.
110
                -Do not add phrases like "This is the summary of .. " or "Summary: " or "Here is a summary of the conversation".
111
                -Within ${summaryLength} words, include the relevant infromation for the use case in the summary as possible.
112
                -Do not include any introductory or concluding remarks.
113
                -Do not include any line breaks in the summary.
114
                -Provide your answer in **English** only
115
116
                ${customOutputFormat}
117
118
```

119

Requirements



Batches





Batches "Your goal is to classify the input data into meaningful categories for the given use case"





Large Language Model

(i.e. GPT-4o-mini, etc.)



Batches "Your goal is to classify the input data into meaningful categories for the given use case"





Large Language Model

(i.e. GPT-4o-mini, etc.)



Seed List



Electronics



Sports



Programming

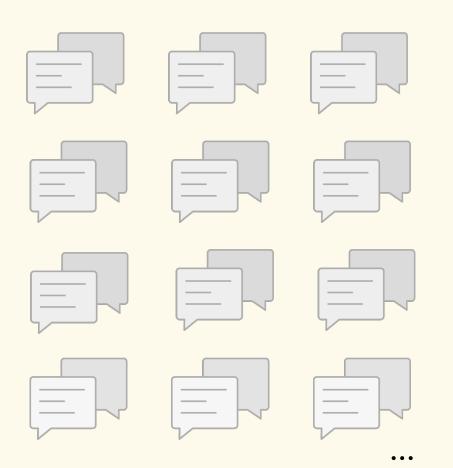


Travel & tourism





Batches



Seed List



Electronics



Sports



Programming

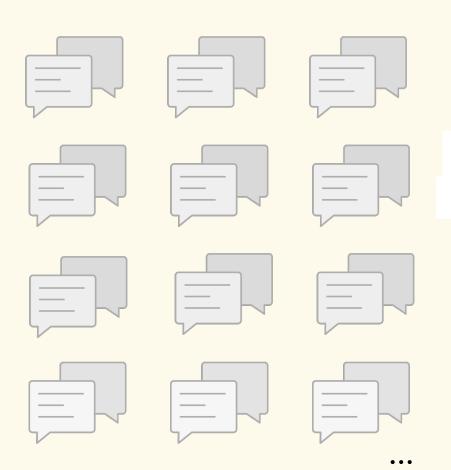


Travel & tourism





Batches







(i.e. GPT-4o-mini, etc.)

Seed List



Electronics



Sports



Programming

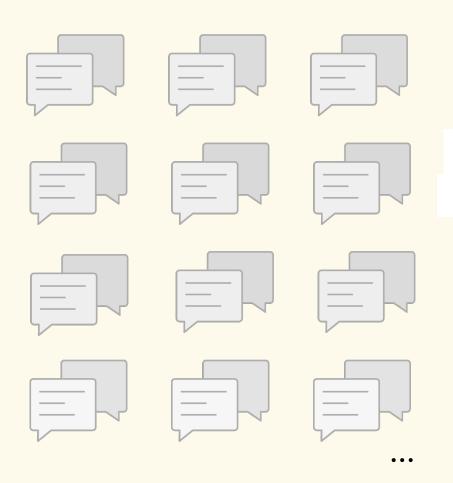


Travel & tourism





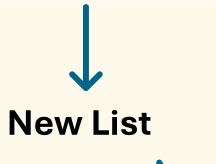
Batches





Large Language Model

(i.e. GPT-4o-mini, etc.)





Seed List



Electronics



Sports



Programming

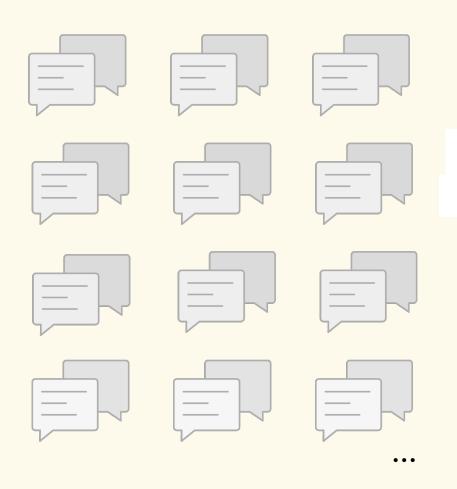


Travel & tourism





Batches







Large Language Model

(i.e. GPT-4o-mini, etc.)





Updated List



Electronics



Sports



Programming



Travel & tourism



Shopping



Education

```
clusters: ClusterTable | InitialClusterTable | string,
 data: ChatRecord[],
 maxClusters: number,
 useCase: string,
 clusterNameLength: number,
                                                                                                                                                    Define goal with
 suggestionLimit: number,
 customOutputFormat: string = '',
): string {
                                                                                                                                                    use case
       # Goal:
       Your goal is to review the given reference table based on the input data for the specified use case, then update the reference table if needed.
         -You will be given a reference cluster table ,which is built on existing data. The reference table will be used to classify new data points.
         -You will compare the input data with the reference table, output a rating score of the quality of the reference table, suggest potential edits, and update the reference table if needed.
       -**Reference cluster table**: The input cluster table is a markdown table with the following fields:
           -id: category identifier
           -name: category name
           -description: category description
       -**Data**: The input data is a markdown table with summaries for a list of human-AI conversations, including the following fields:
           -id: conversation identifier
           -summary: conversation summary
       Use case: ${useCase}
       # Requirements:
       -name: category name should be within ${clusterNameLength} words** It can be either a verb phrase or a noun phrase, whichever is more apropriate.
       -Total number of categories should be **no more than ${maxClusters}**.
       ## Quality
       - **No overlap or contradiction** among categories.
       -Name is concise and clear for the category. Use only phrases that are specific to each category and avoid phrases common to all categories.
       -**Description** differentiates the category from other categories.
       -**Name** and **description** can **accurately
       ** and **consistently** classify new data points **without ambiguity**.
       -**Name** and **description** are consistant with each other.
       -Output clusters match the data as closely as possible, without missing important categories or adding unnecessary ones
       -Output clusters should serve given use case well.
       -Output clusters should be specific and meaningful. Do not invent categories that are not in the data.
       # Reference cluster table:
                                                                Cluster List
       ${clusters}
       # Data:
       ${summariesToMarkdown(data)}
                                                                Batched Summaries
```

export function generateClusterUpdatePrompt(

\${customOutputFormat}

Requirements



CASCADIAJS 2025

```
// Step 4: Iteratively update cluster list with each batch
130
        for (const batch of batched.slice(1)) {
131
          if (!updatedClusterList) {
            console.error('No clusters available for assignment.');
132
133
            break:
134
135
136
          const updatePrompt = generateClusterUpdatePrompt(
137
            updatedClusterList,
138
            batch,
139
            MAX_CLUSTERS,
140
            USE CASE,
141
            CLUSTER NAME LENGTH,
142
            SUGGESTION_LIMIT,
143
          );
144
        const updateResponse = await LLMClient(updatePrompt, UpdatedClusterListSchema, { failover: true });
145
146
          const parsed = safeJSONParse<ClusterTable>(updateResponse);
         if (parsed.ok) {
147
148
            updatedClusterList = parsed.value;
149
         } else {
150
            console.error(`Failed to parse initial cluster JSON: ${parsed.error.message}`);
151
```

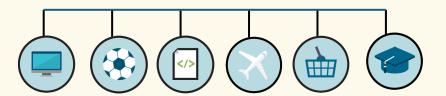


CASCADIAJS 2025

```
// Step 4: Iteratively update cluster list with each batch
130
        for (const batch of batched.slice(1)) {
131
          if (!updatedClusterList) {
            console.error('No clusters available for assignment.');
132
133
            break:
134
135
136
          const updatePrompt = generateClusterUpdatePrompt(
137
            updatedClusterList,
            batch,
138
139
            MAX CLUSTERS,
140
            USE CASE,
141
            CLUSTER NAME LENGTH,
142
            SUGGESTION_LIMIT,
143
          );
144
        const updateResponse = await LLMClient(updatePrompt, UpdatedClusterListSchema, { failover: true });
145
146
          const parsed = safeJSONParse<ClusterTable>(updateResponse);
147
          if (parsed.ok) {
148
            updatedClusterList = parsed.value;
149
          } else {
150
            console.error(`Failed to parse initial cluster JSON: ${parsed.error.message}`);
151
```



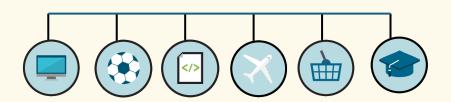
Final label taxonomy





Classifying Data

Final label taxonomy



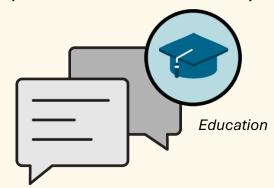


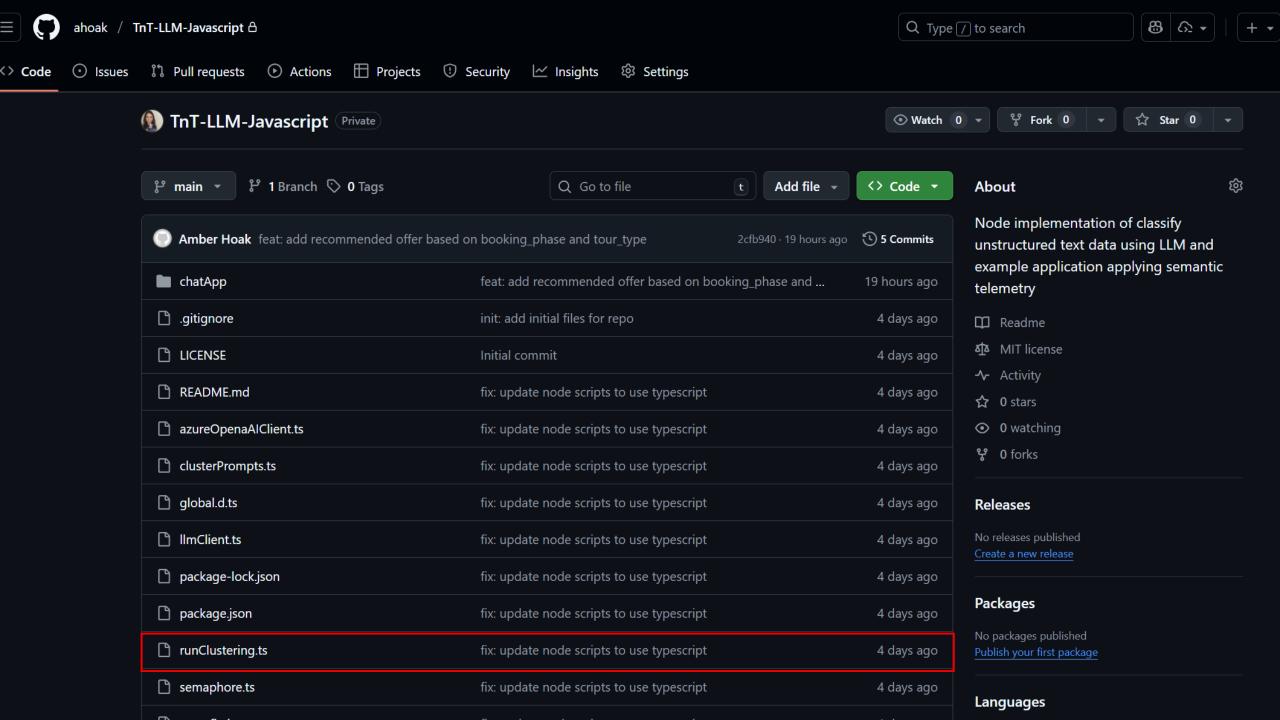
New Chat Data



Large Language Model

(i.e. GPT-4o-mini, etc.)







MSR Blog posts

Semantic Telemetry: Understanding how users interact with Al systems

Published March 10, 2025

By Amber Hoak, Senior Software Development Engineer; Scott Counts, Senior Principal Research Manager; Kate Lytvynets, Software Development Engineer; <u>David Tittsworth</u>, Software Engineer; <u>Siddharth Suri</u>, Senior Principal Researcher; <u>Ben Cutler</u>, Senior Director;

Share this page f X in 🚳 🔊

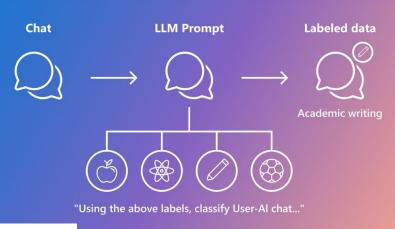








Weiwei Yang, Principal SDE Manager





https://www.microsoft.com/enus/research/blog/semantic-telemetryunderstanding-how-users-interact-with-aisystems/

Engagement, user expertise, and satisfaction: Key insights from the **Semantic Telemetry Project**

Published April 14, 2025

By Amber Hoak, Senior Software Development Engineer; Scott Counts, Senior Principal Research Manager; Kate Lytvynets, Software Development Engineer; David Tittsworth, Software Engineer; Siddharth Suri, Senior Principal Researcher; Nirupama Chandrasekaran, Principal Research Software Engineer; Ben Cutler, Senior Director; Weiwei Yang, Principal SDE Manager

Share this page **f** X **in 6**





https://www.microsoft.com/enus/research/blog/engagement-userexpertise-and-satisfaction-key-insightsfrom-the-semantic-telemetry-project/

Technical approach for classifying human-Al interactions at scale

Published July 23, 2025

By Amber Hoak, Senior Software Development Engineer; <u>David Tittsworth</u>, Software Engineer; <u>Kate Lytvynets</u>, Software Development Engineer; Scott Counts, Senior Principal Research Manager; Weiwei Yang, Principal SDE Manager; Ben Cutler, Senior Director; Jonathan McLean, Senior Software Engineer

Share this page **f** X in **6**















https://www.microsoft.com/enus/research/blog/technical-approach-forclassifying-human-ai-interactions-at-scale/



Thank you!



https://github.com/ahoak/TnT-LLM-Javascript



https://www.linkedin.com/in/amberhoak/



@amber_hoak

