Signal in the Noise: Decoding Airline Service Quality with Large Language Models

WP No. 250

September 2026

This working paper is authored by Ahmed Dawoud, Lead Economist at ECES, and Osama El Shamy, Data Scientist at ECES. Ahmed Habashy, Data Scientist at ECES

Signal in the Noise: Decoding Airline Service Quality with Large Language Models

Osama El-Shamy, Ahmed Habashy, Ahmed Dawoud,

March 6, 2026

Abstract

Traditional measures of airline service quality provide useful indicators of performance, but they offer limited insight into why passengers are dissatisfied. This study uses a Large Language Model (LLM) to analyze unstructured passenger reviews and identify the specific service problems underlying changes in perceived service quality. We analyze 16,622 TripAdvisor reviews of two anonymized airlines, referred to as Airline A and Airline B, collected between 2016 and 2025 and covering 13 languages. The analysis classifies passenger complaints into 36 specific issues grouped into eight broader service categories. The results show sharply different rating trajectories: Airline A's average rating fell from 3.3 in 2019 to 1.6 in 2024, while Airline B remained comparatively stable. The issue analysis provides a clearer explanation of these differences. Poor communication during delays is mentioned 536 times, compared with 690 mentions of flight delays and cancellations, while rude flight attendants are the most frequently reported customer-service complaint, with 591 mentions. The results also show substantial differences across passenger market groups. Overall, the findings suggest that broad service-quality measures can conceal the specific experiences that shape passenger evaluations. LLM-based analysis can complement conventional approaches by organizing large volumes of multilingual, unsolicited passenger feedback and showing which service problems appear most often and how they change over time.

1 Introduction

Airline service quality matters because it affects how passengers evaluate an airline and whether they are likely to use it again. Airlines have traditionally assessed service through operational measures and structured passenger surveys based on models such as SERVQUAL. These measures are useful, but they provide only a broad picture of the passenger experience. Operational data can show that a flight was delayed, for example, but not how passengers experienced the delay. Surveys face a related limitation: respondents are asked to evaluate predefined dimensions, which can make it difficult to capture problems that were not anticipated when the questionnaire was designed.

Online reviews provide a different source of evidence. Passengers describe what happened in their own words, often including details that would not appear in a structured survey response. TripAdvisor therefore provides a large source of unsolicited feedback about specific aspects of the travel experience. The difficulty is that these reviews are unstructured, multilingual, and often expressed in informal language. Earlier approaches to analyzing online reviews have been useful, but they can struggle to distinguish between similar complaints or to work consistently across languages (Brochado et al., 2019).

Large Language Models provide one possible way to address these limitations. They can interpret context, identify topics and sentiment, and recognize different ways of expressing the same complaint. This makes it possible to organize large collections of reviews without reducing them to simple keyword matches or fixed survey categories (Haleem et al., 2022).

This paper uses an LLM-based approach to examine what passengers report in their reviews. Rather than focusing only on overall ratings, the study identifies the specific issues behind those ratings and examines how they vary over time and across passenger market groups. To demonstrate the approach, we compare two anonymized airlines, referred to throughout the paper as Airline A and Airline B.

2 Objectives

The main objective is to assess whether an LLM can organize large volumes of multilingual passenger reviews into a form that can be analyzed systematically. The study focuses on three questions:

1. Methodological Assessment: To assess whether an LLM can identify specific service issues and sentiment patterns in raw passenger reviews at a scale that would be difficult to handle manually.

2. Identifying Service Problems: To identify the specific factors behind passenger dissatisfaction that may be difficult to see in broad service-quality measures.

3. Practical Application: To examine how this type of analysis can help identify recurring service problems and differences across passenger markets.

3 Literature Review

Research on airline service quality has largely relied on quantitative surveys using established frameworks such as SERVQUAL and related airline-specific measures. These studies typically use structured questionnaires to evaluate dimensions such as tangibles, reliability, responsiveness, assurance, and empathy. Statistical analysis is then used to examine the relationship between these dimensions and outcomes such as passenger satisfaction or loyalty. This approach provides comparable measures of service quality, but it also depends on the dimensions included in the questionnaire.

Survey-based methods have several limitations when the aim is to understand the details of passenger experience. First, predefined categories can overlook problems that were not anticipated when the survey was designed. Second, numerical ratings provide limited information about the circumstances that produced a particular score. Third, survey responses may be affected by respondent characteristics, survey fatigue, and the way questions are framed. Finally, averaging responses across broad dimensions can hide specific problems that passengers repeatedly describe in their own words.

Some researchers have therefore turned directly to passenger text. Traditional content analysis, however, usually depends on manual coding and predefined categories, which makes it difficult to apply consistently to large, multilingual datasets such as those generated by international airlines (Kirilenko et al., 2018). It can also blur the distinction between related complaints—for example, a delay itself and the lack of communication about that delay. Multilingual reviews may also require translation before they can be analyzed.

These limitations make it difficult to capture the detail contained in large collections of online passenger feedback. An LLM offers one way to address this problem by working directly with the language used in the reviews and organizing different descriptions of similar problems into common categories. The present study builds on this approach to examine a large multilingual collection of passenger reviews (Haleem et al., 2022).

4 Methodology

4.1 Data Acquisition

We used Python-based web scraping to collect passenger reviews from TripAdvisor between Q1 2016 and Q1 2025. The final dataset contains 16,622 reviews: 5,171 for Airline A and 11,451 for Airline B. The reviews cover 13 languages and provide the text used in the analysis.

4.2 LLM-Based Analysis

We developed a multi-stage pipeline informed by the Clio framework (Tamkin et al., 2024). Rather than relying on keywords alone, the pipeline uses the meaning and context of each review to group different descriptions of similar problems into a common set of categories.

Stage 1: Filtering Negative Reviews

To focus on negative experiences, we first selected reviews with ratings from 1 to 3 stars. For Airline A, this produced 2,857 negative reviews containing direct expressions of dissatisfaction. These reviews were then used to examine the reasons behind the low ratings.

Stage 2: Issue Extraction and Standardization

The LLM processed the reviews in their original languages rather than translating them first. Using structured prompts and contextual information, it assigned each complaint to one of 36 specific issue labels. For example, descriptions such as 'rude staff' and 'bad attitude' were grouped under Customer Service. Table 1 shows how several complaints within a single review can be separated and classified.

Stage 3: Grouping Service Issues

For analysis, the 36 specific issues were grouped into eight broader categories. When a review contained several complaints, the LLM first identified the relevant text for each complaint and then assigned the complaint to the appropriate category. The eight categories are:

Category	Description
Airport Services	Pre/post-flight processes (check-in, security, boarding).
Baggage Handling	Lost, delayed, or damaged luggage issues.
Booking Issues	Reservations, cancellations, website errors, and ticketing.
Cleanliness	Hygiene standards within the cabin and airport facilities.
Customer Service	Staff behavior, responsiveness, and interaction quality.
Flight Disruptions	Schedule adherence (delays, cancellations, connections).
In-Flight Experience	Onboard hardware, catering, entertainment, and comfort.
Safety Concerns	Perceived risks or violations of safety protocols.

Table 1: Service issue categories used in the analysis

Table 2: Example of Unstructured Text to Structured Data Conversion

Raw Passenger Review	Extracted Issue Category
"Finally arrived at the destination. The flight was delayed by 3 hours with zero updates from the gate agents. Once onboard, the seat would not recline, and the food was completely cold."	1. Poor Communication (Delay) — Flight Disruptions 2. Broken Seats — In-Flight Experience 3. Poor Food Quality — In-Flight Experience

5 Findings: What Passengers Report

This section examines the review ratings and complaint patterns for both airlines, with particular attention to the period after the pandemic.

5.1 Diverging Passenger Ratings

The number of reviews increased again after the pandemic for both airlines (Figure 1), but the ratings followed noticeably different trajectories between 2019 and 2025.

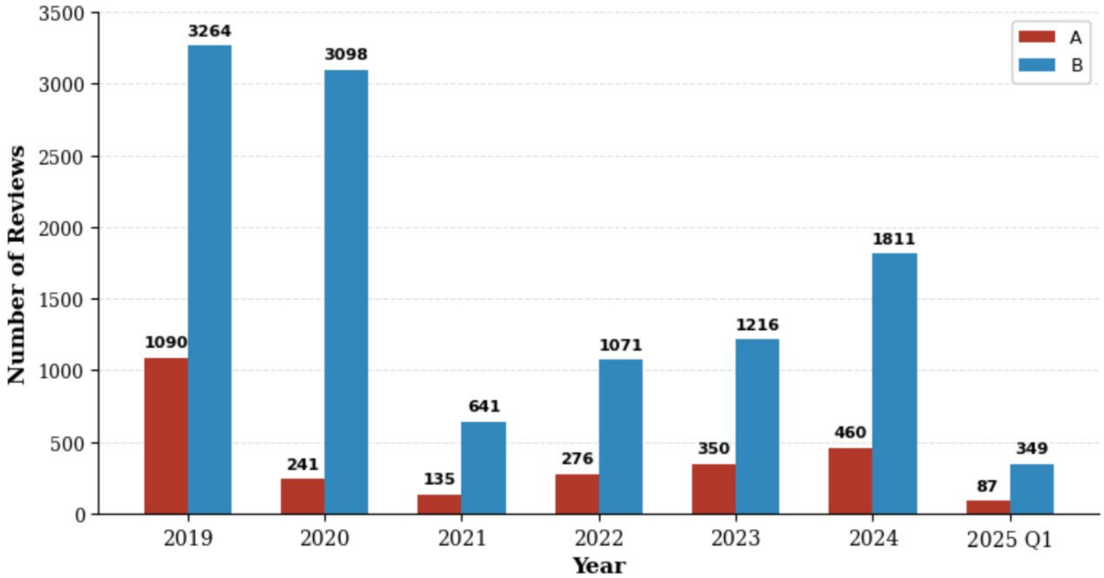

Figure 1: Review Frequency Comparison (2019-2025 Q1)

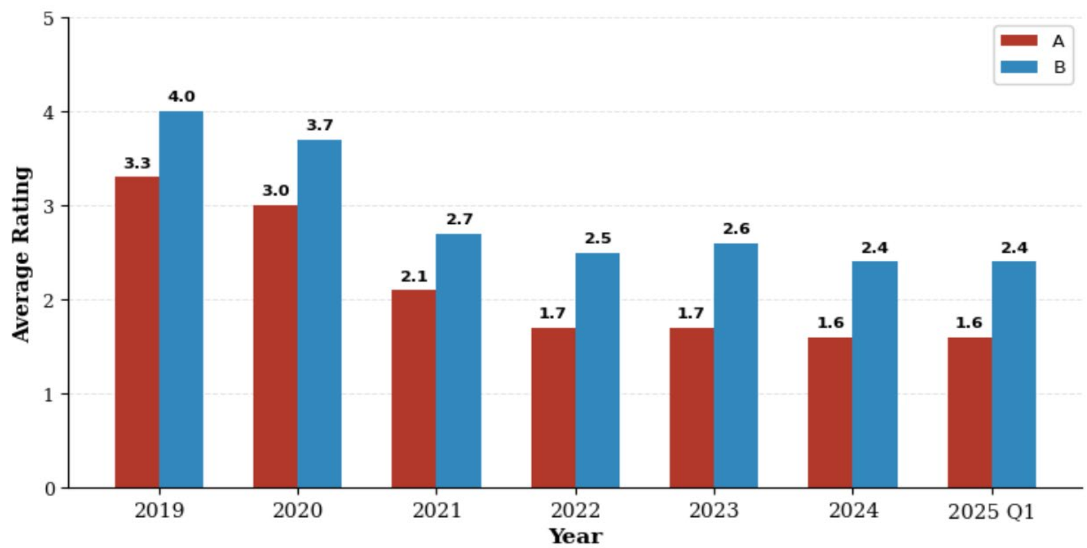

Figure 2: Average Rating Trajectory (2019-2025 Q1)

Airline B remained relatively stable during the recovery period, with average ratings above 3.5/5 in most years. Airline A followed a different pattern. Its average rating was 3.3 in 2019, fell below 2.0 from 2022 onward, and reached about 1.6 in 2024 (Figure 2). The size and persistence of this decline suggest that dissatisfaction extends beyond a small number of isolated complaints.

Looking at individual service categories gives a clearer picture of where these ratings are coming from. For Airline A, complaints are spread across several parts of the passenger experience rather than being concentrated in one area. Customer Service and Value for Money are particularly important detractors, while categories such as Legroom and Cleanliness also moved toward poor ratings by 2025. Airline B also received complaints about areas such as In-Flight Entertainment and Food, but its ratings remained comparatively stable and negative reviews were more concentrated around a smaller number of issues.

5.2 Differences Across Passenger Markets

Breaking the ratings down by passenger origin highlights substantial differences across market groups (Figures 3 and 4). To protect the identities of the airlines and their specific markets, the regional groups are reported using anonymized market-group labels.

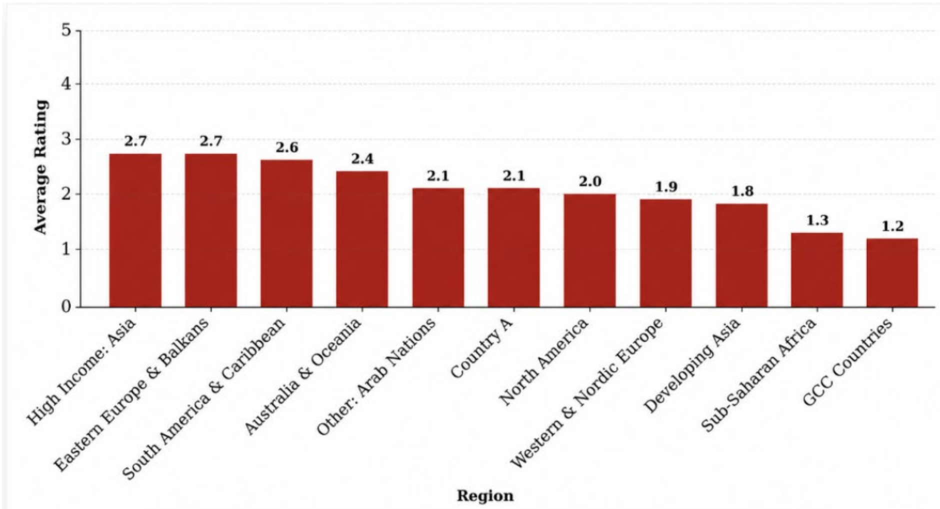

Figure 3: Airline A Average Rating by Origin Region (2024-2025 Q1)

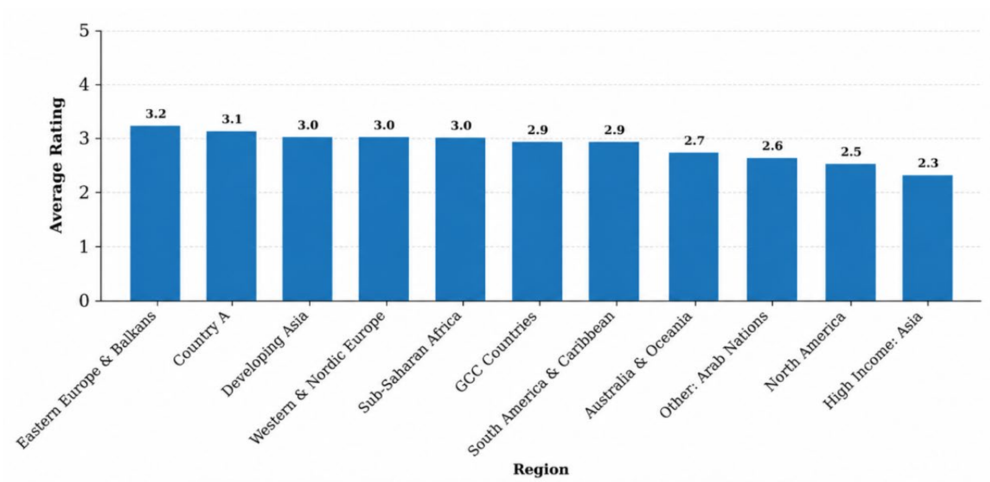

Figure 4: Airline B Average Rating by Origin Region (2024-2025 Q1)

Airline A shows a much wider spread in ratings across market groups, with average ratings ranging from 1.2 to 2.7. Airline B is more consistent, with ratings ranging from 2.3 to 3.2. The difference suggests that passenger experiences vary considerably across markets for Airline A, while Airline B shows a narrower distribution.

5.3 What Drives Dissatisfaction

The 36 issue labels allow us to look beyond the overall score and identify the complaints that appear most often. They also show how the mix of complaints changed between the pre-pandemic and post-pandemic periods.

1. Flight Disruptions and Communication: ‘Poor Communication Regarding Delay’ appears 536 times, close to the 690 mentions of flight delays or cancellations. This suggests that passengers are reacting not only to the disruption itself, but also to the lack of clear information about what is happening.

2. Staff Behavior: ‘Rude Flight Attendants’ is the most frequently mentioned specific complaint within Customer Service, with 591 mentions. The reviews describe interactions involving shouting, dismissiveness, and, in some cases, perceived discrimination.

3. Tangible Service Problems: ‘Poor Food Quality’ and ‘Delayed Baggage’ also appear frequently. The continued presence of baggage-delay complaints suggests that getting luggage to passengers quickly remains a problem even when bags are not ultimately lost.

The temporal pattern in the complaint data shows a sharp decline in complaint volume during 2020–2021, followed by a strong increase as air travel recovered. The composition of complaints also changed. Before the pandemic, complaints were spread more widely across areas such as In-Flight Experience. After 2022, Flight Disruptions and Customer Service account for a larger share of the complaints. This pattern suggests that communication and staff interactions became particularly important sources of dissatisfaction during the recovery period.

Table 3: Frequency of Service Issues by Category (2016–2025)

Main Category	Specific Issue	Count
Flight Disruptions	Flight Delays/Cancellations	690
	Poor Communication Regarding Delay	536
	Unexplained Cancellation	94
	Missed Connection	92
	Excessive Flight Delay	73
	Inadequate Pre-Flight Communication	24
	Unclear Announcements	1
Customer Service	Rude Flight Attendants	591
	Poor Customer Service	475
	Unresponsive Crew	346
	Lack of Assistance	194
	Unhelpful Phone Support	92
In-Flight Experience	Poor Food Quality	555
	Lack of Amenities	442
	Uncomfortable Seating	288
	Seat Assignment Problems	131
	Broken Seats	107
	Broken Entertainment System	81
	In-Flight Experience (General)	66

	Lack of Legroom	12
	Seat Issues	5
	Poor Entertainment	1
Baggage Handling	Lost Baggage	335
	Damaged Baggage	111
	Delayed Baggage	41
	Baggage Handling Fees	28
	Baggage Handling (General)	2
Cleanliness	Dirty Cabin	280
	Unclean Restrooms	102
	Cleanliness (General)	2
Safety Concerns	Lack of Safety Enforcement	130
Airport Services	Disorganized Boarding	72
	Disorganized Airport Staff	44
Booking Issues	Difficult Booking Process	46
	Website Issues	37

Table 4: Comparison with Earlier Research

Feature	This Study	Earlier Literature
Method	Analysis of unsolicited multilingual TripAdvisor reviews using an LLM.	Primarily surveys and manual content analysis.
Service issues	Customer service, flight disruptions, in-flight comfort, baggage, cleanliness, and other specific complaints.	Staff attitude, reliability, responsiveness, and tangible service dimensions.
Level of detail	Separates specific complaints, such as delays and communication about delays, into 36 issue types.	Generally examines broader service dimensions.
Coverage	Examines patterns across years, service categories, and passenger market groups.	Generally based on smaller structured samples.
Geographical analysis	Compares passenger ratings across market groups.	Limited geographical segmentation.
Scale	16,622 reviews across 13 languages.	Typically smaller survey samples.

Figure 5 shows how the main complaint categories changed over time. Total complaint volume fell sharply during 2020–2021 and then rose again as air travel recovered. The composition of complaints also changed. Before the pandemic, complaints were spread more widely across areas such as In-Flight Experience. After 2022, Flight Disruptions and Customer Service account for a larger share of the complaints. The pattern suggests that communication and staff interactions became particularly important sources of dissatisfaction during the recovery period.

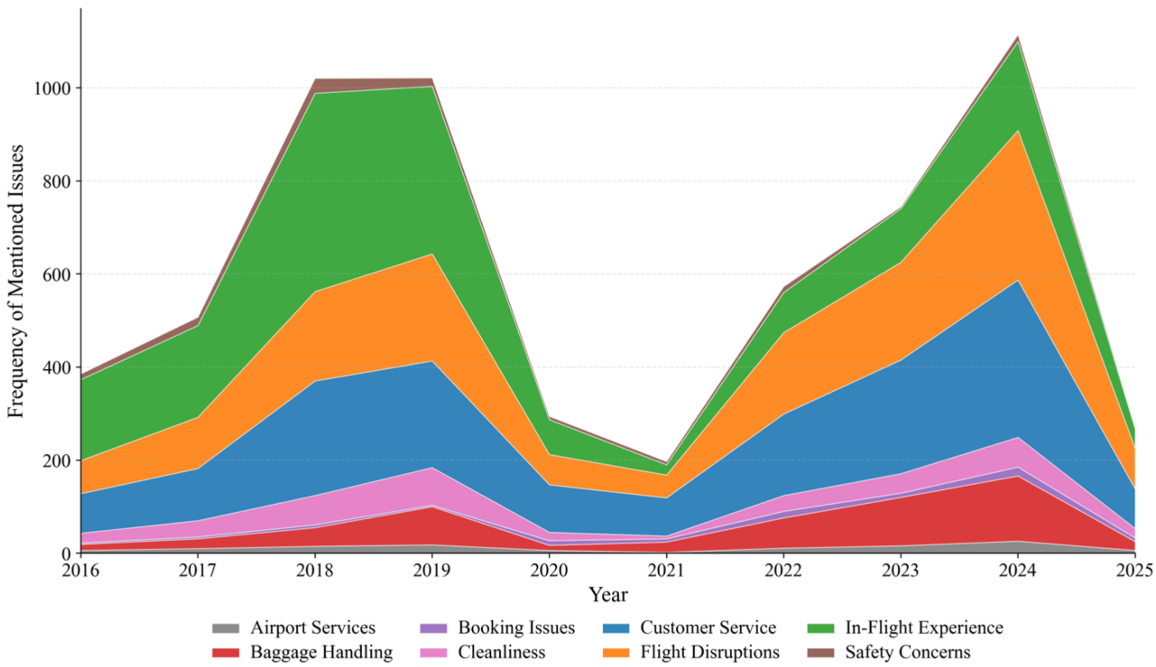

Figure 5: Temporal Evolution of Complaint Themes (2016-2024)

6 Discussion and Conclusion

The findings should be interpreted with some caution. The dataset consists of reviews posted on TripAdvisor and therefore reflects the experiences of passengers who chose to leave a review rather than a representative sample of all passengers. The analysis is also based on issues expressed in the reviews, so the frequency of a complaint should not be interpreted as a direct measure of its prevalence among all passengers.

The results point to a clear difference between overall service ratings and the specific experiences described by passengers. Airline A's ratings fell sharply after 2022, while the most common complaints focused on communication and interactions with staff. This suggests that passengers evaluate the service as a whole rather than on isolated aspects of the journey.

6.1 Interpreting the Findings

The complaints suggest that the way a problem is handled can matter almost as much as the problem itself. Delays are a recurring source of dissatisfaction, but so is the lack of information during those delays. Staff behavior is another major issue. Taken together, these findings point to communication, service recovery, and staff interactions as important areas for improvement, alongside the physical aspects of the flight.

6.2 Implications

The differences across market groups suggest that service problems are not experienced uniformly. For an airline operating across multiple markets, monitoring passenger feedback by market can therefore provide information that an overall satisfaction score cannot. Identifying where ratings are consistently low may help direct attention to particular service processes or passenger-facing operations.

6.3 Future Directions

This study suggests that Large Language Models can be useful for analyzing large collections of passenger reviews. Their main advantage here is not simply automation, but the ability to work with the language passengers actually use. This makes it possible to separate related complaints, such as a delay and poor communication about that delay, and to identify recurring patterns that broad service categories may combine.

Future work could strengthen the approach by comparing LLM classifications with independent human coding, testing the stability of classifications across languages, and examining whether changes in complaint patterns correspond to changes in other service indicators. Such validation would help establish how reliably LLM-based review analysis can be used as part of routine service-quality monitoring.

References

Brochado, A., Rita, P., & Gameiro, C. (2019). Airline passengers' perceptions of service quality: Themes in online reviews. International Journal of Contemporary Hospitality Management, 31(2), 855–873.

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of Artificial Intelligence and its qualitative applications in the service sector. Journal of Open Innovation: Technology, Market, and Complexity, 8(4), 198.

Kirilenko, A. P., Stepchenkova, S., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. Journal of Travel Research, 57(8), 1012–1025.

Tamkin, A., McCain, M., Handa, K., Durmus, E., et al. (2024). Clio: Privacy-Preserving Insights into Real-World AI Use. arXiv preprint arXiv:2412.13678.