Yelp LLM-Assisted Anomaly Detection Project

Beyonce Zhou, Zijun Gao

Overview

We investigated the use of Large Language Models (LLMs) to detect anomalous reviews where textual content and numerical ratings are inconsistent. Using GPT-4.1 models to predict ratings from review text, we developed a statistical framework to identify potential outliers in a sample dataset of 10,000 Yelp reviews. Our preliminary findings reveal systematic patterns in rating inconsistencies and demonstrate feasibility of automated anomaly detection in review systems. Of 10,000 reviews analyzed, 2.4% were flagged as outliers using the Benjamini-Hochberg procedure, with distinct patterns emerging for overrated (1.6%) versus underrated (0.7%) reviews.

1. Introduction

Online review platforms face persistent challenges with manipulated, biased, or inconsistent ratings that can mislead consumers and distort business reputation. Traditional approaches to anomaly detection focus on metadata patterns (reviewer behavior, timing, IP addresses) but rarely examine for semantic consistency between review text and ratings. This study explores whether Large Language Models can effectively identify reviews where the expressed sentiment in text conflicts with the assigned numerical rating.

2. Methodology

2.0 Data Source

We used the publicly available Yelp Open Dataset accessed through Kaggle. The dataset contains over 8 million reviews with associated star ratings, business information, and user metadata. For this study, we focused exclusively on the review text and its corresponding star rating.

From the full dataset, we sampled 10,000 reviews across a variety of businesses to balance computational feasibility with statistical reliability. Reviews were randomly selected, ensuring diversity in ratings (1-5 stars) and business categories. No reviewer- or business-level metadata was used in the analysis, as our study specifically targets the semantic alignment between text and rating.

2.1 Model Selection and Cost Optimization

We evaluated three GPT-4.1 variants (Nano, Mini, and full GPT-4.1) on a test set of 10 reviews to determine the most cost-effective approach:

Initial Test Results for Discrete Predictions:

- Review #1: Minor variation (prediction of 2 vs 3) between runs for both Nano and Mini
- Reviews #2-#10: Identical predictions across all models
- **Conclusion:** No significant performance differences between models

Model Performance Comparison:

Original → Predicted

Model	Review #1	Review #2	Review #3	Review #4	Review #5	Review #6	Review #7	Review #8	Review #9	Review #10
Nano	3→2	5→5	3→4	5→5	4→4	1→1	5→5	5→5	3→2	3→2
Mini	3→3	5→5	3→4	5→5	4→4	1→1	5→5	5→5	3→2	3→2
Full	3→3	5→5	3→4	5→5	4→4	1→1	5→5	5→5	3→2	3→2

GPT-4.1 Nano was selected for discrete predictions as the most cost-effective option without sacrificing accuracy.

Initial Test Results for Probability Vector Predictions:

- The three GPT variants displayed drastically different results for the probability vector prediction of the reviews
- **Conclusion:** Significant difference in performance between models

Model Performance Comparison:

Predicted Probability Vector (P1 P2 P3 P4 P5)

Review #	Original Stars	Nano	Mini	Full
1	3.0	0.05 0.10 **0.20** 0.25 0.40	0.3 0.3 **0.2** 0.15 0.05	0.05 0.15 **0.35** 0.35 0.10
2	5.0	0.01 0.02 0.07 0.2 **0.7**	0.01 0.01 0.03 0.20 **0.75**	0.01 0.01 0.03 0.15 **0.8**

3	3.0	0.05 0.10 **0.30** 0.30 0.25	0.05 0.05 **0.15** 0.40 0.35	0.01 0.04 **0.15** 0.45 0.35
4	5.0	0.05 0.10 0.15 0.20 **0.50**	0.01 0.01 0.03 0.20 **0.75**	0.01 0.01 0.03 0.15 **0.8**
5	4.0	0.05 0.10 0.20 **0.30** 0.35	0.05 0.1 0.3 **0.4** 0.15	0.01 0.03 0.15 **0.55** 0.26
6	1.0	**0.05** 0.10 0.15 0.20 0.50	**0.85** 0.10 0.03 0.01 0.01	**0.65** 0.25 0.07 0.02 0.01
7	5.0	0.05 0.05 0.1 0.2 **0.6**	0.01 0.01 0.03 0.15 **0.80**	0.01 0.01 0.03 0.15 **0.8**
8	5.0	0.05 0.10 0.15 0.20 **0.50**	0.01 0.01 0.03 0.20 **0.75**	0.01 0.01 0.03 0.15 **0.8**
9	3.0	0.05 0.10 **0.20** 0.30 0.35	0.7 0.2 **0.05** 0.03 0.02	0.45 0.3 **0.15** 0.07 0.03
10	3.0	0.05 0.15 **0.3** 0.25 0.25	0.6 0.2 **0.1** 0.05 0.05	0.15 0.30 **0.35** 0.18 0.0

*True star rating bolded

GPT-4.1 Full was selected for probability vector predictions, with accuracy out-weighing cost efficiency.

2.2 Inconsistency Metrics

We developed two complementary inconsistency measures:

1. **Discrete Inconsistency:** |Predicted Rating - True Rating|
2. **Probabilistic Inconsistency:** -log(P(True Rating)) from GPT's probability vector

2.3 Statistical Framework

Following the Benjamini-Hochberg (BH) multiple testing correction procedure:

1. Created a clean inlier dataset (183 reviews / 200 sampled) by manually validating extreme inconsistencies (e.g. highest 10% from sample data)
2. Computed p-values for test reviews based on inlier distributions
3. Applied False Detection Rate control at q=0.2 to identify significant outliers

4. Separated outlier analysis into overrated (true rating > predicted rating) and underrated (true rating < predicted rating) categories

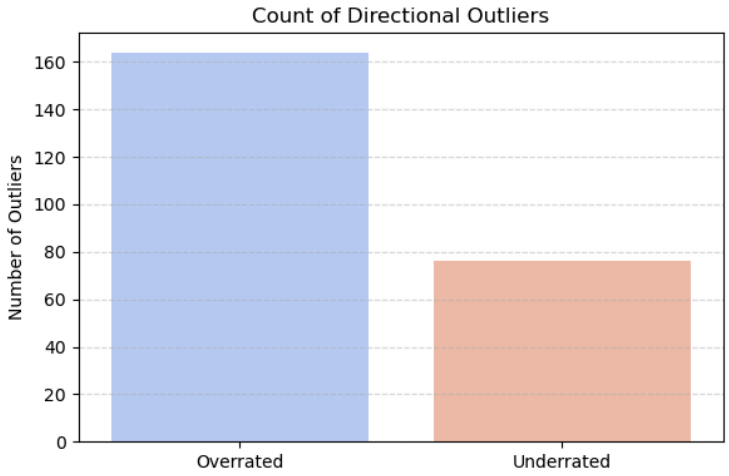
3. Key Findings

3.1 Performance Summary

Primary Results from 10,000 Review Analysis:

- **Total Outliers Detected:** 240 (2.4%)
- **Overrated Outliers:** 164 (1.6%)
- **Underrated Outliers:** 76 (0.8%)
- **Consistent Reviews:** 9760 (97.6%)
- **False Discovery Rate:** Controlled at 20%

Figure 1: Count of Directional Outliers



What does this mean?

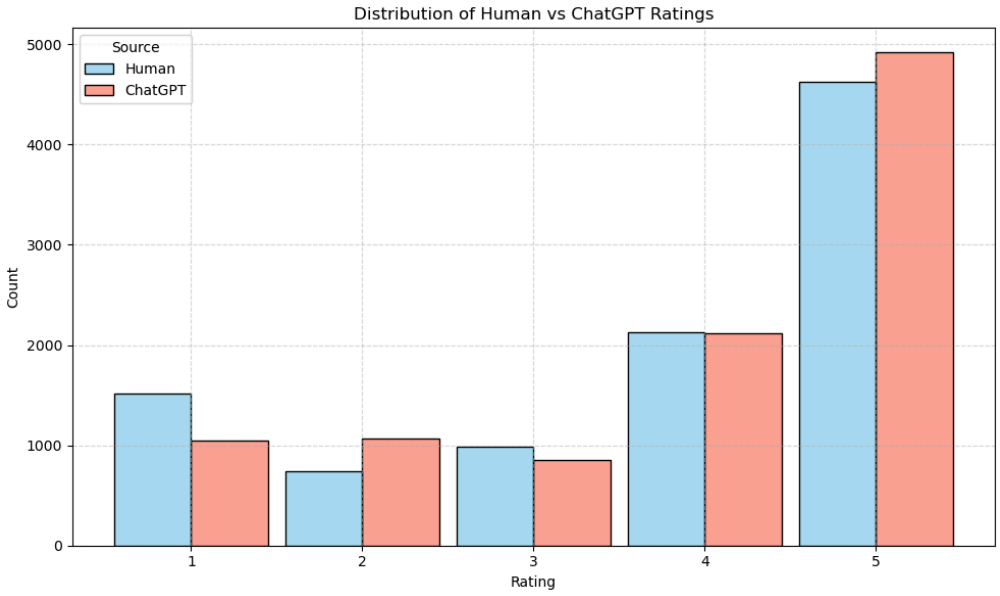
- More than double the amount of overrated as underrated, meaning for outliers, its more likely that the human rated higher than ChatGPT expected it to. This is especially interesting given that ChatGPT is proven to be slightly more optimistic than humans when rating
- Overrated: for overrated, examples show that **polite or soft-negative reviews** still got decent ratings, possibly customer bias or emotional buffering.
- Underrated: These could be **false negatives**, or reflect **sarcasm, misclicks, or sabotage**, useful for cleaning unfair ratings.

- Overall: humans are more emotionally inclined to rate higher than they feel when they have a negative experience, and less inclined to rate lower when they've had a decent/good experience

ChatGPT Vs. Human Rating Stats

- Mean (Human): 3.76
- Mean (ChatGPT): 3.79
- Std (Human): 1.47
- Std (ChatGPT): 1.24
- KS Test Statistic: 0.454, p-value: 0.000

Figure 2: Distribution of Human vs ChatGPT Ratings



What does this mean?

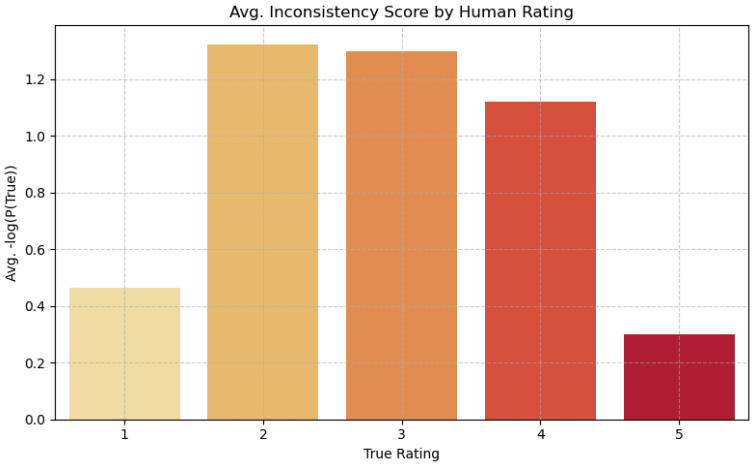
- ChatGPT is slightly more optimistic (+0.03 stars on average)
- ChatGPT is more conservative (-0.23 SD)
- The full shape of the distributions is significantly different (High KS Test)
- The **probability** of seeing a KS statistic as large as 0.454 **if** the two distributions were actually the same is **almost zero**, so the LLM predictions has its own pattern and diverge in a structured way
- Consistent for both Humans and ChatGPT that more reviews are 5s than any other star value
- Least reviews for humans are 2 stars and least reviews for ChatGPT are 3 stars

3.2 The Mid-Range Problem Validated at Scale

Our initial finding about 3-star reviews being problematic was confirmed in the larger dataset:

True Rating	Mean Inconsistency (Across 1,000 Reviews)	Mean Inconsistency (Across 10,000 Reviews)
1	0.465	0.463026
2	1.391	1.323491
3	1.136	1.297922
4	1.145	1.119597
5	0.313	0.301677

Figure 3: Average Inconsistency Score by Human Rating across 10,000 Reviews



We also looked at our inlier dataset to see the distribution of consistent labeling across true ratings:

Rating Distribution in Clean Inliers:

- 5 stars: 106 reviews (57.9%) - Heavily skewed toward positive
- 4 stars: 33 reviews (18.0%)
- 1 stars: 23 reviews (12.6%)
- 3 stars: 15 reviews (8.2%) - Notably underrepresented
- 2 stars: 6 reviews (3.3%) - Most rare in inlier dataset

What does this mean?

- Low inconsistency for 1 and 5 star reviews: ChatGPT tends to agree well with user and finds these easy to interpret
- Mid range, especially 2 and 3 star reviews, are the most ambiguous and ChatGPT has the most trouble with. Differences can come from mixed sentiment, sarcasm or subtle tone shifts, or masked criticism/politeness

3.3 Probabilistic Inconsistency Distribution Analysis

The distribution of probabilistic inconsistency scores (-log(P(True Label)) revealed:

Distribution Characteristics:

- Heavily Right-Skewed: The majority of reviews cluster near very low inconsistency (0.0-0.2).
- Long Tail: Inconsistency scores extend well past 3.0, showing rare but extreme cases.
- Bimodal Pattern: Evidence of two groups – reviews that align closely with predictions (low scores) vs. reviews that diverge significantly (high scores).

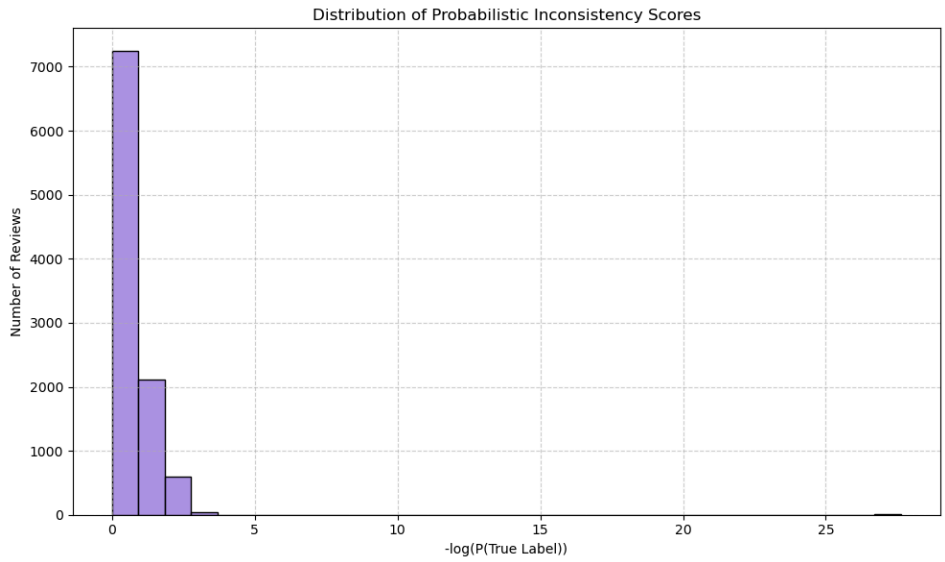
Score Interpretation:

- 0.0-0.5: Generally consistent reviews, with the model aligning well with human ratings.
- 1.0-2.0: Moderate inconsistency; worth closer inspection.
- 2.0+: High inconsistency; strong indication of potential anomalies or outliers.

Statistical Summary (10,000 Reviews)

- Mean: 0.674
- Std. Dev: 0.828
- Min: 0.0
- 25th Percentile: 0.223
- Median: 0.431
- 75th Percentile: 1.050
- Max: 27.63

Figure 4: Distribution of Probabilistic Inconsistency Scores



Extreme Outliers:

Five reviews exhibited -log(P(True)) > 25, indicating the model assigned essentially 0 probability to the true human rating:

--- Review #102 ---

Human Rating: 3

ChatGPT Expected Rating: 4.00

-log(P(true)): 27.6310

Review Text:

I had my first ever massage yesterday, and took advantage of Richel D'Ambra's Spa Week specials.

The massage itself was fantastic, and my masseuse was great. Before he began, I told him about some back pain I'd had and he made sure to pay special attention to the area. I left feeling like a big bowl of Jell-O.

However, I was (apparently) the only customer in the spa, as the other employees were having a very loud conversation right outside the room I was in. It was very difficult to relax while having to listen to a group of women cackling and laughing. I could barely hear the new-age music that was playing in the room.

I'd have given 4 stars if it wasn't for the noise.

--- Review #279 ---

Human Rating: 5

ChatGPT Expected Rating: 3.90

-log(P(true)): 27.6310

Review Text:

I don't ever give 5 stars. But I have to restrain myself from eating here too often. Let's get to it. The duck fries are delicious. The burger is good. Their brunch is awesome, (love places that will blanch a potato before frying it for home fries - makes a huge difference). Great omelettes, great biscuits. They do a maple pork belly hash special that is really tasty.

Every different style of mussels they cook is outrageous, favorite being the Provençal (skip it if you don't like lots of garlic and sun dried tomatoes). They had a duck breast special that was

probably the best thing I ate all month (including $30/lb prime rib for Christmas). It had a parsnip purée and I think a red wine sauce (AND a pear ginger coulis?) and it was better than the duck at Morimoto that cost twice as much. If you like creamy (heavier) pastas, the cavatelli with short rib is delicious. They had a ground pork burger special, that again, was delicious.

I'm really going to miss this spot if I leave the neighborhood.

--- Review #2543 ---

Human Rating: 3

ChatGPT Expected Rating: 4.00

-log(P(true)): 27.6310

Review Text:

I have been to this place only two times. First time the chicken was perfect, second time it was raw. I really like the flavor of the chicken and the nice selection of side orders. I am removing one star because of the raw chicken on the second visit.

--- Review #3397 ---

Human Rating: 3

ChatGPT Expected Rating: 4.00

-log(P(true)): 27.6310

Review Text:

I really like Dr. Bardach, however, I am very disappointed in regards to the inefficiencies related to scheduling. On two separate occasions I have called to schedule an appointment (during normal business hours). I left detailed messages with my information and requested a call back, and never received a call back. As a healthcare practice manager myself, this is a big NO NO and quite upsetting. Taking 2 stars off for that.

--- Review #3880 ---

Human Rating: 3

ChatGPT Expected Rating: 4.00

-log(P(true)): 27.6310

Review Text:

This is my neighborhoods grocery store. The ENTIRE neighborhoods grocery store. In fact, it is right smack in the middle of my neighborhood, which makes it always jam packed. I have never not waited in line for less than 20 minutes.

A lot of people are talking about the staff negatively. I have been shopping here for years, and recognize a lot of the employees who work here. I give them credit because it appears that the store is grossly under-staffed and more than likely, under paid. I love the one cashier, she is an older lady with dark hair. She may be nicest lady alive. And the football player looking guy who always helps me find things. And the teenage girl with the freckles, who is obviously smart but just stuck in this job. They might be miserable, but they know who their regulars are, and they always give me my gas card and pot stickers. The store may have its issues, but the staff has been around forever, and they deserve some credit. My 3 stars goes to the staff alone.

Patterns:

1. Explicit Star Mentions & Self-Adjustment

- Reviews #102 and #3397 explicitly narrate their star assignment rationale:
    - *"I'd have given 4 stars if it wasn't for the noise"* (#102).
    - *"Taking 2 stars off for that"* (#3397).
- These cases show transparent human reasoning for penalizing ratings based on contextual negatives (noise, scheduling inefficiency).
- Why it breaks the model: The model weights overall sentiment and expected rating distribution more heavily than explicit user-declared rating adjustments. As a result, it predicts higher ratings (~4) while the human settles lower (~3).

2. Strongly Positive Content with Harsh Deduction

- Review #2543: Mostly positive, but one severe incident (raw chicken) drove the rating down to 3.
- Review #102 (massage): Great experience, but noise ruined the relaxation.

- Pattern: Humans often overweight “single salient negatives”, even when the review is predominantly positive.
- Model mismatch: The model interprets the bulk of sentiment as positive → expects 4 stars.

3. Inflated Positivity & Rare 5-Star Behavior

- Review #279:
    - Reviewer claims *“I don’t ever give 5 stars”*, but then awards 5.
    - Content is overflowing with superlatives (best duck, favorite mussels, “outrageous” dishes).
- Pattern: Humans sometimes overcorrect their own bias against giving 5 stars → when they finally do, the language may read closer to a 4 in distributional terms.
- Model mismatch: Model interprets this as ~3.9 (very strong but not absolute), failing to account for the reviewer’s personal rating style.

4. Community & Contextual Loyalty

- Review #3880:
    - Complaints about long lines and understaffing dominate, but the rating is softened to 3 stars out of loyalty to staff.
    - Reviewer explicitly says: *“My 3 stars goes to the staff alone.”*
- Pattern: Humans adjust ratings based on meta-factors (community loyalty, empathy for staff), not just review text sentiment.
- Model mismatch: Model weights the negative descriptions more heavily and expects a harsher score (~4 vs 3).

4. Statistical Validation and Dashboard Analysis

4.1 P-Value Distribution Validation

The p-value distributions for both overrated and underrated categories showed appropriate statistical behavior:

- **Healthy right-skew** with concentration near zero for true outliers
- **Clear separation** at FDR threshold (dashed red line)
- **Appropriate tail behavior** confirming statistical assumptions

4.2 Expected vs True Rating Scatter Analysis

The scatter plot of LLM Expected vs True Ratings revealed:

- **Strong diagonal correlation** for most reviews (clustered near perfect agreement line)
- **Clear outlier separation**: Blue dots (underrated) above the line, red dots (overrated) below
- **Systematic bias**: Slight tendency for LLM to predict lower ratings than humans assign

5. Conclusion and Future Directions

5.1 Final Conclusion

This study demonstrates the feasibility of using Large Language Models (LLMs) to detect anomalies in online review systems by comparing semantic content with numerical star ratings. By applying GPT-4.1 models across a 10,000-review sample from the Yelp Open Dataset, we identified systematic inconsistencies where human ratings diverged from model-predicted sentiment. Our framework, combining discrete prediction mismatches, probabilistic inconsistency scores, and statistical validation through Benjamini-Hochberg correction, flagged 2.4% of reviews as anomalies, with a higher incidence of human overrated (1.6%) to human underrated (0.8%) cases.

The findings reveal several patterns:

- Low inconsistency in extreme ratings (1 and 5 stars): Humans and the model generally align on clear-cut cases.
- Higher inconsistency in mid-range ratings (2 and 3 stars): these remain the most ambiguous, often influenced by mixed sentiment, subtle tone, or user-specific rating habits.
- Human heuristics: Extreme outlier cases show reviewers explicitly narrating rating logic, overweighting single salient negatives, or adjusting ratings based on loyalty, empathy, or personal rating philosophy, which are factors that the LLM does not capture.

Overall, our work validates that LLMs can provide a structured, scalable approach to anomaly detection, complementing existing metadata-based methods with semantic consistency checks.

5.2 Future Applications

- Fraud and sabotage detection: LLM-based inconsistency metrics could flag suspicious reviews where text sentiment sharply diverges from star ratings, signaling manipulation attempts.
- Platform integrity: Online platforms like Yelp, Amazon, or TripAdvisor could integrate these tools to automatically surface reviews for moderation, improving trust and reducing noise.
- Business intelligence: Businesses could use anomaly flags to identify unfairly low ratings (sabotage, misclicks) or overly generous reviews that may bias reputation.

5.3 Next Steps

To deepen this research, future work may explore:

1. Cross-Referencing Metadata:
    a. Map review inconsistency and anomaly analysis with reviewer behavior (e.g. prolific low ratings, location/IP metadata).
    b. Analyze at the business level to detect whether certain types of businesses attract more inconsistent reviews (e.g. cultural rating norms).
2. Model Refinement:
    a. Explore prompt engineering refinement, especially for mid-range cases.
    b. Benchmark against alternative anomaly detection methods (traditional NLP sentiment models, embeddings-based clustering).

5.4 Closing Note

Our results illustrate that text–rating alignment is a meaningful and underutilized dimension of anomaly detection. With further integration of user and business metadata, as well as refinement of linguistic cues, LLM-assisted frameworks could significantly improve the reliability and fairness of online review ecosystems.