Human-AI Tools for Contextualizing Differences: Bridging Data-Driven Insights with Real-World Interpretability

Elara Liu, Medini Chopra

March 2025

Abstract

Understanding and quantifying contextual differences in human experiences across locations is a fundamental challenge in computational social science, human-computer interaction, and information retrieval. While large-scale textual datasets such as Yelp reviews provide rich opportunities for comparative analysis, existing methods struggle with category granularity, document length biases, and ranking inconsistencies. Traditional frequency-based approaches like TF-IDF often fail to capture conceptual shifts effectively, leading to unreliable cross-location comparisons.

This paper presents a human-AI system for analyzing and visualizing contextual differences across locations. By leveraging BM25, our approach mitigates document length distortions while improving term weighting for robust ranking. To enhance interpretability, we introduce a threshold-based relevance framework, which categorizes conceptual differences dynamically based on score distributions. The system integrates interactive visualizations to make complex ranking results accessible to both technical and non-technical users.

Our evaluation demonstrates that this approach effectively quantifies contextual differences, addressing key challenges in AI-assisted comparative analysis. The findings contribute to broader discussions on ranking techniques, human-centered AI, and information retrieval for contextual understanding.

1 Introduction

Understanding the contextual, conceptual, and cultural differences in user experiences across regions is crucial for enhancing the relevance and personalization of data-driven insights. Imagine visiting a coffee shop in another country, expecting to work there as you would back home, only to find that it's not the norm in this new place. Perhaps coffee shops in this region are more suited for socializing than working. This mismatch highlights how user expectations and experiences are shaped by local customs and context, underscoring the need for systems that adapt to these regional variations.

Existing information retrieval and recommendation systems often assume that user behavior and preferences are homogeneous, failing to account for the nuanced differences influenced by location, climate, and social norms. They attempt to capture these conceptual differences and typically focus on high-level statistical summaries, sentiment analysis, or assumptions about the *intended* use of a place by humans. These differences fail to capture how people are actually using those places, what activities they do, what emotions they feel, and what experiences they seek there. For instance, if you wanted to try an authentic cheesesteak, you would naturally expect a sandwich shop in Philadelphia to offer a better experience than one in Illinois. This reflects a difference rooted in cultural and geographical factors between the two places. While we intuitively understand these distinctions, the challenge lies in capturing and quantifying them effectively.

Recent advances in natural language processing and machine learning have improved the ability to analyze large datasets for patterns. However, most systems remain limited in their capacity to contextualize these findings, especially when it comes to understanding why certain experiences resonate differently across regions. While some approaches attempt to incorporate metadata or user demographics, they often treat these elements as secondary to the data itself, rather than integral to the explanation process. The challenge lies in building a system that not only extracts patterns from data but also explains the underlying contextual factors driving those patterns in a human-interpretable manner. For example, while existing systems like language models might generate a list of sandwich shops to get a cheese-steak at from both Philadelphia and Illinois, they fail to capture the underlying conceptual difference. A user without prior knowledge about the significance of Philadelphia cheese-steaks could mistakenly assume that both locations offer an equally authentic experience. These systems are not designed for this level of contextual understanding and would require significant time, effort, and resources to fine-tune for such a task.

In this paper, we address this challenge with a system that contextualizes differences in user experiences across regions. Our system leverages BM25 scoring techniques to identify significant themes within user-generated content, i.e. Yelp reviews. By comparing relevance scores across regions for activities in categories, our system surfaces patterns and generates contextual differences that help users understand how certain experiences, such as sunbathing or seafood preferences, vary between locations. The assumption is that if people prefer an activity at a place, there will be more people talking about that activity in the place, leading to more reviews.

The core idea behind our system is to combine data-driven insights with human-interpretable explanations. It looks for latent factors that influence user experiences in geographical locations. This approach bridges the gap between statistical analysis and real-world understanding of human behavior.

To build our system, we conducted a large-scale analysis of Yelp reviews, comprising 6,990,280 reviews and 150,346 businesses across 11 metropolitan areas. The system successfully identified regional differences in user experiences, demonstrating improved relevance and interpretability over baseline models.

Our contributions are as follows:

- System Contribution: We propose a novel approach to combining data-driven analysis with human-interpretable visualizations of conceptual differences (geographical, cultural, and social diversities), advancing the field of context-aware computing.
- Technical Contribution: We introduce a system that identifies and contextualizes differences in user experiences across regions, places, and activities using BM25 scoring of Yelp reviews.
- Conceptual Contribution: We demonstrate that user experiences are inherently shaped by geographic and social contexts, offering a framework for incorporating this insight into information retrieval systems.

The rest of this paper is organized as follows: Section 2 reviews related work in context-aware recommenders, biases in AI and information retrieval systems. Section 3 outlines our design goals, highlighting the challenges and opportunities in extracting location-specific relevance. Section 4 describes the implementation of our system, including the user interface, data processing, our algorithm for difference computation, and our scoring thresholds. Section 5 discusses our findings, limitations, and future directions.

2 Related Work

2.1 Context-Aware Recommendations

[1] examines how environmental factors like location and time shape users' emotions and preferences. For example, people tend to visit open spaces on sunny days, engage with local news, and connect with others based on shared interests. This motivates the need to tap into such aspects when thinking about conceptual differences across locations, providing evidence of the same. [2] analyzes advancements in e-commerce recommender systems, highlighting how cultural, geographical, and socio-economic factors shape consumer preferences. It argues that current systems struggle to capture these nuances and adjust to changing socio-economic conditions, limiting their ability to reflect user preferences accurately. This underscores the limitations of current technology in capturing subtle contextual differences, let alone effectively incorporating them into industry-grade recommender models. [3] developed three graph-based models for location-based social networks using a random walk with restart algorithm on a user-location graph. The models combined user-related aspects, location attributes, and environmental conditions. This approach outperformed traditional methods which were based on popularity, collaborative filtering, and content-based recommendations. This paves the way for innovation in using traditional models in tandem with newer technology to yield accurate results.

2.2 Bias in AI

[4] raised concerns about training data quality and quantity, noting that pretrained language models often reflect biases related to gender, race, sexuality, and minority groups. They also point out that language use varies globally, and training models primarily on English data reinforces a Western cultural bias since it misses out on cultural nuances that stem from language. This brings out the limitations of relying on language models without considering the impact of training data on conceptual differences. On the training data, [5] also addresses that Internet-based data is not curated, and LLMs inherit "stereotypes, misrepresentations, derogatory and exclusionary language, and other denigrating behaviors that disproportionately affect already-vulnerable and marginalized communities". LLM can often amplify biases too, such as negative sentiment and toxicity towards some minority social groups. Further, [6] mention that researchers have observed LLMs often retrieving information that is inaccurate and skewed toward LLM-generated content. Once again, purely relying on language models for identifying conceptual differences may propagate biases. While [7] begins to tackle this challenge by proposing a cost-effective solution to integrate cultural differences into LLMs, its ability to grasp the subtleties of differences remains limited.

2.3 Information Retrieval and BM25

[8] claims that from the different information retrieval techniques, BM25 is one of the most influential lexical search algorithms, and despite the growing focus on text embedding-based semantic search, BM25 maintains high efficiency and a strong generalization ability. It is used in many domains of information retrieval including medicine [9], regulation [10], and search engines [11]. Nowadays, it is extensively used in retrieval-augmented generation systems, serving as a fundamental framework for context-aware ranking by combining term frequency, inverse document frequency, and document-specific refinements. These features of BM25 made it a compelling algorithm to employ for our use case.

3 Design Space

3.1 Design Problem

Understanding contextual differences across locations is inherently complex for both humans and AI, yet it is crucial for making informed design decisions and ensuring inclusivity. People often struggle to make decisions about unfamiliar contexts, and those with technical expertise may attempt to build a system from scratch using available data. However, for individuals without a technical background, reasoning about contextual variations is significantly more challenging. Even with technical implementation, processing, interpreting, and building a system based solely on existing data can be time-consuming and may introduce inclusivity biases depending on the strength and representativeness of the dataset.

With the advancement of Large Language Models (LLMs) and their increasing generalization toward Artificial General Intelligence (AGI), LLMs offer potential assistance in understanding contextual differences. However, LLMs often exhibit decision biases when reasoning about location-specific contexts, making Human-AI collaboration unreliable in this domain.

Conceptual differences are nuanced and not purely quantitative. It is impractical to define a single threshold that universally captures contextual variations across geographical, cultural, and social dimensions. Human interpretation of contextual differences is further complicated by distinguishing true contextual variations from mere data prevalence or activity prevalence, which can shift based on location. Understanding these contextual differences is only the first step—equally important is the effective visualization of these differences.

Making contextual differences interpretable for users is a significant challenge. Traditional data visualizations are often difficult for non-technical users to comprehend. Storytelling-based approaches have been shown to improve accessibility for the general public [12], while machine learning-based methods can enhance interpretability through Explainable AI (XAI) techniques. However, XAI methods are often opaque and difficult to understand for those unfamiliar with machine learning concepts [13]. Addressing these challenges requires a multi-faceted approach that balances technical robustness, interpretability, and inclusivity.

3.2 Design Goals

The primary objective of our system is to make contextual differences easily understandable for users without prior technical knowledge or familiarity with the target locations. Simply presenting numerical values or computed scores is not sufficient for a general audience. A key challenge lies in defining what constitutes interpretability within the system or algorithm [14].

From a user perspective, the specific model or algorithm used behind the scenes is irrelevant; what matters is whether the system helps users understand contextual differences effectively [15]. Whether the system employs statistical data measurement or machine learning methods involving similarity and loss calculations, these details should remain hidden from users—especially those without a technical background. Instead, the focus should be on delivering clear, meaningful insights that directly answer user inquiries.

Furthermore, balancing accuracy with usability is critical. A system that can compute highly accurate results but requires several minutes or hours for processing is impractical. Similarly, an overly complex algorithm may be challenging to explain to users who seek to understand their results. Striking a balance between computational efficiency and interpretability is essential for the success of the system [16].

3.3 Design Space

When making design choices and decisions, we encountered several key research and design questions that shaped the system's architecture:

How can we ensure inclusivity across different contextual, cultural, and social backgrounds? Ensuring inclusivity in AI-based systems is challenging, as LLMs and AI models often struggle with bias when reasoning across geographical, cultural, and social dimensions. For example, consider a scenario where users seek affordable first-date restaurants. If we define "affordable" as below $15 per person, the system may exclude many restaurants that are typically considered ideal for first dates by the majority of users. This introduces a bias, as minority preferences may not be adequately represented. Prior research has demonstrated the importance of embedding diversity, equity, and inclusion (DEI) principles in AI system design [17]. To address this, we chose to use BM25, a purely statistical retrieval method, instead of LLM-based decision-making. BM25 allows us to compute contextual differences without relying on AI-driven inference, reducing the risk of implicit biases.

How should the system define and measure "conceptual difference"? Conceptual differences are not binary but exist along a spectrum. Using a frequency-based approach (e.g., counting how many people within a given location prefer a particular activity) aligns better with user needs than arbitrary threshold-based classification. However, controlling for data prevalence issues is essential, as raw frequency counts may reflect dataset limitations rather than actual contextual differences. To mitigate this, our system restricts comparisons to data that is available within our database to ensure meaningful interpretation.

How should the system present contextual differences in an interpretable way? Raw numerical scores alone are unintuitive for non-expert users, making it necessary to explore different visualization and explanation strategies. We considered bar charts and graphical comparisons, which provide a straightforward visual representation, textual explanations that offer detailed insights but may be difficult to interpret for users with reading difficulties or unfamiliarity with specific terminologies, and narrative storytelling approaches that enhance engagement but may introduce ambiguity in conveying quantitative differences. To ensure broad accessibility, we opted for a graph-based visualization approach that allows users to compare categories across different locations visually. The target category is highlighted using distinct, noticeable colors, and our color scheme adheres to WCAG 2 contrast and perceptual accessibility standards, ensuring content remains accessible for users with color vision deficiencies and low vision [18].

4 System Description

4.1 User Interface

Considering the focus and contribution of our system, it is essential to provide a convincing and intuitive user interface that allows the general public to effectively query and understand contextual differences across locations.

4.1.1 Overview of the System and User Workflow

This system primarily targets users interested in exploring and comparing contextual differences across locations. Many users in this category may have limited knowledge of data science, AI, or computational models, making it challenging for them to interpret conceptual differences based solely on raw data. At

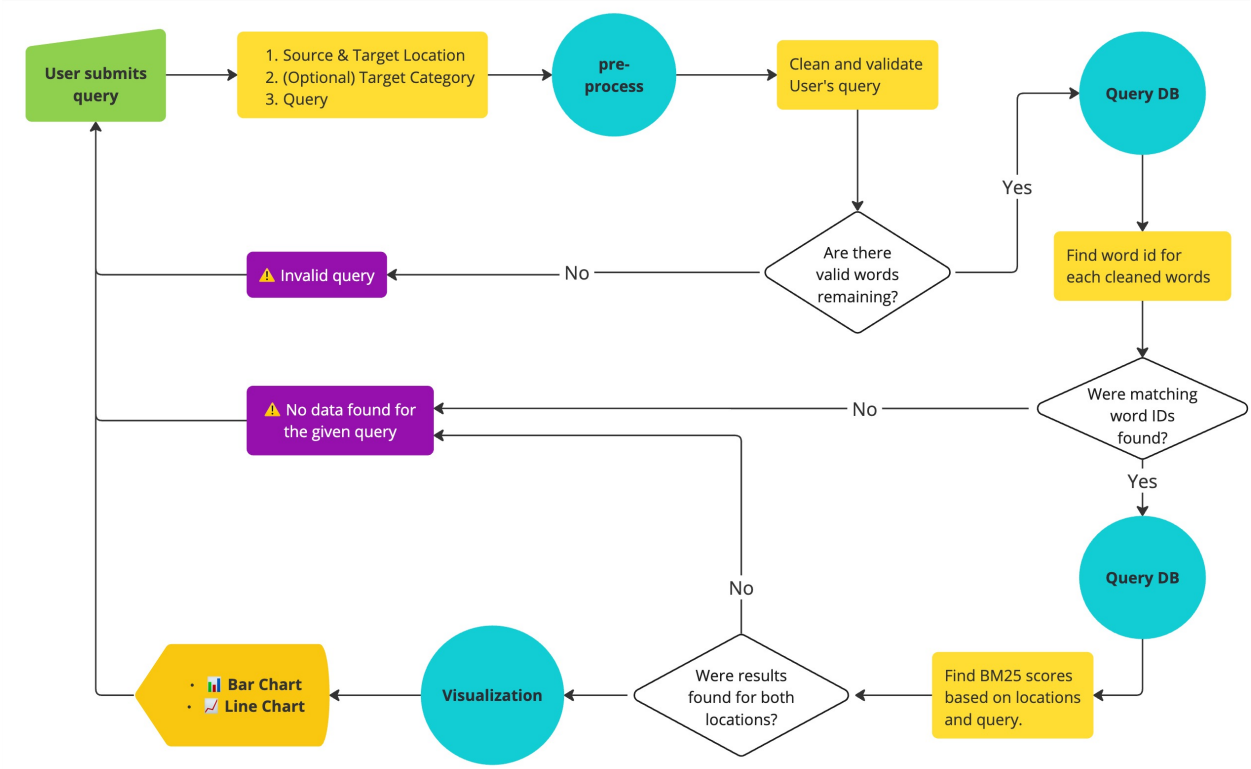

Figure 1: User workflow within the system.

the same time, the system is also designed to accommodate users with technical backgrounds, who may have sufficient expertise to build custom solutions for analyzing geographic, cultural, or social differences. However, even for technically proficient users, unfamiliarity with a specific target context can make it difficult to identify which differences should be prioritized for investigation.

For example, consider a user seeking to understand differences in Christmas celebrations between Japan and the United States [19]. Without background knowledge of these cultural variations, they may struggle to determine which aspects of the celebration to focus on when analyzing available data. Even if they are familiar with the general context, developing a one-time system to explore such differences is not easily generalizable for future use cases.

Our system provides a structured and automated workflow that enables users to query and visualize contextual differences in an intuitive and interpretable manner. Users are only required to select comparison locations and input a query, after which the system executes a comprehensive backend process, as shown in Figure 1. The system then presents the results in an easily understandable format, allowing users to explore differences without the need for technical expertise.

4.1.2 Query Input and User Interaction

Our system follows a user-centered design approach, ensuring that query input is straightforward and accessible. The input interface is shown in Figure 2. Users begin by selecting the source and target locations, which can be specified at either the state or city level. Additionally, users can refine their search by selecting a target category, which is chosen from the list of available categories in the source location.

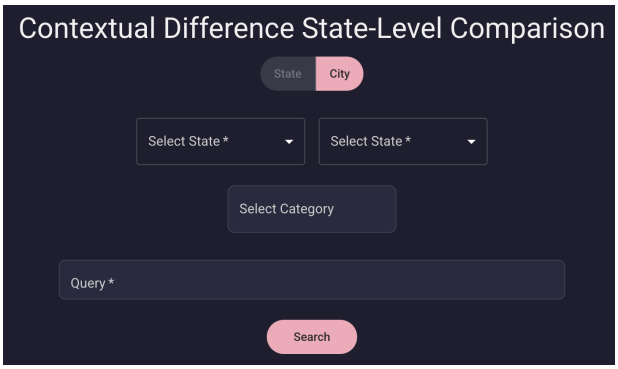

Figure 2: User selection and input interface.

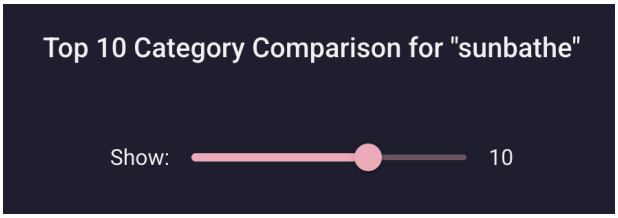

Figure 3: Bar chart slider to adjust the number of displayed categories.

Finally, they input a query, which is best formatted as a single-word (unigram) query. The query represents an activity or concept that the user wishes to compare between the source and target locations.

For example, if a user wishes to compare `sunbathing in Florida (FL) versus Pennsylvania (PA) at beaches`, the system will process the input through a word-cleaning pipeline, which aligns with the BM25 scoring mechanism. If any words remain after preprocessing, the system proceeds to query the database using the cleaned input.

The query process consists of the following steps:

1. The system attempts to find a word ID for each cleaned word to determine whether it exists in the dictionary.
2. If at least one valid word ID is found, the system queries the database separately for the source and target locations.
3. If scores exist for both locations, the system retrieves them directly from the PostgreSQL database.
4. The system uses BM25 scoring to ensure ranking consistency while preserving category granularity (further details in Sections 4.2 and 4.3).

4.1.3 Visualizing Contextual Difference

To enhance interpretability, the system automatically generates two types of visualizations:

1. Bar charts for categorical comparison.
2. Line charts for raw score distribution.

Raw numerical scores alone are difficult for non-technical users to interpret. Even for users with technical expertise, the scores can vary significantly based on different statistical models, hyperparameter configurations, and normalization approaches. Thus, using graphical representations provides a clearer and more intuitive way to understand contextual differences. If the user selected a target category, it is highlighted in both charts to facilitate comparison and validation.

`Bar charts for categorical comparison.` The bar chart (Figure 4) defaults to displaying the top 10 categories from the source location, with corresponding target location scores presented alongside them.

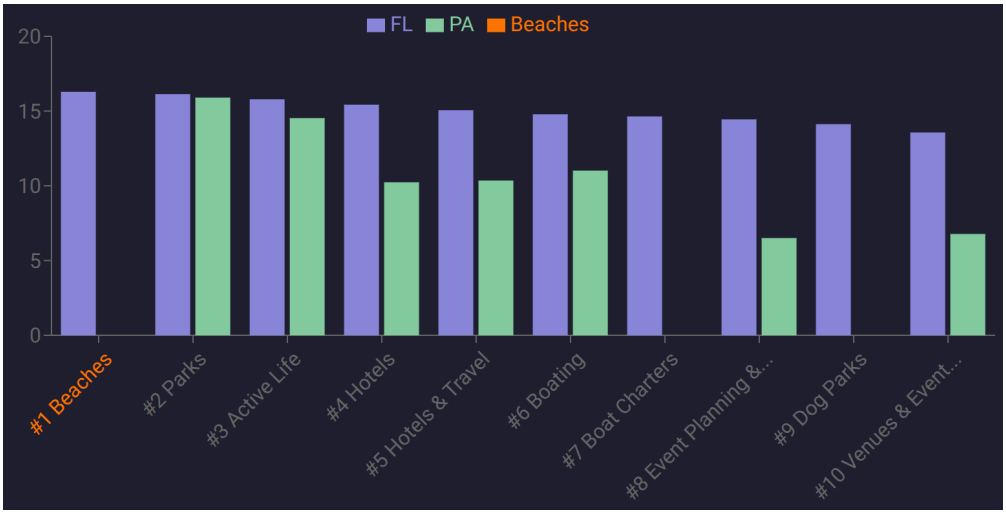

Figure 4: Bar chart visualization for "sunbathing" from FL to PA at beaches.

Each category has two bars—one representing the source and the other the target. If the target score is zero or undefined, it indicates one of two possibilities: either the category exists in the target location but the query word does not, or the category does not exist in the target location at all. Users can adjust the number of displayed categories from a minimum of 2 to a maximum of 15 using the slider control (Figure 3).

`Line charts for raw score distribution.` For line chart visualizations (Figure 5), data points are plotted along a continuous scale, with two key reference markers: the "Relevance Threshold" (yellow dashed line) and the "No Engagement Threshold" (red dashed line). Each location has its own line chart, where the relevance status of the user-selected category is indicated as one of four levels: relevant, somewhat relevant, not relevant, or not found. If a target category exists, it is highlighted in the chart and included in the legend. Users can zoom in and out of the line chart by adjusting the brush control at the bottom of the graph. To improve clarity, bar charts are sorted in descending order while line charts follow an ascending order, making trends more interpretable. Additionally, the system adheres to WCAG 2 contrast and perceptual accessibility standards [18] to ensure readability for users with color vision deficiencies and low vision.

4.2 Technical Details

Our system's functionality relies on a robust preprocessing pipeline for computing BM25 scores and contextual thresholds, enabling intuitive frontend visualizations. The integration of these backend computations with a scalable frontend interface ensures that the system remains efficient and interpretable.

4.2.1 Data Processing and Storage

The system is built using the Yelp Open Dataset, which consists of 6,990,280 reviews and 150,346 businesses across 11 metropolitan areas. Processing such a large dataset requires an efficient and scalable approach. The raw dataset is provided in JSON format, including approximately 5GB of review data and 300MB of business metadata. To optimize data access and processing speed, we preprocess and store the data in parquet format, enabling batch-based parallel computation.

To ensure consistency across locations, we preprocess business data by standardizing state and city names.

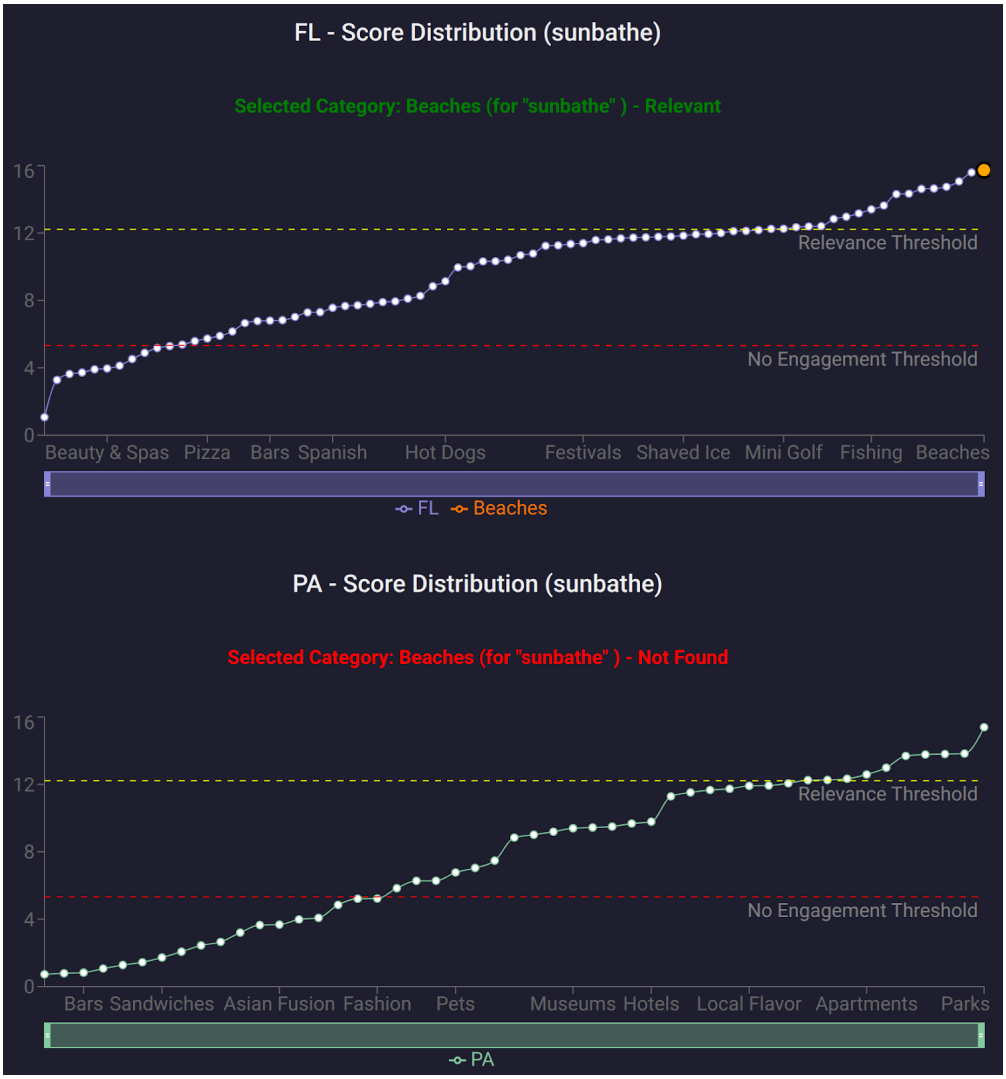

Figure 5: Line chart visualization for ”sunbathing” from FL to PA at beaches.

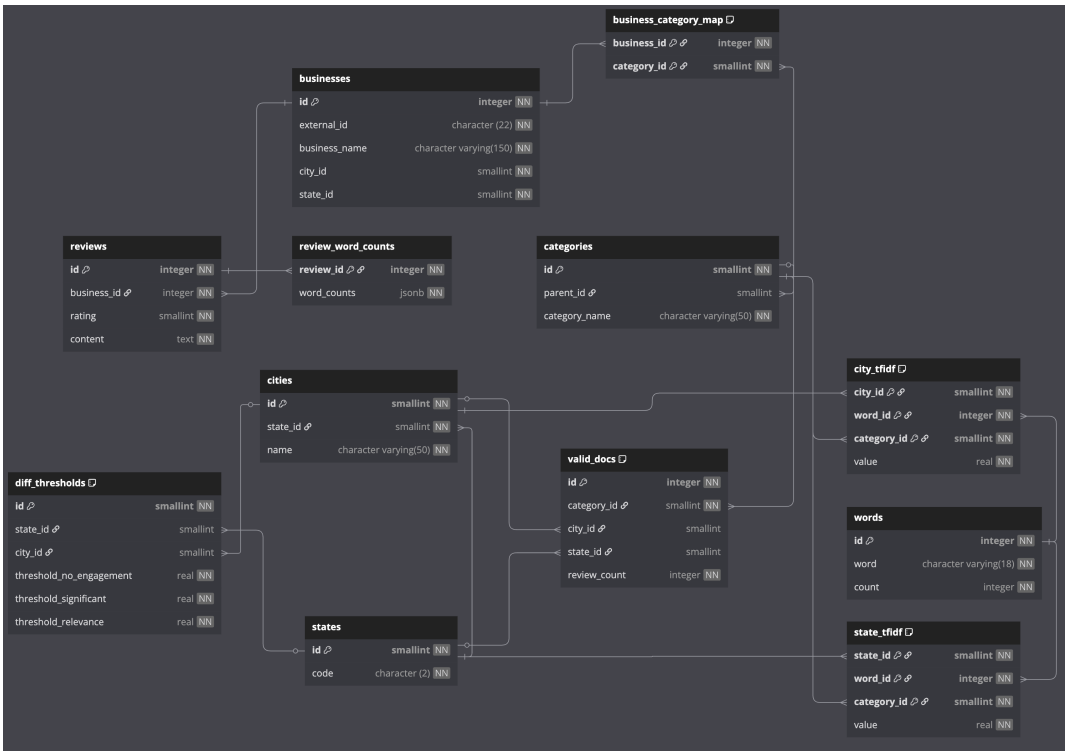

Figure 6: Postgres database schema visualization

This includes resolving inconsistencies such as abbreviations (e.g., "St. Louis" $\rightarrow$ "Saint Louis") and merging duplicate city representations (e.g., "Ventura" and "Ventura County"). Additionally, business categories are extracted from their original comma-separated format into structured lists.

The system is implemented using `PostgreSQL`, with 12 relational tables storing processed businesses, categories, reviews, and precomputed scores. Unique integer identifiers replace raw string identifiers to improve storage efficiency and query performance. While category hierarchies and granularity mismatches are not explicitly corrected in preprocessing, they are effectively addressed through BM25 ranking, which ensures that category-specific importance is appropriately weighted.

For review data processing, we map each review to its corresponding business ID and apply text normalization techniques. The preprocessing pipeline includes regular expression-based text cleaning, lemmatization using `NLTK`, spell correction via `SymSpell`, and stopword removal. This step ensures that all textual data is normalized before frequency-based computations. To handle the large-scale dataset efficiently, reviews are processed in parallel batches using `Python`'s `concurrent.futures`.

Following preprocessing, we construct a word dictionary mapping unique word IDs to their corresponding cleaned terms. Each review's word distribution is stored in `JSONB` format within `PostgreSQL`, allowing efficient retrieval for BM25-based ranking computations. Once all preprocessing steps are completed, we precompute BM25 scores for each (location, category) pair and store them in the database for optimized querying.

Algorithm	Scoring Method	Rank of "Beaches"	Rank of "Parks"	Top Ranked Category
TF-IDF	L2 Norm	#6	#7	Active Life
TF-IDF	Sublinear TF	#6	#7	Active Life
TF-IDF	L2 Norm + Sublinear TF	#6	#7	Active Life
BM25	$b = 0.75$, $k_1 = 1.2$	#1	#2	Beaches
BM25	$b = 0.6$, $k_1 = 1.3$	#1	#2	Beaches

Table 1: Comparison of TF-IDF and BM25 ranking methods for "sunbathing" from FL to PA.

State	"sunbathe" Count	Total Word Count	Frequency per Million Words	BM25 IDF
FL	71	49,847,610	1.424	2.049
PA	29	76,126,414	0.355	2.428

Table 2: Effect of document length on term frequency and BM25 IDF for "sunbathe" in FL vs. PA.

4.2.2 Contextual Difference Computation

To quantify contextual differences, we evaluate two primary statistical methods: TF-IDF and BM25. One of the key challenges in using TF-IDF is the category granularity issue caused by Yelp's hierarchical category structure. Since businesses often belong to multiple categories at varying specificity levels, review frequency counts can overlap significantly, leading to inconsistencies in score distributions.

As shown in Table 1, we compare TF-IDF and BM25 for the query "sunbathing" when comparing Florida (FL) and Pennsylvania (PA) in the Beaches category. Even when applying L2 normalization, sublinear TF scaling, or IDF smoothing, TF-IDF ranks "Beaches" and "Parks" lower in comparison, failing to emphasize their relevance. This occurs because TF-IDF's linear term frequency weighting does not account for document length variations across categories. The TF-IDF score used in our system is computed as:

$$TF - IDF(q_i, d) = \frac{(1 + \log(TF(q_i, d))) \cdot \ln\left(\frac{N}{DF(q_i)+1} + 1\right)}{\sqrt{\sum_{j=1}^{m} TF(q_j, d)^2}}$$

where $TF(q_i, d)$ is the term frequency of query term q_i in document d, $DF(q_i)$ is the document frequency, and N is the total number of documents. The numerator applies sublinear TF scaling, the denominator applies L2 normalization, and the smoothed IDF function ensures numerical stability.

However, TF-IDF struggles with document length differences between categories. Table 2 demonstrates this issue with "sunbathing" in FL vs. PA. While Florida has more occurrences of "sunbathe" (71), Pennsylvania has fewer instances (29) despite a larger total word count. This skews the TF-IDF ranking, as the raw term frequency per million words is 4× higher in FL than in PA, but TF-IDF fails to correctly weigh this impact.

In contrast, BM25, which incorporates non-linear frequency scaling and document length normalization, effectively differentiates contextual differences between locations. By adjusting the b and k_1 hyperparameters, BM25 assigns higher relevance to infrequent yet important terms, allowing a more robust comparison between source and target locations. As seen in Table 1, BM25 correctly identifies "Beaches" and "Parks" as the most relevant categories, demonstrating its advantage over TF-IDF in handling varying document lengths and term distributions.

BM25 scores are computed using:

$$BM25(q,d) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{(TF(q_i,d) \cdot (k_1+1))}{TF(q_i,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{\bar{d}})}$$

where $TF(q_i, d)$ represents the term frequency of query term q_i in document d, $|d|$ is the document length, and $\bar{d}$ is the average document length. The hyperparameters k_1 and b control term saturation and document length normalization, respectively.

Unlike TF-IDF, BM25 mitigates document length bias by weighting term importance relative to document size rather than relying on absolute term frequency. The IDF calculation in BM25 further improves ranking stability:

$$IDF(q_i) = \ln\left(\frac{N - DF(q_i) + 0.5}{DF(q_i) + 0.5} + 1\right)$$

where $DF(q_i)$ is the number of documents containing term q_i, and N is the total number of documents. This formulation penalizes frequently occurring terms while boosting rare but meaningful terms, ensuring more stable ranking across categories.

To quantify conceptual shifts, we compute the contextual difference using:

$$\Delta_{context} = |BM25_{source} - BM25_{target}|$$

where larger values of $\Delta_{context}$ indicate greater conceptual differences between the source and target locations.

4.2.3 Threshold-Based Interpretation

To improve interpretability, we define dynamic thresholds based on the BM25 score distribution for each location (state or city). Instead of presenting raw scores, which may be difficult to interpret, we categorize relevance using three key thresholds:

$$Threshold_{\text{no engagement}} = 5\text{th percentile}$$

$$Threshold_{\text{significant}} = (\max \text{ value} + \min \text{ value}) \times 0.15$$

$$Threshold_{\text{relevance}} = (\max \text{ value} + \min \text{ value}) \times 0.6$$

These thresholds are computed independently for each location, ensuring an adaptive interpretation that reflects the local distribution of scores.

- **No Engagement Threshold**: Categories below this threshold are considered not relevant to the query.
- **Significant Difference Threshold**: If the absolute change in score from source to target exceeds this threshold, the difference is considered significant, regardless of whether the value increases or decreases.

```
 Differences From FL to PA for query: sunbathe
--------------------------------------------------------------------------------
Category                       Prev Rank    Curr Rank    Prev Score   Curr Score   Status
--------------------------------------------------------------------------------
Beaches                        1            Removed      2.3695       0.0000       Became Inactive
Dog Parks                      3            Removed      2.3201       0.0000       Became Inactive
Boat Charters                  6            Removed      2.2856       0.0000       Became Inactive
Fishing                        10           Removed      2.1898       0.0000       Became Inactive
Parks                          2            1            2.3604       2.4293       Slight Change
Active Life                    4            5            2.3001       2.2531       Slight Change
Boating                        5            6            2.2923       2.2298       Slight Change
Hotels                         7            21           2.2633       1.8031       Slight Change
Landmarks & Historical Buildings 8            4            2.2537       2.2823       Slight Change
Hotels & Travel                9            24           2.2211       1.7766       Slight Change
--------------------------------------------------------------------------------
```

Figure 7: Command line interface (CLI) in Python.

- **Relevance Threshold**: Only categories above this threshold are considered relevant to the query.

By implementing these thresholds, the system translates numerical BM25 scores into meaningful insights, allowing users to easily identify highly relevant, marginally relevant, and irrelevant conceptual differences. This approach ensures greater interpretability of location-based variations, reducing cognitive load and making comparisons more intuitive.

4.2.4 Design Iterations

Our initial approach for frontend visualization involved a Python CLI-based output, which displayed only raw BM25 scores and absolute differences, as shown in Figure 7. However, early user testing revealed that interpreting numerical outputs was highly challenging, particularly when analyzing abstract contextual differences. Users found it difficult to determine the significance of numerical variations without additional guidance.

To enhance usability, we transitioned to a graphical comparison model, providing intuitive visualizations that facilitate comparative reasoning across locations. This shift significantly improved user comprehension and interaction, making the system more accessible to both technical and non-technical users.

4.3 Implementation

To develop a scalable and efficient system, we utilized a combination of frontend and backend technologies. The implementation stack includes:

- **Programming Languages**: Python for backend processing, TypeScript for frontend development.
- **Frontend Framework**: Next.js with MUI for UI components.
- **Data Fetching and ORM**: TanStack React Query for efficient data fetching, integrated with Prisma as an ORM layer.
- **Visualization Library**: Recharts for bar chart and line chart visualizations.

- **Data Processing**: Custom BM25 implementation using Numba and NumPy for JIT-optimized computations.
- **Database**: PostgreSQL for structured data storage and retrieval.
- **Package Management**: UV for backend dependencies, pnpm for frontend dependencies.

5 Discussion

This study presents a system that enables data-driven contextual comparisons across geographic locations, making complex conceptual differences interpretable for users with varying technical backgrounds. By Integrating BM25-based information retrieval, dynamic thresholding, and human-centered visualizations, the system effectively quantifies and presents conceptual differences while addressing challenges related to document length bias, category granularity, and interoperability. In this section, we reflect on the key techniques and design elements that contributed to the system's effectiveness, discuss the challenges encountered during deployment, and explore the broader implications of this research. Additionally, it outlines the limitations of the current implementation and highlights direction for future work.

5.1 Generalizable Techniques and Design Insights

One of the key innovations in this system is the adaptive thresholding mechanism, which dynamically sets engagement, relevance, and significance thresholds based on the local distribution of BM25 scores. Unlike traditional approaches that rely on fixed numerical cutoffs, this method ensures that comparisons remain meaningful across diverse contexts. The 5th percentile engagement threshold effectively filters out categories that are not relevant to the concept or activity, while the significance threshold (15% of the score range) identifies meaningful contextual shifts in activity prevalence. This approach can be generalized to another ranking-based system where context-aware interpretation is essential, such as consumer behavior analysis, public health trend detection, and cultural studies.

To further enhance interpretability, the system employs human-centered visualization to translate BM25 scores into easily digestible insights. Rather than presenting raw numerical output, the system visualizes differences through bar charts and line charts, highlighting target categories with relevant, and nonrelevant thresholds using dashed lines. This approach makes the information accessible even to non-technical users, facilitating intuitive comparisons between locations. The inclusion of color schemas aligned with WCAG 2 accessibility guidelines ensures that users with color vision deficiencies can effectively engage with the system. This method of interactive AI-assisted visualization holds potential applications beyond contextual analysis, such as in personalized recommendation systems, dynamic market segmentation, and real-time policy evaluation, where transparent and interpretable decision-making is crucial.

5.2 Challenges and Observations from Deployment

Despite the effectiveness of these techniques, the deployment of the system revealed several challenges that provide important insights for future development. One major limitation stems from BM25's lack os semantic understanding, as it operates purely on statical term weighting without incorporating contextual meaning. This results in challenges when dealing with polysemous words or idiomatic expressions, where word frequency alone is insufficient to determine conceptual relevance. For example, the system may

fail to differentiate between ”hot dog” (food) and ”hot dog” (enthusiastic person) unless strong category data is available to disambiguate the context. Even the example used in prior sections, '*sunbathing on beaches*,' may introduce ambiguity: does the review actively recommend the location for sunbathing, or is it cautioning against it? Furthermore, does it refer to suitability for people, or is it primarily describing a space designed for dogs, such as a dog park? These nuances highlight the limitations of purely frequency-based retrieval. Addressing this issue requires integrating BM25 with contextual embeddings, such as BERT-based re-ranking, to refine retrieval results based on semantic similarity.

Another challenge is the system's dependence on Yelp's dataset, which, while extensive, may introduce sampling biases due to uneven review distributions across locations. Some activities or businesses might be underrepresented in the dataset, leading to weaker comparisons for certain concepts and activities. For example, if fewer reviews discuss kayaking in Pennsylvania compared to Florida, this does not necessarily mean that kayaking is less relevant in Pennsylvania. Instead, it may reflect differences in reviewing behaviors, regional engagement with Yelp, or variations in how users describe similar activities. While BM25 can still highlight kayaking's significance in Pennsylvania if it appears consistently across relevant documents, extreme disparities in dataset size may distort relative rankings. This highlights the need for complementary normalization techniques or multi-source data integration to ensure more reliable contextual comparisons.

5.3 Limitations

While the system successfully addresses key challenges in contextual comparison, several limitations remain. First, its reliance on keyword-based retrieval rather than deep sematic understanding means that some queries requiring nuanced interpretation may not be fully captured. Additionally, data sparsity issues in certain categories may affect result reliability, particularly when comparing locations with significantly different review volumes. Moreover, the system does not currently offer personalized ranking adjustments, which could enhance user experience by allowing custom filtering based on individual preference or domain expertise. These limitations, while important, do not diminish the system's contributions but rather highlight areas for future improvement.

5.4 Future Work

Building on the system's successes and addressing its limitations, several key directions for future research and development emerge. One significant improvement would be the integration of BM25 with transformer-based models, such as BERT or SBERT, to enhance semantic understanding. A hybrid approach, where Bm25 serves as an initial retriever and an embedding-based model refine rankings, would improve the system's ability to handle ambiguous or multi-meaning terms.

Expanding the dataset scope is another crucial step. Currently, the system is limited to Yelp reviews, which may not fully capture regional differences in behavior or preferences. Future iterations should integrate multiple online sources, such as Google Reviews, Reddit discussions, and social media platforms like Instagram. This would provide a more comprehensive view of contextual differences, reducing dataset biases and improving the result diversity.

Additionally, enabling real-time data updates through incremental indexing would enhance the system's adaptability. Instead of relying on batch processing, implementing dynamic indexing techniques would allow rankings to continuously evolve as new data becomes available. This would be particularly beneficial for tracking emerging trends in different cultural behaviors.

Finally, conducting user studies to evaluate the system's interpretability is essential. While the visualization

techniques employed are designed for accessibility, formal evaluation is needed to assess usability, effectiveness, and decision-making support. Future research should compare how technical and non-technical users interact with the system, measuring cognitive load, comprehension speed, and accuracy of interpretation.

References

[1] A. B. Suhaim and J. Berri. Context-aware recommender systems for social networks: Review, challenges and opportunities. *IEEE Access*, 9:57440–57463, 2021.

[2] Kelley Ann Yohe. Toward global, socio-economic, and culturally aware recommender systems, 2023.

[3] Alimohammadi A Khazaei, E. Context-aware recommender systems for social networks: Review, challenges and opportunities. *ISPRS International Journal of Geo-Information*, 2019.

[4] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2), June 2023.

[5] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024.

[6] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models, 2024.

[8] Xianming Li, Julius Lipp, Aamir Shakir, Rui Huang, and Jing Li. Bmx: Entropy-weighted similarity and semantic-enhanced lexical search, 2024.

[9] Hristidis V. Weiner M. Cheng, S. Leveraging user query sessions to improve searching of medical literature. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, page 214–223, 2013.

[10] Jhon Rayo, Raul de la Rosa, and Mario Garrido. A hybrid approach to information retrieval and answer generation for regulatory texts, 2025.

[11] Dr Vishal Ratansing patil Nirali Arora, Dr Harsh Mathur. Towards contextual search optimization: A unified ranking approach for relevance prioritization. *Journal of Information Systems Engineering and Management*, 2025.

[12] Hongbo Shao, Roberto Martinez-Maldonado, Vanessa Echeverria, Lixiang Yan, and Dragan Gasevic. Data storytelling in data visualisation: Does it enhance the efficiency and effectiveness of information retrieval and insights comprehension? In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[13] Courtney Ford and Mark T Keane. Explaining classifications to non-experts: an xai user study of post-hoc explanations for a classifier when people lack expertise. In *International Conference on Pattern Recognition*, pages 246–260. Springer, 2022.

[14] Owen Lahav, Nicholas Mastronarde, and Mihaela van der Schaar. What is interpretable? using machine learning to design interpretable decision-support systems. *arXiv preprint arXiv:1811.10799*, 2018.

[15] Joachim Diederich. Methods for the explanation of machine learning processes and results for non-experts. 2018.

[16] André Assis, Douglas Véras, and Ermeson Andrade. Explainable artificial intelligence-an analysis of the trade-offs between performance and explainability. In *2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6. IEEE, 2023.

[17] Sarika Kondra, Supriya Medapati, Madhuri Koripalli, Sri Rama Sarat Chandra Nandula, and Julie Zink Zink. Ai and diversity, equity, and inclusion (dei): Examining the potential for ai to mitigate bias and promote inclusive communication. *Journal of Artificial intelligence and Machine Learning*, 3(1):1–8, 2025.

[18] World Wide Web Consortium et al. Web content accessibility guidelines (wcag) 2.0. 2008.

[19] Junko Kimura and Russell Belk. Christmas in japan: Globalization versus localization. *Consumption Markets & Culture*, 8(3):325–338, 2005.