Invisible Walls in Cities: Designing LLM Agent to Predict Urban Segregation Experience with Social Media Content

Anonymous Submission

Abstract

Understanding experienced segregation in urban daily life is crucial for addressing societal inequalities and fostering inclusivity. The abundance of user-generated reviews on social media encapsulates nuanced perceptions and feelings associated with different places, offering rich insights into segregation. However, leveraging this data poses significant challenges due to its vast volume, ambiguity, and confluence of diverse perspectives. To tackle these challenges, we propose a novel Large Language Model (LLM) Agent to automate online review mining for segregation prediction. Specifically, we propose *reflective LLM coder* to digest social media content into insights consistent with real-world feedback, and eventually produce a codebook capturing key dimensions that signal segregation experience, such as *cultural resonance and appeal*, *accessibility and convenience*, and *community engagement and local involvement*. Guided by the codebook, LLMs can generate both informative review summaries and ratings for segregation prediction. Moreover, we design a **REasoning-and-EMbedding (RE'EM)** framework, which combines the reasoning and embedding capabilities of language models to integrate multi-channel features for segregation prediction. Experiments on real-world data demonstrate that our agent substantially improves prediction accuracy, with a 22.79% elevation in R^2 and a 9.33% reduction in MSE. The derived codebook is generalizable across three different cities, consistently improving prediction accuracy.Moreover, our user study confirms that the codebook-guided summaries provide cognitive gains for human participants in perceiving POIs' social inclusiveness.Our study marks an important step toward understanding implicit social barriers and inequalities, demonstrating the great potential of promoting social inclusiveness with Web technology.

CCS Concepts

• **Applied computing → Sociology**; • **Information systems → Information retrieval**.

Keywords

Segregation, Social Media, Large Language Model, Human Mobility

ACM Reference Format:
Anonymous Submission. 2018. Invisible Walls in Cities: Designing LLM Agent to Predict Urban Segregation Experience with Social Media Content. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid advancement of web technologies has enabled fine-grained records of social interactions through diverse media, such as textual posts [34], photos [11], and short videos [5]. As a result, web platforms have become extensive social sensing systems, generating rich digital traces that reflect social dynamics in the physical world [13, 29]. Such data offers a unique opportunity to uncover subtle patterns or preferences in individuals' everyday life, allowing researchers to study long-standing social issues from a new perspective [21]. Among the most enduring of these issues is segregation: the systematic separation of individuals or groups based on certain characteristics. However, identifying cues of segregation experience from massive, multimodal web content poses significant challenges, requiring advanced methods to capture nuanced feelings, and there is no established procedure to guide this process.

Historically, research on segregation has focused primarily on residential patterns [24], where individuals of the same racial or ethnic group are more likely to reside in the same neighborhoods [28]. This type of spatial segregation has been linked to negative social outcomes, such as limited upward social mobility and increased crime rates [12, 15]. Recent studies have shifted focus to experienced segregation—the dynamic segregation individuals experience in daily movements [21]. Despite the seemingly free human movements in most modern urban spaces, certain demographic groups continue to face barriers to equal access and social interactions. In other words, there seem to be "invisible walls" in cities that prevent sufficient social interactions between groups. These walls are not merely the result of physical proximity but are influenced by cultural, social, and economic factors that drive segregation on a more nuanced level. Accurately predicting segregation experiences can inform individuals to avoid potentially uncomfortable situations and support policymakers in fostering social inclusiveness. Moreover, research suggests that segregation experiences can intensify in larger urban areas [23], underscoring the need to explore the underlying mechanisms for growing more equitable cities admist ever-increasing urbanization. However, existing studies only present retrospective and holistic measurements, providing limited insights and prediction capabilities.

To address these challenges, we propose to leverage the reasoning power of LLMs for automated social media content analysis. We propose a *reflective LLM coder*, featuring a strategic agentic workflow that integrates two key components: a *reflective attributor* and a *code summarizer.* The *reflective attributor* employs an abductive reasoning approach: it prompts the LLM to estimate a location's appeal to different demographic groups from its social media content, and reconcile discrepancies through iterative reflection on observed segregation patterns. Subsequently, the *code*

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

summarizer applies chain-of-thought reasoning [33] to iteratively merge insights into a structured codebook. This codebook guides LLMs to transform free-text reviews into structured summaries, highlighting segregation-related factors like cultural resonance and community engagement. Complementing this qualitative approach, we propose a **REasoning and EMbedding (RE'EM) framework** for quantitative segregation prediction, which combines LLMs' reasoning capabilities with the representational power of pre-trained embedding models. First, we prompt the LLM to provide structured ratings of a place's appeal to different groups based on the codebook, allowing the reasoning outcomes of LLMs to be vectorized and easily integrated with other channels of features. Concurrently, we finetune an embedding model to learn global representations optimized for segregation prediction. Finally, we fuse structured ratings, global embeddings, and population information using a neighbor-aware multi-view predictor.

We validate our approach through qualitative user studies and quantitative experiments on four US cities. In our user study with 75 researchers, participants' prediction accuracy greatly improved when provided with LLM-generated summaries. Furthermore, as many as 80% of participants prefer the codebook-guided summaries over vanilla LLM outputs. In quantitative experiments, the RE'EM framework improves the predictive R^2 by 22.79% and reduces the MSE by 9.33% compared to baseline models relying solely on local racial composition, and is generalizable across cities.

Our contributions are summarized as follows:

- We are the first to explore the use of social media content for predicting POI experienced segregation.
- We design a reflective LLM coder to effectively summarize online reviews and identify cues of segregation.
- We propose a REasoning-and-EMbedding (RE'EM) framework that combines the reasoning and embedding capabilities of language models to predict segregation.
- We validate our approach through comprehensive qualitative and quantitative evaluations, demonstrating its effectiveness in extracting operationable insights and generalizing segregation prediction improvement.

2 Related Work

Mining Web Data with LLMs. The digital footprints and rich textual contents people generate on Web platforms have proven to be informative for a wide array of social phenomena, including health outcome [22], mental well-being [20], crime rates [10], and neighborhood disparities [13, 26]. However, these studies often rely on pre-calculated indices or pre-defined word lists to engineer and extract features. This is not only labor-intensive but also lacks adaptability to evolving research questions and contexts. Given the remarkable language understanding and reasoning capabilities of LLMs, recent research explored the potential of leveraging LLMs to mine Web data for social good, such as revealing food-related social prejudice [18], evaluating public accessibility [16], and capturing urban perception [27]. Nevertheless, these studies often lack an examined framework for insight extraction from massive social media content, which limits their effectiveness in capturing the complexity of Web data. In contrast, our work presents the first step toward unlocking the reasoning power of LLMs to code free-form online reviews, extracting human-comprehensible, informative insights for segregation prediction.

Experienced Segregation. The study of segregation traces its roots back to the early 20th century, when sociologists examined the division of urban spaces into "ecological niches", each inhabited by distinct social groups [24]. Researchers revealed the detrimental effects of enduring segregation even in the absence of "legalized racial segregation", on education, income, housing, and crime [3, 14, 19]. Albeit offering valuable insights, these studies are highly constrained by data availability, overlooking the reality that individuals spend significant time and engage in numerous interactions beyond the confines of their residential neighborhoods [32]. With the rapid proliferation of smart mobile devices, the ability to track people's intricate movements in urban spaces has emerged [35]. This has paved the way for a novel research avenue aimed at quantifying and understanding the type of segregation individuals actually *experience* in their daily movements [2, 7, 21], including how it may change during disasters [4, 36] and scale with city sizes [23]. Nevertheless, existing studies primarily focus on static spatial distribution of demographic groups and physical movement, overlooking the complex socioeconomic factors behind the segregation phenomenon, such as cultural resonance and local involvement. Thus, these works can only provide retrospective measurement but have limited predictive power. In contrast, our work establishes an LLM-based method to automatically extract nuanced features from online reviews for experienced segregation prediction.

3 Preliminary

3.1 Problem Formulation

Inspired by [21], we compute the proportion of visitors at POI i by each racial group q, denoted as τ_{qi}. We then quantify the segregation at POI i as the deviation of the visitor proportion τ_{qi} from the city's residential proportion T_q (k is a constant that normalizes S_i to range between 0 and 1):

$$S_i = k \sum_q (|\tau_{iq} - T_q|). \tag{1}$$

Our task can be formulated into a POI segregation prediction problem. Given a set of POIs $I = \{i_1, i_2, ..., i_N\}$ along with their corresponding social media content $C = \{c_1, c_2, ..., c_N\}$ and the racial composition of local population $P = \{p_1, p_2, ..., p_N\}$, the objective is to learn a function $f(i, c, p)$ to minimize the discrepancy between the predicted segregation and the real segregation S_i:

$$min \text{ diff}(f(i, c, p), S_i). \tag{2}$$

3.2 Data

We use three datasets to obtain social media, demographic, and mobility information. Social media and demographic data serve as input features, while mobility data is used to calculate the ground truth segregation at POIs, i.e., our prediction targets. We select cities based on the sufficiency of available POIs within the overlapping scope of the three datasets. As a result, we identify four cities for analysis: Philadelphia (5,360 POIs), Tucson (2,703 POIs), Tampa (2,222 POIs), and New Orleans (2,392 POIs).

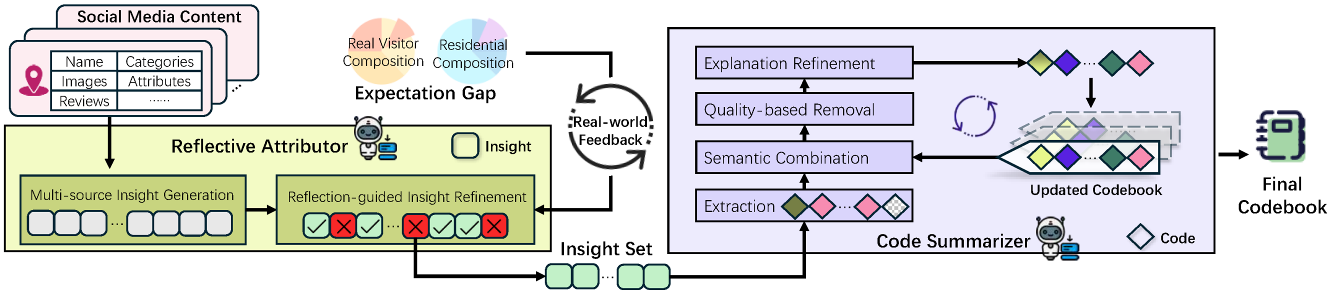


Figure 1: Overview of the *reflective LLM coder*. It consists of a *reflective attributor* that integrates multi-source review and image signals and refines insights using real visitation patterns, and a *code summarizer* that consolidates these insights into a structured, generalizable codebook capturing key factors shaping experienced segregation.

Demographic data. We obtain the demographics of census block group (CBG) residents from the American Community Survey 5-year Estimates (ACS)[1]. A CBG is the smallest spatial unit with available demographics, typically consisting of 600-3,000 residents. We categorize CBG residents into five distinct groups: *hispanic*, *black*, *asian*, *white*, and *others*, and calculate their respective ratios by dividing the population of each group by the CBG population.

Social media data. We use the Yelp Open Dataset[2] for social media data. For each POI, we obtain its name, location, stars, categories, service attributes, and multi-modal user-generated content. The category captures its business domain, such as “food” and “department stores”. The attributes reflect its specific business traits, such as price range, delivery options, and parking availability. The user-generated content includes textual reviews and images uploaded by visitors. Each POI is associated with at least 5 reviews.

Mobility data. To derive the ground truth of experienced segregation, we utilize the Safegraph Patterns Dataset (accessed through Advan)[3]. This dataset records monthly visits to each POI originating from various CBGs. We aggregate one year’s visitation records to construct a robust dataset with extensive observations. To address the distinct POI indexing systems between Yelp and SafeGraph, we match POIs by ensuring identical names and a location deviation within 200 meters. In Appendix A, we show the variability of POI features, highlighting the importance of our study.

We use demographic and mobility data from 2019 for temporal consistency. To leverage historical POI information, we include social media content from 2019 (16%) and the preceding eight years (84%). Each review is timestamped, enabling LLM to reason about the temporal relationship between the content and the current analysis period.

4 Reflective LLM Coder

The vast volume of raw social media content creates a significant cognitive barrier to revealing the encoded segregation experience. Therefore, an effective approach is needed to process and extract insights. Building on research demonstrating the consistency and stability of LLM-based coding [6, 30], we design a *reflective LLM coder* to generate an insightful codebook from a small set of POI data, capturing the multi-faceted factors influencing segregation and offering a structured framework for deeper analysis. As shown in Figure 1, it contains two modules, *reflective attributor* and *code summarizer*.

4.1 Reflective Attributor

We design a *reflective attributor* to identify factors influencing a POI’s appeal to different groups. Specifically, we design a two-step Chain-of-Thought (CoT) scheme [33] that guides an LLM to integrate multiple perspectives from texts and images, forming a reflection-driven process that refines insights based on real-world feedback.

Multi-source insight generation. In this step, the LLM generates insights by integrating multiple sources of information: the POI’s name, user-posted reviews, and images. It first evaluates the name for potential cultural, racial, ethnic, or socioeconomic associations that might influence the appeal to specific groups. Next, it analyzes the reviews and images to identify key aspects like atmosphere, pricing, service quality, and cultural relevance. These analyses are then synthesized into a concise set of insights, each identifying factors that could attract or repel particular racial/ethnic groups, ensuring comprehensive coverage without redundancy.

Reflection-guided insight refinement. Importantly, we introduce real-world data to guide LLM’s reflections upon these insights. We measure the difference between the demographic composition of visitors to each POI and that of nearby residents, using it as a ground-truth signal to identify which racial/ethnic groups it attracts or repels. This signal guides the LLM in refining the existing insights in an abductive manner, retaining only those consistent with real-world evidence. This step not only enhances the relevance of the insights, but also provides a grounded explanation for observed visitation preferences.

Combining these two steps, the *reflective attributor* links POI attributes with patterns in visitation behavior, offering a data-driven understanding of segregation experiences.

4.2 Code Summarizer

The *code summarizer* module aggregates outputs from the *reflective attributor* into a comprehensive and orthogonal codebook that encapsulates the diverse factors influencing POI appeal. We design another four-step CoT scheme to guide the LLM reasoning process.

[1] https://www.census.gov/programs-surveys/acs/
[2] https://www.yelp.com/dataset
[3] https://www.deweydata.io/data-partners/advan

Table 1: Automatically constructed codebook. Each dimension represents a distinct driver of POI attractiveness or deterrence for different demographic groups, forming the conceptual "bricks" underlying experienced segregation patterns. The explanations summarize the thematic meaning of each code.

Index	Name	Detail
1	**Cultural Resonance and Appeal**	Culturally themed offerings, such as Italian-American or South Indian cuisine, attract visitors seeking authentic or familiar experiences, influencing visitation based on cultural representation and resonance.
2	**Price Sensitivity and Economic Accessibility**	Moderate pricing, coupons, and cost-effective policies like BYOB appeal to budget-conscious visitors, impacting visitation patterns based on affordability and economic considerations.
3	**Service Quality and Customer Experience**	Professional and attentive service, despite occasional inconsistencies, attracts visitors valuing high service standards and personal interactions, influencing demographics based on service expectations.
4	**Atmosphere and Social Environment**	Lively, trendy, or family-friendly settings attract visitors prioritizing social and communal experiences, impacting visitation based on social and family-oriented preferences.
5	**Accessibility and Convenience**	Central locations, parking availability, and delivery services attract visitors prioritizing efficiency and accessibility, influencing patterns based on transportation and convenience.
6	**Visual and Aesthetic Appeal**	Modern, chic, and historically themed environments attract visitors who appreciate aesthetic and immersive experiences, influencing demographics based on visual and cultural preferences.
7	**Cultural and Social Inclusivity**	Inclusive, diverse, and culturally sensitive environments attract a broad demographic by catering to varied identities and preferences, influencing visitor composition based on inclusivity and cultural representation.
8	**Product Variety and Quality**	Diverse and high-quality offerings, including visually appealing and culturally themed products, attract visitors prioritizing variety and quality, influencing visitation based on product expectations.
9	**Community Engagement and Local Involvement**	Establishments with strong community ties and neighborhood vibes attract visitors valuing local engagement and communal experiences, influencing demographics based on community integration and involvement.

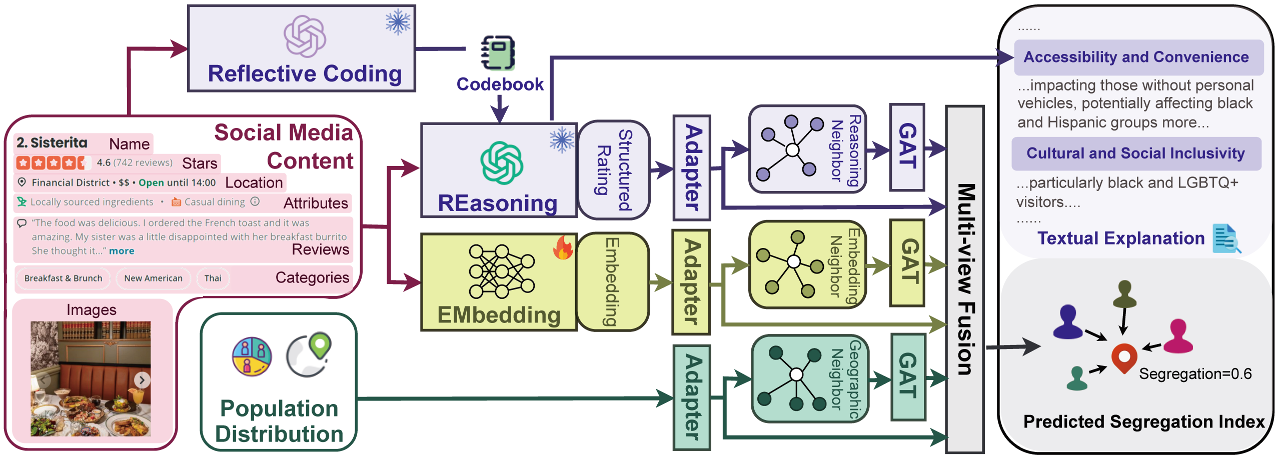

Reflective Coding
Codebook
Social Media Content
2. Sisterita
Name
4.6 (742 reviews)
Stars
Financial District • $$ • Open until 14:00
Location
Locally sourced ingredients • Casual dining
Attributes
"The food was delicious. I ordered the French toast and it was amazing. My sister was a little disappointed with her breakfast burrito She thought it..." more
Reviews
Breakfast & Brunch
New American
Thai
Categories
Images
Population Distribution
REasoning
Structured Rating
EMbedding
Embedding
Adapter
Adapter
Adapter
Reasoning Neighbor
Embedding Neighbor
Geographic Neighbor
GAT
GAT
GAT
Multi-view Fusion
......
Accessibility and Convenience
...impacting those without personal vehicles, potentially affecting black and Hispanic groups more...
Cultural and Social Inclusivity
...particularly black and LGBTQ+ visitors....
......
Textual Explanation
Segregation=0.6
Predicted Segregation Index

Figure 2: Reasoning-and-Embedding (RE'EM) framework. RE'EM integrates three complementary channels (reasoning, embedding, and population) with a neighbor-aware multi-view fusion to predict POIs' experienced segregation.

Extraction. Taking the insights extracted by the *reflective attributor*, this step extracts phrases representing the analysis perspectives for each insight. It ensures that previously-identified features are captured for further refinement.

Semantic combination. This step combines insights with similar semantics to reduce redundancy, e.g., merging "Cultural Relevance and Appeal" and "Inclusivity and Cultural Representation" into a unified and coherent code.

Quality-based removal. This step filters out insights that are less distinguishable or less generalizable. It ensures that the final codebook focuses on the most relevant and impactful factors, reducing the impact of noise.

Explanation refinement. Finally, this step refines the one-sentence explanation for each code, highlighting broader, thematic insights to ensure flexibility and interpretability across contexts while preserving analytical rigor.

Trough the above coding process (formalized in Appendix B), we obtain **a codebook identifying 9 distinct "bricks" that form the invisible walls**, shown in Table 1. Each brick represents a key aspect of the visitor experience, allowing a nuanced understanding of how places appeal to diverse racial demographics. For example, *Cultural Resonance and Appeal* captures how culturally themed offerings such as Italian-American cuisines attract visitors seeking authentic experiences, while *Price Sensitivity and Economic*

Accessibility captures the impact of affordability on visitation, with moderate pricing and cost-effective policies like BYOB appealing to budget-conscious visitors.

5 REasoning-and-EMbedding (RE'EM) Framework

To effectively integrate different features for accurate segregation prediction, we design a **REasoning-and-EMbedding (RE'EM)** framework composed of four key components: a *reasoning channel*, an *embedding channel*, a *population channel*, and a *neighbor-aware multi-view predictor*, as shown in Figure 2.

5.1 Reasoning Channel

We prompt an LLM to assess place attractiveness to different racial/ethnic groups under the guidance of our codebook (Table 1). Specifically, we instruct the LLM to simulate the perspective of each racial/ethnic group, and rate the POI along the codebook-identified dimensions. To unleash the LLM's reasoning power, we design a two-step CoT scheme.

Codebook-guided summary. For each POI i, we denote its associated social media content as c_i, which includes the POI's name, review text, and images (if any). We instruct the LLM to analyze c_i and generate a structured summary u_i of the POI's characteristics along the 9 codebook-defined dimensions, focusing on features that may attract or repel specific racial/ethnic groups.

Structured rating. Based on the summary u_i, we ask the LLM to imagine itself as a member of each of the 5 racial/ethnic groups and assign ratings to the POI across the 9 dimensions. Ratings range from 0 (strong repulsion) to 10 (strong attraction), with 5 indicating neutrality. This process yields a 45-dimensional vector r_i with corresponding textual explanations e_i, providing both quantitative scores and qualitative insights into the POI's social inclusiveness.

$$r_i, e_i = \text{LLM}(u_i, \text{codebook}). \tag{3}$$

To integrate the ratings into our predictive framework, we employ a *reasoning adapter* consisting of a multi-layer perceptron (MLP). It transforms the ratings into a vector representation v_i^r:

$$v_i^r = \text{ReasonAdapter}(r_i). \tag{4}$$

5.2 Embedding Channel

The *embedding channel* extracts deep semantic representations v_i^e from the review corpus associated with each POI i. Due to the token length constraints, we apply data augmentation, i.e., sample different subsets of c_i for input to the embedding model. We fine-tune an open-source text embedding model, GTE-base [17], to enhance its capability of extracting higher-level representations while preserving the pre-trained semantic knowledge. Although both the reasoning and embedding channels take c_i as input, they extract complementary aspects: structured attribute reasoning in the former, and holistic semantic embedding in the latter. The resulting embedding is then passed through an *embedding adapter* (MLP) to produce v_i^e.

$$v_i^e = \text{EmbeddingAdapter}(\text{GTE}(c_i)). \tag{5}$$

5.3 Population Channel

For each POI i, we compute a 5-dimensional feature vector p_i representing the racial composition of the surrounding population. Specifically, we identify all CBGs whose centroids fall within a 0.5km radius around the POI, and compute a population-weighted average of their racial compositions. The resulting p_i reflects the static demographic context of the POI's geographic location. We apply a *population adapter* (another MLP) to transform p_i into a latent representation v_i^p for subsequent prediction.

$$v_i^p = \text{PopulationAdapter}(p_i). \tag{6}$$

5.4 Neighbor-Aware Multi-View Predictor

This predictor primarily performs two key operations: neighbor aggregation and multi-view fusion. For each POI i, we identify its five nearest neighbors in three separate spaces (views), respectively: the reasoning space (based on similarity to r_i), the embedding space (similarity to e_i), and the geographic space (physical proximity). This forms three neighbor sets $N_i^r = \{n_1^r, ..., n_5^r\}$, $N_i^e = \{n_1^e, ..., n_5^e\}$ and $N_i^p = \{n_1^p, ..., n_5^p\}$. We use a graph attention network (GAT) for each view to aggregate the features of the respective neighbor set, producing neighbor-aware representations $v_{n_i}^r$, $v_{n_i}^e$, and $v_{n_i}^p$:

$$v_{n_i}^* = \text{GAT}_*(v_i^*, \{v_n^* | n \in N_i^*\}) \tag{7}$$

Finally, a fully connected fusion module integrates both the POI's representations v_i^* and its neighrbors' representations $v_{n_i}^*$ from all three channels. This fusion process ultimately outputs the predicted segregation $\hat{S}_i$:

$$\hat{S}_i = \text{Fuse}(v_i^r, v_{n_i}^r, v_i^e, v_{n_i}^e, v_i^p, v_{n_i}^p). \tag{8}$$

6 User Study

We conduct a user study to evaluate the cognitive gain to humans provided by our generated codebook. We recruit participants via snowball sampling among graduate students and researchers with hands-on experience or sufficient knowledge in social media studies: approximately 17.3% from urban planning, 24% from social science, and 58.7% from data mining. To ensure that participants fully understood the survey tasks, we provided clear explanations of the geographic and racial concepts required for the study. Our questionnaire can be accessed [4].

We randomly sample four POIs whose visitor demographics significantly differ from local residents. For each POI, we generate review summaries using a vanilla method (directly prompting LLMs without the codebook) and our codebook-guided method (see Appendix Figure 6). Participants answer four questions per POI, progressively incorporating different levels of contextual information. (1) **Baseline estimation**: Participants predict whether each racial group is over- or underrepresented at the POI based solely on its basic attributes (e.g., name, category, and service details). (2) **Review-informed prediction**: Participants reassess their estimates after being provided with a random sample of user reviews. (3) **Summary-enhanced prediction**: Participants make a final prediction after reviewing a summary of the user reviews. (4) **Summary preference evaluation**: Participants determine whether

[4] https://drive.google.com/file/d/1-NIjIPiVi_m_ONBFruYYuh4nvyhe4PF-/view?usp=sharing

Table 2: Segregation prediction performances across different method. Compared with population-only and embedding baselines, RE'EM achieves significant improvements across MSE, RMSE, MAE, and R².

Model	MSE↓	Improv.	RMSE↓	Improv.	MAE↓	Improv.	R^2↑	Improv.
Population	0.0075±0.0002	——	0.0864±0.0012	——	0.0683±0.0013	——	0.3164±0.0204	——
GTE	0.0073±0.0002	2.67% (p=0.7623)	0.0855±0.0011	1.04% (p=0.7887)	0.0685±0.0010	-0.29% (p=0.9713)	0.3249±0.0151	2.69% (p=0.8905)
BERT	0.0072±0.0002	4.00% (p=0.0147)	0.0846±0.0011	2.08% (p=0.0181)	0.0680±0.0010	0.44% (p=0.3708)	0.3394±0.0176	6.78%(p=0.0549)
GloVE	0.0072±0.0002	4.00% (p=0.0242)	0.0848±0.0011	1.85% (p=0.0290)	0.0685±0.0008	-0.29% (p=0.6118)	0.3357±0.0140	5.75% (p=0.0776)
Qwen3-Embedding-8B	0.0072±0.0002	4.00% (p=0.0196)	0.0848±0.0011	1.85% (p=0.0292)	0.0683±0.0009	0.00% (p=0.4789)	0.3349±0.0181	5.85% (p=0.0812)
RE'EM	**0.0068±0.0003**	**9.33% (p=0.0080)**	**0.0823±0.0021**	**4.75% (p=0.0071)**	**0.0662±0.0021**	**3.07% (p=0.0597)**	**0.3885±0.0278**	**22.79% (p=0.0022)**

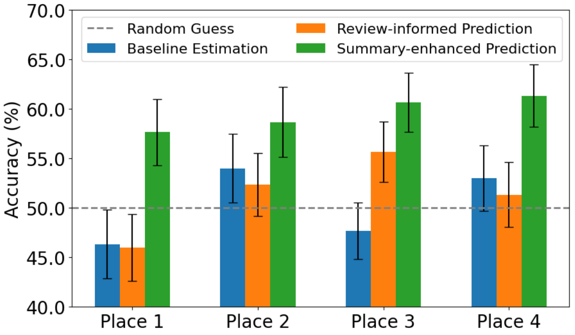


Figure 3: Human prediction accuracy with different information availability. Accuracy remains near random (50%) when only POI metadata (Baseline Estimation) or raw reviews (Review-informed Prediction) are available, but increases substantially once participants receive codebook-guided LLM summaries (Summary-enhanced Prediction), demonstrating the cognitive benefit of structured review distillation.

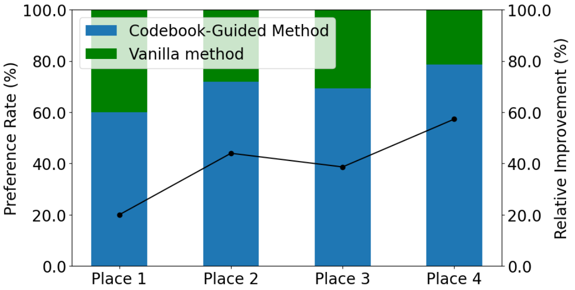


Figure 4: User preference between codebook-guided summary and vanilla summary. Across all sampled POIs, a strong majority of participants favor codebook-guided summaries, indicating that structured summaries better support human understanding of POI inclusiveness.

the codebook-guided summary or the vanilla summary is more informative.

Figure 3 presents the human prediction accuracies with different information availability. When relying solely on basic POI attributes, accuracy fluctuates around 50%, indicating that visitor preferences cannot be reliably inferred from POI metadata alone. Moreover, introducing randomly sampled user reviews does not lead to a notable improvement in prediction accuracy. This is likely due to the sparsity and unstructured nature of raw review data, which can be difficult to process and may even mislead human judgment. In contrast, providing participants with the review summary consistently enhances prediction accuracy across all sampled POIs. This result underscores the effectiveness of the codebook-guided summarization, which condenses extensive social media content into concise yet highly informative text snippets, enhancing human understanding of POIs' social inclusiveness.

Figure 4 presents the results of the summary preference evaluation. Across all sampled POIs, 60%-80% of participants prefer the codebook-guided summary over the vanilla one, with an average improvement of 40%, showing that our approach more effectively distills useful information to support human judgment. As Appendix E illustrates, the codebook-guided summaries capture nuanced insights on the potential deterrants for certain minority groups, which are often absent or less explicit without codebook guidance.

7 Prediction Experiments

7.1 Experiment Settings

We compare our model against five baselines. The first baseline is an MLP making use of only the population information. The other four baselines process the social media content with a frozen pre-trained text embedding model, and fuse it with population information. We adopt four widely-used powerful embedding models: GTE-base [17], BERT-base [9], GloVE-330B [25] and Qwen3-Embedding-8B [38]. The structures of the population and embedding MLPs are identical to those in our model. We use four metrics to evaluate model performances: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2).

We employ GPT-4o for reflective coding on a small subset (n=190) of POIs, and GPT-4o-mini for codebook-guided reasoning on all POIs for economic considerations. All models are trained using the Adam optimizer to minimize the MSE loss, with hyperparameters (e.g., learning rate, weight decay) tuned based on validation performances. POIs in each city are randomly split into training, validation, and test sets with a ratio of 6:2:2. For further implementation details, please refer to Appendix C and our released code[5]. All LLM prompts used in this work are also provided in the repository.

7.2 Experiment Results

Table 2 presents the performances of our RE'EM model alongside four baselines, with results reported as the mean and standard deviation over five repetitions on 50% randomly sampled test sets. A key insight is that **all models incorporating social media data**

[5] https://anonymous.4open.science/r/REEM-73DB/

Table 3: Ablation study of RE'EM components. Using a single channel (row 1-3) or removing the codebook (row 4) leads to performance degradation. The full model consistently achieves the best results, highlighting the complementary contributions of reasoning, embedding, and population features.

Model	MSE↓	RMSE↓	MAE↓	R^2↑
Population	0.0075±0.0002	0.0864±0.0012	0.0683±0.0013	0.3164±0.0204
Embedding	0.0096±0.0002	0.0980±0.0011	0.0814±0.0008	0.1118±0.0052
Rating	0.0104±0.0003	0.1021±0.0013	0.0837±0.0010	0.0384±0.0117
w/o Codebook	0.0072±0.0002	0.0849±0.0016	0.0685±0.0018	0.3499±0.0219
Full Model	**0.0068±0.0002**	**0.0823±0.0021**	**0.0662±0.0021**	**0.3885±0.0278**

Table 4: Cross-city generalization performances using codebook derived from Philadelphia. RE'EM outperforms all baselines across Tucson, Tampa, and New Orleans, demonstrating strong transferability.

City	POI Num	Model	MSE↓	RMSE↓	MAE↓	R^2↑
Tucson	2,221	GTE	0.0048±0.0001	0.0692±0.0011	0.0565±0.0009	0.2117±0.0134
		BERT	0.0049±0.0002	0.0698±0.0012	0.0572±0.0011	0.1983±0.0165
		GloVE	0.0047±0.0002	0.0684±0.0011	0.0560±0.0011	0.2287±0.0175
		Qwen3-Embedding-8B	0.0042±0.0001	0.0645±0.0012	0.0532±0.0010	0.3147±0.0244
		RE'EM	**0.0039±0.0001**	**0.0625±0.0009**	**0.0506±0.0011**	**0.3744±0.0290**
Tampa	2,703	GTE	0.0049±0.0003	0.0697±0.0018	0.0578±0.0018	0.2905±0.0300
		BERT	0.0052±0.0002	0.0719±0.0016	0.0594±0.0016	0.2455±0.0255
		GloVE	0.0049±0.0002	0.0704±0.0015	0.0582±0.0016	0.2762±0.0268
		Qwen3-Embedding-8B	0.0049±0.0002	0.0702±0.0017	0.0581±0.0017	0.2800±0.0240
		RE'EM	**0.0044±0.0002**	**0.0668±0.0014**	**0.0525±0.0017**	**0.3499±0.0224**
New Orleans	2,391	GTE	0.0100±0.0006	0.1002±0.0028	0.0814±0.0029	0.1958±0.0135
		BERT	0.0110±0.0006	0.1051±0.0029	0.0858±0.0030	0.1154±0.0101
		GloVE	0.0099±0.0006	0.0996±0.0029	0.0804±0.0029	0.2045±0.0209
		Qwen3-Embedding-8B	0.0103±0.0005	0.1013±0.0028	0.0827±0.0029	0.1771±0.0106
		RE'EM	**0.0087±0.0005**	**0.0929±0.0028**	**0.0745±0.0024**	**0.2781±0.0424**

outperform the population-only baseline. This result underscores the predictive power of social media content, which provides crucial insights beyond static population distributions. It supports our hypothesis that segregation patterns are shaped not only by demographic distributions but also by the social and cultural perceptions reinforcing the "invisible walls" in cities. Further, **RE'EM consistently and significantly surpasses all embedding-based baselines**. It achieves a 9.33%, 4.75%, and 3.07% improvement in MSE, RMSE, and MAE, respectively, over the population-only baseline. R^2 increases to 0.3885, marking a substantial 22.79% improvement. These results indicate that while social media data encapsulates rich and valuable information, a naive embedding strategy is insufficient to fully capture the interplay between population distribution, POI attributes, and visitor dynamics. In contrast, the structured multi-view approach of RE'EM effectively integrates heterogeneous information sources, enabling accurate predictions of segregation. We provide further analyses in Appendix D, showing that RE'EM remains stable across different base LLMs, and direct LLM prompting performs substantially worse than RE‘EM architecture.

To validate our design of different model components, we perform an ablation study. As Table 3 shows, each channel (reasoning, embedding, population) exhibits certain predictive capabilities, but none achieves optimal performance. Without the codebook, relying solely on vanilla LLM rating yields markedly worse performance. The full RE'EM model, which strategically combines all three channels and leverages the codebook for structured reasoning, consistently delivers superior results. Thus, the integration of multi-view signals enables complementary information flow, allowing the model to refine its predictions beyond what any single feature source can achieve.

To assess the generalization capability across cities, we use the codebook obtained in Philadelphia to train models in the other three cities. As shown in Table 4, RE'EM consistently achieves the best performance in all three cities along all metrics. Compared to the strongest baseline, RE'EM achieves substantial improvements, with MSE reduced by as much as 12.12%, MAE reduced by as much as 9.17% and R^2 increasing by as much as 35.99%. These findings indicate that RE'EM, and specifically, our constructed codebook, captures fundamental segregation mechanisms that extend beyond city-specific characteristics, making it a promising tool across diverse geographic and socio-economic settings.

8 Discussion

Our work highlights the transformative potential of LLMs in uncovering segregation experiences encoded in social media content. By revealing the nuanced and often invisible barriers that shape social interactions from online reviews, our work exemplifies the ethical and socially beneficial application of AI, demonstrating how advanced technologies can be harnessed to foster inclusivity and promote more equitable urban environments. Policymakers, urban planners, and community leaders can leverage these insights to better understand and address patterns of social exclusion, ultimately working toward the development of more cohesive and diverse communities.

From a technical perspective, our framework advances the application of LLMs in tackling complex societal challenges. With reflective coding and integration of LLMs' reasoning and embedding capabilities, our framework not only significantly improves prediction accuracy but extends beyond conventional predictive tasks to offer explanations and actionable insights. Furthermore, our work demonstrates how prompt-guided analytical processes can mitigate biases and hallucinations, paving the way for more reliable and socially responsible AI applications.

Our work has several limitations. Despite rigorous efforts to filter POIs with adequate review coverage, inherent imbalances in demographic representation across social media platforms may skew corpus distributions. While our work demonstrates the overall effectiveness of using social media data to predict segregation, future work could further explore how and to what extent these data biases impact prediction accuracy across different POIs. Besides, our analysis does not account for temporal variations in segregation, such as those driven by POI updates, policy shifts, or acute events like pandemics and economic crises. Future work could integrate temporal modeling to capture these dynamics.

9 Conclusion

In this paper, we pioneer the use of social media data to predict experienced segregation, designing a reflective LLM coder to generate insightful summaries and a REasoning-and-EMbedding (RE'EM) framework to integrate reasoning and embedding for accurate predictions. Our approach is validated through a qualitative user study and quantitative experiments, demonstrating its effectiveness in producing human-comprehensible review summaries and reliable segregation predictions. This work not only advances the understanding of experienced segregation but also provides a foundation for leveraging social media data to address broader societal challenges, such as fostering inclusivity and mitigating social inequalities.

10 Ethics Statement

All datasets in this study were sourced from publicly available or academically licensed repositories, with privacy protections in place. The review data was obtained from the Yelp Open Dataset under its Data Licensing agreement. The data was collected under users' consent and fully anonymized. The Safegraph mobility data is currently accessible through the Dewey Data platform[6] for academic purposes. This data was aggregated to the CBG level by month and processed with differential privacy techniques to safeguard against individual identity leakage. The ACS demographic data is publicly available[7]. As such, no Institutional Review Board (IRB) approval was required by the authors' institutions.

[6] https://www.deweydata.io/

[7] https://www.census.gov/programs-surveys/acs/

References

[1] Meta AI. 2025. llama4. https://www.llama.com/models/llama-4/.

[2] Susan Athey, Billy Ferguson, Matthew Gentzkow, and Tobias Schmidt. 2021. Estimating experienced racial segregation in US cities using large-scale GPS data. *Proceedings of the National Academy of Sciences* 118, 46 (2021), e2026160118.

[3] Camille Zubrinsky Charles. 2003. The dynamics of racial residential segregation. *Annual review of sociology* 29, 1 (2003), 167–207.

[4] Lin Chen, Fengli Xu, Qianyue Hao, Pan Hui, and Yong Li. 2023. Getting Back on Track: Understanding COVID-19 Impact on Urban Mobility and Segregation with Location Service Data. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 126–136.

[5] Zhilong Chen, Peijie Liu, Jinghua Piao, Fengli Xu, and Yong Li. 2024. Shorter Is Different: Characterizing the Dynamics of Short-Form Video Platforms. *arXiv preprint arXiv:2410.16058* (2024).

[6] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 9993–10001.

[7] Àlex G de la Prada and Mario L Small. 2024. How people are exposed to neighborhoods racially different from their own. *Proceedings of the National Academy of Sciences* 121, 28 (2024), e2401661121.

[8] DeepSeek-AI. 2025. DeepSeek-V3.2-Exp: Boosting Long-Context Efficiency with DeepSeek Sparse Attention.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:52967399

[10] Masoomali Fatehkia, Dan O'Brien, and Ingmar Weber. 2019. Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas. *PloS one* 14, 2 (2019), e0211350.

[11] Eric Gilbert, Saeideh Bakhshi, Shuo Chang, and Loren Terveen. 2013. " I need to try this"? a statistical overview of pinterest. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2427–2436.

[12] Ian Gordon and Vassilis Monastiriotis. 2006. Urban size, spatial segregation and inequality in educational outcomes. *Urban Studies* 43, 1 (2006), 213–236.

[13] Waleed Iqbal, Vahid Ghafouri, Gareth Tyson, Guillermo Suarez-Tangil, and Ignacio Castro. 2023. Lady and the Tramp Nextdoor: Online Manifestations of Real-World Inequalities in the Nextdoor Social Network. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 399–410.

[14] A Thomas King and Peter Mieszkowski. 1973. Racial discrimination, segregation, and the price of housing. *Journal of political economy* 81, 3 (1973), 590–606.

[15] Michael R Kramer and Carol R Hogue. 2009. Is segregation bad for your health? *Epidemiologic reviews* 31, 1 (2009), 178–194.

[16] Lingyao Li, Songhua Hu, Yinpei Dai, Min Deng, Parisa Momeni, Gabriel Laverghetta, Lizhou Fan, Zihui Ma, Xi Wang, Siyuan Ma, et al. 2024. Toward satisfactory public accessibility: A crowdsourcing approach through online reviews to inclusive urban design. *arXiv preprint arXiv:2409.08459* (2024).

[17] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281* (2023).

[18] Yiwei Luo, Kristina Gligorić, and Dan Jurafsky. 2024. Othering and Low Status Framing of Immigrant Cuisines in US Restaurant Reviews and Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 985–998.

[19] Douglas S Massey, Gretchen A Condran, and Nancy A Denton. 1987. The effect of residential segregation on black social and economic well-being. *Social Forces* 66, 1 (1987), 29–56.

[20] Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M Danforth. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one* 8, 5 (2013), e64417.

[21] Esteban Moro, Dan Calacci, Xiaowen Dong, and Alex Pentland. 2021. Mobility patterns are associated with experienced income segregation in large US cities. *Nature communications* 12, 1 (2021), 4633.

[22] Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, and Ming Wen. 2016. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR public health and surveillance* 2, 2 (2016), e5869.

[23] Hamed Nilforoshan, Wenli Looi, Emma Pierson, Blanca Villanueva, Nic Fishman, Yiling Chen, John Sholar, Beth Redbird, David Grusky, and Jure Leskovec. 2023. Human mobility networks reveal increased segregation in large cities. *Nature* 624, 7992 (2023), 586–592.

[24] Robert E. Park and Ernest W. Burgess. 1925. *Understanding Policy-Based Networking* (1st ed.). The University of Chicago Press, Chicago, IL.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics,

Doha, Qatar, 1532–1543. doi:10.3115/v1/D14-1162

[26] Daniele Rama, Yelena Mejova, Michele Tizzoni, Kyriaki Kalimeri, and Ingmar Weber. 2020. Facebook ads as a demographic tool to measure the urban-rural divide. In *Proceedings of The Web Conference 2020*. 327–338.

[27] Frances A Santos, Thiago H Silva, and Leandro A Villas. 2024. REAL-UP: Urban Perceptions From LBSNs Helping Moving Real-Estate Market to the Next Level. In *Companion Proceedings of the ACM on Web Conference 2024*. 1071–1074.

[28] Thomas C Schelling. 2006. *Micromotives and macrobehavior*. WW Norton & Company.

[29] Andrew J Stier, Kathryn E Schertz, Nak Won Rim, Carlos Cardenas-Iniguez, Benjamin B Lahey, Luís MA Bettencourt, and Marc G Berman. 2021. Evidence and theory for lower rates of depression in larger US urban areas. *Proceedings of the National Academy of Sciences* 118, 31 (2021), e2022472118.

[30] Robert H Tai, Lillian R Bentley, Xin Xia, Jason M Sitt, Sarah C Fankhauser, Ana M Chicas-Mosier, and Barnas G Monteith. 2024. An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods* 23 (2024), 16094069241231168.

[31] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2017. Graph Attention Networks. *ArXiv* abs/1710.10903 (2017). https://api.semanticscholar.org/CorpusID:3292002

[32] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. 2018. Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences* 115, 30 (2018), 7735–7740.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[34] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. 2014. Twitter and society: An introduction. *Twitter and society [Digital Formations, Volume 89]* (2014), xxix–xxxviii.

[35] Fengli Xu, Yong Li, Huandong Wang, Pengyu Zhang, and Depeng Jin. 2016. Understanding mobile traffic patterns of large scale cellular towers in urban environment. *IEEE/ACM transactions on networking* 25, 2 (2016), 1147–1161.

[36] Takahiro Yabe, Bernardo García Bulle Bueno, Xiaowen Dong, Alex Pentland, and Esteban Moro. 2023. Behavioral changes during the COVID-19 pandemic decreased income diversity of urban encounters. *Nature communications* 14, 1 (2023), 2310.

[37] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

[38] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

A Variability of POI features

Taking Philadelphia as an example, Figure 5 illustrates the distributions of the variations for POI features within a single CBG, including stars, prices, income segregation, and racial segregation. We can observe that POIs exhibit considerable variability even when examined within the CBG locality, which underscores the significance of our study.

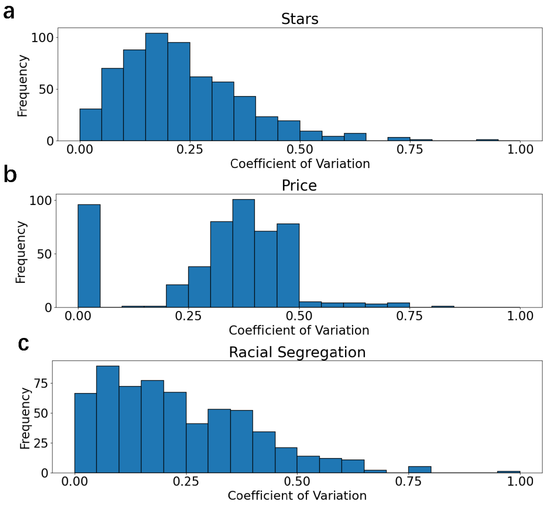


Figure 5: The Coefficient of Variation distributions for POIs features within the same CBG highlight considerable variability among POIs. Figures a-c demonstrate the distributions of stars, price range, and racial segregation, respectively.

B Reflective Coding Algorithm

We formalize the process of reflective coding in Algorithm 1.

C Implementation Details

To obtain a representative subset for the coding process, we categorize all POIs into 36 types base on different combinations of visitor-residential composition gaps (e.g., higher *black* and *asian* ratios with lower *white*, *hispanic*, and *other* ratios) and perform stratified sampling within each type to ensure diversity.

For the *reasoning* channel, we train an MLP with hidden dimensions of 512, 128, and 64, removing the output layer to form the Rating Adapter. For the *embedding* channel, we fine-tune the last two layers of the pre-trained GTE-base model and extend it with three fully connected layers (512, 256, 128) to form the Embedding Adapter. For the *population* channel, we train an MLP with dimensions of 100, 30, and 10, extracting the weights of the first two layers to form the Population Adapter. For the *neighbor-aware multi-view predictor*, we adopt Graph Attention Networks (GATs) [31] as the neighbor aggregators. The Reasoning GAT maps feature from 64 to 128, the Embedding GAT maintains a 128 to 128 mapping, and the Population GAT maintains a 30 to 30 mapping. The multi-view fusion module comprises three fully connected layers with dimensions 512, 128, and 64. During training, we first optimize the three Adapters, then freeze them before training the predictor.

Algorithm 1 Reflective Coding Algorithm

```
Require: Set of POIs $\mathcal{I}$ with for each POI $i \in \mathcal{I}$:
    - Social media content $c_i = (\mathcal{N}_i, \{\text{review}_k, \text{image}_k\}_{k=1}^{m_i})$
    - Observed visitor composition $\boldsymbol{\tau}_i = [\tau_{iq}]_{q=1}^{Q}$ (Q racial groups)
    - Local residential composition $\mathbf{p}_i = [p_{iq}]_{q=1}^{Q}$
Ensure: Codebook $\mathcal{B} = \{(d_j, e_j)\}_{j=1}^{M}$ with dimensions $d_j$ and explanations $e_j$
// Phase 1: Reflective Attributor
Initialize insight set $\mathcal{L} \leftarrow \emptyset$
for each POI $i \in \mathcal{I}$ do
    // Multi-source insight generation
    $\mathcal{L}_i^{\text{name}} \leftarrow \text{LLM}_{\text{analyze}}(\mathcal{N}_i)$    ▷ Extract name signals
    $\mathcal{L}_i^{\text{text}} \leftarrow \text{LLM}_{\text{analyze}}(\{\text{review}_k\})$    ▷ Analyze textual reviews
    $\mathcal{L}_i^{\text{img}} \leftarrow \text{LLM}_{\text{analyze}}(\{\text{image}_k\})$    ▷ Analyze visual content
    $\mathcal{L}_i^{\text{sum}} \leftarrow \text{LLM}_{\text{synthesize}}(\mathcal{L}_i^{\text{name}}, \mathcal{L}_i^{\text{text}}, \mathcal{L}_i^{\text{img}})$
    // Reflection-guided refinement via abductive reasoning
    $\Delta\boldsymbol{\tau}_i \leftarrow \boldsymbol{\tau}_i - \mathbf{p}_i$    ▷ Compute demographic discrepancy
    $\mathcal{L}_i^{\text{refined}} \leftarrow \text{LLM}_{\text{reflect}}(\mathcal{L}_i^{\text{sum}}, \Delta\boldsymbol{\tau}_i)$
    $\mathcal{L}[i] \leftarrow \mathcal{L}_i^{\text{refined}}$
end for
// Phase 2: Code Summarizer
Partition $\mathcal{L}$ into $K$ stratified subsets: $\{\mathcal{L}_{\text{sub}}^{(k)}\}_{k=1}^{K}$
Initialize codebook $\mathcal{B} \leftarrow \emptyset$
for each subset $\mathcal{L}_{\text{sub}}$ do
    // Extract candidate codes
    $\mathcal{C}_{\text{sub}} \leftarrow \emptyset$
    for each insight $\ell \in \mathcal{L}_{\text{sub}}$ do
        $(c, e) \leftarrow \text{LLM}_{\text{extract}}(\ell)$
        $\mathcal{C}_{\text{sub}} \leftarrow \mathcal{C}_{\text{sub}} \cup \{(c, e)\}$
    end for
    // Iterative refinement on combined set
    $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{C}_{\text{sub}}$ ▷ Merge candidates with existing codebook
    $\mathcal{B} \leftarrow \text{LLM}_{\text{semantic}}(\mathcal{B})$ ▷ LLM merges redundant dimensions
    $\mathcal{B} \leftarrow \text{LLM}_{\text{quality}}(\mathcal{B})$    ▷ LLM filters low-generalizability codes
    $\mathcal{B} \leftarrow \text{LLM}_{\text{refine}}(\mathcal{B})$ ▷ LLM generates thematic descriptions
end for
return $\mathcal{B}$
```

We conduct a compact grid search on the Philadelphia dataset, exploring learning rates in $[5 \times 10^{-6}, 5 \times 10^{-3}]$, weight decays in $[10^{-5}, 5 \times 10^{-4}]$, early-stopping patience between 10 and 25. The width of every MLP hidden layer is tuned by doubling or halving around the final choices (approximately 64–512 units). The configuration achieving the lowest validation MSE is selected and is released as the default value in our open-source code.

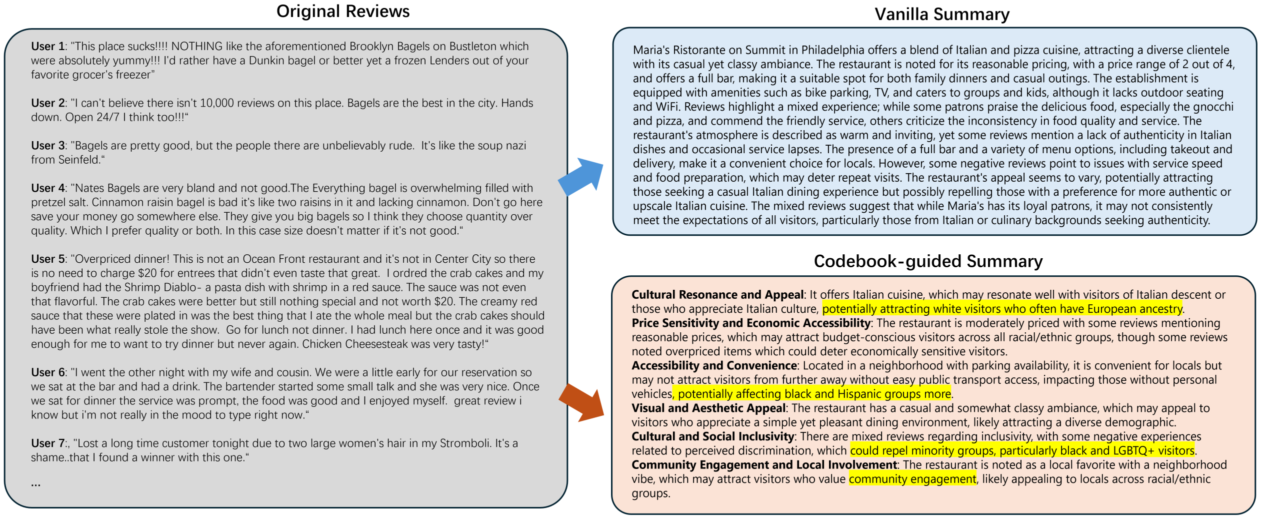
All experiments run on a Linux server with an Intel Xeon Platinum 8358 CPU, 4×RTX 4090 and 2×RTX 3090 GPUs (CUDA 12.4), and 503 GB RAM.

Table 5: Evaluation of different base LLMs in the RE'EM framework. Replacing GPT-4o-mini with other SOTA LLMs yields nearly identical results, showing that RE'EM is robust to the choice of base model.

Base LLM	MSE↓	RMSE↓	MAE↓	R^2↑
GPT-4o-mini	0.0068±0.0003	0.0823±0.0021	0.0662±0.0021	0.3885±0.0278
Qwen-plus	0.0068±0.0003	0.0822±0.0021	0.0661±0.0021	0.3905±0.0289
Llama4-scout	0.0068±0.0004	0.0823±0.0022	0.0660±0.0021	0.3896±0.0305
DeepSeek-v3.2-exp	0.0068±0.0004	0.0823±0.0022	0.0662±0.0021	0.3883±0.0301

Table 6: Comparison with direct LLM prompting baseline. Removing RE'EM pipeline leads to severe performance degradation, confirming the necessity of RE'EM.

Model	MSE↓	RMSE↓	MAE↓	R^2↑
Qwen-plus	0.1281±0.0039	0.3579±0.0054	0.3202±0.0068	-10.7420±0.4472
Llama4-scout	0.0586±0.0020	0.2420±0.0043	0.1916±0.0042	-4.3700±0.2091
DeepSeek-v3.2-exp	0.0689±0.0022	0.2625±0.0042	0.2127±0.0030	-5.3191±0.2724
RE'EM	**0.0068±0.0003**	**0.0823±0.0021**	**0.0662±0.0021**	**0.3885±0.0278**



Original Reviews

User 1: "This place sucks!!!! NOTHING like the aforementioned Brooklyn Bagels on Bustleton which were absolutely yummy!!! I'd rather have a Dunkin bagel or better yet a frozen Lenders out of your favorite grocer's freezer"

User 2: "I can't believe there isn't 10,000 reviews on this place. Bagels are the best in the city. Hands down. Open 24/7 I think too!!!"

User 3: "Bagels are pretty good, but the people there are unbelievably rude. It's like the soup nazi from Seinfeld."

User 4: "Nates Bagels are very bland and not good.The Everything bagel is overwhelming filled with pretzel salt. Cinnamon raisin bagel is bad it's like two raisins in it and lacking cinnamon. Don't go here save your money go somewhere else. They give you big bagels so I think they choose quantity over quality. Which I prefer quality or both. In this case size doesn't matter if it's not good."

User 5: "Overpriced dinner! This is not an Ocean Front restaurant and it's not in Center City so there is no need to charge $20 for entrees that didn't even taste that great. I ordred the crab cakes and my boyfriend had the Shrimp Diablo- a pasta dish with shrimp in a red sauce. The sauce was not even that flavorful. The crab cakes were better but still nothing special and not worth $20. The creamy red sauce that these were plated in was the best thing that I ate the whole meal but the crab cakes should have been what really stole the show. Go for lunch not dinner. I had lunch here once and it was good enough for me to want to try dinner but never again. Chicken Cheesesteak was very tasty!"

User 6: "I went the other night with my wife and cousin. We were a little early for our reservation so we sat at the bar and had a drink. The bartender started some small talk and she was very nice. Once we sat for dinner the service was prompt, the food was good and I enjoyed myself. great review i know but i'm not really in the mood to type right now."

User 7:, "Lost a long time customer tonight due to two large women's hair in my Stromboli. It's a shame..that I found a winner with this one."

...

Vanilla Summary

Maria's Ristorante on Summit in Philadelphia offers a blend of Italian and pizza cuisine, attracting a diverse clientele with its casual yet classy ambiance. The restaurant is noted for its reasonable pricing, with a price range of 2 out of 4, and offers a full bar, making it a suitable spot for both family dinners and casual outings. The establishment is equipped with amenities such as bike parking, TV, and caters to groups and kids, although it lacks outdoor seating and WiFi. Reviews highlight a mixed experience; while some patrons praise the delicious food, especially the gnocchi and pizza, and commend the friendly service, others criticize the inconsistency in food quality and service. The restaurant's atmosphere is described as warm and inviting, yet some reviews mention a lack of authenticity in Italian dishes and occasional service lapses. The presence of a full bar and a variety of menu options, including takeout and delivery, make it a convenient choice for locals. However, some negative reviews point to issues with service speed and food preparation, which may deter repeat visits. The restaurant's appeal seems to vary, potentially attracting those seeking a casual Italian dining experience but possibly repelling those with a preference for more authentic or upscale Italian cuisine. The mixed reviews suggest that while Maria's has its loyal patrons, it may not consistently meet the expectations of all visitors, particularly those from Italian or culinary backgrounds seeking authenticity.

Codebook-guided Summary

Cultural Resonance and Appeal: It offers Italian cuisine, which may resonate well with visitors of Italian descent or those who appreciate Italian culture, potentially attracting white visitors who often have European ancestry.
Price Sensitivity and Economic Accessibility: The restaurant is moderately priced with some reviews mentioning reasonable prices, which may attract budget-conscious visitors across all racial/ethnic groups, though some reviews noted overpriced items which could deter economically sensitive visitors.
Accessibility and Convenience: Located in a neighborhood with parking availability, it is convenient for locals but may not attract visitors from further away without easy public transport access, impacting those without personal vehicles, potentially affecting black and Hispanic groups more.
Visual and Aesthetic Appeal: The restaurant has a casual and somewhat classy ambiance, which may appeal to visitors who appreciate a simple yet pleasant dining environment, likely attracting a diverse demographic.
Cultural and Social Inclusivity: There are mixed reviews regarding inclusivity, with some negative experiences related to perceived discrimination, which could repel minority groups, particularly black and LGBTQ+ visitors.
Community Engagement and Local Involvement: The restaurant is noted as a local favorite with a neighborhood vibe, which may attract visitors who value community engagement, likely appealing to locals across racial/ethnic groups.

Figure 6: Case study comparing the vanilla summary and codebook-guided summary for the same venue. The codebook-guided summary captures more nuanced factors influencing demographic appeal, which are absent in the vanilla summary, illustrating the interpretability and analytical advantage of structured reasoning.

D Additional Evaluation on Different LLMs

To further examine whether RE'EM depends on a particular language model, we conduct two additional evaluations involving both open- and closed-source state-of-the-art LLMs. First, we replace the base model (GPT-4o-mini) used in the reasoning channel with three strong alternatives: Qwen-plus [37], Llama4-scout [1], and DeepSeek-v3.2-exp [8], keeping all other components unchanged. As shown in Table 5, the performance across all metrics remains nearly identical, indicating that RE'EM is highly robust to the choice of the base LLM.

Second, we assess whether the performance gain comes from the RE'EM architecture rather than the inherent ability of LLMs. We directly prompt the same LLMs to predict segregation from social media content and population distribution, removing the RE'EM pipeline. Table 6 shows a dramatic drop in performance for all models, confirming that naive LLM prompting is insufficient and that the structured multi-channel design of RE'EM is essential for reliable quantitative prediction.

E Case study

In Figure 6, we present a case study to compare the original reviews, the vanilla method-generated summary, and our codebook-guided summary.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009