Gendered Perceptions of Places in LLMs vs. Human Reviews

Whose Perspective Does AI Reflect?

MSc Data Science
Department of Computer Science
Middlesex University of London

Kostanca Kovaci

Supervisor: Dr.Giovanni Quattrone

A thesis submitted in fulfillment of the requirements
for the degree of Data Science
in the
Middlesex University

Hendon, July 2025

Abstract

This dissertation investigates whether large language models (LLMs), specifically GPT-5 and Claude Sonnet 4, reflect gendered perceptions of safety in London neighborhoods when compared to human-authored Airbnb reviews and actual crime statistics. The study examines if LLMs favor specific gendered perspective over another and whether their safety assessments align with real-world experiences. Findings reveal that LLMs often exaggerate risk and introduce systematic gender bias, particularly labeling female-associated prompts as "safe" more frequently than male ones, a pattern absent in Airbnb data. Unlike human reviews, which align more closely with crime statistics, LLM outputs demonstrate weak connections to ground truth. These findings outstand the implications of using LLMs for social intentions, such as travel suggestions, where biased or misrepresented outputs could reinforce stereotypes and affect communities. The dissertation contributes to AI fairness research by raising awareness of gendered and geographical biases in LLMs and highlights the need for responsible, transparent and fair AI deployment. This research is from the few ones to examine gendered safety perceptions in AI-generated neighborhood descriptions, addressing a gap in existing literature.

Content Table

List of Figures

List of Tables

1 Introduction

1.1 Background

Artificial Intelligence (AI) models have rapidly increased in recent years, changing the way people nowadays search for information and make decisions. Large Language Models (LLMs), such as GPT-5 and Claude Sonnet 4, are increasingly used as tools for decision-making, answering questions and generating travel suggestions. While these systems are distinguished for their accessibility and comprehensibility, their training on vast and boundless amount on unfiltered datasets raises vital concerns. Training data may include gendered stereotypes and geographical biases, which LLMs risk reproducing or amplifying, influencing users perceptions. This is especially problematic in socially sensitive domains, such as neighborhood safety, where misrepresentation can affect user trust and community perceptions.

1.2 Problem Statement

Although LLMs are powerful and useful, their outputs often lack personalization, fairness and alignment with real-world conditions. This investigation shows that LLM-generated neighborhood descriptions exaggerate risks and exhibit systematic gender bias. For instance, LLMs label neighborhoods as 'safe' for female-travelers more frequently than for male-travelers, an disparity detail that is absent on human-authored Airbnb reviews. Additionally, while Airbnb reviews align more closely with actual crime statistics, LLM outputs show weaker correlation to ground truth. These findings highlight the risks of relying on AI-generated content for travel decisions. In that way, LLMs may reinforce gender stereotypes, distort perceptions of urban safety, and misrepresent the lived experiences of minority groups.

1.3 Aim and Objectives

Aim of this dissertation is to investigate whether LLM-generated reviews reflect gendered safety nuances similar or different from human-authored reviews and actual crime data. Specifically, this study compares whether LLMs tend to favor specific gendered perspectives over others and investigate whether they are sensitive to gendered safety concerns. Explores how LLMs describe London neighborhoods across male, female and neutral traveler prompts and evaluates the alignment of LLM and Airbnb content with official crime statistics. Exploring the implications for fairness and trust in AI systems.

1.4 Research Questions

Based on the investigation stated, the following research questions have been developed to address our researches accomplishment:

Q1: Do LLMs reflect gendered nuances in neighborhood descriptions, and how are they compared to human-authored reviews?

Q2: To what extend are LLMs sensitive to gendered safety concerns when recommending London neighborhoods, and how this align with crime data and human-perceptions?

Q3: Do LLMs recommendations favor specific gendered perspectives over another, and what

are the implications on travel content?

1.5 Contribution

This dissertation contributes to AI fairness, by investigating gendered safety perceptions in LLM-generated neighborhood descriptions. It reveals that LLMs amplify and misrepresent gendered perspectives, exaggerating safety differences between genders, an absent detail in human-authored reviews. In exploration of this project, Airbnb reviews exhibited stronger alignment with crime statistics, while LLM outputs diverge. This finding underlines the risks of biased AI suggestions and decisions in socially influenced domains. The insights raise awareness among AI designers, researchers and policymakers of the emergent need for bias detection and mitigation strategies. Moreover, this research introduces the transparent and responsible deployment needed in LLMs, ensuring that AI systems provide equitable and trustworthy information.

1.6 Dissertation Structure

The dissertation is structured as follows:

- **Introduction** - Developing the background, resolution, aim of the project, research questions and the contribution of this study.
- **Literature Review** - LLMs behavior in gender bias, travel recommendation, occupation bias, gender characterization and research gaps.
- **Methodology** - Data collection, preprocessing, feature-engineering, LLM prompts and analytical methods.
- **Results** - Findings and insights from LLM-generated, Airbnb human reviews and actual Crime statistics, highlighting gendered patterns and safety sensitivity.
- **Discuss** - Results and findings interpretation, discuss of research questions, AI fairness and implications.
- **Conclusion** - Summaries of findings, contributions, gaps, limitations and directions for future research.

2 Literature Review

2.1 Background

Large Language Models (LLMs) are investigated on several domains, including **biases across gender, occupation** and **geography**. Multiple investigations inform that LLMs can reinforce inherited social stereotypes, in occupational context, where certain professions are described as 'female-associated' (e.g., *nurse, fashion manager*) or 'male-associated' (e.g., *doctor, pilot*). [5][8][16]. Similarly, other studies demonstrated that LLMs prefer to describe females with 'supportive' traits (e.g. *'supportive'*, *'good to work with'*), instead of the 'powerful' and 'dominant' descriptions on males (e.g. *'dominant'*, *'stand out in the industry'*) [9][17][18].

LLMs also exhibit **geographical and cultural biases** in their outcomes, often favoring Western-centric concepts and popular places, while marginalizing less-represented regions [6][10][11]. Such biases and disparities emerge through unbalanced training data in LLMs, leading to over-representation of privileged places and under-representation of others. This can develop real-world consequences and influence humans opinion about specific locations. For instance, places affected by disasters may be underrepresented, influencing tourism and investment decisions [4].

Research on human-authored reviews, such as those on Airbnb, reveals similar patterns. Studies indicate gendered dynamics in trust and safety perceptions, with male hosts often receiving higher ratings and users showing preferences for hosts of the same gender [3]. Women, in particular, report stronger safety concerns when choosing travel destinations, with crime perceptions influencing their willingness to travel [15]. While these insights are important, relatively few studies examine whether LLMs reflect such gendered safety concerns in neighborhood descriptions, leaving a critical gap.

Related studies on human-authored reviews, such as Airbnb, illustrate gendered dynamics in trust and safety perceptions, in users descriptions. Studies show that male hosts often receive higher-rated reviews and users tend to prefer accommodations of hosts of the same gender [3]. In particular, women report stronger safety concerns when choosing travel destinations, leaving potential crime perceptions influencing their desire to travel [15]. Although these findings are important, relatively few studies examine how LLMs reflect gendered perspectives and safety concerns in neighbourhood suggestions, leaving a critical gap.

2.2 Methodologies in Related Work

Previous related studies of gender and geographical biases have displayed multiple approaches on how those biases are explored and exhibited, including:

- **Prompt-based experiments** comparing LLM-generated outcomes with human-authored content [2][5][16].
- **Lexical and sentiment analysis** to identify stereotypical language biases in text [5][17].
- **Benchmark datasets**, like WinoBias and WinoGender, to test and measure gender bias in LLMs text [8].
- **Bias mitigation strategies**, including fine-tuning, debias tuning and hyperparameter adjustment methods [5][9].

These approaches have provided valuable insights, but most of them focus on occupations or general textual biases. While, few studies address gendered perceptions of places or safety concerns, specifically in the context of travel suggestions.

2.3 Key Contributions & Comparisons

This project equips important contributions, by investigating a prior work, addressing an underexplored research gap. This gap refers to LLMs behavior and role in reflecting gendered perceptions and safety concerns in neighbourhood descriptions. Most studies focus on the exploration of gender and geographical bias in isolation, but this project investigates LLMs generated outputs with human-authored reviews and crime statistics in gender and geographical combination. Moreover, by examining how LLMs differ in the way they reflect gendered perceptions of places, our research aware for the included biases in LLM-suggestions and contribute on understanding sensitive safety concerns. This project addresses interactions between gender and geography in LLMs, providing a rich evaluation of whether LLMs differ or resemble from human perceptions and real-world data.

2.4 Strengths & Limitations of Related Work

Existing investigations examine several strengths of LLMs for gender and geographic biases, studied in fields of occupation, generated text, place recommendations and review analysis. These studies offered strong background of quantitative approaches like: prompt-based testing, sentiment analysis, word embeddings, etc, providing essential and robust tools for measuring bias [5]. Mitigation strategies in related work, including debias and data augmentation, provide practical approaches and guidance for improving LLMs behavior [9][18]. Cross-domain investigations further highlight the extensive implications of LLMs and the potential social biases across contexts, from occupations to travel [5][15]. Moreover except the important strengths that were found in related studies, some important limitations still remain. Several studies focus on occupations or textual biases of LLMs, leaving gendered perceptions of urban areas unexplored [1][15]. Methods like template-based prompting, while common, can miss implicit biases or complex linguistic indicators, such as human sarcasm [16]. Moreover, Western-centric training limit cultural and geographical generalization [8]. Additionally, most studies address gender and geography exploration separately, ignoring their intersection, which is critical in travel suggestions, where perceptions of safety are included.

2.5 Research Gaps

Despite the notable progress in bias mitigation and measurement, several critical key gaps remain. For example, lack of studies in how LLMs reflect gendered perceptions of neighbourhood safety, or how LLMs align with human-authored reviews and crime data. Direct comparison between these sources in rare and develops and important gap in research domain. Moreover, the intersection of gender and geographical bias, still remains underexplored, specifically for LLM travel suggestions. This project aims to address these gaps by investigating how LLMs reflect male and female perspectives in London neighborhoods, evaluating alignment with human reviews and actual crime data, and assessing whether LLMs favor on reflecting one gender over another.

3 Methodology

3.1 Overview

This research methodology focuses and investigates on how LLMs-generated travel suggestion content illustrates gendered perceptions of London's neighbourhoods and whether they favor specific gender perspectives over another. It applies NLP approaches -like sentiment analysis, topic modeling and safety lexical content- to explore emotional patterns and topic distinctions between LLM-generated content and human-authored reviews. By comparing inferred Airbnb human reviews with LLM generated-content and actual crime statistics, his study identifies potential biases associated with gender and neighbourhood terminology. Establishing the framework for exploring the intersection of gendered language, perceived safety, and AI-generated content, forming the foundation for the subsequent methodological steps in data collection, preprocessing and analysis.

3.2 Data Sources

This Section demonstrates an understanding of the dataset that has been used to implement project's investigation and the steps included to answer the research questions. The project employs three main datasets.

3.2.1 Airbnb Dataset

Human-authored Airbnb reviews provide vital information for comprehending authentic perceptions of neighborhoods from a gendered perspective. It helps identify how different gender groups experience safety within London neighbourhoods. By comparing LLM-generated content against Airbnb reviews, the analysis identifies discrepancies between AI-mediated interpretations and real-world human experiences.

The Airbnb dataset, taken from *Inside-Airbnb.com* [7], specifically for London, contains two key files:

- **reviews.csv.gz**

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	13913	80770	2010-08-18	177109	Michael	My girlfriend and I hadn't known Alina before ...
1	13913	367568	2011-07-11	19835707	Mathias	Alina was a really good host. The flat is clea...
2	13913	529579	2011-09-13	1110304	Kristin	Alina is an amazing host. She made me feel rig...
3	13913	595481	2011-10-03	1216358	Camilla	Alina's place is so nice, the room is big and ...
4	13913	612947	2011-10-09	490840	Jorik	Nice location in Islington area, good for shor...

Figure 1: Structure of Airbnb reviews dataset

- **listings.csv.gz**

	id	neighbourhood_cleansed	latitude	longitude
0	264776	Lewisham	51.44306	-0.01948
1	264777	Lewisham	51.44284	-0.01997
2	264778	Lewisham	51.44359	-0.02275
3	264779	Lewisham	51.44355	-0.02309
4	264780	Lewisham	51.44333	-0.02307

Figure 2: Structure of Airbnb listings dataset

3.2.2 Crime Rate Dataset

Official crime statistics were collected from UK Home Office portal **data.police.uk** [14], providing reliable and ground-truth references for safety comparison across London neighbourhoods. Crime data are vital for the evaluation on whether LLM safety statements align or not with actual crime incidents.

	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type
0	NaN	2024-05	Metropolitan Police Service	Metropolitan Police Service	0.140127	51.588913	On or near Beansland Grove	E01000027	Barking and Dagenham 001A	Anti-social behaviour
1	NaN	2024-05	Metropolitan Police Service	Metropolitan Police Service	0.145722	51.594296	On or near Providence Place	E01000027	Barking and Dagenham 001A	Anti-social behaviour
2	NaN	2024-05	Metropolitan Police Service	Metropolitan Police Service	0.140576	51.583419	On or near Rams Grove	E01000027	Barking and Dagenham 001A	Anti-social behaviour
3	NaN	2024-05	Metropolitan Police Service	Metropolitan Police Service	0.135924	51.587353	On or near Gibbfield Close	E01000027	Barking and Dagenham 001A	Anti-social behaviour
4	db4b2ce47048772fb4a60142c317b0d5ecb653eb13be0b...	2024-05	Metropolitan Police Service	Metropolitan Police Service	0.135924	51.587353	On or near Gibbfield Close	E01000027	Barking and Dagenham 001A	Burglary

Figure 3: Structure of Crime Rate dataset

3.2.3 LLM-generated prompts

Travel suggestion prompts were extracted from two LLMs: ChatGPT-5 (OpenAI) and Claude Sonnet 4 (Anthropic). To ensure reproducibility the exact prompt templates are listed below. For each prompt, the placeholders [**neighbourhood**] and [**gender**] were substituted as follows:

- **neighbourhood** = one of 21 selected London neighborhoods present in all three datasets (Airbnb, crime, LLM).
- **gender** = 'female', 'male' or 'neutral/general'.

Prompt templates:

1. 'Describe [neighbourhood] life, is it safe for a [gender] traveler?'
2. 'What is like to visit [neighbourhood] in London?'
3. 'As a [gender] traveler is [neighbourhood] good neighbourhood to stay?'

Each template was created three times per model to reduce variability, producing 315 responses per model and 630 responses in total. The same model versions were used throughout the study (GPT-5, Claude Sonnet 4), with default settings (temperature and max tokens left at system defaults).

3.3 Data Preparation

Airbnb human authored-reviews and crime datasets were preprocessed to ensure data quality, relevant and unbiased outcomes. This section implements tasks of data cleaning and preparation, transformation and feature-engineering and preparation for subsequent NLP and statistical analyses.

3.3.1 Airbnb Data Preparation

In Airbnb dataset only essential columns were retained to employ the projects objectives (Figure 2 and 3). Data preparation involved three main stages: cleaning, merging, and feature extraction.

In **Data cleaning and merging** processes duplicate entries, null values and records with missing fields were removed. The reviews and listings datasets were merged on their common 'listing_id' column, providing a combined dataset including review text, geographical coordinates and neighborhood information.

In **Feature-engineering** steps, were included:

1. **Gender extraction**. From *reviewer_name* column were extracted the first names after removing non-English characters and symbols. Applied the **Gender-Guesser** python library for classifying first names of users into: 'male', 'female' or 'unknown'. Validated accuracy using two approaches: **Gender.io API** on a 100 name sample, achieving **98.21%** accuracy, and **manually check** of the same sample, achieving **97%** accuracy.

2. **Neighbourhood Extraction**. For London neighbourhood details the *neighbourhood_cleansed* column from the original Airbnb data, includes London boroughs. We extracted ward-level neighborhoods using GeoPandas and ONS Ward boundaries (GeoJSON) [13], to create neighbourhood_wards column for fine-grained spatial analysis, containing London neighbourhoods.

3. **Text Preprocessing**. For the *comments* column, text normalization approach of links-tags removal and non-standard characters was implemented. All text was converted to lowercase English characters, ensuring compatibility with sentiment analysis, topic modeling, and semantic processing in further investigation sections.

Column	Description
listing_id	Unique identifier of the listing
id	Review identifier
date	Date of the review
reviewer_name	Name of the reviewer
comments	Original review text
latitude	Latitude of the listing
longitude	Longitude of the listing
neighbourhood_cleansed	Normalized neighbourhood name
clean_comments	Preprocessed review text
clean_reviewer_name	Cleaned reviewer name
first_name	Extracted first name
last_name	Extracted last name
gender	Inferred gender of the reviewer_name

Table 1: Info of preprocessed and cleaned Airbnb dataset.

id	first_name	gender	longitude	latitude	neighbourhood	...	clean_comments
80770	Michael	male	-0.112700	51.568610	Islington	...	my girlfriend and i...
367568	Mathias	male	-0.112700	51.568610	Islington	...	alina was a really...
529579	Kristin	female	-0.112700	51.568610	Islington	...	alina is an amazing...
595481	Camilla	female	-0.112700	51.568610	Islington	...	alinas place...

Table 2: Schema sample of transformed Airbnb dataset.

The final merged dataset contained 1,625,171 reviews, with 688 neighborhoods and 33 boroughs.

3.3.2 Crime Data Preparation

In crime dataset was employed similar preprocessing to ensure relevant compatibility with Airbnb data.

Preprocessing Steps:

- Removed records with missing coordinates or incomplete location information.
- Filtered data only to London records using borough boundaries (longitude, latitude).
- Extracted neighborhood information using the same ONS Ward GeoJSON file for consistency.
- Normalized the crime data by population category.

The processed crime dataset contained 693 neighborhoods and 33 boroughs, perfectly aligned with the Airbnb geographical boundaries. Table 3 shows the final schema, with Table 4 providing sample records.

Column	**Description**
Crime ID	Unique identifier of the crime incident
Month	Month when the crime was reported
Reported by	Police authority that reported the crime
Longitude	Longitude of the crime location
Latitude	Latitude of the crime location
Location	Street-level description of the crime location
LSOA code	Code of the Lower Super Output Area (LSOA)
LSOA name	Name of the Lower Super Output Area (LSOA)
Crime type	Category of the crime (e.g., burglary, violence, etc.)
Last outcome category	Outcome status of the crime investigation

Table 3: Info of preprocessed and cleaned Crime dataset.

Crime ID	**Month**	**Longitude**	**Latitude**	**Location**	...	**Crime type**	**Last outcome category**
db4b2ce...	2024-05	0.135924	51.587353	Gibbfield Close	...	Burglary	Investigation complete; no suspect identified
6c27050...	2024-05	0.135924	51.587353	Gibbfield Close	...	Burglary	Awaiting court outcome
1fbac9c...	2024-05	0.140194	51.582356	Hatch Grove	...	Criminal damage and arson	Unable to prosecute suspect

Table 4: Schema sample of transformed Crime dataset.

With this approaches we ensure the quality of data from both datasets. Generated consistent neighborhood label conventions, with accurate and valid coordinate ranges for London (latitude:

51.2-51.7, longitude: -0.5-0.3). Preprocessing Airbnb and crime datasets ensures relevant and efficient analyses, including gendered perception evaluation, sentiment scoring and spatial crime correlation.

3.4 Data Preprocessing

This project exhibits a variety of distinctive approaches, by employing a combination of Airbnb reviews, LLM-generated content and crime statistics. This section combines data preprocessing steps with exploratory data analysis to comprehend sources characteristics and guide further through analytical choices.

3.4.1 Gender Extraction

Gender extraction, as mentioned in previous section, was performed using the Gender-Guesser python library on first names, from reviewer_name column. The classification approach separates classes into 'male', 'female' or 'unknown'. Names labeled as 'unknown' were either short or invalid (e.g. Jo, Aan, Zzz), or sometimes neutral names.

gender	count
female	711164
male	655673
unknown	258334

Table 5: Amount of gender distribution counts included in Airbnb dataset.

The accuracy was validated by using two methods:

1. **Gender.io API** on a small sample of 100 names: 98.21% accuracy.
2. **Manual check** for the same sample: 97% precision/recall.

The dataset is dominant by female reviewers (44%), followed by males (40%) and unknown (16%), indicating reliable gender extraction for analysis. The high accuracy reveals reliable gender extraction between names and genders.

3.4.2 Neighbourhood Extraction

Airbnb Dataset

Spatial alignment across dataset sources required consistent neighborhood mapping. Using GeoPandas and ONS Ward boundaries (GeoJSON) [12], geographical coordinates were mapped to:

- **Borough level:** 33 London boroughs
- **Neighbourhood level:** 688 neighborhoods for deeper analysis

Column	Unique Values
neighbourhood_cleansed	33
neighbourhood_wards	688

Table 6: Boroughs and Neighbourhoods counts in Airbnb dataset.

	neighbourhood_cleansed	neighbourhood_wards
count	1625171	1624496
unique	33	688
top	Westminster	West End
freq	218026	37749

Table 7: Borough and Neighbourhood Airbnb info statistics.

Crime Dataset

For the **Crime statistics**, we implemented similar neighbourhood extraction approach, for efficient and accurate comparison. Crime dataset contained 827 neighborhoods and 55 boroughs, reflecting UK-wide coverage, so we implemented filtering only for London area. Dataset after mapping, included:

- **Borough level:** 33 London boroughs
- **Neighbourhood level:** 693 London neighbourhoods

Column	Unique Values
Neighbourhood	693
Boroughs	33

Table 8: Boroughs and Neighbourhoods counts in Crime dataset

3.4.3 Airbnb Data Characteristics

After the gender and neighbourhood extraction, we employ **Exploratory Data Analysis (EDA)** on Airbnb dataset to identify trends, distributions and characteristics. The processed Airbnb dataset contains **1,625,171** reviews.

Gender Distribution

- **Female:** 44% (711,164)
- **Male:** 40% (655,673)
- **Unknown:** 16% (258,334)

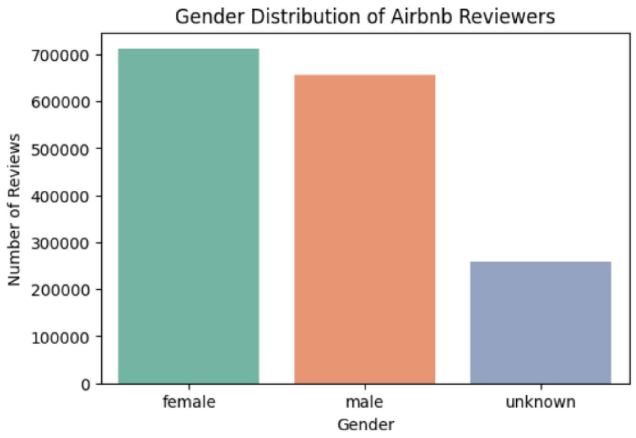


Figure 4: Shows the Distribution of reviews by gender.

Temporal Patterns of Reviews

By analyzing the temporal distribution of Airbnb reviews across years, were exhibited volume peaks in **2018-2019** and **2022-2024**, with female reviewers more active during those peak periods.

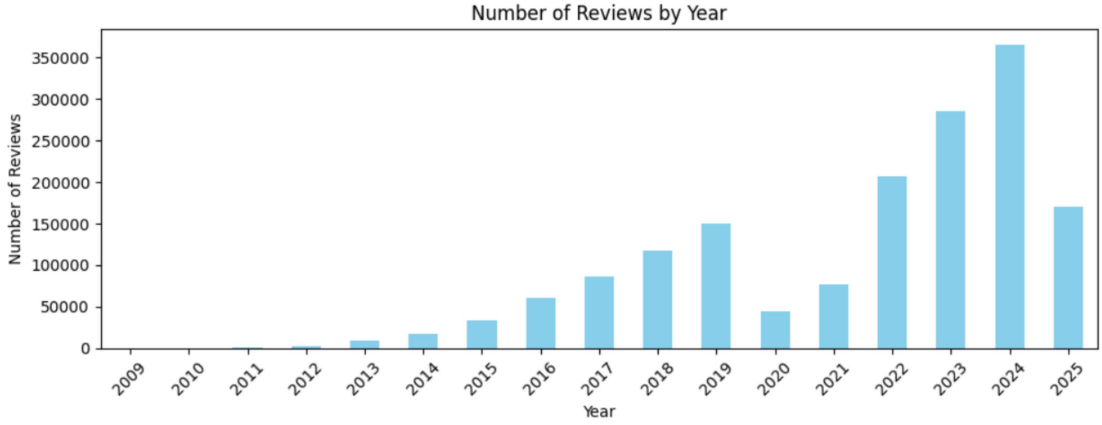


Figure 5: Reviews counts over time in Airbnb data.

The analysis reveals that for the gender dynamics over time (Figure 6):

- **Female** reviewers are more dominant in 2018, 2019 and between 2022-2024.
- **Male** reviewers show similar activity over 2015-2016, 2020-2021 and 2025.
- **Unknown** gender reviewers follow similar trends, but in lower volumes.

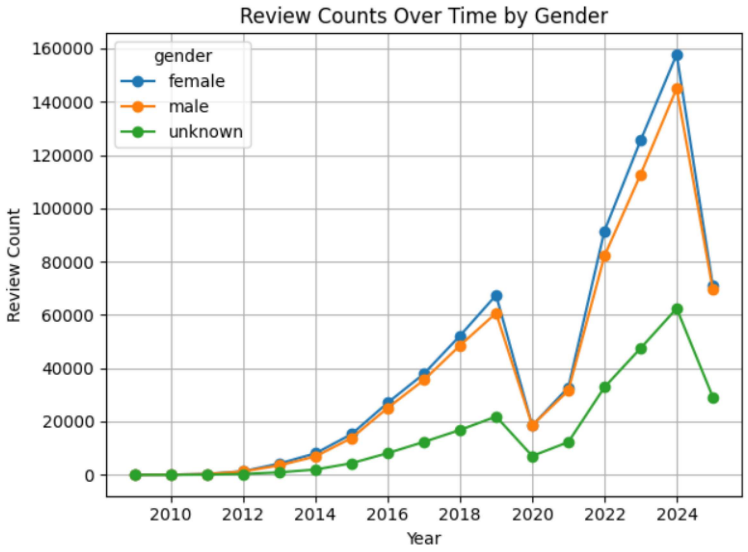


Figure 6: Reviews Count over time by gender.

This exploration provides context for gender-specific experiences and highlights periods when gender bias in the dataset might influence results.

Content Analysis

Most frequent words used in Airbnb, included: 'great', 'stay', 'location', indicating focus on accommodation quality and neighborhood features rather than detailed safety discussions. Figure 7 presents the 30 most frequently used words across all reviews.

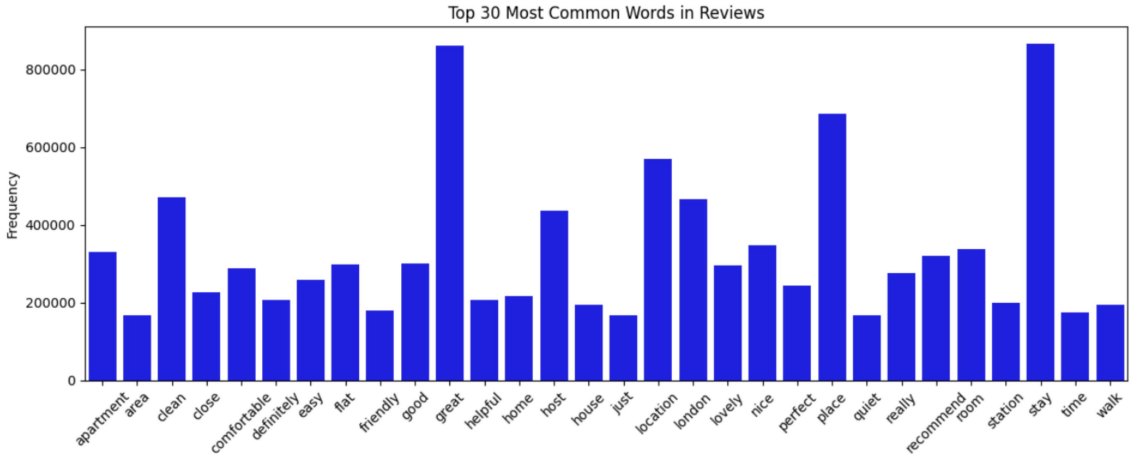


Figure 7: Most frequent Words included in Airbnb reviews

These key-words reveal what reviewers seem to describe in their review-experiences, such as the quality of the accommodation and location, rather than broader neighborhood experiences. Providing useful insights for the comparison of LLM-generated content with human-authored reviews.

Review Length by Gender

Analysis of word count density, reveals information of how much reviewers write on their

reviews.

1. Both genders seem to write short reviews, where most range between 0 to 50 word counts.
2. Female reviewers seem to write slightly longer reviews than males do, illustrating a broader density distribution peak.
3. Male reviewers show a sharper peak at lower word counts, suggesting shorter descriptions.

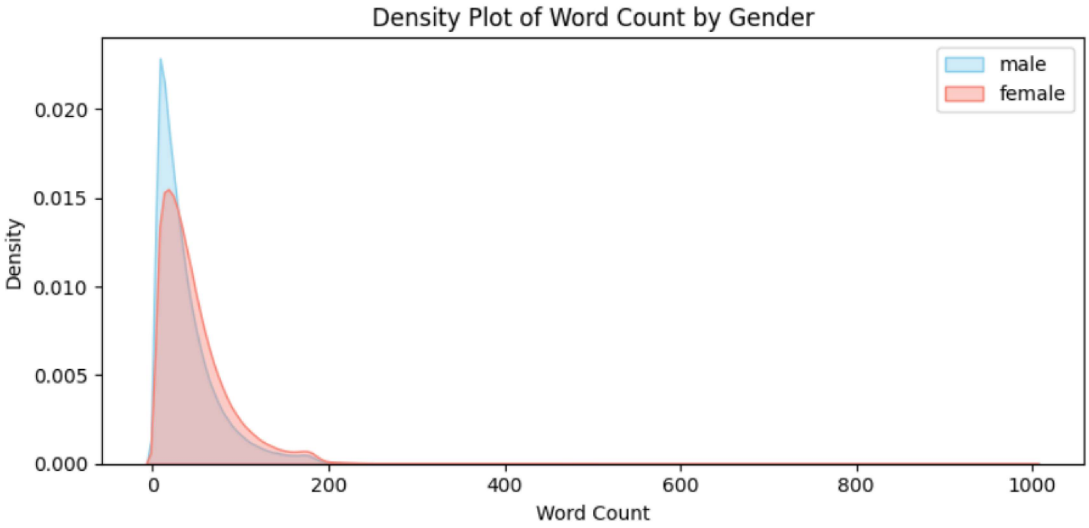


Figure 8: Density of word counts per gender in Airbnb data.

These patterns suggest that while both genders contribute almost similarly to the dataset, females tend to provide more detailed descriptions, which may influence the extraction of sentiment and topical information.

3.4.4 Crime Patterns Analysis

Comprehending safety in London neighbourhoods is vital to be observed beyond raw numbers, but through patterns that define neighbourhoods and boroughs. By using official crime records normalized per 1,000 residents, we discover important insights across London between central and outer neighbourhoods.

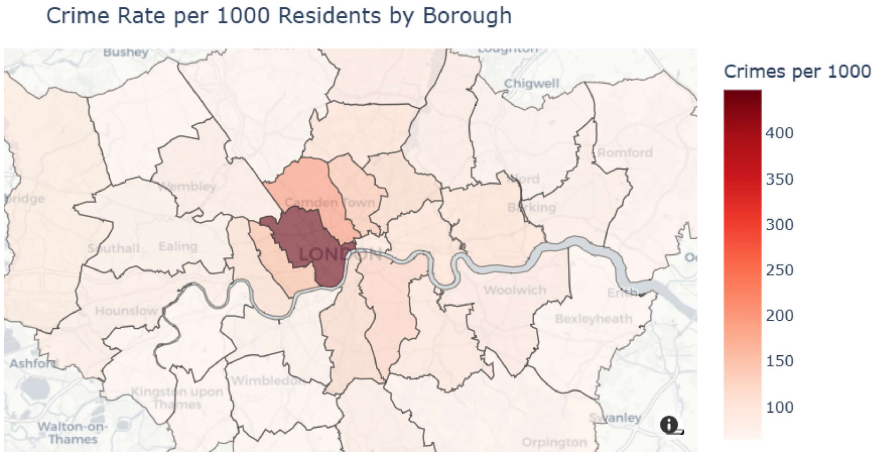
Boroughs-Level Analysis

At borough-level analysis, reveals that crime is equally distributed. Westminster emerged as a notable outlier with 448 crimes per 1,000 residents, more than double of any other borough (Table 9). Camden (205/1,000) and Kensington & Chelsea (162/1,000) followed as high crime recorded areas. These central boroughs share common features of high tourism, nightlife, and commercial activity that increase crime opportunities. Additionally, outer boroughs like Sutton (65/1,000) and Harrow (64/1,000) showed substantially lower crime rates, suggesting that their suburban and residential character is the cause of the low counts. A Choropleth visualization (Figure 9) further demonstrates the spatial differences between inner and outer London neighbourhoods, where the central ones reports higher crime densities. The borough analysis establishes a baseline safety hierarchy for comparison with human perceptions and LLM outputs.

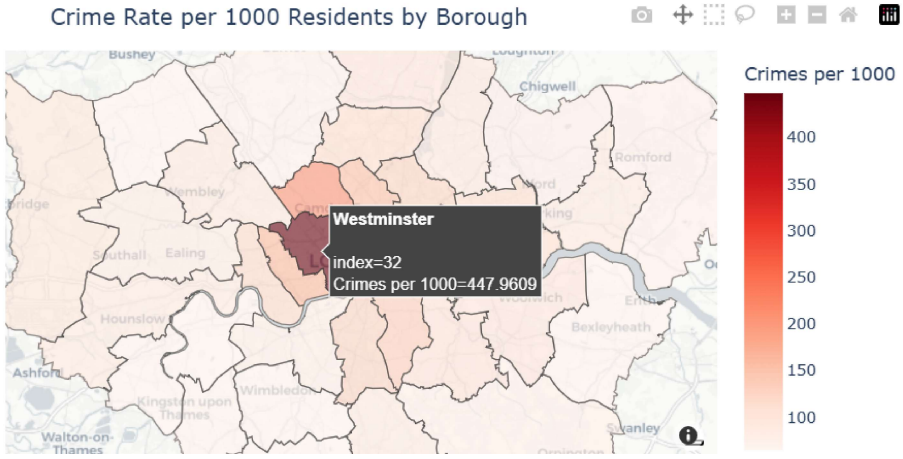
Interestingly, the distribution of crime patterns in each neighbourhood, is not function of its population. Some areas with similar population sizes exhibit different crime rates, revealing that each 'unsafe' area include different factors for its dangerous profile. For example, Lambeth and Hackney have comparable population, but diverge in crime rates per 1000 residents. This suggest that other factors such as tourism, nightlife, and commercial activity influencing crime patterns.

ID	LAD24CD	Name	Crime Count	Population	Crimes per 1000
32	E09000033	Westminster	94070	209996	447.96
6	E09000007	Camden	44385	216943	204.59
19	E09000020	Kensington and Chelsea	23419	144518	162.05
18	E09000019	Islington	34015	223024	152.52
27	E09000028	Southwark	42383	314786	134.64
11	E09000012	Hackney	34084	266758	127.77
21	E09000022	Lambeth	40367	316920	127.37
12	E09000013	Hammersmith and Fulham	22382	188687	118.62
24	E09000025	Newham	42682	374523	113.96
29	E09000030	Tower Hamlets	37694	331886	113.58
13	E09000014	Haringey	29812	263850	112.99
22	E09000023	Lewisham	31439	301255	104.36
0	E09000001	City of London	1556	15111	102.97
10	E09000011	Greenwich	29597	299528	98.81
9	E09000010	Enfield	32034	327434	97.83
16	E09000017	Hillingdon	31997	329185	97.20
4	E09000005	Brent	34155	352976	96.76
30	E09000031	Waltham Forest	26243	279737	93.81
17	E09000018	Hounslow	27443	299424	91.65
7	E09000008	Croydon	37321	409342	91.17
1	E09000002	Barking and Dagenham	21162	232747	90.92
8	E09000009	Ealing	33757	385985	87.46
31	E09000032	Wandsworth	27074	337655	80.18
25	E09000026	Redbridge	25245	321231	78.59
5	E09000006	Bromley	25664	335319	76.54
15	E09000016	Havering	20799	276274	75.28
2	E09000003	Barnet	29650	405050	73.20
20	E09000021	Kingston upon Thames	12074	172692	69.92
3	E09000004	Bexley	17418	256434	67.92
23	E09000024	Merton	14752	218539	67.50
26	E09000027	Richmond upon Thames	13102	196678	66.62
28	E09000029	Sutton	13977	214525	65.15
14	E09000015	Harrow	17423	270724	64.36

Table 9: Crime records across London boroughs, normalized by 1000 residents.



(a) Choropleth Map by Borough



(b) Choropleth Map by Borough labeled

Figure 9: Choropleth Map by Borough of 1000 residents crime counts

According to Figure 10, ten most and least safe boroughs are highlighted, observing extreme high crime concentrations in certain areas.

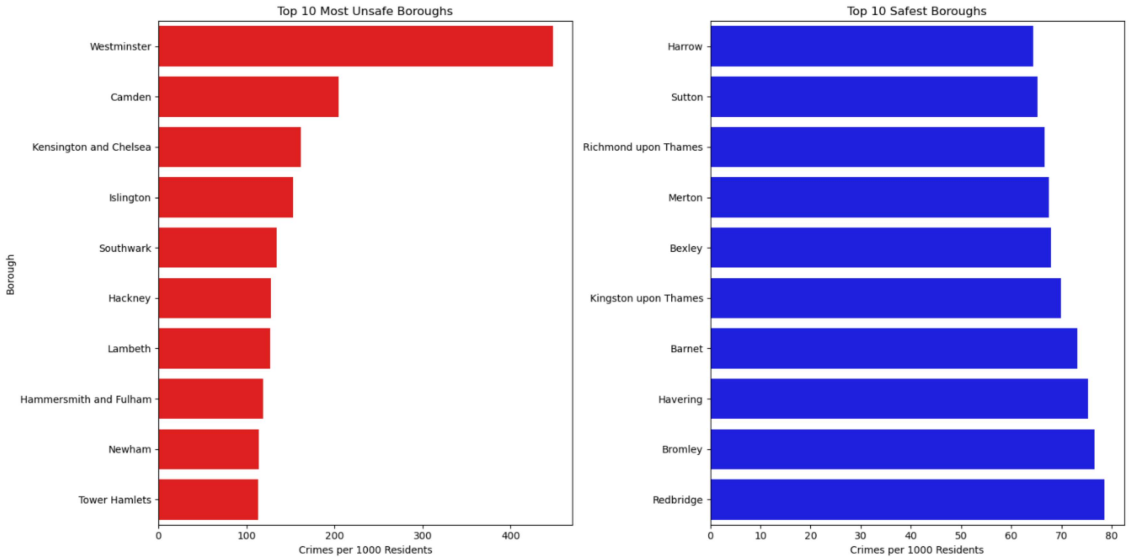


Figure 10: Top 10 Unsafe & Safe Boroughs by crimes per 1000 residents.

Figure 11, illustrates dominant crime types in the top 5 unsafe boroughs, showing that different boroughs are affected by distinct crime factors, suggesting tailored mitigation strategies are required.

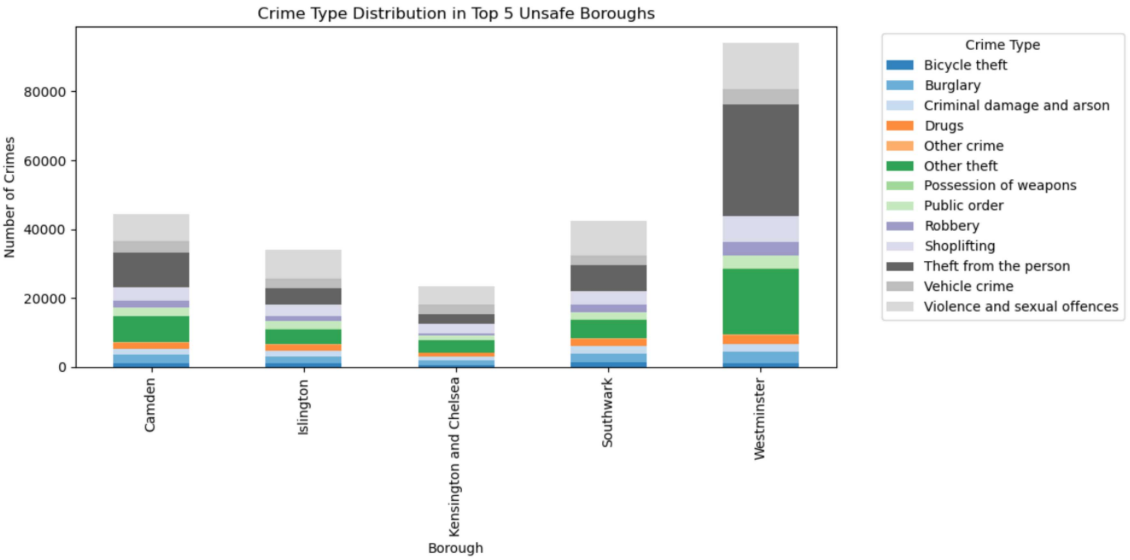


Figure 11: High crime type factors on top 5 Unsafe Boroughs.

3.4.5 Neighborhood-Level Crime Patterns

Neighborhood-level analysis suggested more granular safety variations, instead of general analysis by borough boundaries. Neighbourhood-level analysis include:

- **High risk neighbourhoods:** West End (186/1,000), St James's (115/1,000) and Blooms-

bury (38/1,000).

- **Low risk neighbourhoods:** Multiple outer areas with records lower than 2 crimes per 1,000 residents.

Neighbourhood	Population	Crime Count	Crimes per 1000
Abbey	227219.61	1915	8.43
Abbey Road	210214.41	827	3.93
Abbey Wood	299503.99	1795	5.99
Abingdon	144518.00	992	6.86
Addiscombe East	409342.00	900	2.20
...	...	...	...
Worcester Park North	214525.00	354	1.65
Worcester Park South	214525.00	248	1.16
Wormholt	188687.00	554	2.94
Yeading	329313.99	1321	4.01
Yiewsley	329185.00	1711	5.20

Table 10: Crime statistics across London neighbourhoods of 1000 residents.

As illustrated in Figure 12, top safe (right chart) and unsafe (left chart) neighbourhoods, reveal Rickmansworth Town, Lime Street, Queenhithe, etc. as safe areas with almost 0 crime counts per 1,000 residents, while shows West End, St Jame's and Bloomsbury as dangerous ones with 186 crime counts per 1,000 residents.

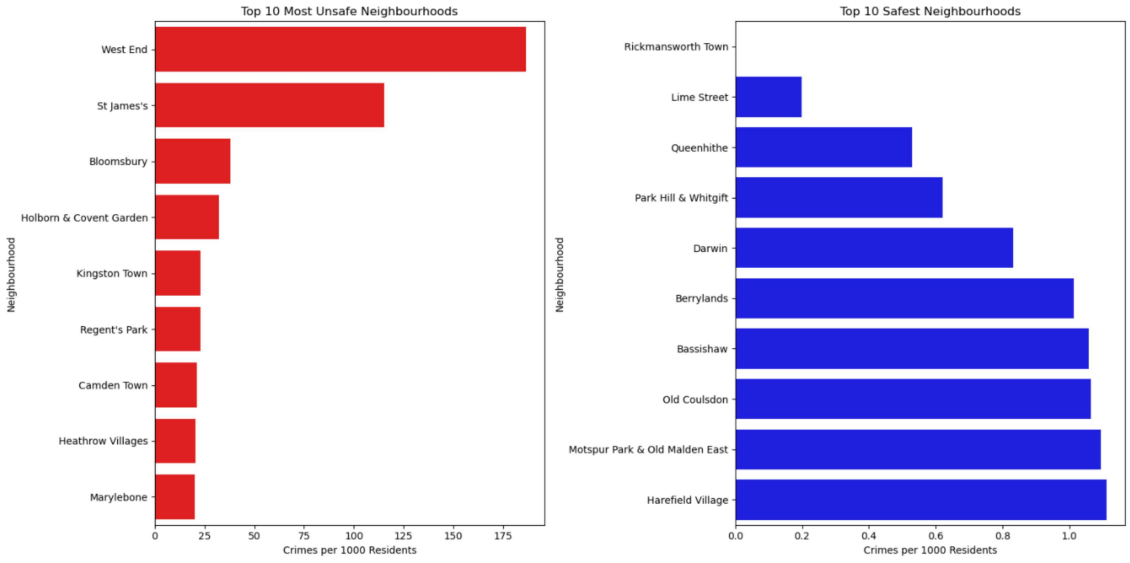
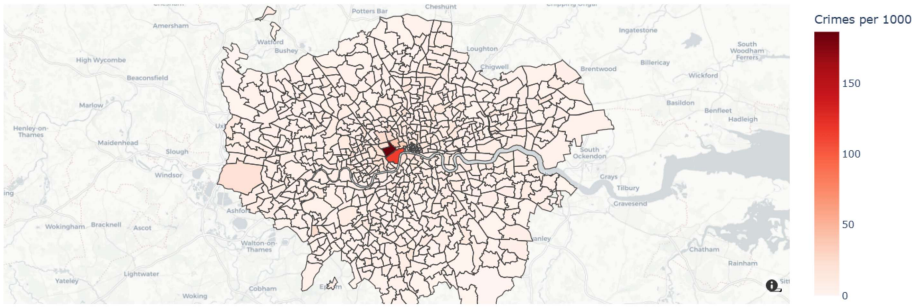
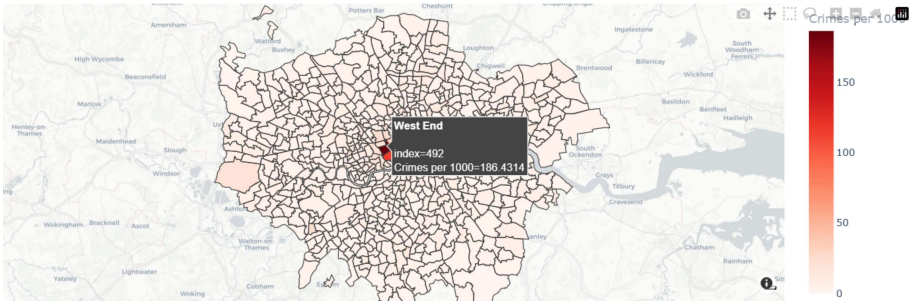


Figure 12: Top unsafe and safe neighbourhoods in London



(a) Choropleth Map by Neighbourhood



(b) Choropleth Map by Neighbourhood labeled

Figure 13: Choropleth Map crime counts by Neighbourhood normalized per 1000 residents.

Crime Types in Unsafe Neighborhoods

Figure 14 demonstrates that ”theft from the person” and ”other theft” crime types dominate the crime profile of the West End and Bloomsbury, while ”violence and sexual offenses” are more prominent in areas like Kingston Town. This indicates that different unsafe areas are unsafe for different reasons, emphasizing that uniform safety strategies are inadequate.

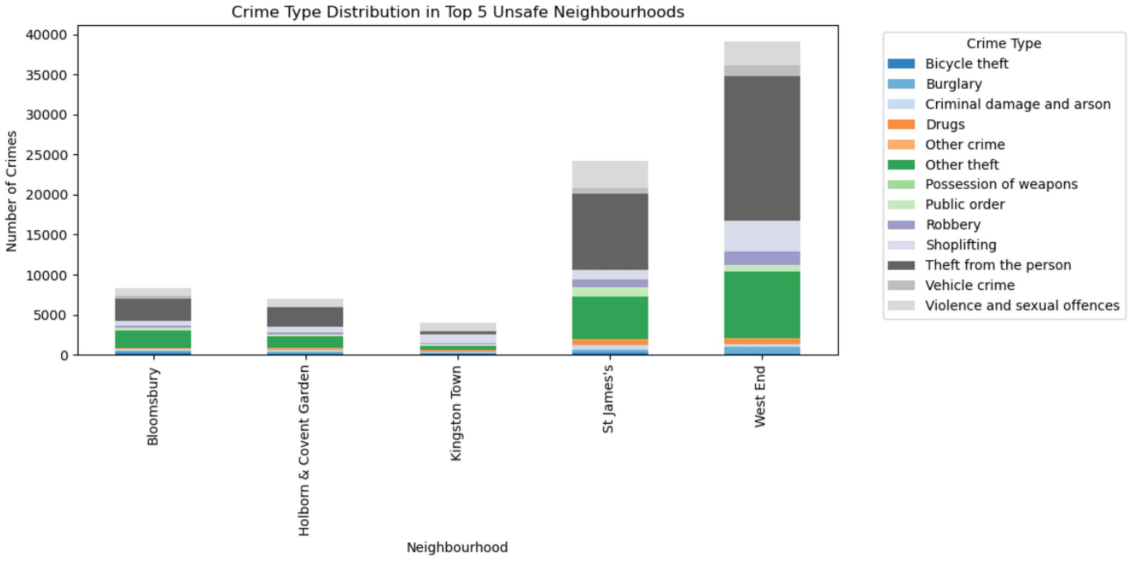
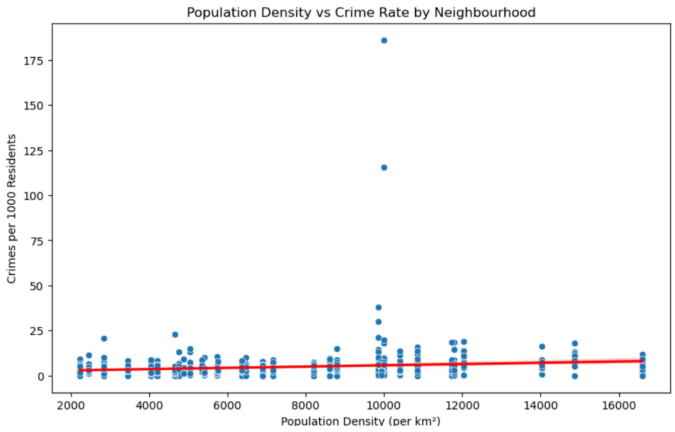


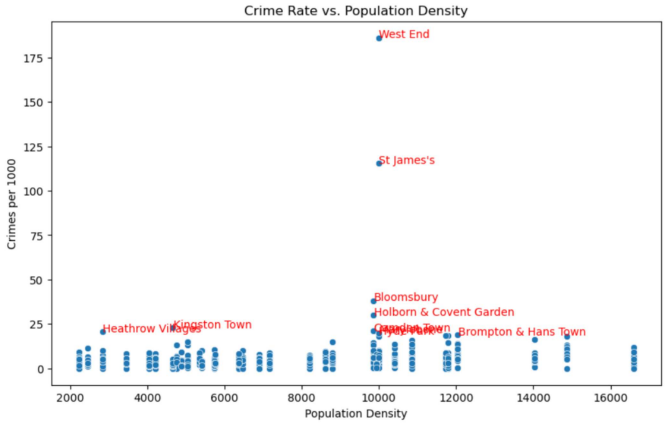
Figure 14: Crime type distribution of 5 top unsafe neighbourhoods.

Population Density Correlation

Analysis exhibits a positive correlation between population density and crime rates, with most areas clustering between 2,000-8,000 residents/km² and 60-120 crimes per 1,000 residents (Figure 15). However, notable outliers like West End and St James's show disproportionately high crime rates relative to their density (Table 11), indicating that factors beyond population density—such as commercial activity and tourism—significantly influence crime patterns.



(a) Population density vs crime rate with outliers



(b) Crime rate vs Population Density with named outliers

Figure 15: Correlation between population density and crime rate by neighbourhood

Neighbourhood	Crimes per 1000	Population Density
West End	186.13	9,999.81
St James's	115.36	9,999.81
Bloomsbury	38.09	9,861.05
Holborn & Covent Garden	30.03	9,861.05
Kingston Town	22.94	4,667.35
Camden Town	21.07	9,861.05
Heathrow Villages	20.58	2,837.80
Marylebone	20.04	9,999.81
Hyde Park	19.16	9,999.81
Brompton & Hans Town	18.77	12,043.17

Table 11: Top Neighbourhood Crime Rate Outliers in London.

3.4.6 Temporal and Demographic Trends

Seasonal Crime Fluctuations

Monthly crime data from May 2024 to May 2025 show seasonal fluctuations (Figure 16), peaking in October 2024 ($80,500$) and reaching the lowest point in February 2025 ($67,500$). These patterns reflect social variations from weather conditions, tourism to social activity, suggesting crime opportunities and public exposure.

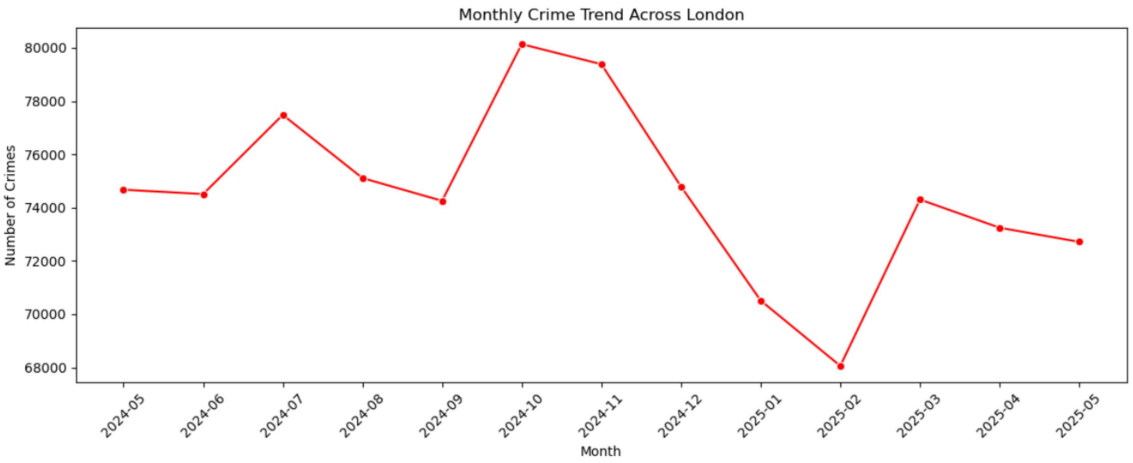


Figure 16: Seasonal fluctuation of reported crime in London, with high and low fluctuation peaks.

Population Density and Crime

In correlation analysis, were revealed a positive association between borough-level population density and crime rates (Figures 17–18). Most boroughs aggregate between 60–120 crimes per 1,000 residents at densities of 2,000–8,000 residents/km^2. **Outliers** such as Westminster and Camden reveal disproportionately high crime, possibly reflecting the touristic, nightlife, and social activity of the central areas.

In neighborhood-level, were exhibited similar nuances: dense central areas like West End and St James's experience elevated crime, while others (e.g., Haggerston) remain relatively safe despite similar density levels. These findings indicate that while density contributes to crime risk, socioeconomic and spatial contexts also play a significant role.

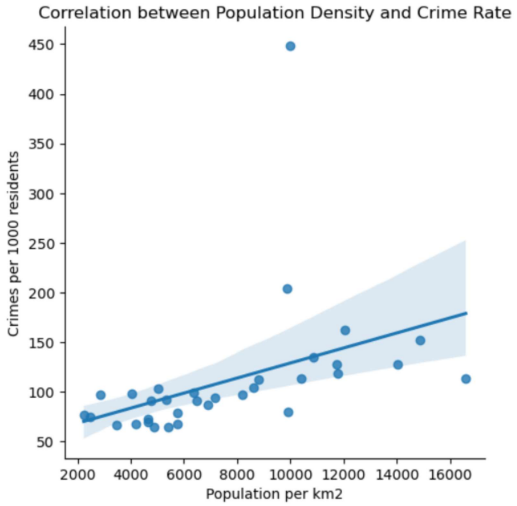


Figure 17: Correlation between borough population density and crime rates. Density (km^2) representation per 1,000 residents.

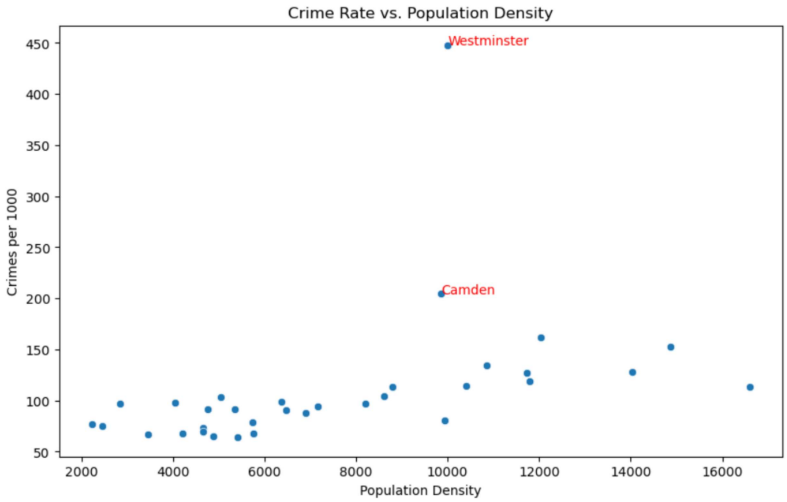


Figure 18: Borough-level fluctuation in crime records relative to population density.

3.5 Analytical Methods

This section highlights the analytical framework employed to investigate gendered perceptions of London neighborhoods between human-authored Airbnb reviews and LLM-generated travel content. With a combination of topic detection, topic modeling and sentiment analysis, applied to explore how reviewers describe London neighbourhoods and how these patterns vary across gender groups and geographic contexts.

The methodology structure integrates **Natural Language Processing (NLP)** approaches, including **prototype-based semantic embeddings, BERTopic** and **sentiment classification** pipelines. Each approach was selected to capture fully detailed aspects of the data such as: classification for thematic segmentation, topic modeling for latent structure discovery, lexical analysis for descriptive enrichment and sentiment analysis for safety perception evaluation.

3.5.1 Topic Detection on Airbnb Data

Airbnb reviews were preprocessed at the sentence level using NLTK to separate accommodation- (property) and neighbourhood- (location) related content. To distinguish between property and location content, a prototype-based semantic embedding approach was applied, to classify sentences using cosine similarity against example keywords. Example keywords or prototypes for *property* (e.g., 'apartment', 'room', 'clean') and for *location* (e.g, 'neighborhood', 'quiet', 'station') were encoded using the SentenceTransformers library. Sentences were compared to prototypes using cosine similarity, with non-matching cases defaulting to 'unknown'. A threshold of 0.30 (optimized using macro F1-score) achieved 69% accuracy.

Classification Outcomes

Category	Average Proportion (%)
Property	36.92%
Location	29.87%
Unknown	33.20%

Table 12: Classification outcomes percentage proportion per separated class.

Classification outcomes revealed that 36.9% of sentences described property features, while 29.9% described locations and 33.2% were categorized as unknown (Table 12). Gender differences showed men slightly more focused on property (38%), while women balanced property and location. Geographical variation also emerged: central districts like West End emphasized location, while others such as Weavers focused on property.

Word cloud for location sentences (Figure 19) highlights frequent phrases describing neighbourhoods. We can observe through the word cloud, what are the phrases or word that are displayed more frequent in location-class of Airbnb reviews.

Figure 19: Location class word cloud from sentence-level.

3.5.2 Topic Modeling on Location-Related Data

After the classification task, we isolate location-related sentences for the next investigation step of topic modeling to uncover recurrent themes in neighborhood descriptions.

Location-related sentences were analyzed with BERTopic to uncover thematic clusters and ex-

plore the way users talk or experience those places. Sentence embeddings (model: all-MiniLM-L6-v2) and bag-of-words features were clustered and automatically labeled through cosine similarity to predefined categories (e.g., Transportation, Amenities, Safety/Noise). Results showed that Amenities and Transportation dominated across all groups, but women more frequently highlighted Safety/Noise Neighbourhood (15%), compared to men (10%). Lexical distinctions reinforced this: women used affective descriptors (quiet, safe, unsafe), while men emphasized practical ones (station, flat, good).

	Topic	Top Words	Category
0	0	quiet, safe, noise, noisy, neighborhood, night...	Safe/Noisy Neighborhood
1	1	tube, station, stations, walk, stops, close, b...	Transportation
2	2	flat, located, lovely, location, cute, matt, g...	Location
3	3	time, london, trip, stayed, days, family, spen...	Transportation
4	4	come, definitely, im, time, return, visit, ill...	Location
...	...	...	...
241	241	friendly, don, lady, people, recommendations, ...	Recommendations
242	242	night, sleep, ill, lone, nofrills, normal, cha...	Else
243	243	space, spacey, subways, fantastically, bigger,...	Transportation
244	244	mayfair, ridges, beat, oxford, self, selfridge...	Neighborhood
245	245	spot, abba, overnight, heart, position, northe...	Location

Table 13: Topics and categories with matched frequent words.

Key Topic Categories

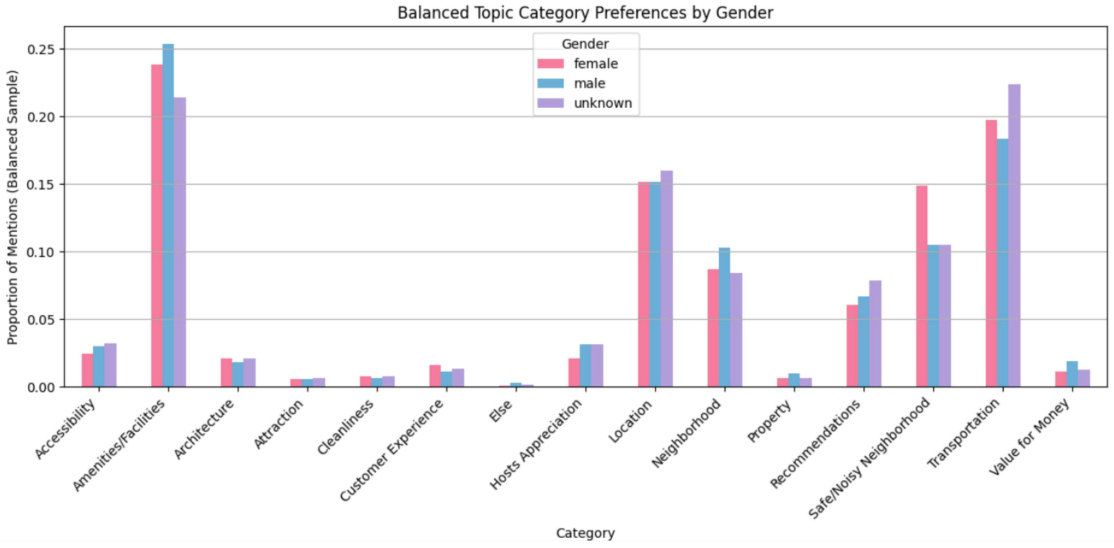


Figure 20: Topic category preferences by gender.

Topic	Category	Top Words (Weight)
0	Safe/Noisy Neighborhood	quiet (0.0260), safe (0.0201), noise (0.0188), noisy (0.0177), neighborhood (0.0167), night (0.0136), peaceful (0.0133), neighbourhood (0.0126), street (0.0117), loud (0.0111)
30	Safe/Noisy Neighborhood	quiet (0.0316), la (0.0240), quieter (0.0154), london (0.0124), una (0.0119), rare (0.0118), perfecta (0.0111), far (0.0110), reasonably (0.0107), parts (0.0106)
68	Safe/Noisy Neighborhood	flat (0.0580), quiet (0.0326), noise (0.0297), road (0.0211), ground (0.0198), floor (0.0197), despite (0.0188), busy (0.0163), street (0.0146), means (0.0145)
103	Safe/Noisy Neighborhood	staying (0.2244), highly (0.1353), recommend (0.1072), mold (0.0451), chantals (0.0451), resolved (0.0451), mireilles (0.0451), marilias (0.0402), future (0.0378), id (0.0293)
159	Safe/Noisy Neighborhood	sarahs (0.2432), emmas (0.1490), lauras (0.0847), calm (0.0433), convenientin (0.0407), fronts (0.0407), delivers (0.0407), capital (0.0363), colorful (0.0363), sandras (0.0337)
224	Safe/Noisy Neighborhood	hustle (0.3151), bustle (0.2889), oasis (0.0724), core (0.0672), tucked (0.0652), hustling (0.0598), refuge (0.0501), rush (0.0482), quit (0.0482), calm (0.0476)

Table 14: Words of Topics in the Safe/Noisy Neighborhood Category.

Gendered Topic Preferences

In Figure 21, from word cloud result comparison suggests lexical distinctions, where male reviewers use practical and diverse descriptors (good, station, flat), while female reviewers employ affective language (quiet, safe, felt unsafe). These differences settle the foundation for comparing gender-specific human experiences with LLM-generated safety content.

(a) Word Cloud for Female

(b) Word Cloud for Male

Figure 21: Gendered lexical differences between females and males.

3.5.3 Sentiment Analysis of Safety-Related Data

In this part we focus on safety perceptions across genders and neighbourhoods.To examine that a sentiment classification was applied, as showed in previous location-related Airbnb review sentences.

Classification Pipeline

For sentiment analysis implementation we encoded sentences into numerical vectors, using *"all-MiniLM-L6-v2"* SentenceTransformers model. After that we classified sentences using Lo-

gistic Regression classifier, chosen specifically for its interpretability and efficiency with high-dimensional vectors ability. The classifier was trained through manual labeling of safety-related data to distinguish safe, unsafe, and neutral expressions.

- **Safety classification:** safe vs. unsafe.
- **Sentiment polarity:** positive, negative, neutral.

This dual-layer method allowed us to evaluate whether a sentence reflected safety and the emotional tone of its expression.

Results Overview

From the analysis, safety classifier labeled 114,684 sentences as safe and 3,096 as unsafe. Also, sentiment polarity analysis displayed higher amount of sentences clustered in safety positively (85,475), with less negative or unsafe experiences captured (8,655). Although positive framing dominates, unsafe perceptions were strongly gendered and context-specific.

Safety Sentences Class	Count
Safe	114,684
Unsafe	3,096

Table 15: Predicted class distribution for safety sentences classification.

Sentiment Class	Count
Safe	85,475
Unknown	23,650
Unsafe	8,655

Table 16: Distribution of sentiment classification across review sentences.

Gender and Neighborhood Comparisons

Aggregated results showed clear gender contrasts, with female reviewers describing in higher proportion of neighborhoods as unsafe. In contrast, male reviewers tended to characterize the same areas as safe or neutral, leading to different gender place-descriptions.

For the lexical analysis results reinforce few differences on phrases used by gender. Female reviewers were captured to use affective expressions such as 'felt unsafe', 'late night' and 'traffic noise', while male reviewers highlighted more neutral descriptors such as 'street noise' and 'road noise'. According to that, females safe phrases often emphasize reassurance ('felt safe', 'quiet neighborhood'), while males emphasized more on general descriptions ('quiet area', 'great location').

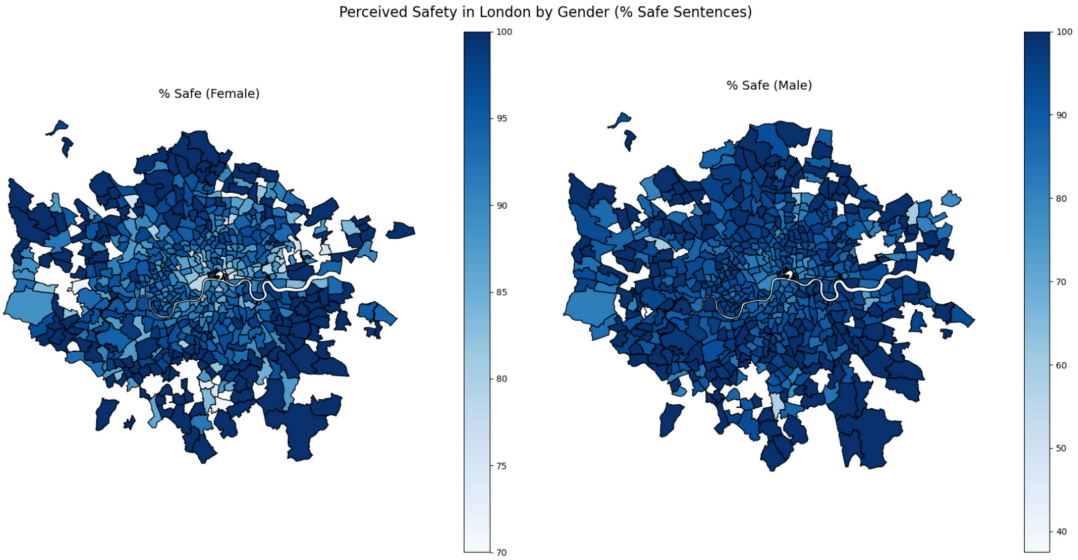


Figure 22: Distribution of London neighborhoods described as safe by gender.

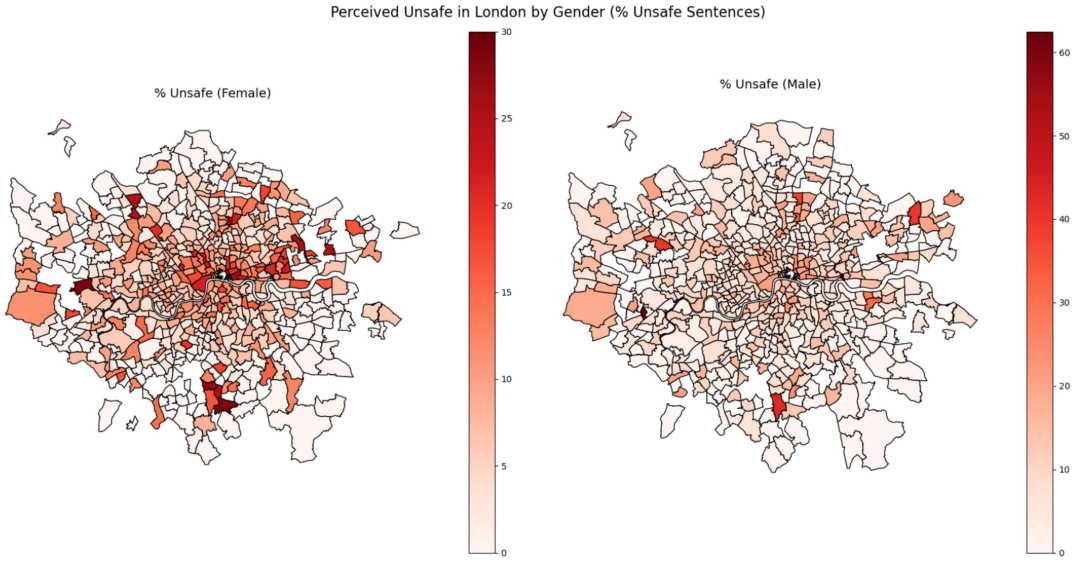


Figure 23: Distribution of London neighborhoods described as unsafe by gender.

Note: Both Figures 25 and 26 illustrate how genders (males and females), describe similar London neighbourhoods through their Airbnb reviews content. As it is observed in a general look, some main differences are captured on safe and unsafe areas, with females describing more areas as unsafe than males do.

Female (Safe)			Female (Unsafe)		Male (Safe)		Male (Unsafe)	
Rank	**Bigram**	**Freq**	**Bigram**	**Freq**	**Bigram**	**Freq**	**Bigram**	**Freq**
1	quiet neighborhood	2489	felt safe	176	quiet area	1999	street noise	67
2	quiet area	2472	safe walking	88	quiet neighborhood	1981	felt safe	47
3	quiet street	2403	street noise	84	quiet street	1793	traffic noise	37
4	felt safe	2337	late night	72	nice quiet	1773	late night	35
5	great location	1972	noise street	46	great location	1514	noise street	30
6	nice quiet	1963	walking around	45	quiet location	1330	walking around	24
7	neighborhood quiet	1710	traffic noise	44	location quiet	1266	light sleeper	22
8	location quiet	1596	felt unsafe	38	neighborhood quiet	1021	main road	22
9	quiet safe	1551	feel safe	34	quiet neighbourhood	1017	noise night	19
10	area quiet	1504	busy street	29	quiet place	989	safe walking	18
11	quiet location	1482	light sleeper	28	quiet peaceful	965	hear traffic	17
12	quiet peaceful	1379	could hear	27	area quiet	947	road noise	17
13	quiet neighbourhood	1296	didn't feel	26	located quiet	939	busy street	17
14	located quiet	1114	safe area	26	quiet safe	921	could hear	17
15	quiet residential	1088	main road	26	place quiet	852	police station	16

Table 17: Female and Male review phrases exhibited into Safe and Unsafe categories.

3.6 LLMs Analysis

Large Language Models (LLMs) such as GPT-5 and Claude Sonnet 4 are rapidly increased and used for decision-making tasks and question-answering, including travel suggestions. While these models produce fluent and detailed content, their unchecked use risks reproducing or amplifying biases, particularly around sensitive dimensions such as gender and safety. This section examines how LLMs portray London neighborhoods when prompted with gendered travel queries, comparing their outputs with human-authored Airbnb reviews and official crime statistics. The analysis employs sentiment classification, lexical and spatial comparisons, topic modeling, semantic similarity measures and fairness evaluations.

3.6.1 Dataset Overview

To evaluate LLM outputs, a dataset of 630 generated reviews was collected, from GPT-5 and Claude Sonnet 4, covering 21 selected London neighborhoods across 11 boroughs. Each neighborhood was described on three prompt type questions, across three gender perspectives (female, male and general/neutral). This design ensured balanced representation between both genders from using two gender-level prompts and one prompt generated for neutral gender. In that way we have a gender variation prompts approach focused mainly to female and male responses.

Prompt Structure and Reproducibility

To ensure reproducibility, all LLM queries were generated using a specific schema:

---

prompt_id | neighbourhood | borough | prompt_type | gender | model | response | question

---

Neighbourhood Coverage:

The following 21 London neighborhoods were included:

'Knightsbridge & Belgravia', 'Camden Town', 'Greenwich Peninsula', 'Tottenham Central', 'Chelsea Riverside', 'Waterloo & South Bank', 'Stratford', 'Whitechapel', 'Marylebone', 'Hyde Park', 'Blackheath', 'Kennington', 'Primrose Hill', 'Notting Dale', 'Highgate', 'Brixton Rush Common', 'Bloomsbury', 'Canary Wharf', 'South Richmond', 'Hendon' and 'King's Cross'.

Prompt templates:

Three question templates were used, with **[neighbourhood]** and **[gender]** generated accordingly:

1. "Describe **[neighbourhood]** life, is it safe for a **[gender]** traveler?"
2. "What is it like to visit **[neighbourhood]** in London?"
3. "As a **[gender]** traveler, is **[neighbourhood]** a good neighborhood to stay?"

Repetition scheme:

The repetition and substitution scheme was as follows:

- Q1: repeated 3× for female and 3× for male (6 total per neighborhood).
- Q2: repeated 3× for unknown (no gender specified, 3 total per neighbourhood).
- Q3: repeated 3× for female and 3× for male (6 total per neighborhood).

This produced 15 prompts per neighborhood × 21 neighborhoods = 315 prompts per model, totally 630 prompts across GPT-5 and Claude Sonnet 4. Each question was posed multiple times to reduce variability. Responses were constrained by instruction: 'Keep the response short and explain each safe or unsafe part in different sentences, do not mix meanings'.

	prompt_id	neighbourhood	borough	prompt_type	gender	model	response	question
0	1	Knightsbridge & Belgravia	Kensington and Chelsea	Safety	female	GPT-5	Knightsbridge & Belgravia is lively and friend...	Describe Knightsbridge & Belgravia life, is it...
1	2	Knightsbridge & Belgravia	Kensington and Chelsea	Safety	female	GPT-5	Knightsbridge & Belgravia is safe for women in...	Describe Knightsbridge & Belgravia life, is it...
2	3	Knightsbridge & Belgravia	Kensington and Chelsea	Safety	female	GPT-5	Daytime in Knightsbridge & Belgravia feels sec...	Describe Knightsbridge & Belgravia life, is it...
3	4	Knightsbridge & Belgravia	Kensington and Chelsea	Safety	male	GPT-5	Knightsbridge & Belgravia is busy and generall...	Describe Knightsbridge & Belgravia life, is it...
4	5	Knightsbridge & Belgravia	Kensington and Chelsea	Safety	male	GPT-5	The market area in Knightsbridge & Belgravia i...	Describe Knightsbridge & Belgravia life, is it...

Figure 24: Overview of LLM dataset structure.

Feature	Unique Values
Neighbourhoods	21
Boroughs	11
Prompt Types	4
Genders	3
Models	2

Table 18: Statistic summary of LLM dataset, across all included features.

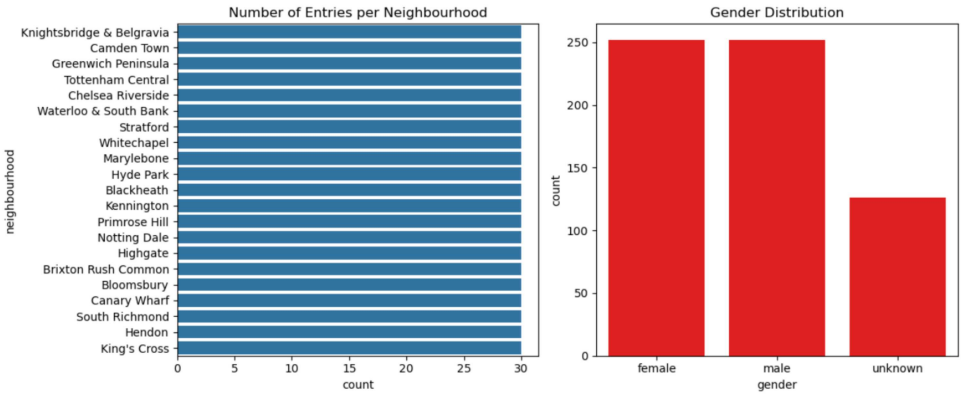


Figure 25: LLMs dataset features statistics.

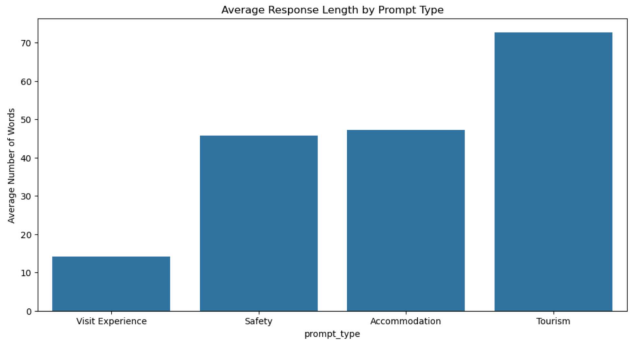


Figure 26: LLMs models prompt type categories statistics.

3.6.2 LLMs Sentiment Analysis

Sentiment analysis was used to provide critical insights into how LLMs describe neighborhood safety and to compare their outputs with Airbnb reviews. Each response was split into sentences and classified as safe, unsafe or neutral. Logistic Regression with class weighting was used, mirroring the Airbnb pipeline, to handle class imbalance while maintaining interpretability.

Results by Model

From results Claude Sonnet 4 generated higher neutral classification sentences (1,514), with fewer safe (253) or unsafe (119). In contrast, GPT-5 exhibited a more balanced distribution between groups, with higher counts of both safe (164) and unsafe (252) sentences. This suggests GPT-5 provides stronger polarity in safety perceptions, while Claude favors neutrality. To avoid bias due to unequal representation of male, female or neutral prompts, all analyses were normalized by neighbourhood, ensuring that findings reflect proportions rather than absolute counts.

Model	Neutral	Safe	Unsafe
Claude Sonnet 4	1514	253	119
GPT-5	235	164	252

Table 19: Distribution of predicted labels across different models.

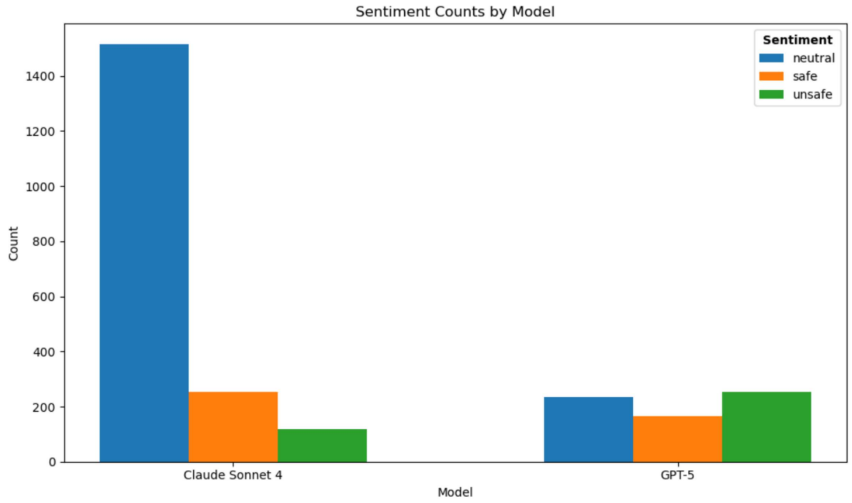


Figure 27: Sentiment predictions by model, showing GPT-5's stronger safe/unsafe polarity vs. Claude's neutral bias.

Neighborhood-Level Trends

At the neighborhood-level distributions show variation in model-generated perceptions. For example, Kennington and King's Cross received more unsafe classifications, while Blackheath was distinguished with neutral/safe. Analyses were normalized by neighborhood to avoid gender or prompt imbalance.

Neighbourhood	Neutral	Safe	Unsafe
Blackheath	89	18	14
Bloomsbury	87	22	12
Brixton Rush Common	79	17	26
Camden Town	80	19	22
Canary Wharf	87	22	12
Chelsea Riverside	79	29	13
Greenwich Peninsula	78	22	21
Hendon	90	16	15
Highgate	88	21	12
Hyde Park	87	21	13
Kennington	82	13	26
King's Cross	79	15	27
Knightsbridge & Belgravia	77	25	13
Marylebone	84	25	12
Notting Dale	86	12	23
Primrose Hill	88	21	12
South Richmond	88	21	12
Stratford	79	18	24
Tottenham Central	81	19	21
Waterloo & South Bank	85	23	14
Whitechapel	76	18	27

Table 20: Distribution of predicted labels across different neighbourhoods.

Gendered Patterns

Further examination of gender conditioning, revealed clear differences in how male and female perspectives are reflected in LLM predictions. For Blackheath, female prompts showed slightly higher unsafe proportions, while male prompts displayed toward safety. Lexical analysis reinforced these trends:

- **Female prompts:** 'solo travel', 'felt unsafe', 'late night', 'traffic noise'

- **Male prompts:** 'late-night safety', 'street noise', 'backstreet risks'

These patterns suggest that LLMs encode gendered linguistic nuances when describing neighbourhood safety.

Neighbourhood	Label	Female Prop.	General Prop.	Male Prop.
Blackheath	neutral	0.370	0.278	0.361
Blackheath	safe	0.132	0.056	0.130
Blackheath	unsafe	0.164	0.000	0.176

Table 21: Proportions of predicted labels by gender in Blackheath neighbourhood.

Top Phrases for Female			Top Phrases for Male		
Safe	**Neutral**	**Unsafe**	**Safe**	**Neutral**	**Unsafe**
female travelers	female travelers	late night	safe men	male travelers	late night
solo female	female visitors	avoid isolated	male travelers	travelers seeking	trouble nightlife spots
streets feel safe	travelers enjoy	unsafe backstreets	generally safe	male travelers seeking	streets noisy late
lively friendly day	female travelers enjoy	uncomfortable night	generally safe men	area provides	streets noisy
friendly day busy	travelers seeking	secluded areas dark	streets day	travelers appreciate	tense late night
busy streets feel	female travelers seeking	areas dark safe	safe men day	local businesses	tense late
streets feel	area provides	areas dark	secure men main	male travelers appreciate	noisy late night
feel safe	local establishments	streets night	secure men	neighborhood offers	noisy late
busy streets	transport connections	spots avoid	busy generally safe	male visitors	noisy tense late
day busy	central areas	streets night secluded	main streets	local establishments	trouble nightlife

Table 22: Comparison of top phrases for Female and Male labels across Safe, Neutral, and Unsafe categories.

Spatial Patterns

To examine spatial trends, aggregated proportions of safe, unsafe and neutral labels were mapped with Choropleths visualizations (Figures 28–29). Results highlight both overlaps and divergences in how GPT-5 and Claude illustrate neighborhood safety across genders, providing a vital framework for assessing potential spatial biases in model outputs and enable comparisons between gendered perceptions.

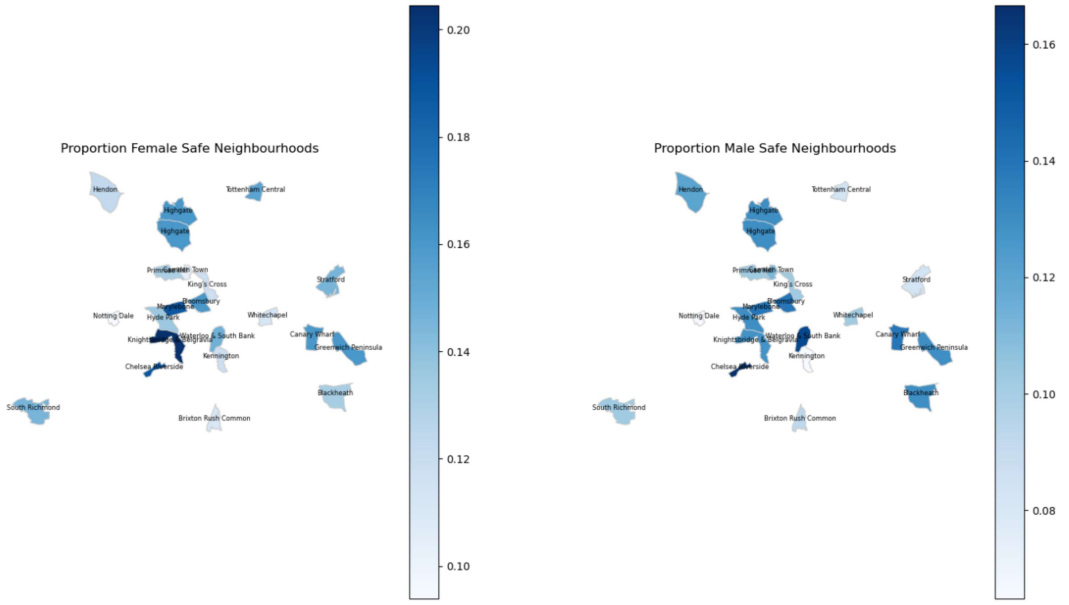


Figure 28: Choropleth map of safe classification across London neighborhoods.

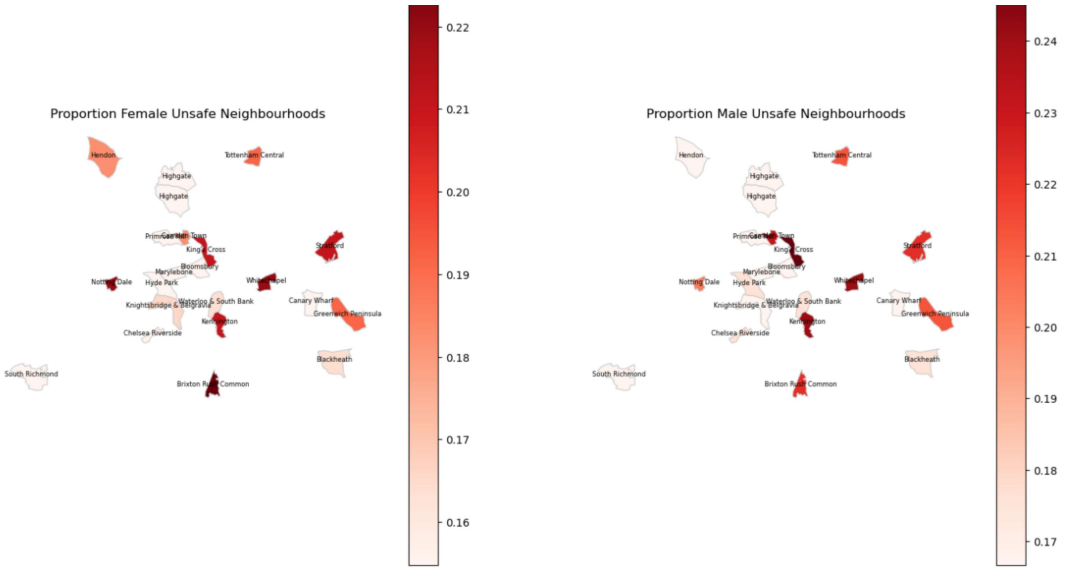


Figure 29: Choropleth map of unsafe classification across London neighborhoods.

Note: Choropleth maps of unsafe and safe descriptions in LLMs reveal important findings on how it illustrates results for both genders. We can observe few differences in how similar places are considered as safe or unsafe for LLM-generated results. However, similarities are also captured in LLMs responses rather than the human descriptions. Allowing a good comparison for further finding in similar neighbourhoods of Human-authored vs LLM descriptions.

Model Comparison

Further analysis shows that GPT-5 generates higher proportions of labels, such as 31–33% unsafe, compared with Claude's 5%. Claude emphasizes neutrality (48–52%), relative to GPT-5 (19–20%). These results indicate that GPT-5 captures more explicit human-like safety patterns, while Claude underrepresents extremes. These comparisons, suggest that LLMs differ substantially, with GPT-5 providing higher safe/unsafe assessments, whereas Claude tends toward neutrality. Gendered prompts further reveal biases, with women's queries exhibiting more unsafe descriptions. These findings underscore the need to account for model behavior and demographic framing when evaluating AI fairness in travel content.

Model	Category	Female Prop.	Male Prop.
Claude Sonnet 4	Safe	13.1%	8.7%
GPT-5	Safe	15.4%	14.3%
Claude Sonnet 4	Neutral	48.6%	52.5%
GPT-5	Neutral	20.3%	19.0%
Claude Sonnet 4	Unsafe	5.0%	5.5%
GPT-5	Unsafe	31.0%	33.3%

Table 23: Proportions of predicted labels by gender for each model and category.

Figure 30: Female Safe Breakdowns

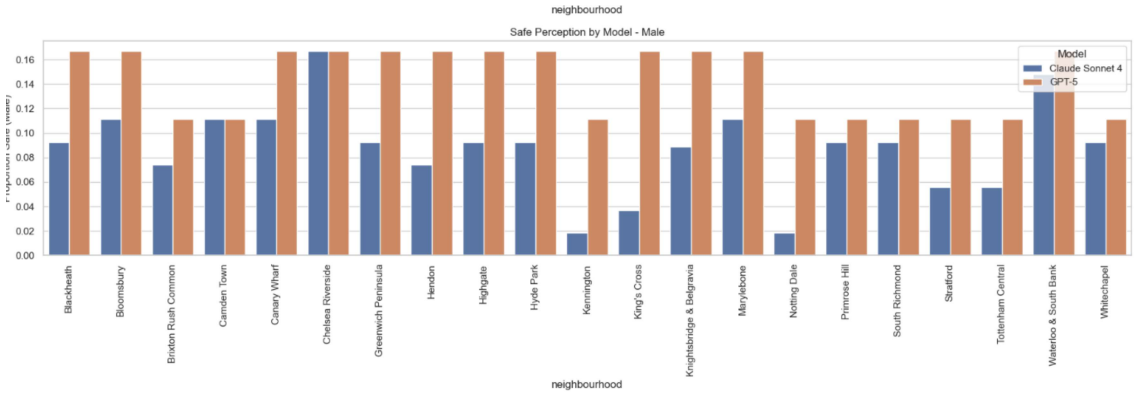


Figure 31: Male Safe Breakdowns

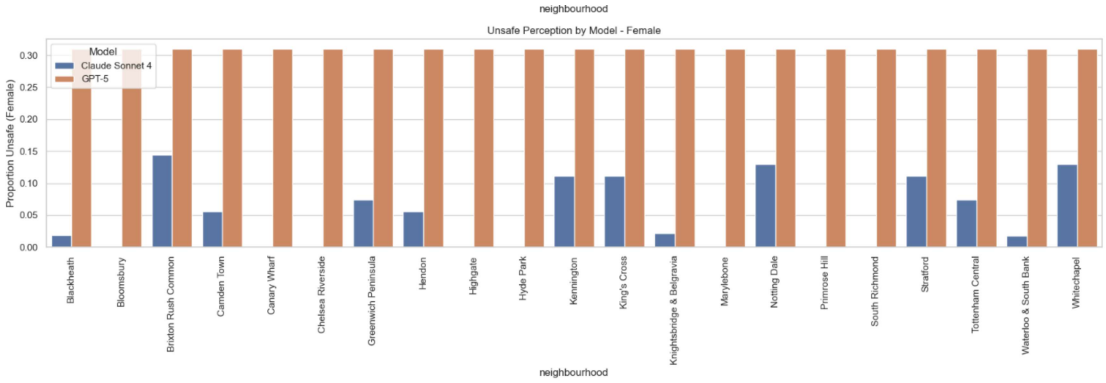


Figure 32: Female Unsafe Breakdowns

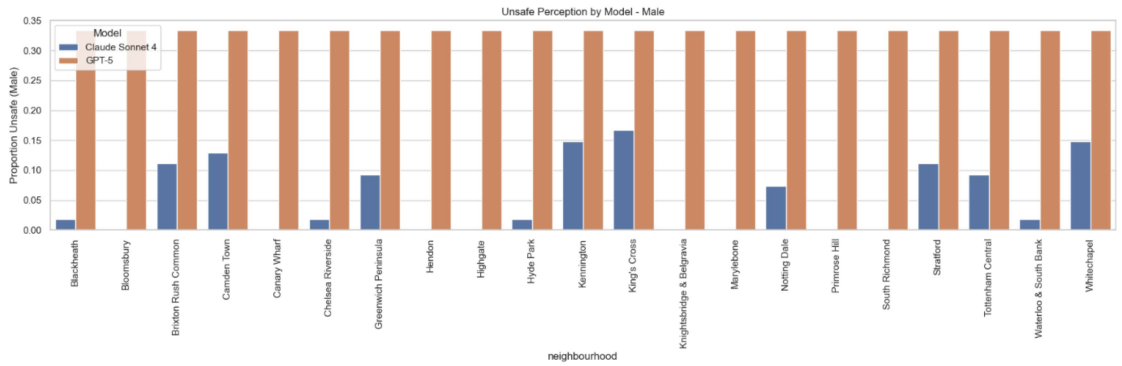


Figure 33: Male Unsafe Breakdowns

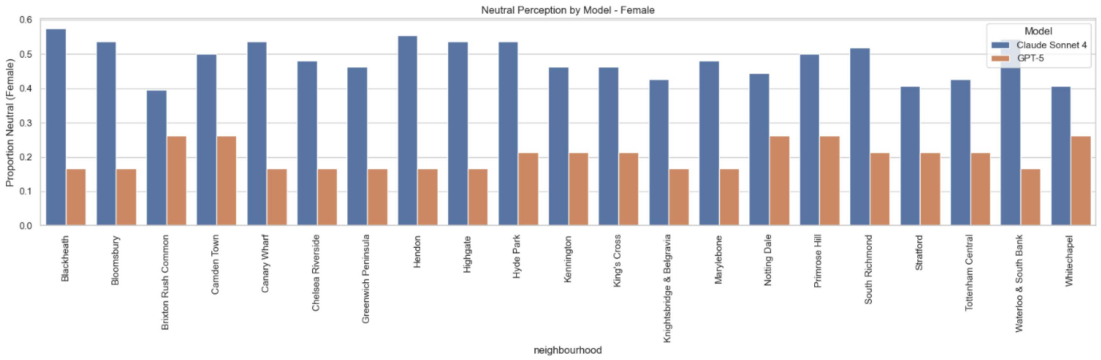


Figure 34: Female Neutral Breakdowns

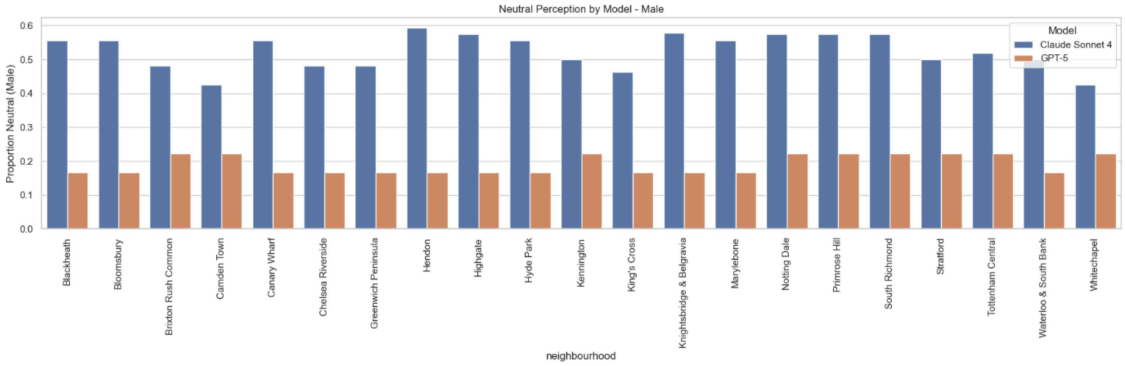


Figure 35: Male Neutral Breakdowns

3.6.3 LLM models Exploratory Analysis

After preprocessing and sentiment classification, examined on previous sections, the dataset contained slightly more female prompts (1,028) compared to male (1,005) and neutral (504) (Table 24). Similarly to Airbnb dataset, exhibiting imbalance counts of gender sizes, developing possible bias results. To mitigate this, all analyses were standardized by neighborhood aggregation and reported in proportions rather than raw counts. This ensured comparability across genders and models, regardless of sentence frequency.

For stable and accurate results, agreement analyses were conducted across both gender and

model dimensions. For gender analysis, safe/unsafe proportions were compared between male and female prompts within the same neighborhoods and models, then aggregated to estimate the degree of gender alignment. For model analysis, female safety predictions from GPT-5 and Claude Sonnet 4 were compared to evaluate cross-model consistency.

Gender	Count
Female	1028
Male	1005
General	504

Table 24: Distribution of LLM prompts across gender categories.

This method exhibits a quantitative measure of alignment, providing us information on whether models predict similar outcomes across genders or they develop frequent safety assessments for the same gender. By analyzing this, we identify areas of convergence and divergence, offering insights into the reliability and potential biases embedded in model predictions. Results showed (Table 28):

- **Unsafe predictions:** perfect gender agreement (1.000) across both models.
- **Safe predictions:** higher gender agreement in Claude (0.984) than GPT-5 (0.841).
- **Cross-model female safe predictions:** much lower agreement (0.372), revealing divergence between models in how they classify women's safety.

Results suggests that while unsafe outputs are consistent, safe classifications model-exclusive dependent, with Claude generating more gender-balanced outputs.

For further evaluation of reliability on predictions between ChatGPT-5 and Claude Sonnet 4, we compute the **mean absolute differences** for safe, unsafe and neutral labels, separately for female and male groups. Mean absolute differences further highlighted this:

- **Safe predictions:** low (0.055 female, 0.049 male), suggesting closer alignment across models.
- **Unsafe predictions:** higher (0.190 female, 0.204 male), highlighting weaker consistency.

These findings suggest that while both models generally agree on what is 'safe', they diverge more strongly on what is 'unsafe', especially for female prompts. Table 26 illustrates neighborhood-level proportions, showing where Claude and GPT-5 differ or resemble most in safety classifications.

Model	Safe Agreement (Gender)	Unsafe Agreement (Gender)
Claude Sonnet 4	0.984	1.000
GPT-5	0.841	1.000
Cross-Model Agreement (Female, Safe): 0.372		

Table 25: Gender and model agreement proportions for safe and unsafe predictions.

Neighbourhood	Claude Female	GPT-5 Female	Claude Male	GPT-5 Male
Blackheath	0.111	0.308	0.139	0.250
Bloomsbury	0.194	0.308	0.167	0.250
Brixton Rush Common	0.189	0.154	0.111	0.167
Camden Town	0.167	0.154	0.167	0.167
Canary Wharf	0.194	0.308	0.167	0.250
Chelsea Riverside	0.278	0.308	0.250	0.250
Greenwich Peninsula	0.194	0.308	0.139	0.250
Hendon	0.083	0.308	0.111	0.250
Highgate	0.194	0.308	0.139	0.250
Hyde Park	0.194	0.231	0.139	0.250
Kennington	0.139	0.231	0.028	0.167
King's Cross	0.139	0.231	0.056	0.250
Knightsbridge & Belgravia	0.303	0.308	0.121	0.250
Marylebone	0.278	0.308	0.167	0.250
Notting Dale	0.139	0.154	0.028	0.167
Primrose Hill	0.250	0.154	0.139	0.167
South Richmond	0.222	0.231	0.139	0.167
Stratford	0.222	0.231	0.083	0.167
Tottenham Central	0.250	0.231	0.083	0.167
Waterloo & South Bank	0.162	0.308	0.222	0.250
Whitechapel	0.194	0.154	0.139	0.167

Table 26: Neighborhood safety proportions per model and gender for all 21 neighbourhoods. Claude Sonnet 4 and GPT-5 results are shown side by side for female and male populations.

3.6.4 LLMs Topic Modeling Analysis

To examine the thematic structures of LLM-generated reviews, BERTopic was applied to sentence-level embeddings (using 'all-MiniLM-L6-v2'). Each prompt sentence was converted into a dense vector representation, capturing semantic similarity on frequent words. After clustering, representative keywords were extracted and grouped into interpretable categories. Predefined semantic labels (e.g., Restaurants, Security, Nightlife & Social Life) were assigned using cosine similarity matching. This approach provides flexible and interpretable method for exploring the underlying content of textual responses across different neighbourhoods, genders and model outputs.

Topic	Top Words
0	reflect, restaurants, businesses, traditional, shops, markets, local, incredible, independent, cuisine
1	insight, london, development, showcases, genuine, social, communities, dynamics, authentic, neighborhood
2	hotels, boutique, amenities, premium, guesthouses, comprehensive, apartments, offers, personalized, serviced
3	cater, understand, international, requirements, establishments, executives, tourists, discerning, specific, professional
4	research, specific, displaying, items, valuable, situational, avoid, awareness, maintain, male
5	value, affordable, exploration, safety, considerations, improving, safer, evaluation, tours, guided
6	surveillance, affluent, engagement, active, benefits, invest, residential, attracts, natural, professionals
7	cross, king, regeneration, transport, urban, improvements, convenience, significant, connectivity, men
8	chelsea, riverside, luxury, exceptional, experiences, waterfront, seeking, exceptionally, men, physic
9	low, crime, rates, maintains, reputation, established, effective, measures, consistently, neighborhood

Table 27: Top 10 representative words for each topic generated using BERTopic on LLM dataset.

Topic	Representative Words	Category
0	reflect, restaurants, businesses	Restaurants
1	insight, london, development	Neighborhood
2	hotels, boutique, amenities	Accommodation
3	cater, understand, international	Accommodation
4	research, specific, displaying	Experiences
5	value, affordable, exploration	Experiences
6	surveillance, affluent, engagement	Security
7	cross, king, regeneration	Neighborhood
8	chelsea, riverside, luxury	Accommodation
9	low, crime, rates	Security
⋮	⋮	⋮
85	pinnacle, represents, integration	Experiences
86	hospitality, knowledge, genuine	Accommodation
87	rare, incidents, extremely	Security
88	star, personalized, service	Accommodation
89	address, appreciate, sophisticated	Neighborhood

Table 28: Topics generated by BERTopic with 3 representative words and assigned categories. Only selected words are shown for brevity; full lists are available in the appendix.

Following the BERTopic analysis of topic identification, we generated the top words that align within each cluster and excluding the outlier topic (-1). All topic clusters were used to extract meaningful semantic categories, by implementing predefined labeling approach based on semantic similarity. According to the results, we found that most frequent categories were: 'Nightlife & Social Life' (18.8%) and 'Accommodation' (13.6%). Moreover, 'Safety' and 'Security' together account for nearly 18%, highlighting the profile of neighborhood safety in LLM-generated responses. Less frequent categories were 'Noise' (2.6%) and 'Outdoors' (3.4%), suggesting these are minor concerns in generated texts.

Category	Count	Proportion
Nightlife & Social Life	416	0.188
Accommodation	300	0.136
Security	235	0.106
Neighborhood	226	0.102
Experiences	190	0.086
Safety	162	0.073
Restaurants	157	0.071
Transportation	150	0.068
Affordability	84	0.038
Attraction	79	0.036
Architecture	76	0.034
Outdoors	75	0.034
Noise	58	0.026

Table 29: Overall counts and proportions of categories across the dataset as determined by BERTopic.

Gender-specific patterns displayed that female prompts exhibit 'Safety' (10.6%) and 'Experiences' (10.3%). While, male prompts prioritize Security (12.8%) and Accommodation (16.7%). This inform us about the different genders perspectives when they describe similar areas, or what are the characteristics that both genders prioritize the most according to LLM-prompts. For the neutral prompts often displayed preference on 'Restaurants' (19.3%) and 'Neighborhood' descriptions (16.9%). All these details, reveal that LLMs reflect gendered framing: women's prompts stress subjective safety, while men's highlight structural security and housing quality.

Category	Female (%)	Male (%)	Unknown (%)
Accommodation	12.9	16.7	8.7
Affordability	5.8	2.8	1.4
Architecture	1.6	1.5	11.6
Attraction	2.7	5.2	2.2
Experiences	10.3	9.0	4.1
Neighborhood	7.3	10.1	16.9
Nightlife & Social Life	21.0	19.4	13.0
Noise	0.0	4.8	3.9
Outdoors	4.2	2.0	4.6
Restaurants	5.8	2.6	19.3
Safety	10.6	6.6	1.7
Security	10.3	12.8	7.0
Transportation	7.6	6.6	5.6

Table 30: Proportion (%) of each category across Female, Male, and Unknown gender groups.

Note: This presents the distribution of thematic categories across female, male, and unknown gender groups. The results inform that Nightlife & Social Life is the most dominant category for both female (21.0%) and male (19.4%) gender groups, revealing a strong agreement on social experiences, according to LLMs. Notable trends are illustrated on Accommodation and Security categories for males compared to females, while females have higher amounts of sentences in Safety and Experiences. Such results indicate that while males are concerned about the Security of a neighbourhood, females focus on the safety of that place. However, both genders according to LLM responses are cautious on the safety of London neighbourhoods, with high proportions. Categories such as Noise and Outdoors are represented as not so important across genders, with low sentences amount, suggesting that these topics are not discussed that much. Overall, these patterns provide insights into gender-specific thematic priorities within the dataset and can inform subsequent analyses of model predictions and neighborhood safety perceptions.

Gender	Category	Proportion	Count
Female	Nightlife & Social Life	0.210	194
Male	Nightlife & Social Life	0.194	168
Unknown	Restaurants	0.193	80

Table 31: Most referred category per gender, showing proportion and absolute count of sentences.

Further in neighborhood-level distributions, illustrated how dominant categories vary geographically. For example, Marylebone is strongly associated with 'Accommodation' (46%), Knightsbridge & Belgravia with 'Nightlife & Social Life' (37%), and Whitechapel with 'Architecture' (31%). These results show that LLMs reproduce stereotypical associations, such as luxury and higher-class neighbourhoods highlight nightlife, central areas tourism and residential or suburban neighborhoods to housing quality.

Neighborhood	Category	Proportion	Count
Blackheath	Neighborhood	0.291	32
Bloomsbury	Nightlife & Social Life	0.245	27
Brixton Rush Common	Experiences	0.370	40
Camden Town	Neighborhood	0.220	20
Canary Wharf	Experiences	0.320	33
Chelsea Riverside	Experiences	0.333	36
Greenwich Peninsula	Safety	0.227	25
Hendon	Neighborhood	0.352	38
Highgate	Neighborhood	0.292	31
Hyde Park	Outdoors	0.321	35
Kennington	Nightlife & Social Life	0.316	31
King's Cross	Transportation	0.365	38
Knightsbridge & Belgravia	Nightlife & Social Life	0.372	35
Marylebone	Accommodation	0.463	50
Notting Dale	Affordability	0.333	34
Primrose Hill	Attraction	0.287	31
South Richmond	Attraction	0.280	30
Stratford	Nightlife & Social Life	0.398	43
Tottenham Central	Accommodation	0.184	18
Waterloo & South Bank	Safety	0.324	36
Whitechapel	Architecture	0.308	33

Table 32: Frequent category per neighborhood with proportion and count of sentences.

3.7 Cross-Comparison Research

3.7.1 Gender Nuances across LLMs and Airbnb

After completing all the individual research parts for Airbnb reviews, crime dataset and LLM prompts, we focus on question answering and the implementation of cross-comparison and synthesis stage.

This section presents a cross-comparison of Airbnb reviews and LLM-generated neighborhood descriptions, focusing on whether LLMs reflect gendered safety perceptions similarly to human-authored content. All the data from 21 London neighborhoods were aggregated by gender and normalized to account for unequal or biased review counts. Sentiment labels (safe, unsafe, neutral) were calculated as percentages to enable fair comparisons across datasets and models. To compute and identify the agreement, we computed the **Mean Absolute Error (MAE)** between both LLM predictions (GPT-5 and Claude Sonnet 4) and Airbnb review sentiment proportions. Higher MAE indicates stronger divergence from human perceptions.

Metric	Value (% points)
Female_MAE_Safe	50.94
Male_MAE_Safe	54.59
Female_MAE_Unsafe	12.65
Male_MAE_Unsafe	12.87

Table 33: Mean Absolute Differences (percentage points) across neighbourhoods for safe and unsafe classifications.

The MAE is revealing significant disagreement in 'safe' predictions (51–55%), while 'unsafe' classifications are more aligned (12–13%). Key neighborhoods with the largest divergences include Primrose Hill, Stratford, and Kennington, showing that differences cluster in specific areas rather than being random.

Neighborhood-level tables highlight the top five areas with the largest discrepancies.

Neighbourhood	Female Safe Difference
Primrose Hill	-67.18
Notting Dale	-66.56
Greenwich Peninsula	-60.65
Brixton Rush Common	-59.53
Kennington	-58.63

Table 34: Top 5 neighbourhoods by absolute Female Safe Difference.

Neighbourhood	Male Safe Difference
Stratford	-69.59
Canary Wharf	-63.45
Kennington	-62.10
Greenwich Peninsula	-62.09
Tottenham Central	-61.94

Table 35: Top 5 neighbourhoods by absolute Male Safe Difference.

Neighbourhood	Female Unsafe Difference
Notting Dale	24.05
Kennington	23.99
Greenwich Peninsula	22.11
Brixton Rush Common	21.34
Stratford	20.03

Table 36: Top 5 neighbourhoods by absolute Female Unsafe Difference.

Neighbourhood	Male Unsafe Difference
Kennington	27.21
Stratford	22.51
Greenwich Peninsula	20.73
Brixton Rush Common	20.63
Camden Town	19.06

Table 37: Top 5 neighbourhoods by absolute Male Unsafe Difference.

While the mean absolute error values provide an aggregate picture of disagreement between models, the neighbourhood-level tables reveal its uneven distribution. For safe classifications, large divergence were observed in areas such as Primrose Hill (–67.2 percentage points for females) and Stratford (–69.6 for males), which substantially inflate the overall MAE values. Unsafe classifications are displayed more stable on average (12–13 percentage points), still divergent, where most visible in Kennington and Notting Dale. These results exhibit that disagreements are not randomly distributed but instead cluster around specific neighbourhoods, reflecting context-specific sensitivities in how the models capture safety and risk. Figures 36–39 illustrate these gendered divergences in safe and unsafe classifications.

- **Figure 36.** Female 'Safe' proportions: LLMs underestimate safety compared to Airbnb.

- **Figure 37.** Male 'Safe' proportions: fluctuation is slightly higher for male-rated neighborhoods.
- **Figure 38.** Female 'Unsafe' proportions: consistent alignment between human and LLM perceptions.
- **Figure 39.** Male 'Unsafe' proportions: LLMs capture unsafe cases more consistently, though some neighborhoods (e.g., Kennington) show place divergence.

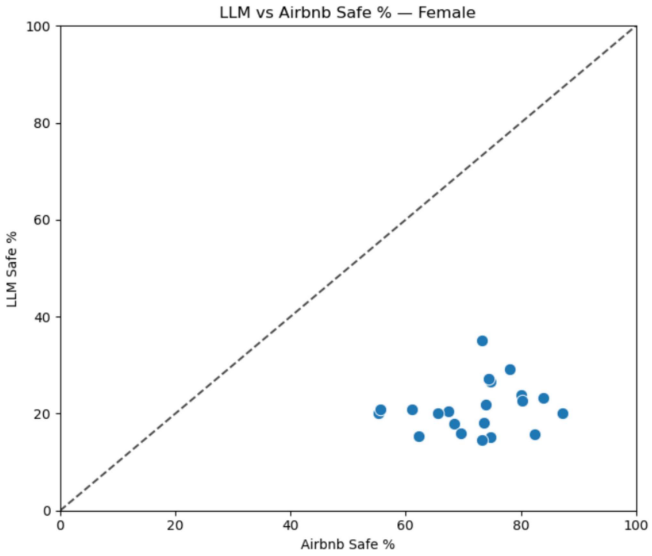


Figure 36: LLM vs. Airbnb Safe % of Female

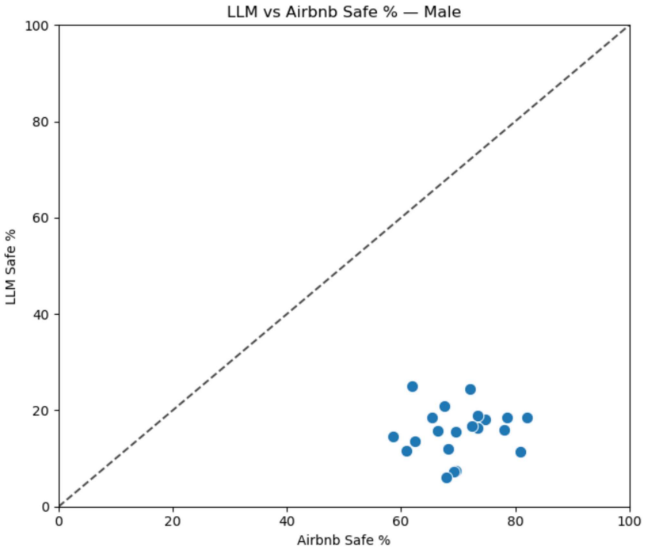


Figure 37: LLM vs. Airbnb Safe % of Male

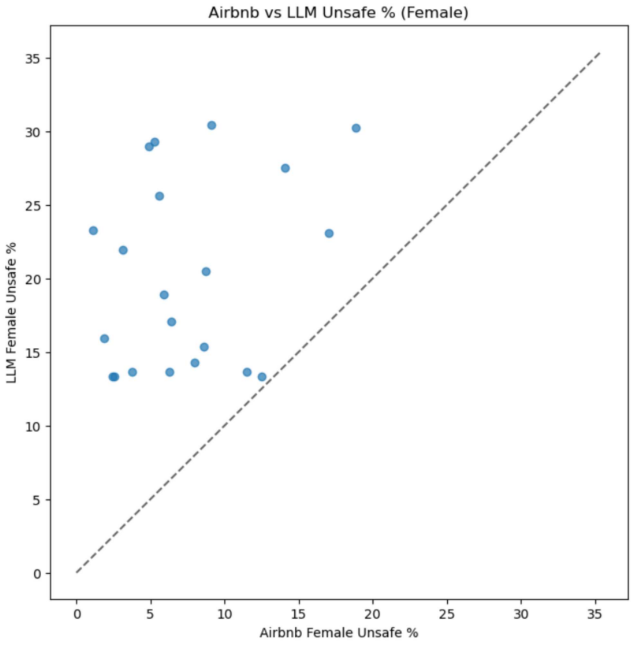


Figure 38: LLM vs. Airbnb Unsafe % of Female

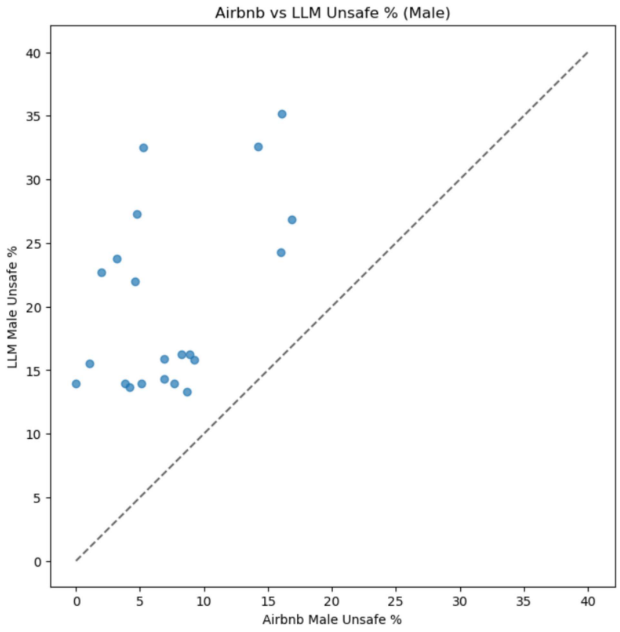


Figure 39: LLM vs. Airbnb Unsafe % of Male

The scatter plot analysis exhibit differences in how human reviewers and LLMs capture neighborhood safety. The data suggest that LLM alignments consistently fall below the diagonal reference line across both gender categories, revealing that LLMs capture neighbourhoods as safer than human Airbnb reviewers. This nuance suggests that LLMs exhibit bias when identifying unsafe cases. Gender analysis shows that male assessments illustrate greater variability, up to 35% unsafe ratings, compared to female assessments of 30% for Airbnb and 20% for LLMs. The safe percentage plots reinforce the bias, with Airbnb ratings ranging in the 60-80%

scale, while LLM alignment scale around 15-25%, indicating LLMs apply different values on labeling areas as safe. These findings define how human perceptions could lead to misaligned risk evaluations.

Topic Similarity and Lexicon Analysis

The topic similarity analysis heatmap reveals how relationships vary across neighbourhoods and genders, through LLM and Airbnb data. Results reflect how LLM models and Airbnb topic resemble or differ from each other, exhibiting the high or low matches between how genders write their experiences through both datasets.

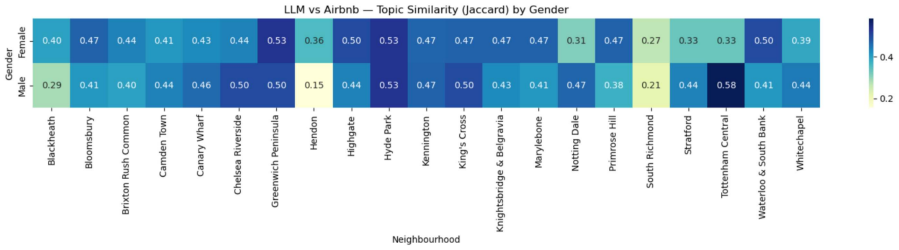


Figure 40: LLM vs. Airbnb topic similarity by gender

Model	Gender	Airbnb Safe	Airbnb Unsafe	LLM Safe	LLM Unsafe	Diff Safe	Diff Unsafe
Claude Sonnet 4	Female	0.904	0.096	0.518	0.482	-0.386	0.386
Claude Sonnet 4	Male	0.906	0.094	0.443	0.557	-0.463	0.463
GPT-5	Female	0.904	0.096	0.518	0.482	-0.386	0.386
GPT-5	Male	0.906	0.094	0.443	0.557	-0.463	0.463

Table 38: Safe vs. Unsafe proportions from Airbnb and LLM predictions with differences.

In Table 38 results present a comparative analysis of neighborhood safety perceptions between Airbnb and LLM predictions from Claude Sonnet 4 and GPT-5, disaggregated by gender. The Airbnb data display high safety perceptions, with female and male respondents reporting approximately 90–91% of neighborhoods as 'safe' and only 9–10% as 'unsafe'. However, LLM models predict lower proportions of safe neighborhoods and higher unsafe proportions. For example, Claude Sonnet 4 estimates 51.8% of neighborhoods as safe for females and only 44.3% for males, while GPT-5 shows similar patterns, this shows that both models agree on their average sentiment proportions. The difference analysis (*Diff Safe* and *Diff Unsafe*) exhibit the difference between LLM predictions and Airbnb perceptions, revealing that LLMs tend to underestimate safety (negative *Diff Safe*) and overestimate risk (positive *Diff Unsafe*). The differences for male-rated neighborhoods, suggesting that model predictions variate more from human perceptions for male-related assessments. Overall, these results highlight that while LLMs capture general safety patterns, they differ from human experiences on safety, reflecting potential model bias and the need for careful interpretability when applying LLMs to social perception tasks.

The comparison between Airbnb and LLM perceptions, exhibits both Claude Sonnet 4 and GPT-5 underestimating London's neighborhood safety while overestimating unsafe references. This differentiation is reflected more for male-rated neighborhoods, indicating that model predictions differ more from human perceptions in such cases. Despite differences in magnitude, both LLM models reveal similar trends, suggesting a consistent pattern of safety references. These results highlight potential model biases and emphasize the need for cautious interpretation when using LLMs to evaluate social perceptions of safety.

Top and bottom neighborhoods for topic overlap reveal that LLMs capture some themes well (e.g., Tottenham Central) but diverge significantly in others (e.g., Primrose Hill, Hendon).

Lexicon-based analysis of gendered safety words (Table 41) and normalized frequencies show that GPT-5 and Claude Sonnet 4 emphasize different safety descriptors than Airbnb reviews.

Neighbourhood	Gender	Model	Topic Similarity
Tottenham Central	male	Claude Sonnet 4	0.500
Hyde Park	male	Claude Sonnet 4	0.467
Hyde Park	female	Claude Sonnet 4	0.467
Knightsbridge & Belgravia	male	Claude Sonnet 4	0.462
Waterloo & South Bank	female	Claude Sonnet 4	0.462
Stratford	female	Claude Sonnet 4	0.462
Chelsea Riverside	male	Claude Sonnet 4	0.438
King's Cross	male	Claude Sonnet 4	0.438
Highgate	female	Claude Sonnet 4	0.438
Greenwich Peninsula	male	Claude Sonnet 4	0.429

Table 39: Top 10 Highest Topic Overlap Across Neighborhoods, Gender, and Model

Neighbourhood	Gender	Model	Topic Similarity
Hendon	male	Claude Sonnet 4	0.077
Primrose Hill	male	GPT-5	0.077
South Richmond	male	GPT-5	0.111
Whitechapel	male	GPT-5	0.125
Kennington	male	GPT-5	0.133
Marylebone	male	GPT-5	0.133
Chelsea Riverside	male	GPT-5	0.143
Brixton Rush Common	male	GPT-5	0.143
South Richmond	male	Claude Sonnet 4	0.143
Notting Dale	male	GPT-5	0.154

Table 40: Top 10 Lowest Topic Overlap Across Neighborhoods, Gender, and Model

To examine what words and content of safety-related themes genders use through Airbnb and LLM reviews, we defined a lexicon block of safe and unsafe descriptors (e.g., safe, unsafe, dangerous, etc.) and gender specified terms (e.g., solo female, family-friendly, etc.), implemented on normalized data. The counts were normalized by the total number of words in sentences, expressed as mentions per 1,000 words, ensuring that results are comparable across datasets of different sizes. Analyzing words and phrases related to safety terms across genders in Airbnb reviews and LLM-generated text, allows us to compare how LLMs and human reviewers discuss gendered safety concerns in 21 London neighborhood descriptions. The results help to identify how safety across genders is discussed between human reviews and LLM-generated sentences. By exploring these nuances, we can assess whether LLMs capture gendered patterns in the same way as human-authored reviews, addressing RQ1.

Lexicon Word	Airbnb	GPT-5	Claude Sonnet 4
quiet	87.426	25.516	1.018
safe	13.359	28.149	3.206
busy	3.830	16.201	0.356
avoid	0.068	17.011	2.239
female	0.079	4.050	14.552
men	0.000	18.429	0.000
male	0.000	4.253	12.314
noisy	6.526	8.505	0.000
secure	0.555	10.733	2.951
dark	0.136	12.758	0.051
women	0.034	11.746	0.000
empty	0.011	8.505	0.051
peaceful	4.770	0.000	1.577
friendly	1.156	3.848	0.458
isolated	0.057	4.253	0.712
comfortable	4.170	0.000	0.509
unsafe	0.283	4.253	0.000
uncomfortable	0.023	4.253	0.051
crime	0.000	0.000	3.409
well-lit	0.000	0.000	1.526

Table 41: Normalized frequency (per 1,000 words) of gendered-safety lexicon across Airbnb and LLM outputs.

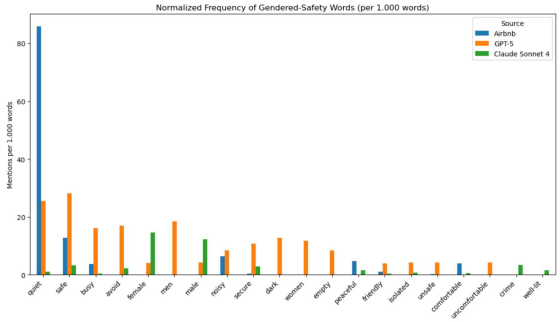


Figure 41: Normalized frequency of gendered-safety words (per 1,000).

Sentiment Pattern Analysis

Comparisons of sentiment proportions highlight that LLMs often generate higher unsafe sentiment than human reviews, especially in neighborhoods such as Waterloo & South Bank and Whitechapel. Chi-square tests and t-tests (Table 44) indicate no statistically significant differences in term usage across gender (Airbnb) or between LLM models, suggesting that while frequency patterns vary, the overall distribution is similar.

Gender	Neighbourhood	Sentiment	Count	Source	Proportion %
Female	Blackheath	Neutral	13	Airbnb	0.245
Female	Blackheath	Safe	39	Airbnb	0.736
Female	Blackheath	Unsafe	1	Airbnb	0.019
Female	Bloomsbury	Neutral	126	Airbnb	0.230
Female	Bloomsbury	Safe	375	Airbnb	0.684
...	...	...	...	...	...
Male	Waterloo & South Bank	Safe	3	GPT-5	0.250
Male	Waterloo & South Bank	Unsafe	6	GPT-5	0.500
Male	Whitechapel	Neutral	4	GPT-5	0.333
Male	Whitechapel	Safe	2	GPT-5	0.167
Male	Whitechapel	Unsafe	6	GPT-5	0.500

Table 42: Sentiment proportions across gender, neighbourhood, and source (Airbnb vs. GPT-5).

The sentiment analysis shows that Airbnb reviews are exhibiting bias to safe classes, with unsafe mentions being lower (e.g., 1 unsafe review in Blackheath, while 39 safe). Model GPT-5 outputs often presented higher proportions of unsafe sentences, such as Waterloo & South Bank (50% unsafe) or Whitechapel (50% unsafe). This divergent results highlight differences on human-authored reviews emphasizing more at safety, while LLMs generate more unsafe tones. This finding is important for RQ1, as it suggests that LLMs may not capture or reflect the gendered sentiment similar to human reviews, raising questions about how AI might shape perceptions of safety when used in real-world travel suggestions.

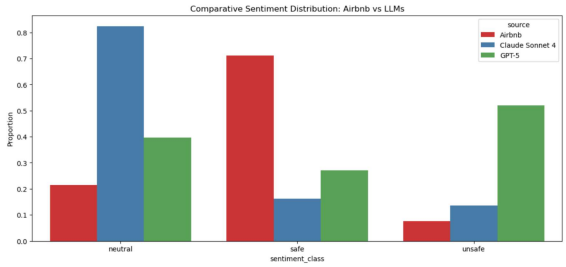


Figure 42: Sentiment Distribution between Airbnb vs LLM models.

Safety-Related Analysis Terms

Safety-related analysis in terms such as safe, unsafe, danger and secure, is essential to evaluate whether LLMs resemble to human safety perceptions in London's neighbourhood descriptions. Examining in both Airbnb reviews and LLM-generated content (ChatGPT-5 and Claude Sonnet 4) safety terms that might illustrate similar or different patterns. Term themes were normalized per 1,000 words to ensure comparability across datasets of different sizes. This approach reveals which safety descriptors are emphasized by humans and whether AI models reflect similar patterns.

Term	Frequency per 1,000 Words	Source
comfort	0.021	Airbnb
peaceful	4.017	Airbnb
quiet	37.691	Airbnb
risk	0.011	Airbnb
safe	10.091	Airbnb
secure	0.288	Airbnb
unsafe	0.256	Airbnb
quiet	12.758	GPT-5
safe	15.391	GPT-5
secure	6.480	GPT-5
comfort	0.153	Claude Sonnet 4
peaceful	1.577	Claude Sonnet 4
quiet	0.509	Claude Sonnet 4
safe	3.002	Claude Sonnet 4
secure	2.748	Claude Sonnet 4

Table 43: Normalized Frequency of Safety-Related Terms per 1,000 Words Across Sources

Test	Comparison	Test Statistic	p-value	Significance
Chi-square	Airbnb: Gender vs. Safe/Unsafe	0.0065	0.936	Not Significant
T-test	GPT-5 vs. Claude Sonnet 4	3.677	0.058	Not Significant

Table 44: Statistical Tests of Safety-Related Term Usage

The chi-square test for Airbnb reviews gave $x^2 = 0.0065$ with p=0.936, suggesting insignificant difference in safety-related term usage between female and male reviewers. Similarly, the t-test

comparing GPT-5 and Claude Sonnet 4 produced t=3.677 and p=0.058, slightly above the usual threshold for significance. This results indicate on: while GPT-5 shows a slightly higher frequency of safety-related terms, the difference between GPT-5 and Claude is not statistically meaningful. Overall, these findings inform that both LLM and human reviews exhibit similar distributions of safety language, supporting the notion that LLMs largely capture the patterns present in human-authored content.

Topic Modeling Content Analysis

In this section, we examine topic modeling content analysis and find the amount of mentions across Airbnb and LLMs reviews. Each dataset was categorized according to predefined content topics (e.g., Accessibility, Accommodation, Neighborhood, Safety). The frequency of each topic was calculated as raw counts per dataset to provide a direct comparison between human-authored reviews and LLM-generated content. This approach allows us to identify which topics are emphasized or marginalized by LLMs compared to human reviewers, providing insight into potential gaps or biases in LLM-generated neighborhood descriptions.

Category	Airbnb	Claude Sonnet 4	GPT-5
Accessibility	29	0	0
Accommodation	444	266	23
Affordability	0	130	40
Amenities/Facilities	775	0	0
Architecture	0	45	15
Attraction	0	91	35
Cleanliness	259	0	0
Experiences	0	137	64
Hosts Appreciation	23	0	0
Location	347	0	0
Neighborhood	1997	199	47
Nightlife & Social Life	269	180	161
Noise	1533	0	57
Outdoors	150	39	34
Restaurants	354	123	11
Safety	299	134	67
Security	54	118	21
Transportation	514	100	35
Value for Money	28	0	0

Table 45: Topic Modeling Absolute Counts by Source

The analysis revealed that certain topics, such as Neighborhood and Safety, were heavily represented across all data sources, suggesting that all pay higher attention to safety areas. However, topics like Affordability and Experiences were more developed in LLM outputs, while Airbnb reviews emphasized on Amenities and Noise. These differences highlight how LLMs illustrate and prioritize content differently from human reviewers, which is crucial for understanding the comparability of LLM outputs with real-world perceptions, informing RQ1 about content alignment between human and AI-generated neighborhood descriptions.

Semantic Similarity Scores

Semantic alignment between human-authored Airbnb reviews and LLM-generated sentences, was employed through sentence embeddings, using the Sentence-BERT model *("all-MiniLM-L6-v2")*, and computed by cosine similarity scores. In Airbnb review sentences comparison against GPT-5 and Claude Sonnet 4 outcomes, we quantify how closely LLM-generated content reflects human perceptions of 21 selected London neighborhoods. Average cosine similarity scores indicated moderate alignment, with GPT-5 resulting a slightly higher similarity (0.283) compared to Claude Sonnet 4 (0.232). These results inform that both LLMs capture some of the semantic content present in human reviews, but there are some differences in phrasing.

This analysis is important for assessing the precision of LLM-generated descriptions relative to human-authored experiences, comprehending content resemblance between AI and human-authored reviews.

Model	Average Cosine Similarity to Airbnb
GPT-5	0.2831
Claude Sonnet 4	0.2320

Table 46: Average semantic similarity between LLM-generated sentences and Airbnb reviews.

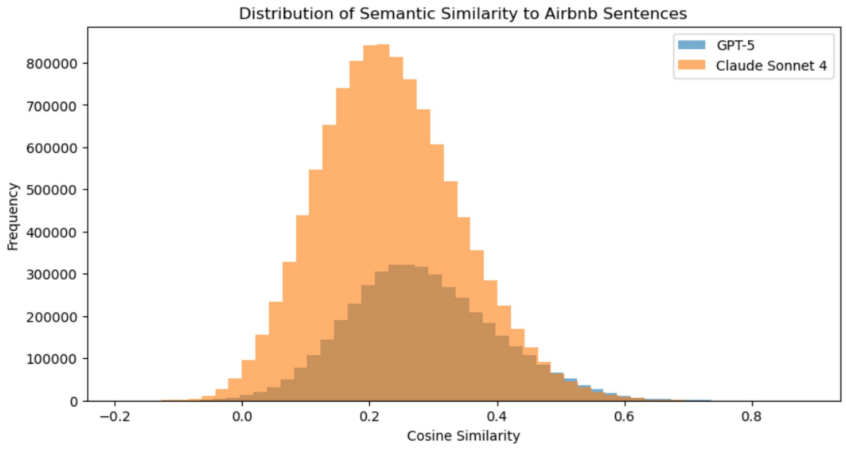


Figure 43: Distribution of Semantic Similarity of LLMs to Airbnb sentences.

3.7.2 Safety Sensitivity and Crime Alignment Analysis

This section investigates whether LLMs reflect gendered safety concerns in their travel suggestions and whether their outputs align with human-authored reviews and crime data. In that way, we address our two research questions RQ2-RQ3. **RQ2** asks whether LLMs capture gendered safety perceptions in neighborhood descriptions and how these relate to crime statistics. While **RQ3** explores whether LLMs exhibit gender bias in predicting 'safe'-related neighborhoods and how such biases compare to real-world evidence, from Airbnb reviews and crime data.

To evaluate how well LLM-generated London neighborhood descriptions and human-authored Airbnb reviews align or differ with ground-truth safety results, we developed a correlation analysis. Safety perception scores, extracted separately from Airbnb reviews and LLM prompts, were compared with actual crime data for London neighborhoods. The **non-parametric Spearman** rank correlation was chosen (appropriate for ordinal or non-linear relationships), to analyze whether perceived safety differed or matched its relationship to crime rates across genders perspectives. This approach is significant, because it reveals whether LLMs and Airbnb reflect safety concerns or differ from crime nuances. To compare perceptions of safety, qualitative text was transformed into numeric sentiment scores and safety-related terms were converted to classes (*safe = 1, neutral = 0, and unsafe = −1*). Both Airbnb reviews and LLM-generated outputs were aggregated by neighborhood and gender to produce mean safety scores. Crime data was normalized into a crime score by dividing incident counts by neighborhood population, providing an objective measure of safety, without biased results. **Spearman's rank correlation** was applied to examine the relationship between perceived safety of Airbnb, LLMs and crime rates disaggregated by gender.

The results show a clear divergence between human-authored reviews and model outputs. Airbnb reviews demonstrated a negative correlation with crime, especially for male reviewers (r = –0.519, p = 0.016), suggesting that human perceptions align with real-world risk. In contrast, LLM outputs showed weak and statistically insignificant correlations, with GPT-5 and Claude Sonnet 4 failing to capture neighborhood-level crime patterns. This indicates that neither model demonstrates genuine sensitivity to crime data, even though Airbnb users often report perceptions consistent with actual risk levels.

Gender	LLM–Crime Corr	LLM p-value	Airbnb–Crime Corr	Airbnb p-value
Female	0.194	0.399	-0.412	0.064
Male	0.058	0.801	-0.519	0.016

Table 47: Correlation of Crime Rates with Safety Perceptions by Gender and Source.

When comparing the two models, Claude Sonnet 4 exhibited slightly higher correlations with Airbnb reviews (female r = 0.275, male r = 0.267) than GPT-5 (female r = 0.328, male r = 0.211), suggesting marginally closer alignment with human perception. However, both remained largely unresponsive to crime statistics, reinforcing their limitations in reflecting empirical realities (Table 48).

Model	Gender	LLM–Crime Corr	LLM–Crime p-value	LLM–Airbnb Corr	LLM–Airbnb p-value
Claude Sonnet 4	Female	0.208	0.365	0.275	0.228
Claude Sonnet 4	Male	0.118	0.609	0.267	0.243
GPT-5	Female	-0.116	0.617	0.328	0.147
GPT-5	Male	0.038	0.869	0.211	0.358

Table 48: Correlation of LLMs with Crime and Airbnb by Gender

Beyond correlation, the frequency and framing of safety mentions reveal important differences. Airbnb reviews overwhelmingly described neighborhoods as safe, with 85% of female-authored sentences coded as positive. By contrast, GPT-5 displayed a marked tendency to frame female-oriented responses in unsafe terms (67.7% of female sentences), while male prompts produced a more neutral tone (58.3%). Claude Sonnet 4, on the other hand, leaned heavily toward neutral framing for both genders, with over 84% of sentences categorized as neither explicitly safe nor unsafe. Figures 47 to 48 illustrate these contrasting distributions, highlighting GPT-5's gendered caution and Claude's overall neutrality.

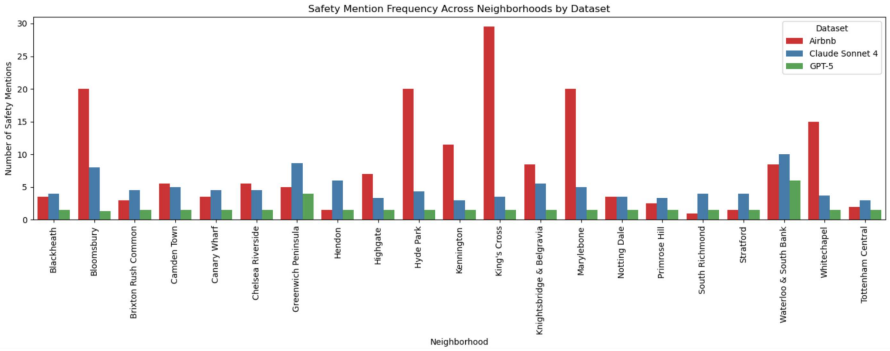
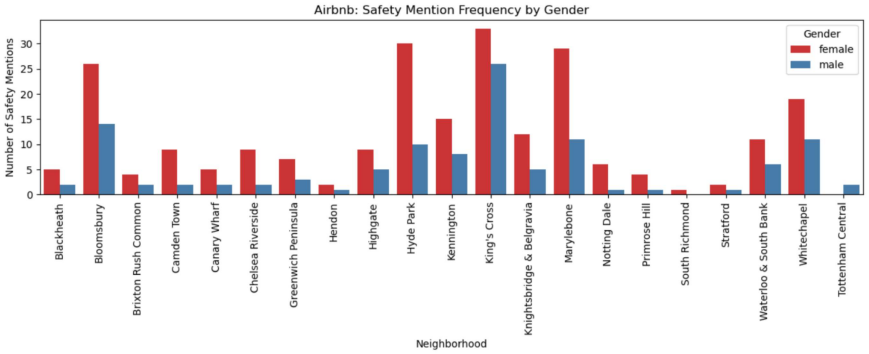
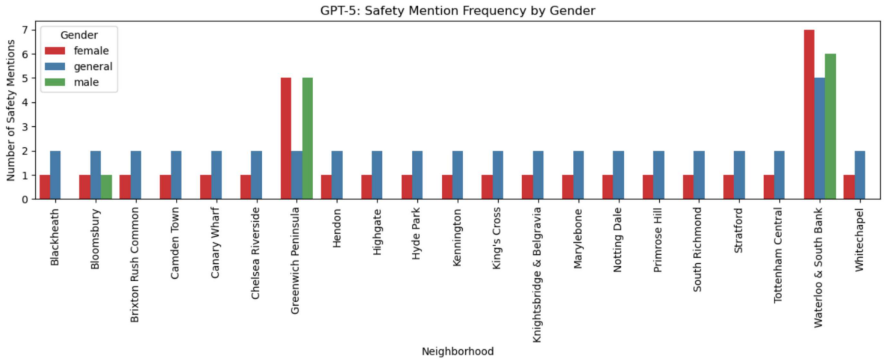


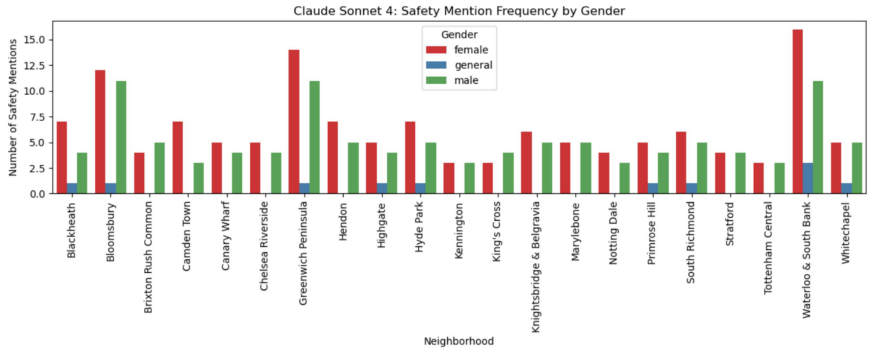
Figure 44: Safety mention frequency across neighbourhoods by data sources.



(a) Airbnb safety mention frequency by gender.



(b) GPT-5 safety mention frequency by gender.



(c) Claude safety mention frequency by gender.

Figure 45: Safety mention frequency across neighbourhoods by gender.

By investigating how safety concerns are expressed differently by gender across Airbnb reviews and LLM-generated content from GPT-5 and Claude Sonnet 4, we capture vital findings for the addressed research questions. This approach exhibits how safety concerns are framed differently for both gender groups. Only safety-related comments were employed, with categories such as 'safety', 'security', 'crime' and 'harassment', these subsets were then analyzed separately for each LLM model. For each gender within Airbnb and LLM datasets, we calculated the proportional distribution of safety-related sentiments—unsafe, neutral, and safe—allowing a direct comparison of how safety is contextualized differently between human-authored reviews and model-generated content.

Sentiment Source	Unsafe	Neutral	Safe
Airbnb Female	0.025	0.126	0.849
Airbnb Male	0.070	0.157	0.774
GPT-5 Female	0.677	0.194	0.129
GPT-5 Male	0.000	0.583	0.417
Claude Sonnet 4 Female	0.045	0.865	0.090
Claude Sonnet 4 Male	0.074	0.843	0.083

Table 49: Proportional distribution of safety-related sentiments by gender for Airbnb and LLM outputs.

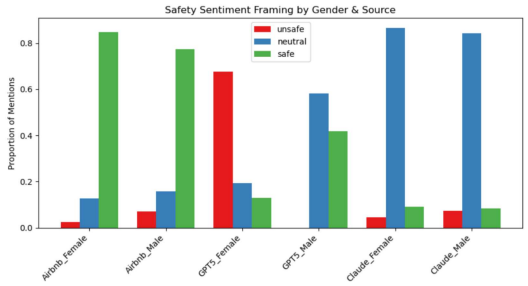
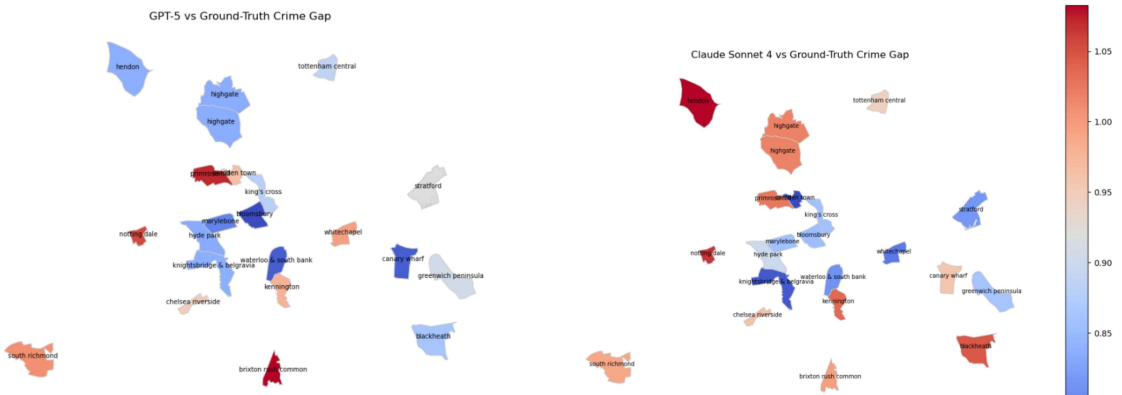


Figure 46: Safety sentiment framing by gender and data source.

Mapping model-generated safety predictions onto actual high- and low-crime neighborhoods reinforced these findings. GPT-5 frequently misclassified low-crime areas as unsafe, reflecting an overly cautious bias, particularly for female prompts. Claude Sonnet 4 achieved greater alignment with real-world labels, correctly identifying several high-crime areas while avoiding the systematic pessimism evident in GPT-5 outputs. Nevertheless, both models successfully captured the high-safety reputations of areas such as Knightsbridge & Belgravia and Marylebone (Table 50, Figures 47 a–d).

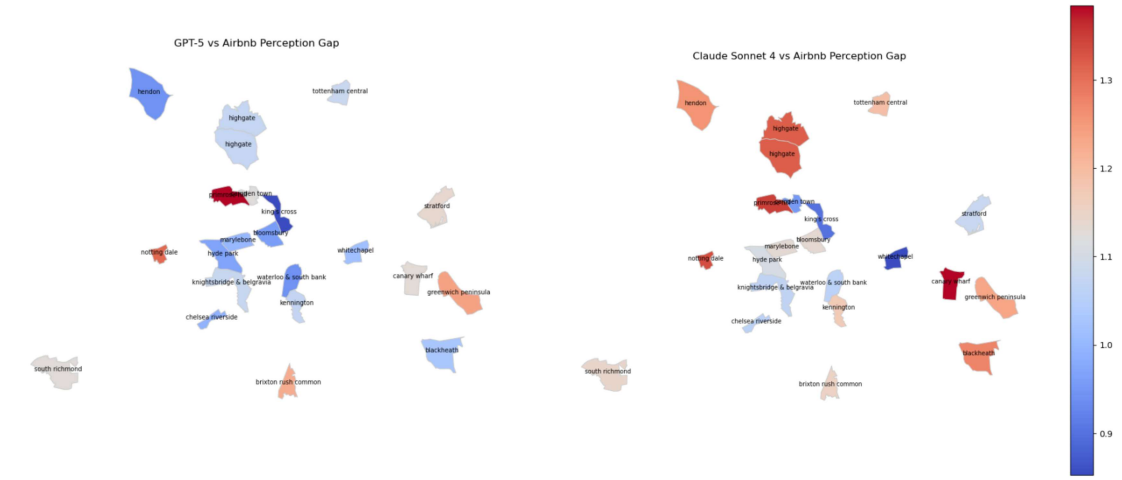
Neighbourhood	Crime Level	GPT-5 Pred	GPT-5 Correct	Claude Pred	Claude Correct
Chelsea Riverside	Low	High	False	High	False
Tottenham Central	Low	High	False	High	False
Blackheath	Low	High	False	Low	True
Camden Town	High	Low	False	High	True
Hendon	Low	High	False	Low	True
Highgate	Low	High	False	Low	True
King's Cross	High	Low	False	High	True
Stratford	High	Low	False	High	True
Whitechapel	High	Low	False	High	True
Bloomsbury	High	High	True	Low	False
Brixton Rush Common	Low	Low	True	High	False
Canary Wharf	High	High	True	Low	False
Greenwich Peninsula	Low	Low	True	High	False
Hyde Park	High	High	True	Low	False
Kennington	Low	Low	True	Low	True
Knightsbridge & Belgravia	High	High	True	High	True
Marylebone	High	High	True	High	True
Notting Dale	Low	Low	True	Low	True
Primrose Hill	Low	Low	True	Low	True
South Richmond	Low	Low	True	Low	True
Waterloo & South Bank	High	High	True	High	True

Table 50: Neighbourhood Safety Prediction Accuracy: GPT-5 vs Claude Sonnet 4



(a) GPT-5 vs Ground-truth Crime Gap.

(b) Claude vs Ground-truth Crime Gap.



(c) GPT-5 vs Airbnb Perception Gap.

(d) Claude vs Airbnb Perception Gap.

Figure 47: Illustration of 21 neighbourhoods, revealing diverge comparison gaps between models, crime data and Airbnb.

In summary, the analysis shows that while Airbnb reviews reflect both perceived and actual safety, LLMs do not. Claude Sonnet 4 demonstrates slightly closer alignment with human perceptions, but both models remain largely insensitive to real-world crime. GPT-5, in particular, exaggerates unsafe framings for female-oriented queries, while Claude defaults to a neutral narrative.

Gendered Recommendation Bias

While the previous subsection focused on sensitivity to safety and alignment with crime, this subsection evaluates whether LLMs exhibit systematic gender bias in their safe neighborhood predictions. Gender bias was assessed by comparing the proportion of 'safe' classifications generated for male versus female prompts across neighborhoods.

The results reveal notable disparities. Claude Sonnet 4 consistently predicted higher safe rates for female prompts than for male ones, with differences as large as +0.24 in Knightsbridge &

Belgravia and +0.22 in Tottenham Central. GPT-5 exhibited smaller but still meaningful gaps, most notably in Tottenham Central (+0.27). Across all neighborhoods, Claude's predictions had a higher mean female–male difference (0.064) and greater variability (standard deviation = 0.077) than GPT-5 (mean = 0.037, standard deviation = 0.069). This suggests that Claude's outputs are not only more biased on average but also more inconsistent across neighborhoods (Tables 51–52, Figures 48–49).

	Neighbourhood	Model	male_safe_pct	female_safe_pct	diff_safe_pct_llm	Crime Neighbourhood	crime_per_1000	diff_safe_pct_airbnb
0	Blackheath	Claude Sonnet 4	0.161	0.129	-0.032	Blackheath	6.921	-0.050
1	Blackheath	GPT-5	0.250	0.308	0.058	Blackheath	6.921	-0.050
2	Bloomsbury	Claude Sonnet 4	0.194	0.188	-0.006	Bloomsbury	38.093	0.020
3	Bloomsbury	GPT-5	0.000	0.143	0.143	Bloomsbury	38.093	0.020
4	Brixton Rush Common	Claude Sonnet 4	0.100	0.152	0.052	Brixton Rush Common	3.714	0.065
5	Brixton Rush Common	GPT-5	0.167	0.154	-0.013	Brixton Rush Common	3.714	0.065
6	Camden Town	Claude Sonnet 4	0.148	0.179	0.030	Camden Town	21.065	-0.002
7	Camden Town	GPT-5	0.100	0.091	-0.009	Camden Town	21.065	-0.002
8	Canary Wharf	Claude Sonnet 4	0.161	0.207	0.046	Canary Wharf	9.090	-0.021
9	Canary Wharf	GPT-5	0.250	0.308	0.058	Canary Wharf	9.090	-0.021
10	Chelsea Riverside	Claude Sonnet 4	0.242	0.250	0.008	Chelsea Riverside	5.536	0.026
11	Chelsea Riverside	GPT-5	0.250	0.308	0.058	Chelsea Riverside	5.536	0.026
12	Greenwich Peninsula	Claude Sonnet 4	0.152	0.226	0.074	Greenwich Peninsula	8.517	0.059
13	Greenwich Peninsula	GPT-5	0.182	0.250	0.068	Greenwich Peninsula	8.517	0.059
14	Hendon	Claude Sonnet 4	0.121	0.097	-0.024	Hendon	4.809	0.000
15	Hendon	GPT-5	0.250	0.308	0.058	Hendon	4.809	0.000
16	Highgate	Claude Sonnet 4	0.156	0.194	0.037	Highgate	9.214	0.054
17	Highgate	GPT-5	0.250	0.308	0.058	Highgate	9.214	0.054
18	Hyde Park	Claude Sonnet 4	0.161	0.194	0.032	Hyde Park	19.158	0.019
19	Hyde Park	GPT-5	0.250	0.231	-0.019	Hyde Park	19.158	0.019
20	Kennington	Claude Sonnet 4	0.034	0.138	0.103	Kennington	3.824	0.037
21	Kennington	GPT-5	0.182	0.167	-0.015	Kennington	3.824	0.037
22	King's Cross	Claude Sonnet 4	0.065	0.185	0.121	King's Cross	14.686	-0.055
23	King's Cross	GPT-5	0.250	0.231	-0.019	King's Cross	14.686	-0.055
24	Knightsbridge & Belgravia	Claude Sonnet 4	0.160	0.400	0.240	Knightsbridge & Belgravia	18.005	-0.001
25	Knightsbridge & Belgravia	GPT-5	0.250	0.250	0.000	Knightsbridge & Belgravia	18.005	-0.001
26	Marylebone	Claude Sonnet 4	0.194	0.258	0.065	Marylebone	20.038	0.069
27	Marylebone	GPT-5	0.250	0.308	0.058	Marylebone	20.038	0.069
28	Notting Dale	Claude Sonnet 4	0.034	0.160	0.126	Notting Dale	11.016	0.131
29	Notting Dale	GPT-5	0.167	0.154	-0.013	Notting Dale	11.016	0.131
30	Primrose Hill	Claude Sonnet 4	0.161	0.219	0.057	Primrose Hill	6.149	0.138
31	Primrose Hill	GPT-5	0.167	0.154	-0.013	Primrose Hill	6.149	0.138
32	South Richmond	Claude Sonnet 4	0.167	0.212	0.045	South Richmond	8.201	-0.068
33	South Richmond	GPT-5	0.167	0.167	0.000	South Richmond	8.201	-0.068
34	Stratford	Claude Sonnet 4	0.094	0.226	0.132	Stratford	11.468	-0.198
35	Stratford	GPT-5	0.167	0.167	0.000	Stratford	11.468	-0.198
36	Tottenham Central	Claude Sonnet 4	0.077	0.300	0.223	Tottenham Central	6.542	0.101
37	Tottenham Central	GPT-5	0.000	0.273	0.273	Tottenham Central	6.542	0.101
38	Waterloo & South Bank	Claude Sonnet 4	0.250	0.179	-0.071	Waterloo & South Bank	18.336	0.121
39	Waterloo & South Bank	GPT-5	0.250	0.308	0.058	Waterloo & South Bank	18.336	0.121
40	Whitechapel	Claude Sonnet 4	0.138	0.233	0.095	Whitechapel	9.392	-0.030
41	Whitechapel	GPT-5	0.167	0.154	-0.013	Whitechapel	9.392	-0.030

Table 51: Comparison of safe prediction percentages (male vs. female) across neighbourhoods and models, alongside crime rates and Airbnb disparities.

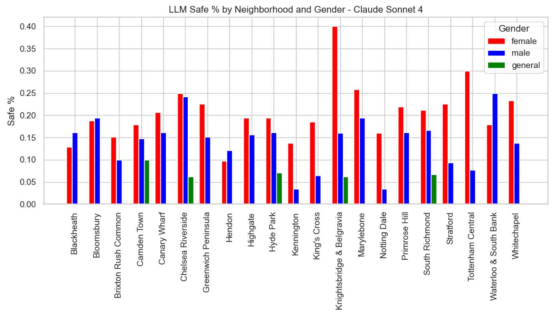


Figure 48: LLM safe % by neighbourhood and gender- Claude Sonnet 4.

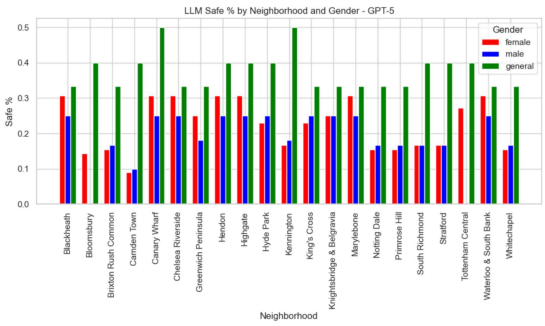


Figure 49: LLM safe % by neighbourhood and gender- GPT-5.

The analysis of gendered safe prediction percentages across neighborhoods illustrates regular patterns in LLMs behavior. For each model, the difference between female and male safe percentages was computed and found positive values, indicating higher safe rates for female prompts. The statistical summary shows that Claude Sonnet 4 exhibits a higher average gender bias (mean = 0.064) with a wider spread (std = 0.077) compared to GPT-5 (mean = 0.037, std = 0.069). Both models display some neighborhoods with negative differences, indicating occasional higher male safe rates, but overall, female prompts tend to be predicted as safer. These results suggest that while gender bias exists in both LLMs, Claude Sonnet 4 demonstrates more pronounced and variable preference patterns across neighborhoods. On this investigation we comprehend whether LLMs treat male and female prompts differently or similar across neighborhoods, by computing the difference in safe prediction percentages. In that way we can see systematic patterns of gender bias, where Claude Sonnet 4 shows a higher average bias and more variation across neighborhoods than GPT-5, suggesting that some models are more sensitive to gendered prompts. Understanding these differences is important because it highlights potential fairness issues and helps guide model evaluation, improvement, and responsible deployment in real-world applications.

Statistical testing confirmed the presence of gender bias in both models. One-sample t-tests showed significant differences between female and male safety predictions for Claude Sonnet 4 ($t = 3.85$, $p = 0.001$) and GPT-5 ($t = 2.44$, $p = 0.024$). By contrast, Airbnb reviews revealed no significant gender differences ($t = 1.16$, $p = 0.261$), with distributions across female and male authors nearly identical (Table 56; Figures 59–60). This indicates that the gendered bias detected in LLM outputs does not stem from patterns present in the real-world dataset.

Model	Count	Mean	Std	Min	25%	50%	75%	Max
Claude Sonnet 4	21	0.064	0.077	-0.071	0.030	0.052	0.103	0.240
GPT-5	21	0.037	0.069	-0.019	-0.013	0.000	0.058	0.273

Table 52: Descriptive statistics of female–male safe prediction differences across neighborhoods for each LLM. Positive values indicate higher safe percentages for female prompts.

To assess whether LLMs and Airbnb data display regular gender bias, we develop a one-sample t-test performance on the female–male safe percentage differences per neighborhood. The null hypothesis assumed no gender difference (mean difference = 0), and the test evaluated whether the observed mean deviations were statistically significant. This approach quantifies whether models or real-world data frequently favor one gender in predicted safety.

Dataset	t-statistic	p-value	Significance
Claude Sonnet 4	3.85	0.0010	Yes
GPT-5	2.44	0.0243	Yes
Airbnb	1.16	0.2610	No

Table 53: One-sample t-test results for female–male safe percentage differences across neighborhoods. "Significance" indicates whether the mean difference is significantly different from zero ($\alpha = 0.05$).

Fairness was further evaluated using demographic parity and equalized odds. In terms of demographic parity, Claude Sonnet 4 exhibited the strongest female skew (0.072), followed by GPT-5 (0.049), while Airbnb reviews were almost perfectly balanced (0.003). Equalized odds analysis showed that Claude produced higher false positive rates for female prompts, suggesting a tendency to overstate safety for women, while GPT-5 exhibited smaller but still present disparities. Airbnb data again demonstrated nearly equal treatment across genders, with both true positive and false positive rates showing negligible variation.

Hypothesis testing reveals that both LLMs exhibit a significant gender bias, with Claude Sonnet 4 ($t = 3.85$, $p = 0.001$) and GPT-5 ($t = 2.44$, $p = 0.024$) producing higher safe percentages for female prompts on average. However, Airbnb data doesn't show a significant gender difference ($t = 1.16$, $p = 0.261$). Comparing LLM and Airbnb patterns across neighborhoods reveals a very low correlation ($r = 0.08$), indicating that LLM gender preferences are largely independent of real-world patterns observed on Airbnb. Visualizations, including scatter plots and distribution histograms, further highlight that LLMs display consistent gendered tendencies across neighborhoods, with variability between models, while real-world data shows weaker or negligible gender differences

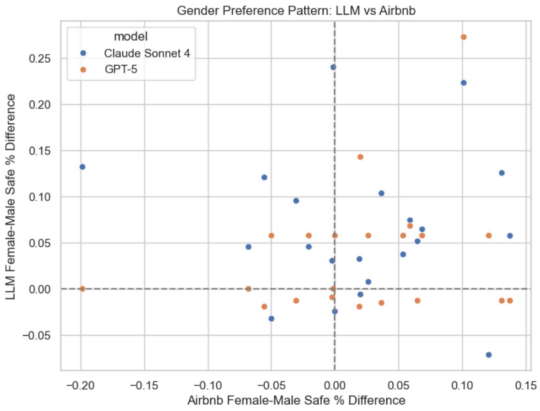


Figure 50: Gender preference patterns.

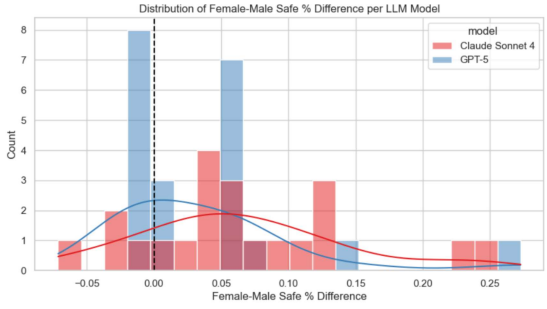


Figure 51: Gender distribution % difference per LLM-model.

The statistical analysis illustrates a critical disconnect between AI model bias and real-world patterns. Both Claude Sonnet 4 and GPT-5 demonstrate crucial gender bias, frequently rating neighborhoods as safer for females than males ($p < 0.05$), while actual Airbnb user data shows no statistically significant gender difference in safety perceptions (p = 0.261). The low correlation of r = 0.08 on Airbnb patterns and LLM predictions relationship, suggest that AI models generate artificial gender bias rather than reflecting human behavior patterns. That informs that AI models have learned false gender associations during training that don't correspond to actual gendered differences in safety assessment, making their bias problematic since they seem to exhibit bias by themselves, rather than derived from real-world data. On Figure 50 the results show that both LLM models GPT-5 and Claude Sonnet 4 are heavily biased towards female-male differences, with GPT-5 showing a more exaggerated bias range around 0.05-0.10.

Bias quantification

Bias quantification evaluates whether models make predictions fairly across London neighbourhoods. Fairness metrics of bias quantification study are: demographic parity, which measures whether each gender receives positive predictions at similar rates, and equalized odds, which assesses whether predictive accuracy is consistent across genders. By comparing LLM predictions with real-world Airbnb data, we can identify systematic gender biases in model outputs relative to observed outcomes. Demographic parity estimates whether different genders are equally reflected on 'safe' predictions, regardless of the true safety. For LLMs, female prompts are predicted as safe more often than male prompts, with differences of 0.072 for Claude Sonnet 4 and 0.049 for GPT-5. In contrast, Airbnb data displays low gender disparity (dp diff = 0.003), suggesting that illustrated differences in the LLMs are not reflective of real-world gender patterns. This suggests that LLMs may regularly favor female prompts in safe classifications, even when ground truth safety is balanced.

Dataset / Model	Female Safe Rate	Male Safe Rate	DP Difference
Claude Sonnet 4 (LLM)	0.748	0.675	0.072
GPT-5 (LLM)	0.308	0.259	0.049
Airbnb	0.875	0.872	0.003

Table 54: Demographic parity analysis: average safe prediction rates by gender and difference (female - male).

Dataset / Model	Gender	TPR	FPR
Claude Sonnet 4 (LLM)	Female	0.424	0.774
Claude Sonnet 4 (LLM)	Male	0.488	0.687
GPT-5 (LLM)	Female	0.326	0.307
GPT-5 (LLM)	Male	0.290	0.257
Airbnb	Female	0.920	0.874
Airbnb	Male	0.929	0.870

Table 55: Equalized odds analysis: true positive rate (TPR) and false positive rate (FPR) by gender for LLMs and Airbnb data.

To further the investigation part of this project we develop a model comparison evaluation approach, examining on whether different LLMs exhibit distinct patterns of fairness and predictive accuracy. In the study we use demographic parity and equalized odds to assess how each model treats female and male prompts in predicting 'safe' outcomes relative to neighborhood crime data. This allows identification of model-specific biases and their potential impact on decision-making. Demographic parity measures whether female and male prompts are equally likely to

receive safe predictions. Claude Sonnet 4 predicts higher safe percentages for females (0.748) than males (0.675), with a difference of 0.072. GPT-5 shows a smaller difference (0.049), indicating less pronounced gender bias. These results demonstrate that model choice can influence the degree of gender disparity in predictions.

Model	Female Safe Rate	Male Safe Rate	General	DP Difference
Claude Sonnet 4	0.748	0.675	1.000	0.072
GPT-5	0.308	0.259	1.000	0.049

Table 56: Demographic parity for each LLM model: average safe prediction rates by gender and the difference (female and male).

Equalized odds examines whether predictive performance is consistent across genders. Claude Sonnet 4 shows a higher false positive rate for females (0.894) than males (0.807) and slightly lower true positive rate for females (0.639) compared to males (0.545). GPT-5 shows smaller differences between genders in both TPR and FPR. These findings indicate that both models exhibit gender-dependent variations in predictive accuracy, with Claude Sonnet 4 showing more pronounced differences.

Model	Gender	TPR	FPR
Claude Sonnet 4	Female	0.639	0.894
Claude Sonnet 4	Male	0.545	0.807
GPT-5	Female	0.323	0.293
GPT-5	Male	0.277	0.240

Table 57: Equalized odds for each LLM model: true positive rate (TPR) and false positive rate (FPR) by gender.

Taken together, these results demonstrate that both LLMs exhibit systematic gender bias in their recommendations. Claude Sonnet 4 not only produces greater disparities but also does so inconsistently across neighborhoods. GPT-5 is somewhat more stable but still produces context-dependent bias. Importantly, such bias is not grounded in real-world patterns, where Airbnb reviews display no evidence of systematic gender differences.

3.7.3 Summary

The findings between both RQ2 and RQ3 highlight significant limitations in the capacity of LLMs to provide accurate and relevant travel safety information. Both models exhibit weak sensitivity to crime data, while Airbnb reviews exhibit meaningful alignment with real-world risk. GPT-5 and Claude Sonnet 4 differ in their safety framing, with GPT-5 disproportionately cautious for female prompts and Claude leading to neutrality. Both AI-models tend to favor female prompts in safe predictions, though with different magnitudes, Claude more variable and GPT-5 more stable. Finally, fairness metrics reveal that neither model achieves demographic parity or equalized odds, in contrast to Airbnb reviews which treat male and female safety perceptions almost identically.

These results underscore the ethical challenges of relying on LLMs for travel safety guidance. The observed biases are not only inconsistent with empirical patterns but risk reinforcing gendered stereotypes or creating misleading assurances. Such limitations point to the need for fairness-aware evaluation frameworks and caution in the deployment of generative AI for socially sensitive applications.

4 Results

In this chapter, are exhibited the findings from the comparative analysis of Airbnb human-authored reviews, actual crime data, and LLM-generated outputs (GPT-5 and Claude Sonnet 4) in relation to neighborhood safety perceptions in London. The results address the research questions by examining differences in perceived safety, semantic and sentiment alignment, and gendered prediction bias, with reference to human-authored reviews and empirical crime rates.

4.1 LLM vs. Human Safety Perceptions

Airbnb reviewers tend to describe London neighborhoods as safe, with over 90% of reviews coded as positive across genders. By contrast, both GPT-5 and Claude predicted significant lower safety levels, with GPT-5 leaning toward caution and Claude toward neutrality. Scatter plots confirmed that LLM predictions systematically underestimated neighborhood safety compared with Airbnb reviews. These results suggest that while human-authored reviews highlight reassurance, LLMs reinforce risk.

4.2 Alignment with Crime and Gendered Sensitivity

When compared with crime data, Airbnb reviews displayed moderate correlations, indicating that user perceptions partly reflect real-world risk. In contrast, LLM predictions showed weak or negative correlations with crime, misclassifying many low-crime areas as unsafe. GPT-5 outputs emphasized unsafe framings for female prompts, while Claude defaulted to neutral responses regardless of gender. Sentiment and lexical analyses reinforced this divergence: Airbnb reviews favored positive descriptors ('quiet', 'peaceful'), GPT-5 introduced caution terms ('unsafe', 'dark') and Claude remained neutral. Overall, LLMs captured some safety-related language but failed to reproduce the gendered nuances or empirical alignment with human-authored data.

4.3 Gender Bias and Fairness in Predictions

Systematic gender disparities were observed in both models, but were absent in Airbnb reviews. Claude consistently predicted higher safety for female prompts, with gaps exceeding 20% in some neighborhoods, while GPT-5 displayed smaller but still significant differences. Fairness metrics confirmed that Claude showed the largest demographic parity differences and higher false positive rates for female prompts, while GPT-5 exhibited more stable but still unequal outcomes. However, Airbnb data revealed insignificant gender differences. This indicates that LLMs generate artificial gender bias not grounded in real-world nuances.

4.4 Summary of Findings

Across all analyses, three main conclusions emerge:

- Human vs. LLM perceptions (RQ1): Airbnb reviewers perceive London neighborhoods highly safe, while LLMs underestimate safety and emphasize risk.
- Crime and sensitivity (RQ2): Airbnb reviews align moderately with crime data, but LLMs fail to capture these dynamics, with GPT-5 overly cautious for females and Claude

revealing neutrality.

- Bias and fairness (RQ3): Both models exhibit gender bias, with Claude showing stronger and more variable disparities and GPT-5 showing smaller but consistent imbalances.

All these findings underscore that LLM outputs differ significantly from human-authored reviews and real-world crime, while introducing gender biases. This highlights the limitations of relying on LLMs for socially consequential safety recommendations.

5 Discussion

This chapter critically interprets the findings of the study, positioning them within existing literature on urban safety perceptions, gender bias and AI fairness. It demonstrates the implications of the results for both research and performance, and addresses the methodological and ethical challenges of applying LLMs, such as GPT-5 and Claude Sonnet 4, in social contexts like neighborhood safety evaluation.

5.1 Divergences in Safety Perceptions

The systematic underestimation of safety by LLMs (44-52% safe classifications) compared to human-authored reviewers (90% safe) represents fundamental differences in how humans and AI systems exhibit safety narratives. Human reviewers seem to apply in their descriptions experiences like direct observation and social knowledge over statistical risk. The dominance of terms like 'quiet', 'peaceful' and 'friendly' in Airbnb reviews suggests safety is constructed through positive indicators rather than threat. In contrast, LLMs appear to weight negative priors more heavily, possibly reflecting training data biases or safety disclaimers in travel suggestions.

These differences affect trust and use of AI safety advice, if AI seems cautious and negative, people might ignore specific travel suggestion completely, even when warnings are valid. On the other hand, relying too much on cautious AI could unnecessarily limit travel, especially for people who are already careful. The 40% point gap between human and AI safety ratings shows that current models aren't ready to make safety suggestions on their own.

5.2 Gendered Biases

The investigation found systematic gender bias in LLMs, with female prompts more often labeled 'safe' than male ones, very different from Airbnb reviews results. Claude displayed stronger disparities, exhibiting structural or training factors encode gendered assumptions. Ethically, such bias risks reinforcing stereotypes and shaping traveler or policy decisions unfairly. One of the most significant findings is that LLMs rate female prompts as safer than male prompts, with Claude rating by 6.4% and GPT-5 by 3.7%, while human reviewers show insignificant difference and crime data doesn't support these patterns.

This gendered bias seems artificial, likely introduced by the AI rather than reflecting reality. Possible causes include could be training data patterns, where safety content often targets women specifically ('safe for solo female travelers'), creating a gap in the model between female-coded language and safety. Moreover, prompt sensitivity could suggest possible drawbacks in models performance, when a prompt explicitly mentions gender, the model may apply stereotyped assumptions, such 'female' towards cautious framing, while 'male' towards neutral framing. Alignment inconsistency caused from learning through human feedback, human reviewers might inherit and reinforce these gendered safety narratives, encoding the bias in the model.

All these possible causes differ on how humans show this systematic bias. While travel studies note some gender differences in perceived risk, Airbnb data shows safety ratings similarly across both genders for the same neighborhoods. This means LLMs are creating a distortion, not reproducing real-world patterns.

From a fairness perspective, this violates demographic parity, suggesting that predictions differ by gender even when real outcomes are the same. LLMs also make more false positives for female prompts, meaning they label places as safe more often for women than they should.

This bias is more than a technical drawback, it exhibits real-world consequences, such as:

- False security for women, overrating safety could lead to risky decisions.
- Perpetuating male stereotypes, underrating safety for male prompts reinforces assumptions that men are naturally more risk-tolerant.

In short, these models aren't just reflecting society, they're amplifying biases in ways humans don't.

5.3 Alignment with Reality

LLM safety statistics illustrated weak connection to real crime data, while Airbnb reviews matched crime patterns more closely, reflecting neighborhood realities better. This gap shows the impact of relying too much on AI for safety or travel decisions suggestions, leading to misinform users, manipulate community perceptions and worsen develop inequalities in already stigmatized areas.

LLMs safety ratings barely match actual crime rates (correlations from -0.16 to 0.21), while Airbnb reviews align stronger with crime data (r = 0.36–0.52, especially for male reviewers). This suggests humans alignment on real world experiences, but LLMs rely on textual associations that aren't grounded in reality. Several factors contribute to this disconnect:

- Different training periods – LLMs are trained on data before 2024, Airbnb contain a good range of real experience between 2009-2025 and crime data from April 2024-April 2025, causing misinformation, because neighborhoods change rapidly so models miss recent trends.
- Different scales of measurement – Crime stats capture specific events (theft, violence, etc.), while LLMs cluster together multiple factors (crime, tourism, social environment). A place may be high-crime but still feel safe because of community presence.
- Language vs. reality – LLMs learn from how places are described, not the actual environment. Neighborhoods labeled as 'dangerous' or 'rough' in text may be flagged as unsafe, even if crime is low.

LLMs are great at generating text, but they cannot effectively predict real-world statistics like crime rates without retraining in recent data. The low similarity scores (0.23–0.28) between model outputs and human descriptions confirm that LLMs often produce reasonable but factually disconnected narratives.

5.4 Implications and Relevance

This study informs for AI fairness and responsibility, highlighting that human experience compared to LLM suggestions are vital for comprehending neighborhood safety. This project notifies us that LLMs inherit artificial gender bias and by leaving them unchecked, could distort perceptions and harm certain communities. Overall, the findings emphasize that AI outputs are not neutral and must be carefully evaluated before use, for transparency and oversight.

Incorrect safety ratings from LLMs can reinforce or create neighborhood stigma, especially if certain areas are consistently labeled unsafe or perceived differently for specific demographic groups. Developing important consequences like lower visitation and reduced local economic

activity, changes property values and investment decisions and create spatial inequality, particularly along racial and socioeconomic lines. This is a form of algorithmic bias inherited in the past and applied in AI-systems. Platforms that are used by LLMs, such as Airbnb or Google Travel, have a critical responsibility to manage AI outputs carefully to avoid reinforcing biases or misinformation. Other mitigation bias measures include ensuring transparency, so users know when recommendations are generated by AI, human reviews or official data.

Beyond practical recommendations, this work contributes to several research domain. In AI fairness, it demonstrates that models can introduce biases even when human data is unbiased, challenging the assumption that AI consistent reflects society. In urban areas, it highlights the limitations of text-based models in reasoning about spatial and empirical realities. Finally, in sociotechnical systems, it shows how technical choices, such as training data and prompt design, can shape social outcomes, influencing travel behaviors and neighborhood reputations.

5.5 Limitations and Directions

In this investigation we had some limitations on the implementation, due to the huge amount of data and by training LLM models. This study is limited tested on GPT-5 and Claude Sonnet 4, so future work should include more models for accurate and good representative outcomes (e.g, Gemini, Llama, etc). There were only two AI-models, due to the difficulty of extracting many accurate prompts and multiple neighbourhoods. Moreover, both Airbnb reviews and crime data cover limited time periods, so longer studies could show changes over time. Another limitation is in LLM outputs that may also inherit global cultural biases that don't match human realities, and small changes in prompts can change results. Future research could try fine-tuning models on local, bias-checked data to improve fairness and alignment. Additionally, several limitations may occur from prompt design, influencing results. Our three-question templates may not capture the full diversity of natural queries. For example, asking 'Is [neighbourhood] safe at night?' or 'What is [neighbourhood] like for families?' could produce different answers.

For future studies will be important to address these limitations and expand the investigation findings. Further studies could explore whether fine-tuning LLMs on balanced, bias-audited datasets reduces gender disparities, or whether retrieval-augmented generation (RAG) systems that incorporate real-time crime data improve alignment with reality. User studies could examine how travelers respond to AI versus human safety suggestions, including whether they notice and discount overly cautious outputs. Cross-cultural replication would test if LLMs exhibit similar biases in non-Western cities or whether training data imbalances favor certain geographies. Finally, explainability research could investigate whether LLMs can provide interpretable justifications for their safety assessments, allowing users to verify outputs against ground-truth data.

5.6 Ethical Considerations

Using LLMs into evaluating travel suggestion might raise safety ethical concerns. LLM suggestions can amplify gender stereotypes, mislabel neighborhoods as unsafe and harm communities through existing stereotypes. Moreover, developers and platforms must ensure transparency and check AI outputs against real data and statistics. LLMs accuracy check could include bias detection tools, combining AI with human inputs or user data, and clear ethical rules for platforms where safety framing shapes decisions. If biased AI suggestions affect user decisions, it raises ethical concerns about responsibility for harm, since it may be unclear whether accountability lies with the model developers, the platforms, or the users themselves.

AI-generated travel suggestion raises significant ethical challenges due to the unspecified re-

sponsibility. Developers, such as OpenAI and Anthropic, create general-purpose models without domain-specific safety guarantees. At the same time, users may not recognize AI-generated content or understand its limitations. When biased suggestions, avoid safe neighborhoods or recommend dangerous areas, assigning responsibility becomes difficult. This underscores the need for clear disclosure and well-defined liability frameworks.

To develop LLM-unbiased travel systems responsibly, several principles should be followed. Hybrid architectures can combine natural language generation with structured data sources, such as crime APIs and verified reviews, to improve accuracy. Continuous bias monitoring ensuring demographic parity across gender, race and other attributes. Human oversight is essential for high-stakes outputs, especially safety assessments, which should undergo review before publication. Even with these checked and bias measurements, AI responsibility is hard to be specified by someone, involving difficult trade-offs.

6 Conclusion

This dissertation examined how LLMs, specifically GPT-5 and Claude Sonnet 4, generate neighborhood safety suggestions and whether they amplify gender biases, comparing outputs with Airbnb reviews and London crime data. The study focused on sentiment, safety-related language, and fairness metrics to assess alignment with human perception and actual safety statistics. The findings reveal that LLMs diverge from human-authored reviews, often exaggerating caution. Both models showed artificial gender bias, rating female prompts as safer than male ones, while Airbnb reviews displayed insignificant differences. LLM outputs also weakly aligned with crime data, highlighting their limited accuracy in reflecting real-world safety. Fairness tests confirmed systematic gender bias, with Claude showing greater variability than GPT-5.

This research contributes empirically by systematically comparing AI terms with human and crime data, methodologically by introducing a framework combining sentiment analysis and fairness metrics, and theoretically by illustrating that LLMs can amplify new forms of bias in sensitive content. Practically, the results reveal that platforms should not exhibit LLM-generated safety suggestions without correction, policy evaluators must consider potential community stigmatization, and AI developers should apply fairness and contextual benchmarking into models. While the study is limited by its traveler-focused dataset, prompt design, and focus on only two LLMs, it points on future research directions including expanded geographical studies, more model testing, cross-cultural comparisons and hybrid human–AI mitigation strategies.

In conclusion, LLMs can generate safety attributes, but differ from human perceptions, misalign with crime data and introduce artificial gender biases. AI outputs are not neutral and require careful evaluation to ensure fairness, accuracy and ethical responsibility in socially sensitive applications.

References

[1] Ashmi Banerjee, Adithi Satish, and Wolfgang Wörndl. Enhancing tourism recommender systems for sustainable city trips using retrieval-augmented generation. *arXiv preprint*, 2024. URL https://doi.org/10.48550/arXiv.2409.18003. Revised apr 12, 2025; accepted at RecSoGood 2024 Workshop (co-located with RecSys 2024).

[2] T. Bas. Assessing gender bias in llms: Comparing llm outputs with human perceptions and official statistics. *arXiv preprint*, 2024. URL https://doi.org/10.48550/arXiv.2411.13738.

[3] E. Choi and E.-Á. Horvát. Airbnb's reputation system and gender differences among guests: Evidence from large-scale data analysis and a controlled experiment. https://par.nsf.gov/servlets/purl/10148340, 2019. National Science Foundation Public Access Repository.

[4] A. Decoupes, P. Martin, and J. Lemoine. Geographical distortions in global agricultural data: A critical review. *Agritrop*, 2023. URL https://agritrop.cirad.fr/611861/1/Geographical_distortions_Decoupes_et_al.pdf.

[5] X. Dong, Y. Wang, P. S. Yu, and J. Caverlee. Disclosure and mitigation of gender bias in llms. *arXiv preprint*, 2024. URL https://doi.org/10.48550/arXiv.2402.11190.

[6] S. Dudy, T. Tholeti, R. Ramachandranpillai, M. Ali, T. Jia-Jun Li, and R. Baeza-Yates. Unequal opportunities: Examining the bias in geographical recommendations by large language models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI'25)*, pages 1–18, Cagliari, Italy, March 2025. ACM. URL https://dl.acm.org/doi/pdf/10.1145/3708359.3712111.

[7] Inside Airbnb. Inside airbnb. https://insideairbnb.com/, 2025. Accessed September 15, 2025.

[8] H. Kotek, R. Dockum, and D. Q. Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference (CI '23)*, 2023. doi: 10.1145/3582269.3615599. URL https://doi.org/10.1145/3582269.3615599.

[9] Z. Liu. Cultural bias in large language models: A comprehensive analysis and mitigation strategies. *Journal of Transcultural Communication*, 2024. doi: 10.1515/jtc-2023-0019. URL https://doi.org/10.1515/jtc-2023-0019.

[10] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint*, 2023. URL https://doi.org/10.48550/arXiv.2310.06213.

[11] R. Manvi, S. Khanna, M. Burke, D. Lobell, and S. Ermon. Large language models are geographically biased. *arXiv preprint*, 2024. URL https://doi.org/10.48550/arXiv.2402.02680.

[12] Office for National Statistics. Estimates of the population for England and Wales, 2024. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/estimatesofthepopulationforenglandandwales. Accessed: October 1, 2025.

[13] Office for National Statistics. Uk census open geography portal dataset. https://geoportal.statistics.gov.uk/datasets/b58c65bdad994ed3a33741eea7bb09ab_0/about, 2025. Accessed September 15, 2025.

[14] Police UK. Open data — police uk. https://data.police.uk/data/, 2025. Accessed September 15, 2025.

[15] Yvette Reisinger and Felix Mavondo. The influence of gender on travel risk perceptions, safety, and travel intentions. *Journal of Travel and Tourism Marketing*, 20(1):97–105, 2006. doi: 10.1300/J073v20n01_09. URL https://doi.org/10.3727/108354210X12645141401269.

[16] V. Thakur. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint*, 2023. URL https://doi.org/10.48550/arXiv.2307.09162.

[17] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint*, 2023. URL https://doi.org/10.48550/arXiv.2310.09219.

[18] L. Wang, M. Song, R. Rezapour, B. C. Kwon, and J. Huh-Yoo. People's perceptions toward bias and related concepts in large language models: A systematic review. *arXiv preprint*, 2024. URL https://arxiv.org/pdf/2309.14504.