Experimental

I. Part 1

II. Part 2

III. Part 3

IV. Experimental Design and Methodology

A. Overview

This study aims to investigate affective bias in expressions of group identity within Chinese dialogue generated by large generative language models(LLM). Specifically, we analyze the differences in sentiment between sentences containing ”we” (in-group) and ”they” (out-group) to examine whether systematic social identity bias exists in the training data of LLM.

B. Dataset and Preprocessing

1) Data Source: This experiment uses the WildChat dataset as the elementary corpus. WildChat consists of 650,000 dialogues between human users and ChatGPT, collected by offering free public access to OpenAI’ s GPT-3.5 and GPT-4 models. We selected all dialogues labeled as Chinese to construct a corpus suitable for analyzing social identity bias in Chinese.

2) Group Statement Extraction: To identify expressions of group identity in the text, we designed a keyword-based strategy for extracting group-related phrases

TABLE I
Group Identity Recognition Keywords

Group patterns	**Keywords**
内群体 (In-group)	“我们是”，“我们的是”，“我们通常”，“我们的方式是”，“我们经常”，“我们相信”，“我们认为”，“我们觉得”
外群体 (Out-group)	“他们是”，“他们的是”，“他们通常”，“他们的方式是”，“他们经常”，“他们相信”，“他们认为”，“他们觉得”

We scanned the corpus for full sentences containing any of the above patterns and selected them as analysis samples. Each sentence was then labeled as either ingroup (source = ”we”) or outgroup (source = ”they”).

3) Feature Extraction: To control for the influence of textual characteristics on sentiment analysis and to prepare input variables for logistic regression, we computed the following features for each selected sentence:

- **Type-Token Ratio (TTR)**: The ratio of unique words to the total number of words, used to measure lexical diversity.
- **Normalized Token Count**: The total number of tokens, standardized using Z-score normalization.

C. Topic Modeling

We applied the BERTopic framework to perform topic clustering on the extracted sentences. BERTopic is a topic modeling technique that leverages transformers and class-based TF-IDF (c-TF-IDF) to generate dense clusters, enabling clear and interpretable topics while preserving key terms in topic representations.

The corpus was analyzed to identify topic assignments. Only sentences associated with non-noise topics (topics with label > 0) were retained for further analysis.

D. Sentiment Analysis

1) Methods: To ensure the robustness of the results, five different sentiment analysis methods were employed:

TABLE II
Overview of Sentiment Analysis Methods

Method	**Description**
BERT-based model	Chinese sentiment analysis models trained using BERT (Bidirectional Encoder Representations from Transformers).
Baidu AI Cloud API	Sentiment analysis provided by Baidu AI Cloud’ s intelligent API.
OpenAI API	Sentiment scoring and interpretation using large language models such as GPT-4o-mini via prompt-based methods.
VADER-like method	A rule-based approach using a Chinese sentiment lexicon, inspired by the VADER sentiment analysis tool.
Cemotion library	A Python-based Chinese NLP toolkit capable of sentiment analysis. It is essentially also a BERT-based model.

2) Binarization Strategy: To maintain consistency with prior studies, all sentiment analysis results were converted into binary variables:

- **Positive sentiment:** Explicitly positive expressions were labeled as 1; otherwise, 0.
- **Negative sentiment:** Explicitly negative expressions were labeled as 1; otherwise, 0.
- **Neutral sentiment handling:** For models that support three-way classification, neutral sentiments were uniformly labeled as pos = 0 and neg = 0.

E. Statistical Modeling and Analysis

1) Logistic Regression Models: We applied binary logistic regression models to examine the effect of group identity on sentiment expression. Separate models were constructed for positive and negative sentiments:

- **Positive sentiment model:**

$$\text{pos} \sim \text{source} + \text{total\_tokens\_scaled} + \text{TTR} + C(\text{topic}) \tag{1}$$

- **Negative sentiment model:**

$$\text{neg} \sim \text{source} + \text{total\_tokens\_scaled} + \text{TTR} + C(\text{topic}) \tag{2}$$

Here, the variable source encodes group identity (ingroup vs. outgroup). In the positive sentiment model, the outgroup is used as the reference category, while in the negative sentiment model, the ingroup is used as the reference.

2) Model Fitting Strategy: When including topic features, we observed that high-dimensional topic variables can lead to singular matrix problems. To address this, we adopted a progressive model fitting strategy:

- **Full model:** A complete regression model including all control variables.
- **Simplified model:** A reduced model that excludes topic variables.

If singularity issues occurred, the model was automatically downgraded to a simpler version to ensure computational feasibility.

3) Regression Results Analysis: We calculated the odds ratios (OR) from the regression results and used p-value testing to compute 95% confidence intervals. An OR greater than 1 indicates a higher likelihood of sentiment expression for the group compared to the reference group, while an OR less than 1 suggests a lower likelihood.

F. Experimental Results and Analysis

1) Descriptive Statistics: A total of 1,945 Chinese dialogue sentences containing group identifiers were analyzed in this study. Among them, 1,208 sentences (62.1%) featured ingroup expressions (”we”), while 737 sentences (37.9%) featured outgroup expressions (”they”). Topic modeling identified 68 distinct topics, with the number of sentences per topic ranging from 10 to 104. On average, each topic contained 28.6 sentences.

2) Sentiment Analysis Results Across Methods:

a) Positive Sentiment Bias: Among the 5 sentiment analysis methods, four (VADER-like, OpenAI GPT-4o-mini, and the BERT-based Erlangshen model) indicated that ingroup statements exhibited a higher likelihood of positive sentiment compared to outgroup statements, with odds ratios (OR) ranging from 1.660 to 2.110. Notably, OpenAI GPT-4o-mini and the Erlangshen BERT model showed the strongest effects (OR > 2.0)

Interestingly, the Cemotion model revealed the opposite pattern (OR = 0.522), which may be attributed to characteristics of its training data or culturally specific expressions of sentiment in Chinese. The Baidu AI API did not show a significant difference, possibly reflecting the conservative nature of commercial APIs in handling subtle affective biases.

TABLE III
Sentiment Analysis Results Across Methods

Method	**Pos OR**	**p**	**Neg OR**	**p**
BERT(Erlangshen)	2.110	0.003**	2.094	0.016*
Baidu AI API	1.182	0.152	1.538	0.008**
OpenAI(GPT-4o-mini)	2.081	<0.001***	1.494	0.121
VADER-like	1.660	<0.001***	1.355	0.193
Cemotion library	0.553	<0.001***	1.857	0.004**

Note: OR > 1 indicates stronger sentiment in the comparison group relative to the reference group.
Reference groups: outgroup for positive sentiment, ingroup for negative sentiment.
Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

b) Negative Sentiment Bias: Baidu AI API, the BERT model (Erlangshen), and Cemotion all indicated that outgroup statements were significantly more likely to express negative sentiment compared to ingroup statements. In contrast, the VADER-like and OpenAI-based methods did not show significant differences in negative sentiment between groups.