Media Studies; Vol.3, No.3; 2026
E-ISSN: 3006-368X P-ISSN: 3006-3671
DOI: https://doi.org/10.70267/ms.v3n3.126147
Published by: Zeus Press

From Free to Paid: Public Attitude Change and Mechanism Evidence around Doubao's Charging Announcement

Rui Zhang*

School of Network Communication, Zhejiang Yuexiu University, Shaoxing 312000, China

**Corresponding author: Rui Zhang.*

Abstract

As generative AI moves from free access to paid subscription, how charging shapes public attitude bears on platform commercialization. Taking Doubao's monthly subscription as the setting, this paper builds a dual-path framework of "product experience - price perception". Using the announcement date (May 4, 2026) as the breakpoint, we collect public Weibo discussions from 30 days before to 30 days after it (2-day periods; the 14 days on each side form the main window), measure attitude with aspect-based sentiment analysis, and test changes at the breakpoint by regression discontinuity in time. Negative evaluation is robust in direction for Payment Willingness ($\tau = +0.751$, $p < 0.001$) and Corporate Brand ($\tau = +0.647$, $p = 0.011$), but rests on only 27 and 7 pre-breakpoint texts, so the magnitudes are read as direction only. The negative share for Price & Charging shows no significant increase ($\tau = -0.161$, $p = 0.123$), so the price-fairness path is not supported. The difference between the two aspect jumps is -0.532 (clustered SE 0.437, $p = 0.245$), opposite to expectation and underpowered given the small pre-breakpoint sample; we report only that no larger negative change was found for the price path than for the feature path. The overall negative share rises by 0.317 ($p = 0.041$) but depends on a pre-breakpoint period with 3 relevant texts and falls to 0.043 ($p = 0.517$) without it, so it is not robustly supported. The conclusions are limited to a marked change in attitude, with no persistent shift and no strict causal claim.

Keywords

generative AI, freemium, paywall, price fairness, public attitude, aspect-based sentiment analysis, regression discontinuity in time

1. Introduction

Since GPT-1 adopted the Transformer architecture in 2018, the business model of generative artificial intelligence has been shifting from free and open access to paid subscription. OpenAI's ChatGPT spread rapidly after its launch at the end of 2022 and, while retaining a free version, introduced a paid subscription, forming a "free plus subscription" tiered service model [1,2]; empirical studies of freemium and subscription models show that the transition from free to paid has become a common path of commercialization for digital products [3,4,5]. Domestic generative AI products follow the same path, as domestic large language models such as ERNIE Bot successively launch value-added services [6]; according to the China Internet Network Information Center, the user base of generative artificial intelligence products in China has reached 249 million

© 2026 The Author(s). Authors retain copyright in their articles.

[7]. Doubao ranks first in monthly active users, reaching 340 million in the first quarter of 2026 [8], and its launch of a monthly subscription triggered extensive public discussion.

The transition from free to paid changes the relationship between users and platforms. At the free stage, users need not consider economic cost; at the paid stage, users move from “users” to “consumers”, and price is superimposed on product experience. User acceptance of information technology depends on perceived usefulness and perceived ease of use [9]; in the paid setting this problem is more complex, and user attitude may change systematically, which in turn affects the sustainability of a platform's business model. Examining the change in public attitude during the paid transition of generative AI and its formation mechanism carries theoretical value and practical meaning for optimizing platform pricing strategies and for the sound development of the industry.

This paper takes Doubao's launch of a monthly subscription as the research setting and answers two progressive research questions. First, after generative AI shifts from free to paid, does public attitude change? Second, if attitude changes, what is the underlying mechanism? The first question is answered through empirical testing, the second mainly through theoretical analysis, supported by auxiliary mechanism-level evidence.

To this end, this paper extends the Technology Acceptance Model (TAM) into a dual-path explanatory framework of “product experience - price perception”, and explains the formation mechanism of attitude change by combining concepts such as freemium, paywall, willingness to pay and price fairness; using public discussion texts on Weibo before and after the breakpoint as data, it applies aspect-based sentiment analysis (ABSA) to measure public attitude, takes the date on which the charging information was announced as the breakpoint, uses regression discontinuity in time (RDiT) to examine changes in public attitude before and after the charging event, and tests auxiliary mechanism-level evidence.

The theoretical significance lies in extending the Technology Acceptance Model from the general context of information technology adoption to the context of business model change in generative AI, providing an analytical framework for the attitudinal responses triggered by “a change in charging method during use”; the practical significance lies in providing empirical evidence for the pricing strategies and charging communication of generative AI platforms.

The remainder of the paper is organized as follows: Section 2 is the literature review, which reviews three bodies of literature, namely the Technology Acceptance Model, freemium and paywalls, and social media and Weibo; Section 3 proposes research hypotheses based on the theoretical framework; Section 4 introduces the data sources, variable measurement and empirical model specification; Section 5 reports the empirical results and hypothesis tests; Section 6 summarizes the research conclusions, discusses the theoretical and practical implications and sets out the limitations of the research.

2. Literature Review

2.1 Technology Acceptance Model

Fishbein and Ajzen proposed the Theory of Reasoned Action (TRA) in 1975, and Ajzen developed the Theory of Planned Behavior from it [10]; Davis proposed the Technology Acceptance Model (TAM) in 1989 on the basis of TRA, taking perceived ease of use and perceived usefulness as its core variables, which jointly shape users' attitude toward use [9]. External variables reach attitude toward use through perceived usefulness and perceived ease of use, and then behavioral intention and actual use. Generative artificial intelligence lowers the threshold of use through web-based and conversational interaction, so perceived ease of use is usually strong.

Later work produced TAM2 [11], UTAUT [12] and TAM3 [13], together with UTAUT2, which is oriented to consumption contexts and introduces the “price value” variable [14]. Price factors thus entered the theoretical framework of the technology acceptance model, giving a basis for explaining attitude change triggered by shifts in business models after a technology has been widely accepted.

In research on generative artificial intelligence, Strzelecki integrated the TAM and UTAUT2 frameworks to examine students' acceptance and use behavior with respect to ChatGPT [1]; Camilleri analyzed the

© 2026 The Author(s). Authors retain copyright in their articles.

influence of performance expectancy and other factors on intentions to use ChatGPT within an information technology acceptance framework [15]; G C et al. examined, on the basis of the UTAUT2 framework, the effects of social influence, learning value and habit on students' intention to use ChatGPT [2]. These studies consistently show that the framework applies across multiple scenarios and angles in generative AI research. TAM and its extended frameworks give this paper the theoretical tools for the product experience path; their explanatory weight, however, falls on the technological decision of "whether to adopt," and price factors are involved only indirectly through the "price value" variable of UTAUT2 [14]. That leaves theoretical space for this paper to extend the price perception path in the payment context.

2.2 Freemium, Paywall, Willingness to Pay and Price Fairness

Freemium and paywall are two typical models through which digital products move from free to paid. Freemium rests on free services and treats paid premium services as its value-added part; it relies on a large-scale user base, and the gap between the functions of the free version and those of the paid version is the lever of conversion [4]. In practice the free version accumulates users and the value-added part delivers monetization [5]; marketing research shows that free products significantly change users' evaluation and choice [3]. The paywall constrains content access through restrictive conditions and has expanded from the traditional newspaper industry to various digital platforms [16]; its introduction changes users' usage behavior and their patterns of word-of-mouth expression [17]. The transformation of generative AI from free to paid is precisely a process in which the two mechanisms act together.

Faced with this transition, users diverge in attitude, which can be explained through willingness to pay and price fairness. Willingness to pay is the maximum price a consumer is willing to pay for a particular item [18], and its formation depends on the combined effect of reference price, perceived value and individual needs [19]; free users use the product to a light extent and have relatively weak perceived value, whereas paid users are the opposite, and the numerous free versions of various AI products and their strong substitutability further weaken consumers' willingness to pay.

Price fairness refers to consumers' perceptual judgment of whether a price is fair. According to equity theory, individuals form fairness judgments by comparing their own input-output relationship and their reference relationship of transactions with others [20]; the principle of dual entitlement further shows that price increases based on rising costs are more readily accepted, whereas price increases attributable to profit motives are regarded as unfair [21]. In the Doubao charging event, if users interpret the charging as profit-seeking behavior by the firm, their acceptance is low; conversely, their acceptance is relatively higher. The above research provides analytical tools for this paper's explanation of the Doubao charging event: the establishment of the paywall constitutes the contextual trigger of attitude change, and price fairness and willingness to pay constitute the core mechanism variables of the price path.

2.3 Social Media and Weibo

The measurement of public attitude has traditionally relied on questionnaire surveys, but questionnaires have items preset by the researcher and are susceptible to social desirability bias and recall bias, and they are cross-sectional snapshots, which makes it difficult to capture the dynamic evolution of an event continuously; social media texts are published autonomously by users, their expression is more authentic and carries continuous time stamps, and they can be produced continuously as an event develops [22]. Although social media data have limitations such as self-selection bias, emotional noise and textual fragmentation, with the development of text mining techniques such as sentiment analysis [23], it has become methodologically feasible to measure public attitude continuously and reproducibly from large-scale social media texts.

This paper chooses the Weibo platform as its data source: at the platform level, Weibo attracts a large user base and is a major platform for discussion of public issues, and Doubao charging once appeared on the Weibo hot search and triggered extensive discussion, which provides sufficient data and clear boundaries for attitude measurement; at the temporal level, Weibo posts carry specific time stamps and constitute a continuous time series, which supports regression discontinuity in time (RDiT) in examining changes in relevant indicators before and after the event on the basis of the continuity of data on the two sides of the breakpoint [24,25]; at the textual level, Weibo comments have clear referents (such as price, features, and whether the product is easy to use), which makes them suitable for aspect-based sentiment analysis to extract evaluations directed at

© 2026 The Author(s). Authors retain copyright in their articles.

a specific aspect and their sentiment polarity [26,27], and existing research on Chinese Weibo sentiment analysis also shows its feasibility [28,29]. Weibo texts can be used to measure public attitude, and the aspect-based sentiment analysis and regression discontinuity in this paper thereby gain a data foundation.

3. Hypotheses

Based on the theoretical framework and the two research questions above, this paper proposes the following hypotheses (the research questions remain open; the hypotheses express only the theoretically expected direction).

At the situational level, Doubao launching a monthly fee is essentially the establishment of a paywall. Under the freemium model, a platform accumulates users with free services and monetizes through value-added services, so users form a stable expectation that basic functions are free; the paywall breaks this expectation and brings previously free resources into the paid scope, an external trigger of attitude change and the direct cause of public attitude reactions.

At the psychological level (the stages of reference perception and trade-off intention), users take the historical free price formed by long-term free use as the reference point and develop loss perception and psychological resistance toward charging [30,31]; they judge price fairness by comparing their own inputs and outputs [20], and perceived unfairness significantly intensifies negative attitudes [32]. Users weigh perceived gains and losses, forming or suppressing willingness to pay [19]; when perceived value does not support the price, willingness to pay declines and negative attitudes are reinforced. Price fairness and willingness to pay are the core variables of the mechanism analysis in this paper.

At the framework level, the classic TAM explains users' acceptance of information technology through perceived usefulness and perceived ease of use [9], but it hardly covers price in a paid context. This paper extends TAM to the paid context: it retains the product experience path, introduces freemium-paywall conversion as a situational variable, and introduces price fairness and willingness to pay as the core variables of the price path, building a dual-path explanatory framework of “product experience - price perception”. The hypotheses follow.

3.1 Hypotheses on Overall Attitude Change

H1: After the charging information is announced, the share of negative attitude in Weibo discussions rises significantly.

3.2 Mechanism Hypotheses

The following hypotheses examine the direction of change of the variables related to the price path, providing auxiliary mechanism-level evidence for explaining attitude change (H2-H4), and test whether the two evaluation paths of price and function diverge (H5); no individual-level mediation test is conducted.

H2 (price fairness): After the charging information is announced, the share of price unfairness discourse (“unreasonable”, “rip-off”, etc.) rises significantly, and the share of fairness discourse falls significantly [20,21,32].

H3 (willingness to pay): After the charging information is announced, the share of discourse refusing to pay and seeking alternatives rises significantly, and the share of willingness-to-pay discourse falls [18,19].

H4 (loss perception): After the charging information is announced, the share of reference-point loss discourse such as “originally free” and “suddenly charged” rises significantly [30,31].

H5 (dual-path divergence): After the charging information is announced, the increase in negative sentiment in the Price & Charging aspect is larger than that in the Feature Experience aspect, that is, the price path and the product experience path diverge at the breakpoint [14].

© 2026 The Author(s). Authors retain copyright in their articles.

4. Research Design

4.1 Data and Sample

4.1.1 Research Event and Breakpoint

The research event is the launch of a monthly subscription by Doubao. On May 4, 2026, Doubao updated its service statement and announced its charging plan on its Apple App Store page (CNY 68 per month for the standard version, among other tiers); that day is the time breakpoint c, and the running variable t is the number of days between a date and the breakpoint (t=0 on the breakpoint day). On the same day, the Weibo topic "Doubao payment" entered the trending topic list (Weibo hot search) for the first time, appearing 3 times in total, peaking at rank 1, with a peak heat score of 26 million and 21 cumulative hours on the list (as of 19:54 on May 5, 2026). The hot search formed after the charging information was released and is a phenomenon of opinion diffusion; the breakpoint is the publication of the charging information, not the hot search.

4.1.2 Data Collection

The platform is Weibo. Collection was self-conducted by day-by-day backtracking through Weibo advanced search: the search interface was accessed with a logged-in cookie and paged day by day on "date + keyword", and the backtracking limit of roughly 50 pages per keyword was mitigated by splitting collection across multiple keyword combinations. The retrieval rule required a text to contain "Doubao" together with any of "charging, subscription, membership, monthly fee, price increase, payment, free". The collection scope spans 30 days on either side of the breakpoint; the main analysis window is 14 days on either side, while the 30-day span serves trend observation and robustness checks. Collection ran from April 4 to June 3, 2026, completing 427 "date x keyword" tasks and yielding 19,188 raw records; for the 19 tasks with unusually high single-day volume, 6,033 further records were collected in 6-hour slices, providing a robustness sample for the truncation bias caused by the result cap. Because Weibo search pages fold long texts behind "expand full text" and carry emoji as images, directly collected texts lost both the second half of folded posts and the emoji; we traced the detail interface for the 21,693 Weibo IDs remaining after deduplication across the two collection routes, completing text and emoji for 21,592 records (99.5%). The other 101 posts had been deleted or were access-restricted, and the text of 3 long posts was unavailable (0.014%). Completion raised the median length of folded posts from 152 to 351 characters (a factor of 2.34) and restored image-borne emoji in 4,381 texts.

4.1.3 Sample Cleaning

Invalid texts were removed by the following rules [22]: (1) deduplication by Weibo ID; (2) removal of mixed-in topics unrelated to pricing (flight refunds, stock market conditions and the like, which also lack pricing words); (3) removal of marketing lotteries and advertising texts; (4) removal of texts shorter than 5 characters; (5) retention of discussions of "free features / the free portion". A text also had to contain both a brand word ("Doubao") and a pricing word (charging, payment, subscription, membership, monthly fee, price increase, pricing, price, free, etc.), so that the corpus stays focused on the charging issue. After cleaning, the main corpus holds 17,635 texts (1,553 removed from the 19,188 originally collected); the complete sample flow appears in Table 1 of Section 5.1.

4.2 Variables and Measurement

4.2.1 Outcome Variables (Attitude Indicators)

Public attitude here means the stance and evaluative discourse the public expresses when discussing the Doubao charging event, covering sentiment polarity (support, neutral, oppose) and judgments about the reasonableness of charging and the fairness of the price. It is measured from Weibo texts through aspect-based sentiment analysis (ABSA), yielding three indicators: (1) overall negative share, the number of texts with "oppose" polarity in a period divided by the number of relevant texts in that period; (2) mean sentiment, with support scored 1, neutral 0 and oppose -1, taken as a weighted mean over 2-day periods; (3) aspect-level negative share, the negative share of each of the four aspects, Price & Charging, Feature Experience, Corporate Brand and Payment Willingness, with the number of texts mentioning that aspect in the period as denominator [26,27]. Price & Charging and Feature Experience are the key comparison for testing H5, and the difference in their jump magnitudes at the breakpoint is tested directly in Section 5.4.

© 2026 The Author(s). Authors retain copyright in their articles.

4.2.2 Mechanism Variables (Discourse Indicators)

On top of aspect-based sentiment analysis, mechanism discourse is identified with a rule dictionary and shares are computed over 2-day periods: (1) price fairness, comprising fairness discourse (reasonable, worth it, good value) and unfairness discourse (ripping users off, greedy, unreasonable, disguised price increase); (2) willingness to pay, comprising willingness-to-pay discourse (willing to top up, worth subscribing to) and refusal-to-pay discourse (uninstall, stop using, substitute, freeloading); (3) loss perception, comprising reference-point discourse (previously free, suddenly, price increase), counted only when it co-occurs with reference-point words such as "previously/suddenly" so as to avoid overlap with price fairness discourse; (4) product experience, comprising discourse on whether the features work well or poorly.

Dictionary matching is checked by cross-agreement: the rank correlation between the daily dictionary series and the daily LLM aspect-level series is close to zero (ρ between -0.143 and 0.000), and directional agreement is 0.639-0.786. Dictionary hits are sparse (most shares fall between 1% and 8%), so mechanism testing takes the LLM aspect-level indicators as the main evidence, with dictionary results serving as a second, independent measurement that offers a directional reference; the cross-check table appears in the separately attached Appendix D.

4.2.3 Measurement Implementation and Validation

Texts were neither segmented nor stripped of stopwords, since the large language model handles raw sentences directly; emotion symbols carried by emoji images were restored to textual form at the collection stage (for example, "laughing cry") and kept in the text, so that the model can detect irony and emotional intensity [23,28,29].

Aspect identification and sentiment classification were performed by a large language model: deepseek-flash from DeepSeek (server-side version DeepSeek-V4.1-Flash, called on September 21, 2026), with thinking mode disabled, temperature set to 0, JSON structured output, and a maximum of 2,000 output tokens per call. A single instruction template directs the model to judge first whether the text concerns Doubao's own charging arrangements, then to identify the evaluation aspect and its polarity, answering in a fixed format; the template and the aspect definitions appear in the separately attached Appendix A and Appendix B. To reduce output randomness, every text was measured independently 3 times and decided by majority vote (majority for "whether relevant"; an aspect was retained only if it appeared in at least 2 measurements, and its polarity was decided by majority).

Measurement validation uses a manually annotated set: 385 texts drawn by stratified sampling (145 before the breakpoint, 240 after it) and annotated independently by 1 researcher following a unified coding protocol. With a single annotator, inter-annotator agreement (Cohen's κ) is unavailable; the limitation is reported as it stands, and reliability is examined through three substitutes: (1) stability of repeated measurement at temperature 0; (2) agreement between two models, deepseek-flash and deepseek-v4-pro; (3) cross-agreement with the dictionary method, an independent measurement. Agreement, accuracy and Macro-F1 between the model and the manual labels appear in Table 2 of Section 5.2, where Macro-F1 no lower than 0.8 marks measurement usability; where it falls short, the prompt and classification rules are revised and validation repeated (the prompt was iterated over two rounds in total). For the 44 samples in which the two models simultaneously contradicted the original manual judgment, the annotator reviewed them and wrote out the reasons, forming a gold standard; Macro-F1 is reported under both specifications.

4.3 Empirical Model Specification

4.3.1 Baseline Model (Regression Discontinuity in Time)

After each attitude indicator is aggregated into 2-day periods, the following regression discontinuity in time (RDiT) model is estimated:

$$Y_p = \alpha + \tau \cdot D_p + \beta 1 \cdot (t_p - c) + \beta 2 \cdot D_p \cdot (t_p - c) + \varepsilon_p$$

Here Y_p is the attitude indicator of period p, p is the period sequence, $D_p = 1\{t_p > c\}$ is the post-breakpoint dummy, t_p is the number of days from the period midpoint to the breakpoint day, c is the breakpoint day (the day the charging information was announced), and τ is the jump at the breakpoint, reflecting the size of the change in public attitude around the charging event [24,25]. Estimation uses local

© 2026 The Author(s). Authors retain copyright in their articles.

linear regression with triangular kernel weights, equal weighting across periods, and HC1 robust standard errors. The main window of ±14 days contains 15 two-day periods in total, but the triangular kernel weight of the outermost later period is 0, so 14 periods actually enter the regression (fewer for aspect-level indicators, because some periods mention no instance of that aspect). Two-day periods replace daily aggregation because discussion volume rose dozens of times on the event day (the first period after the breakpoint contains 2,296 relevant texts, against 3-33 texts in each pre-breakpoint period), so daily aggregation would cause severe zero inflation and variance heterogeneity, whereas 2-day periods stabilize each period's indicator while retaining time-series information. Equal weighting across periods gives a period with very few texts the same weight as an outbreak period: one pre-breakpoint period contains only 3 relevant texts, and its indicator is determined by those few texts and carries high leverage in the fit. Estimates weighted by the number of texts within a period are reported alongside, together with two targeted checks (see items (7) and (8) of the robustness checks in Section 4.3.4).

Testing H5 requires comparing the jump magnitudes of the two aspects, Price & Charging and Feature Experience, rather than examining whether each is significant on its own. The period series of the two aspects are stacked by aspect and a local linear model with an aspect interaction term is estimated:

$$Y = \alpha + \tau \cdot D + \beta 1 \cdot x + \beta 2 \cdot x \cdot D + \delta \cdot A + \lambda \cdot A \cdot D + \mu \cdot A \cdot x + \nu \cdot A \cdot x \cdot D + \varepsilon$$

Here $x = t - c$, A is the aspect dummy (1 for Feature Experience), and λ is the difference in jump magnitudes. Each aspect keeps its own intercept, slope and slope change, so the model is exactly equivalent in point estimates to estimating one local linear regression separately for each aspect (a numerical check confirmed digit-by-digit agreement). The two aspects are measured on the same texts in the same period and their errors are correlated, so the standard error of λ is clustered by 2-day period (14 effective periods, hence a cluster count of G=14); the unclustered HC1 standard error, the normal approximation and the wild cluster bootstrap are reported as well, to show that the conclusion does not hinge on a single specification.

4.3.2 Bandwidth Selection

A bandwidth sensitivity grid is the main approach: estimates are reported for the bandwidths ±8, ±10, ±12 and ±14 days (main window) and ±21, ±28 and ±30 days (extended window) [33]. Data-driven optimal bandwidth selectors such as CCT are not adopted, because with 2-day periods the main window holds only 15 periods and the extended window 31, and the selector's own sampling variation exceeds the efficiency gain it brings. The estimator contains 4 parameters to be estimated (intercept, breakpoint indicator, common slope on both sides, post-breakpoint slope increment), and with fewer than 8 effective periods the residual degrees of freedom are insufficient, leaving both the estimate and the robust standard error unreliable, so results at ±7 days and at narrower bandwidths are marked as insufficient sample and are not reported.

4.3.3 Mechanism Testing Strategy

The same regression discontinuity in time is applied separately to the four mechanism discourse indicators, to examine whether the direction of change of each mechanism indicator at the breakpoint matches the expectations of H2-H5. The data are aggregated over 2-day periods and no individual-level mediation analysis is performed, so these results serve only as an auxiliary mechanism-level test (mechanism-consistency evidence) describing the direction of change of the price-path variables, not as proof of a complete mechanism.

4.3.4 Robustness Checks

(1) Bandwidth sensitivity: repeated estimation across bandwidths; (2) “donut” exclusion: dropping the 1 period (2 days) immediately adjacent to the breakpoint, to remove the influence of the concentrated release of information on the breakpoint day; (3) placebo test: moving the breakpoint successively to the gaps between periods in the window and examining the rank and pseudo p-value of the true-breakpoint jump among the candidate positions; (4) contemporaneous shock control: re-estimation after excluding May 11-14 (the secondary discussion peak triggered by the People's Daily article); (5) specification sensitivity: comparing the two treatments of pure reposts (counted as neutral / excluded); (6) sampling sensitivity: re-estimating the attitude indicators on the hourly-slice deep-collected sample of the 19 high-volume tasks, to test the truncation bias caused by the result cap (the deep-collected sample is a single measurement while the main sample is the majority vote of three repeated measurements, an asymmetry stated in the report); (7) small-sample period exclusion: re-estimation after dropping periods in which the number of relevant texts falls below 10, to gauge

© 2026 The Author(s). Authors retain copyright in their articles.

the influence of extremely small periods on the breakpoint estimate under equal weighting across periods; (8) leave-one-period-out: re-estimating the main specification τ after dropping each period in turn and reporting the range of variation of τ and the most sensitive period, as the main basis for the robustness of the core conclusion. A 2nd-order local polynomial (with only 15 periods in the main window, the local quadratic term is unstable) and a 3-day aggregation specification (which would reduce the number of periods) are not used.

4.3.5 Methodological Limitations

Discussion volume surges around the charging event and posting density jumps markedly at the breakpoint, so the McCrary density test [34] does not carry its conventional manipulation-test meaning in this setting and is not used as a main test. Weibo users are self-selected in composition, and the public attitude discussed here is limited to the attitude reflected in Weibo discussion, so extrapolating the conclusions calls for caution. The discontinuity design characterizes the change in public attitude around the charging event rather than offering strict causal identification, and the conclusions should be read accordingly.

4.4 Research ETHICS

This study uses only public Weibo texts and collects no personal user information; usernames, @ mentions and other identifiable information are removed during analysis; only aggregate statistics are released, and raw texts are not. Aspect-based sentiment analysis requires transmitting texts to a third-party large language model interface; consent for this processing was obtained from the research team in advance, and the transmitted content is limited to public texts and the fields needed for the judgment; interface keys and collection credentials are kept in separate files and stay out of logs, reports and the version repository. The judgment examples in the appendix are all rewritten synthetic examples and contain no real user text. The third-party model service is provided by DeepSeek, and the study did not use the model of the object of study itself (Doubao under ByteDance) for measurement, avoiding the bias that arises when the measurement instrument and the measured object are the same.

5. Empirical Results

5.1 Sample and Descriptive Statistics

The final corpus contains 17,635 Weibo texts (Table 1).

Table 1: Sample flow

Stage	Number of records	Note
Raw collection (search layer)	19,188	7 keywords × 61 days, 427 tasks in total; 50-page ceiling per keyword per day
Deep sampling	6,033	19 high-volume tasks collected additionally in 6-hour slices, used only as a robustness control for truncation bias
Weibo IDs after deduplicating the two	21,693	3,528 records in the intersection; 21,693 = 19,188 + 6,033 - 3,528
Successful text backfilling	21,592 (99.5%)	For the 21,693 deduplicated records, the detail interface was traced by Weibo ID; 101 failures and 3 long Weibo texts unavailable
Corpus after cleaning	17,635	From the 19,188 records of raw collection, 1,553 were removed (deduplication, removal of irrelevant topics and advertisements, removal of fewer than 5 characters, retention of brand-word and pricing-word constraints)
Judged relevant	9,646 (54.7%)	Judged by the large language model, with three repeated measurements taking the majority vote
Manual validation set	385	Stratified sampling of 145 records before the breakpoint and 240 after it; of these, 384 are comparable and had been run by the model

Source: Compiled by the authors. The corpus version is v3 (the text-backfilled version); v2 is the non-backfilled version, used only as a sensitivity control. The deep sample covers only 19 discrete tasks, does not constitute a continuous time series, and is not used for breakpoint estimation.

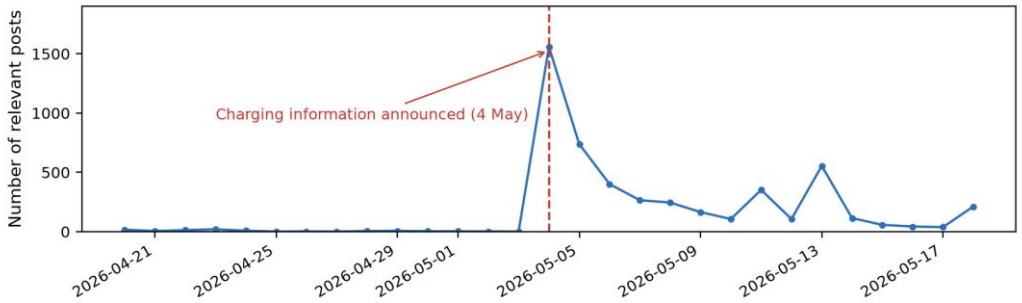
Collection executed 427 “date × keyword” tasks over 7 keywords and 61 calendar days; the 19 tasks with exceptionally high single-day volume were collected again in 6-hour slices to test the truncation bias caused

© 2026 The Author(s). Authors retain copyright in their articles.

by the result ceiling. The search page folds long texts and carries emoji as images, so the collected text had two types of gaps; deduplicated records were then traced to the detail interface by Weibo ID for backfilling, with a success rate of 99.5%. The 101 records not backfilled were deleted original posts or permission-restricted ones, and 3 further long Weibo texts were unavailable (0.014%). Backfilling raised the median length of folded posts from 152 to 351 characters (2.34 times) and recovered image-carried emoji in 4,381 texts.

The first period after the breakpoint (May 4-5, 2026) holds 2,296 texts judged relevant, while the periods before the breakpoint hold only 3-33 texts each (Figure 1); the period of May 2-3, 2026 holds 3. Under the main specification of period-equal weighting, a period with 3 relevant texts carries the same weight as the explosive period with 2,296.

Figure 1: Daily volume of relevant posts in the breakpoint window



Source: Compiled by the authors. The main analysis window covers 14 days on each side of the breakpoint, aggregated into 2-day periods.

5.2 Measurement Validity and Reliability

This paper drew 385 texts from the corpus by stratified random sampling, annotated by 1 researcher under a unified coding protocol. Annotation was completed by a single annotator, so inter-annotator agreement (Cohen's κ) is unavailable; this limitation is reported as it stands, and reliability is tested with three pieces of substitute evidence (Table 2).

Among the 384 comparable records, the judgments fall into three categories: 290 records of complete agreement, 44 records in which both models are opposite to the original manual judgment, and 50 records in which only one model disagrees with the manual annotation. The 44 divergent records were reviewed one by one: 24 were changed to “relevant”, 14 to “not relevant” and 6 kept the original judgment, raising the Macro-F1 for relevance from 0.791 to 0.903; the review referred to the model judgments, so 0.903 carries some optimism and is reported beside 0.791. The original Macro-F1 below 0.8 stems from context-dependent samples: some Weibo texts can be judged only with information from the accompanying image and cannot be distinguished from the text alone, and others carry hashtags inconsistent with their subject; these constitute the main source of measurement error.

Three pieces of substitute evidence show that repeated measurement by the same model under the same parameters agrees at 0.990 ($\kappa = 0.980$), so the judgment is not sampling noise, and that flash and v4-pro agree at 0.858 ($\kappa = 0.719$). Aspect-set stability is clearly lower: the Jaccard is only 0.491 on retest and 0.315 across models, while polarity agreement reaches 0.979; the manual comparison follows the same pattern, with an exact aspect-set agreement rate of 0.343 (Jaccard 0.626) and an overall-stance Macro-F1 of 0.731. “Which aspects a text involves” itself carries measurement noise, so aspect-level indicators should be read as “the negative share among texts that mention that aspect”, not as a decomposition of overall attitude. This limitation is discussed in Section 6.

© 2026 The Author(s). Authors retain copyright in their articles.

Table 2: Measurement validity and reliability tests

Test item	Measure	Result
Agreement rate with manual annotation (before adjudication)	384 records	0.812
Agreement rate with manual annotation (after adjudication)	384 records	0.911
Macro-F1: relevant or not (before adjudication)	384 records	0.791
Macro-F1: relevant or not (after adjudication)	384 records	0.903
Macro-F1: overall stance	95 records (both sides judged relevant and the stance was complete)	0.731
Exact agreement rate of the aspect set / Jaccard	70 records (both sides gave aspects)	0.343 / 0.626
Repeated-measurement stability: relevant or not	300 records retested, same model and same prompt, temperature 0	Agreement rate 0.990, $\kappa = 0.980$
Repeated-measurement stability: overall stance	As above (156 records comparable)	Agreement rate 0.994, $\kappa = 0.989$
Inter-model agreement: relevant or not	1,884 records, flash and v4-pro	Agreement rate 0.858, $\kappa = 0.719$
Aspect polarity consistency	Samples in which both sides annotated that aspect	Agreement rate 0.979
Aspect set consistency	Retest / cross-model	Jaccard 0.491 / 0.315

Source: Compiled by the authors. The criterion is a Macro-F1 no lower than 0.8. The before-adjudication Macro-F1 for relevance (0.791) is below this criterion; the after-adjudication value of 0.903 referred to the model judgments during the review process and therefore carries some optimism, and the two are reported side by side.

5.3 Changes in Public Attitude at the Breakpoint

Table 3 reports the jump estimates at the breakpoint (local linear method in Section 4): 2 days per period, period-equal weighting, HC1 robust standard errors. Four measurement points apply. The ±14-day window contains 15 2-day periods, but the outermost later period (May 18, 2026) has weight 0, so the number of periods entering the regression is 14; aspect-level indicators have fewer if a period does not mention that aspect (13 periods for Feature Experience, 12 for Corporate Brand). The mean columns give "period-equal" means on the same measure as the regression, and the text-weighted measure is given in the table note. τ is a local linear estimate and is not equal to "mean after the breakpoint minus mean before the breakpoint". The pre-breakpoint denominators differ enormously, the smallest period holding only 3 texts at a negative share of 0.000, which gives it high leverage on the fit (Section 5.5). The last row is the direct test of H5.

The last row of Table 3 is the direct test of H5: Price & Charging and Feature Experience are stacked by period and a local linear model with an aspect interaction term is estimated, the interaction coefficient being the difference between the two jumps, so that row's two mean columns and τ all equal the difference between the two rows above it (-0.126 = 0.458 - 0.583, -0.411 = 0.379 - 0.791, -0.532 = -0.161 - 0.372). The standard error of that row is clustered by 2-day period (G = 14); the p-value and confidence interval in the table follow the normal approximation, under t(G-1) (df = 13) the 95% confidence interval is [-1.476, 0.412] with p = 0.245, without clustering the standard error is 0.593 with p = 0.369, the wild cluster bootstrap gives p = 0.622, and under text weighting the difference is -0.629 (clustered standard error 0.380, p = 0.098). Under all four measures and both weighting schemes p exceeds 0.05, and a one-sided test in the direction of H5 gives p = 0.878.

The overall negative share rises by 0.317 at the breakpoint (p = 0.041) and mean sentiment falls by 0.734 (p = 0.041), consistent in direction with H1, but the estimate is highly sensitive to the extremely small periods (Section 5.5). Figure 2 shows the regression discontinuity for the overall negative share.

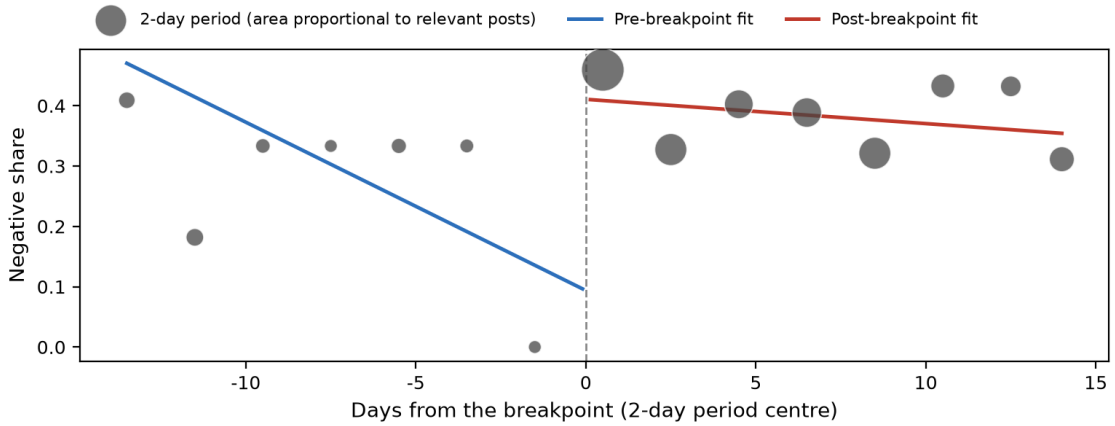
© 2026 The Author(s). Authors retain copyright in their articles.

Table 3: Breakpoint jump estimates for the main attitude indicators

Indicator	Before breakpoint (period-equal)	After breakpoint (period-equal)	Jump τ	Standard error	95% confidence interval	p-value	Effective number of periods
Overall negative share	0.275	0.395	+0.317	0.155	[0.014, 0.620]	0.041	14
Mean sentiment	0.073	-0.282	-0.734	0.359	[-1.437, -0.031]	0.041	14
Price & Charging negative	0.458	0.379	-0.161	0.104	[-0.365, 0.043]	0.123	14
Feature Experience negative	0.583	0.791	+0.372	0.505	[-0.618, 1.361]	0.462	13
Corporate Brand negative	0.200	0.380	+0.647	0.254	[0.150, 1.144]	0.011	12
Payment Willingness negative	0.271	0.605	+0.751	0.074	[0.607, 0.896]	<0.001	14
Loss reference share	0.089	0.090	+0.107	0.054	[0.000, 0.213]	0.050	14
Price & Charging negative - Feature Experience negative	-0.126	-0.411	-0.532	0.437	[-1.389, 0.324]	0.223	14 (Price 14 / Feature 13)

Source: Compiled by the authors. Mean sentiment is the mean with support scored 1, neutral scored 0 and opposition scored -1; the denominator of an aspect-level indicator is the number of texts in the current period that mention that aspect. The two mean columns are the "period-equal" means that use the same measure as the regression; if they were combined by weighting by the number of texts within a period, the overall negative shares before and after the breakpoint would be 0.284 and 0.404 respectively, but that measure is dominated by high-volume periods and cannot be directly subtracted from τ, so it is not listed separately in the table. The numbers of relevant texts in the periods before the breakpoint are 22/33/9/3/12/6/3, among which the minimum within-period n is 3. The total pre-breakpoint denominators of the aspect-level indicators (counting only the periods entering the regression) are: 42 texts for Price & Charging, 18 for Feature Experience, 7 for Corporate Brand and 27 for Payment Willingness, and after the breakpoint they are 3,562, 1,168, 1,190 and 1,395 texts respectively; the pre-breakpoint level of Corporate Brand is in fact determined by only 7 texts (5 periods), and its minimum within-period n is 1. The τ values are local linear (triangular kernel) estimates on both sides of the breakpoint and are not equal to the difference between the two preceding mean columns; because the pre-breakpoint denominators of the aspect-level τ values are extremely small, their absolute magnitudes contain an extrapolation component and should be interpreted only in terms of direction.

Figure 2: Regression discontinuity plot for the overall negative share



Source: Compiled by the authors. Local linear fits on both sides of the breakpoint.
tau = +0.317 (SE 0.155, p = 0.041); effective periods 14; minimum within-period n = 3 (2-3 May 2026: 3 relevant posts).

© 2026 The Author(s). Authors retain copyright in their articles.

5.4 Mechanism-Level Auxiliary Evidence

Two independent sets of measurement supply auxiliary mechanism-level evidence: mechanism discourse identified by a rule-based dictionary (price fairness, willingness to pay, loss perception and product experience), and the large-model aspect-level sentiment indicators. Table 4 reports the jump estimates under the dictionary measure.

Table 4: Breakpoint jump estimates for mechanism discourse indicators (dictionary-based)

Mechanism discourse indicator	Share before breakpoint	Share after breakpoint	Jump τ	Standard error	95% confidence interval	p-value	Effective number of periods
Fairness discourse (positive)	0.041	0.049	-0.007	0.036	[-0.077, 0.064]	0.855	14
Unfairness discourse (negative)	0.024	0.033	+0.020	0.016	[-0.012, 0.052]	0.216	14
Willingness-to-pay discourse	0.019	0.023	+0.007	0.009	[-0.010, 0.024]	0.448	14
Refusal-to-pay discourse	0.058	0.074	+0.037	0.021	[-0.005, 0.078]	0.084	14
Reference-point loss discourse	0.075	0.076	+0.002	0.020	[-0.037, 0.041]	0.913	14
Technical evaluation (positive)	0.084	0.045	-0.087	0.025	[-0.135, -0.038]	<0.001	14
Technical evaluation (negative)	0.011	0.031	+0.005	0.012	[-0.019, 0.029]	0.677	14

Source: Compiled by the authors. Dictionary version v1-20260919, and the item list is given in the separately attached Appendix D; the denominator of each indicator is the number of valid texts in the current period, whose minimum across periods is 83 texts, so the problem of denominators of only single digits in the overall indicators does not exist here.

Under the dictionary measure, the unfairness discourse share rises by 0.020 but is not significant ($p = 0.216$) and fairness discourse shows no significant change ($p = 0.855$), so H2 (a decline in price fairness) is not supported; refusal-to-pay discourse rises by 0.037 ($p = 0.084$), consistent in direction with H3 but only marginally significant; reference-point loss discourse barely changes ($p = 0.913$), so H4 is not supported. Positive technical evaluation falls significantly by 0.087 ($p < 0.001$, 95% confidence interval [-0.135, -0.038]), meaning that positive evaluation of product features declined after the event; this indicator depends on a single period, with a leave-one-period-out minimum of -0.058 ($p = 0.122$), so its direction is negative but its significance unstable and it serves only as directional evidence (Figure 4).

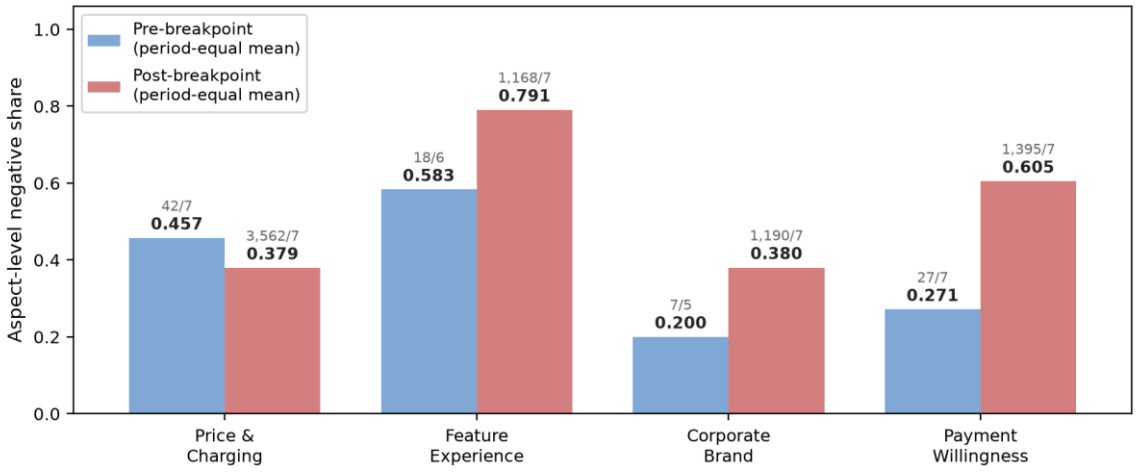
The model indicators partially cross-corroborate the dictionary measure. On the price path, model-measured "Price & Charging negative" showed no statistically significant increase ($\tau = -0.161$, $p = 0.123$) and dictionary-measured fairness discourse also did not change, so no statistically significant increase was observed in negative evaluation at the price level. On the feature path, model-measured "Feature Experience negative" rose by 0.372 (not significant) and dictionary-measured "technical evaluation (positive)" fell significantly, both consistent in direction and both pointing to deteriorating feature evaluation, though neither estimate is robust. On payment willingness, model-measured "Payment Willingness negative" rose significantly by 0.751 ($p < 0.001$) and dictionary-measured "refusal-to-pay discourse" rose in the same direction. The only inconsistency is the loss reference path: the model rose by 0.107 ($p = 0.050$) while dictionary reference-point loss discourse did not change, so the evidence for H4 is not robust.

The rank correlation between the two sets of measurement at the daily level is close to zero (Spearman ρ between -0.143 and 0.000), which does not contradict the directional consistency: dictionary hits are sparse (most shares between 1% and 8%) and their day-to-day swings are mostly the chance of keyword matching. Mechanism conclusions rest mainly on the model's aspect-level indicators, with the dictionary results as a directional reference; the cross-check table is given in the separately attached Appendix D.

© 2026 The Author(s). Authors retain copyright in their articles.

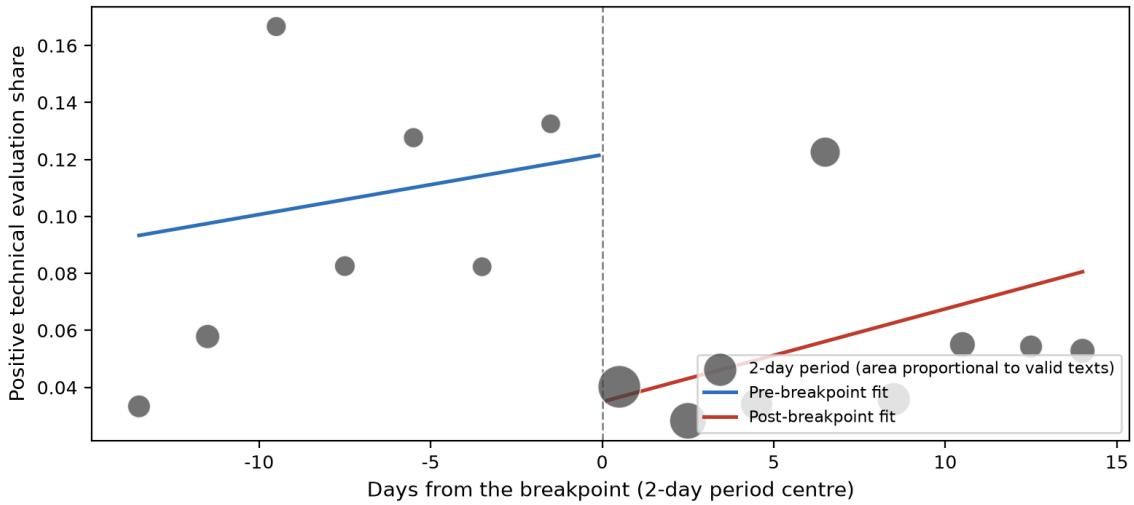
Tables 3 and 4 together show that the negative reaction concentrates in **Corporate Brand** ($\tau = +0.647$, $p = 0.011$) and **Payment Willingness** ($\tau = +0.751$, $p < 0.001$), while no statistically significant increase was observed in **Price & Charging**; these two are also always of the same sign and significant under leave-one-period-out. Their pre-breakpoint sample bases differ sharply: Payment Willingness rests on 27 texts (7 periods) with a raw before-after difference of 0.271 → 0.605, of the same order as $\tau = +0.751$; Corporate Brand rests on only 7 texts (5 periods) with a raw difference of 0.200 → 0.380, while $\tau = +0.647$ (text-weighted +0.879) is clearly larger than that descriptive difference, an extrapolation from an extremely small pre-side sample. Aspect-level τ is read only for direction, not as the magnitude of the rise. Figure 3 compares the four aspects before and after the breakpoint, with the period-equal means as bars and each side's total denominator and period count labeled on them.

Figure 3: Before-after comparison across the four aspects



Source: Compiled by the authors. Bars are period-equal means of the periods entering the regression; the value is shown above each bar, and the denominator (posts/periods) above that. Pre-breakpoint denominators are very small (Corporate Brand: only 7 texts over 5 periods), so aspect-level tau is read as direction only.

Figure 4: Regression discontinuity plot for positive technical evaluation (dictionary-based)



Source: Compiled by the authors. Dictionary-based measure; the denominator is the daily number of valid texts. tau = -0.087 (SE 0.025, p = 0.000); weakest leave-one-period-out estimate -0.058 (p = 0.122).

© 2026 The Author(s). Authors retain copyright in their articles.

Public reaction points to the commercialization motive ("why charge") and to behavioral intention ("should I pay") rather than to the price level ("is it expensive"). Negative evaluation in the Price & Charging aspect was already comparatively high before the event (0.405 before the breakpoint) and did not rise further after it; in the period-by-period removal tests the estimate for this indicator is always negative ([-0.245, -0.071]), so the conclusion that no statistically significant increase was observed is robust.

Path divergence cannot be inferred from whether each aspect is significant on its own; the magnitudes of the two jumps must be compared directly. The interaction estimate is -0.532 (Price & Charging minus Feature Experience), meaning that the rise in the Feature Experience aspect is instead larger, opposite in direction to H5; the difference is not significant under the four standard error measures and both weighting schemes (normal $p = 0.223$, $t(G-1)$ $p = 0.245$, HC1 $p = 0.369$, wild bootstrap $p = 0.622$; measures in the note to Table 3), and a one-sided test in the direction of H5 gives $p = 0.878$. After removing each period in turn the difference estimate moves between [-0.780, +0.243], reaching 5% after one removal (a period in which Feature Experience holds only 1 text), but no removal makes the difference positive and significant.

The comparison test is underpowered: $3.0305 \times 0.437 = 1.324$, so the minimum detectable effect 1.324 exceeds the theoretical upper bound of the share difference of 1 (1.152 under text weighting), given $t(0.975,13) + t(0.80,13) = 3.0305$. The available sample (the pre-breakpoint level of Feature Experience rests on only 18 texts and 6 periods, with a minimum within-period n of 1) cannot resolve a divergence of any magnitude. This paper reports only that "no negative change in Price & Charging larger than in Feature Experience was found", and does not report that "the two paths are the same" or that no divergence exists. The "not supported" verdict for H5 rests on the direction of this direct comparison, not on combining the separate significance tests of the two aspects.

The dictionary measure gives parallel directional corroboration: measuring the price path by unfairness discourse and the feature path by the reverse of "technical evaluation (positive)", the difference between the two jumps is -0.067 (clustered standard error 0.040, $p = 0.117$; -0.067 under text weighting, $p = 0.071$), again with greater deterioration in the feature aspect. This depends on "technical evaluation (positive)": using "technical evaluation (negative)" instead (shares of only 1%-3%) gives +0.015 ($p = 0.253$) and no divergence, and that indicator is itself sensitive to a single period (leave-one-period-out minimum -0.058, $p = 0.122$). The dictionary measure serves only as directional corroboration, not as independent evidence of divergence.

5.5 Robustness Checks

Robustness is checked across nine aspects: bandwidth, weighting scheme, small-sample periods, leave-one-period-out, the placebo test, donut exclusion, contemporaneous shocks, the repost measure and the sampling scheme (Table 5).

Table 5: Summary of robustness checks

Check	Specification	Overall negative jump τ	p-value	Conclusion
Main specification	Period-equal weighting, ±14 days	+0.317	0.041	See below
Leave-one-period-out	Re-estimated after removing each period in turn	Interval [+0.043, +0.399]	-	After removing the period of May 2-3, 2026 it falls to +0.043 ($p = 0.517$)
Small-sample periods	Removing periods with fewer than 10 relevant texts	-0.004	0.964	Falls to zero
Bandwidth sensitivity	±12 days	+0.373	0.012	Direction consistent
Bandwidth sensitivity	±10 days	+0.439	0.001	Direction consistent
Bandwidth sensitivity	±7 days and narrower	Insufficient sample	-	Fewer than 8 effective periods with 4 parameters to be estimated, not reported

© 2026 The Author(s). Authors retain copyright in their articles.

Check	Specification	Overall negative jump τ	p-value	Conclusion
Weighting scheme	Weighting by the number of texts within a period	+0.210	0.151	Direction consistent, significance lower
Placebo test	Moving the breakpoint to each position in the window	-	0.077 (0/12 more extreme)	The true breakpoint is the most extreme; the pseudo p-value has reached the lower bound of this design
Donut exclusion	Removing the 1 period immediately adjacent to the breakpoint	-0.037	0.492	The effect depends on one small-sample period on each side of the breakpoint
Contemporaneous shocks	Removing May 11-14	+0.320	0.052	Direction consistent
Repost measure	Removing pure reposts (robust measure)	+0.381	0.005	Direction consistent, effect stronger
Sampling scheme	Deep-sampled hourly-slice sample	Rank correlation 0.900	-	The two sampling orders are highly consistent

Source: Compiled by the authors. Comparing the deep sample with the main sample on the 19 shared tasks, the rank correlations of the negative share and mean sentiment are 0.900 and 0.872 respectively, and the paired level difference is about 6 percentage points; the deep sample is a single run while the main sample is the majority vote of three repeated measurements, and this asymmetry of measures will produce a level difference, so the difference cannot all be attributed to the sampling ceiling. The leave-one-period-out procedure re-estimates the main-specification τ after removing each period in turn, and the number of texts in each period and the period-by-period estimation results are explained in Section 5.3.

The leave-one-period-out and small-sample-period checks are key. Before the breakpoint a period holds only 3 relevant texts (May 2-3, 2026, negative share 0.000) and has high leverage under equal weighting: removing each period in turn moves the overall negative τ between +0.043 and +0.399, removing that period drops it to +0.043 ($p = 0.517$), and removing periods with fewer than 10 relevant texts gives $\tau = -0.004$ ($p = 0.964$). Mean sentiment and loss reference follow the same pattern (leave-one-period-out minima -0.102, $p = 0.454$ and +0.051, $p = 0.717$). Payment Willingness negative (interval [+0.701, +0.799], $p < 0.001$) and Corporate Brand negative (interval [+0.412, +0.834], $p < 0.001$) always keep the same sign and remain significant. “Robust in direction” differs from “reliable in magnitude”: the directions of these two are cross-validated under both weighting schemes, but their pre-breakpoint denominators are minute (27 texts for Payment Willingness, only 7 for Corporate Brand) and τ contains an extrapolation component, so “robust” refers only to direction; the core conclusion is that negative expression in these two aspects rose after the event.

Two corrections apply to bandwidth sensitivity. The estimator contains 4 parameters, and below 8 effective periods the residual degrees of freedom are insufficient and both the estimate and the standard error are unreliable, so ±7 days and narrower bandwidths are not reported. The overall negative τ rises as the bandwidth narrows (±14 days +0.317, ±12 days +0.373, ±10 days +0.439), which is not evidence of a local jump feature: after removing the period of May 2-3, 2026, all bandwidth estimates fall between +0.043 and +0.109. Bandwidth sensitivity reflects the changing weight of that small-sample period, not the locality of the effect.

The donut test likewise needs reinterpretation. Removing the 1 period immediately adjacent to the breakpoint (May 2-3 and May 4-5, 2026) drops the overall negative τ to -0.037 ($p = 0.492$). Removing only the explosive period (May 4-5) while keeping the pre-breakpoint period gives $\tau = +0.237$ ($p = 0.119$); removing only the pre-breakpoint period while keeping the explosive period drops τ to +0.043 ($p = 0.517$). The effect disappears mainly because of the small pre-breakpoint period and cannot be read as “the effect is concentrated in the first period after the event in an outbreak-decay pattern”, so the temporal pattern is no longer used as a core argument.

© 2026 The Author(s). Authors retain copyright in their articles.

In the placebo test, the jump at the true breakpoint is the most extreme of the 12 placebo positions (0/12 more extreme, pseudo p = 0.077); the lower bound of the pseudo p-value is 1/(number of candidate positions + 1), so this can only be stated as "the jump at the true breakpoint is the most extreme in the window", not as the test having "passed". Removing May 11-14 (the secondary discussion peak following the People's Daily article) leaves the core results intact (+0.320, p = 0.052); the robust measure that drops pure reposts gives a jump of 0.381 (p = 0.005), so the way pure reposts are counted did not drive the result. Re-estimated on the hourly-slice sample of the 19 high-volume tasks, the ceiling of about 50 pages per keyword did not change the direction of the conclusions; the deep sample and the main sample are not aligned in measure, with a paired level difference of about 6 percentage points (the deep sample slightly more negative), which is recorded in the data description as truncation bias.

5.6 Summary of Hypothesis Tests

The test conclusions for each research hypothesis are given in Table 6.

Table 6: Hypothesis test results

Hypothesis	Summary of content	Key results	Conclusion
H1	After the announcement of the charging information, the share of negative attitude rises significantly	Overall negative +0.317 (p = 0.041); weakest in the leave-one-period-out procedure +0.043 (p = 0.517), and -0.004 (p = 0.964) after removing periods with fewer than 10 relevant texts	Not robustly supported
H2	The share of price unfairness discourse rises and fairness discourse falls	Dictionary: unfairness +0.020 (p = 0.216), fairness -0.007 (p = 0.855); model: Price & Charging negative -0.161 (p = 0.123)	Not supported
H3	Refusal-to-pay discourse rises and willingness-to-pay discourse falls	Model: Payment Willingness negative +0.751 (p < 0.001), leave-one-period-out interval [+0.701, +0.799] (27 texts before the breakpoint); dictionary: refusal to pay +0.037 (p = 0.084)	Supported (direction)
H4	The share of reference-point loss discourse rises	Model: loss reference +0.107 (p = 0.050), weakest in the leave-one-period-out procedure +0.051 (p = 0.717); dictionary: loss perception +0.002 (p = 0.913)	Inconsistent evidence, not supported
H5	The rise in negative sentiment in the Price & Charging aspect is larger than in the Feature Experience aspect	Difference (Price & Charging - Feature Experience) -0.532 (clustered standard error 0.437; normal 0.223 / t(13) 0.245 / HC1 0.369 / wild bootstrap 0.622, all not significant; [-0.780, +0.243] after removing each period in turn); Price & Charging -0.161 (p = 0.123), Feature Experience +0.372 (p = 0.462); dictionary same-direction comparison -0.067 (p = 0.117)	Not supported (direction opposite, difference not significant)

Source: Compiled by the authors. "Leave-one-period-out" is the interval of the main-specification τ re-estimated after removing each period in turn. The total pre-breakpoint denominators of the aspect-level indicators are 7-42 texts (7 texts for Corporate Brand, 18 for Feature Experience, 27 for Payment Willingness and 42 for Price & Charging); the directions of their τ values are cross-validated as consistent under the two weighting measures, but the absolute magnitudes contain an extrapolation component and are interpreted only in terms of direction. The key result of H5 is a direct test of the difference between the jumps of the two aspects, and its estimator, standard error measure and power limitation are given in the note to Table 3 and in Section 5.4; limited by the too small pre-breakpoint sample, this comparison test can only conclude that "no negative change in the price path larger than that in the feature path was found", and cannot conclude that "the two paths are the same".

H1 is not robustly supported: the point estimate rises but depends on the pre-breakpoint period holding only 3 relevant texts, and after removing that period or periods with fewer than 10 relevant texts it is not significant (+0.317, p = 0.041; -0.734, p = 0.041; +0.043, p = 0.517; -0.004, p = 0.964), so the overall attitude indicators serve only as descriptive identification, not as inference of a persistent shift or strict causality. H3 is supported (direction): negative expression of Payment Willingness rises significantly and stays significant under leave-one-period-out, with both measurement sets consistent in direction; it has only 27 texts before the breakpoint, so the magnitude of τ contains an extrapolation component. H2 is not supported: neither measure shows a significant rise in negative evaluation of Price & Charging, and the estimate is always negative under leave-one-period-out, so the conclusion that no statistically significant increase was observed is robust. H5 is not

© 2026 The Author(s). Authors retain copyright in their articles.

supported (direction opposite, difference not significant): the difference is -0.532 (clustered standard error 0.437), the rise in the Feature Experience aspect is instead larger, and the difference is not significant under the four standard error measures and both weighting schemes (one-sided test p = 0.878); the pre-breakpoint Feature Experience sample is extremely small (18 texts, 6 periods, minimum within-period n of 1) and the minimum detectable effect 1.324 of this comparison test exceeds the theoretical upper bound of a share difference, so power is insufficient to resolve a divergence of any magnitude and this paper judges only that "no negative change in the price path larger than in the feature path was found", without asserting that "the two paths are the same". This finding differs from the expectation of the existing price fairness literature (Section 6).

For H4 the verdict remains inconsistent evidence, not supported: the model-measured loss reference share rises by 0.107, exactly at the 5% critical level (p = 0.050), falls to +0.051 (p = 0.717) after removing the small pre-breakpoint period, and the dictionary-measured reference-point loss discourse barely changes. The discrepancy may lie in different identification scopes for "reference-point loss": the model identifies implicit comparative expressions (such as "free before, charged now"), whereas the dictionary method requires explicit words such as "originally free" or "suddenly charged". H4 is reported only as suggestive evidence.

6. Conclusion and Discussion

6.1 Main Conclusions

Taking Doubao's announcement of its charging plan as the setting, and drawing on publicly available discussion texts on Weibo (17,635 posts) together with aspect-based sentiment analysis, this paper uses regression discontinuity in time to examine changes in public attitude before and after the event. There are three conclusions.

First, after the charging information was announced, public negative attitude changed noticeably, but the evidence on the overall indicator is fragile. The overall negative share rises by 0.317 at the breakpoint (p=0.041), and the sentiment mean falls by 0.734 (p=0.041); before the breakpoint there is a period containing only 3 relevant texts (May 2-3, 2026), and under "equal weighting across periods" this period carries the same weight as the outbreak period containing 2296 texts, so its leverage is very high. After removing each period in turn, the jump estimate for the overall negative share varies between +0.043 and +0.399, falls to +0.043 (p=0.517) once that period is removed, and is -0.004 (p=0.964) once periods with fewer than 10 relevant texts are removed. This change depends on small-sample periods, so no persistent level shift or strict causal inference is made.

Second, the rise in negative response is concentrated in Corporate Brand and Payment Willingness, while no statistically significant increase was observed for Price & Charging. The negative share rises by 0.647 for Corporate Brand (p=0.011) and by 0.751 for Payment Willingness (p<0.001), and after removing each period in turn the two estimates stay within [+0.412, +0.834] and [+0.701, +0.799] and remain significant; the direction is cross-validated as consistent under the two weighting specifications. The pre-breakpoint samples for both items are very thin: the pre-breakpoint level of Payment Willingness is determined by 27 texts (7 periods), and its raw before-after difference (0.271 to 0.605) is of the same order of magnitude as τ; Corporate Brand is determined by 7 texts (5 periods), with a difference of 0.200 to 0.380 and τ=+0.647 (+0.879 under the text-weighted specification) larger than that difference, which is extrapolation from an extremely small sample. Robustness here refers only to direction, and the absolute values of τ should not be read as the magnitude of the rise. The negative share for Price & Charging is negative and not significant (-0.161, p=0.123), stays negative in the period-by-period removal test (-0.245 to -0.071), and has the largest pre-breakpoint denominator of the four aspects (42 texts), so "no statistically significant increase was observed" is robust. Two independent measures partly corroborate this pattern: refusal-to-pay discourse under the dictionary-based measure rises in the same direction (p=0.084), and positive technical evaluation falls by 0.087 (p<0.001), which becomes -0.058 (p=0.122) after removing a single period and serves only as directional evidence. Public response points mainly to "why charge" (corporate motive) and "should I pay" (behavioral intention), with "is it expensive" (price level) secondary.

Third, H2 and H5 are not supported, the evidence for H4 is inconsistent, and H1 is not robustly supported. The price fairness path is not supported: neither unfairness discourse under the dictionary-based measure nor

© 2026 The Author(s). Authors retain copyright in their articles.

negative evaluation of Price & Charging under the LLM-based measure shows a significant rise. The dual-path divergence hypothesis (H5) tests the difference between the two aspect jumps: the difference is -0.532 (Price & Charging minus Feature Experience; clustered standard error 0.437), Feature Experience rises by more, the direction is opposite to the expectation that the price path deteriorates more, and it is not significant under any specification or weighting. The pre-breakpoint sample for Feature Experience is extremely small (18 texts, 6 periods, minimum within-period n of 1), and the minimum detectable effect of this test (1.324) exceeds the theoretical upper bound of a share difference, so its power is insufficient to distinguish divergence of any magnitude; this paper therefore concludes only that "no negative change larger for the price path than for the feature path was found", not that "the two paths are the same". H1 is significant in its point estimate but depends on a period containing only 3 relevant texts before the breakpoint, is no longer significant under leave-one-period-out, and is thus not robustly supported.

6.2 Theoretical Contributions

First, this paper extends the Technology Acceptance Model to the paid-transition setting of generative artificial intelligence. The classic model explains user acceptance through perceived usefulness and perceived ease of use, but it struggles to characterize the attitudinal response triggered by "a change in the charging method during use". In the charging event public attitude does not unfold along the classic path of price fairness but is concentrated in corporate motive attribution and willingness to pay; in a freemium-to-paid transition, factors beyond the price level, especially a platform's commercialization motives, may have greater explanatory power.

Second, this paper offers a methodological contribution: "full-text completion" of social media texts has a substantial effect on measurement validity. Weibo search pages display long texts in collapsed form and carry emoji as images, so direct collection loses the latter part of the text and the emotional symbols. After completing the text through the detail interface, the median length of collapsed posts rises from 152 characters to 351 characters (2.34 times), the emoji carried as images in 4,381 texts are recovered, and the completion success rate is 99.5%. Text completeness should be reported as a necessary component of measurement validity in Chinese social media research, or conclusions may rest on truncated texts.

Third, this paper provides validity evidence on large language model annotation of Chinese social media. Existing research has tested its usability on large annotation tasks, and the supplementary finding is that the same model is highly stable in repeated measurement at temperature 0 (agreement 0.990, κ=0.980), that agreement between two models is 0.858 (κ=0.719), and that Macro-F1 for overall stance against human annotation is 0.731; the identification stability of the **aspect set** is clearly lower than that of sentiment polarity (retest Jaccard only 0.491, cross-model 0.315, exact agreement with human annotation only 0.343). The model is fairly reliable on sentiment polarity but still noisy on aspect identification; researchers using aspect-level indicators should report their stability and interpret aspect-level shares as conditional shares.

6.3 Practical Implications

For generative artificial intelligence platforms, the findings carry two implications. First, the focus of charging communication should not remain at explaining the price. Negative response points mainly to corporate motive and users' own willingness to pay rather than to the price level; clearly explaining "why charging is necessary" and committing to "which features remain free" may relieve public-opinion pressure more effectively than a price concession. Second, adjustments to the free quota need caution: public discussion of the pricing arrangement was already relatively negative before the event, so tightening the free quota while promoting the paid version is likely to compound into stronger negative perception.

For policy and industry, the paid transition is not only a pricing issue but also a communication and trust issue; the way information is disclosed affects public acceptance.

6.4 Limitations

First, the human annotation was completed by a single annotator, so inter-annotator agreement cannot be computed. This paper reviewed one by one the 44 samples on which the two models disagreed, of which 24 were corrected to "relevant", 14 to "not relevant", and 6 kept the original judgment; after this review Macro-F1 for relevance rose from 0.791 to 0.903. Because the review referred to the model judgments, 0.903 is

© 2026 The Author(s). Authors retain copyright in their articles.

optimistic; this paper also reports 0.791 under the original human annotation that did not refer to the models, and states that the latter falls below the 0.8 criterion.

Second, the identification stability of the aspect set is limited (retest Jaccard 0.491, cross-model 0.315, exact agreement with human annotation 0.343), the denominators of aspect-level indicators carry measurement noise, and their absolute levels should be interpreted with caution. The pre-breakpoint denominators are extremely small: the pre-breakpoint texts entering the regression are 42 for Price & Charging, 18 for Feature Experience, 27 for Payment Willingness, and only 7 for Corporate Brand (the minimum within-period n for Payment Willingness and Corporate Brand is 1), whereas after the breakpoint each runs into the thousands. The pre-breakpoint level of Corporate Brand is in effect determined by 7 texts (5 periods), and the absolute value of its τ is clearly larger than the raw before-after difference, which is extrapolation amplification. Leave-one-period-out can only test whether a result is driven by a single period and cannot remedy the fundamental limitation of too small a pre-breakpoint sample. This paper treats the **direction** of the aspect-level results as evidence and not their magnitude as a conclusion; the comparison test between the two aspects (H5) is subject to the same limitation, and its power is insufficient.

Third, the identification strength of the overall attitude indicator is limited and it is not robust under leave-one-period-out. Discussion volume increased several dozenfold around the event, and before the breakpoint there is a period containing only 3 relevant texts, which under "equal weighting across periods" has very high fitted leverage: after removing that period the jump estimate for the overall negative share falls from +0.317 to +0.043 (p=0.517), and after removing periods with fewer than 10 relevant texts it is -0.004 (p=0.964). This paper restricts the overall indicator to a descriptive reading of "attitude changed noticeably before and after the event", with no persistent shift and no strict causal claim. The surge in discussion volume also means the sample composition jumps at the breakpoint; the two specifications, the placebo test, the donut test and the leave-one-period-out test characterize the severity of this problem but do not fully remove it.

Fourth, the reliability of the mechanism evidence is limited. The dictionary-based method and the model-based aspect-level indicators have no substantial daily correlation (ρ between -0.143 and 0.000), the mechanism conclusions rest mainly on large model measurement, and no individual-level mediation test was conducted; the results serve only as directional consistency evidence.

Fifth, residual gaps remain in the data. Although the text completion rate reaches 99.5%, 101 records could not be completed because the original posts were deleted or access was restricted, and the main texts of a further 3 long Weibo posts are unavailable; users are subject to self-selection bias, and the public attitude referred to here is limited to the attitude reflected in Weibo discussion, so extrapolation requires caution.

6.5 Future Research

Future research can advance along three lines: first, adopt multi-annotator collaboration and multiple rounds of repeated measurement to separate and reduce measurement noise and to test inter-annotator agreement; second, link textual attitude with actual payment data on the platform side to test the association between attitude and behavior; third, apply this design to the charging events of other generative artificial intelligence products to test whether "motive attribution takes precedence over the price level" is robust across products and events.

References

[1] Strzelecki, A. (2024). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*, *32*(9), 5142-5155. https://doi.org/10.1080/10494820.2023.2209881

[2] G. C., S. B., Bhandari, P., Gurung, S. K., Srivastava, E., Ojha, D., & Dhungana, B. R. (2024). Examining the role of social influence, learning value and habit on students' intention to use ChatGPT: The moderating effect of information accuracy in the UTAUT2 model. *Cogent Education*, *11*(1), 2403287. https://doi.org/10.1080/2331186X.2024.2403287

[3] Shampanier, K., Mazar, N., & Ariely, D. (2007). Zero as a special price: The true value of free products. *Marketing Science*, *26*(6), 742-757. https://doi.org/10.1287/mksc.1060.0254

© 2026 The Author(s). Authors retain copyright in their articles.

[4] Wagner, T. M., Benlian, A., & Hess, T. (2014). Converting freemium customers from free to premium—The role of the perceived premium fit in the case of music as a service. *Electronic Markets*, *24*(4), 259-268. https://doi.org/10.1007/s12525-014-0168-4

[5] Günzel-Jensen, F., & Holm, A. B. (2015). Freemium business models as the foundation for growing an e-business venture: A multiple case study of industry leaders. *Journal of Entrepreneurship, Management and Innovation*, *11*(1), 77-101. https://doi.org/10.7341/20151115

[6] Gao, L., Zhu, K., & Zhu, K. (2025). Delineating the legal-business boundary in paid "AI ghostwriting" scenarios: The case of Wenxin Yiyan [in Chinese]. *GSP Social Science Review*, *2*(6), 36-64. https://doi.org/10.62836/ssr.v2n6.0895

[7] China Internet Network Information Center. (2025). *The 55th statistical report on China's internet development* [in Chinese]. https://www.cnnic.cn/n4/2025/0117/c88-11229.html

[8] QuestMobile. (2026). *QuestMobile 2026 Q1 AI application insight: Doubao, Qwen and DeepSeek reach 340 million, 170 million and 130 million MAU respectively; user volume and stickiness rise together as competition enters the second half* [in Chinese]. https://www.questmobile.com.cn/research/report/2046482337382842370

[9] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340. https://doi.org/10.2307/249008

[10] Oestreicher-Singer, G., & Zalmanson, L. (2013). Content or community? A digital business strategy for content providers in the social age. *MIS Quarterly*, *37*(2), 591-616. https://doi.org/10.25300/MISQ/2013/37.2.12

[11] Oh, H., Animesh, A., & Pinsonneault, A. (2016). Free versus for-a-fee: The impact of a paywall on the pattern and effectiveness of word-of-mouth via social media. *MIS Quarterly*, *40*(1), 31-56. https://doi.org/10.25300/MISQ/2016/40.1.02

[12] Wertenbroch, K., & Skiera, B. (2002). Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research*, *39*(2), 228-241. https://doi.org/10.1509/jmkr.39.2.228.19086

[13] Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence. *Journal of Marketing*, *52*(3), 2-22. https://doi.org/10.1177/002224298805200302

[14] Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 267-299). Academic Press. https://doi.org/10.1016/S0065-2601(08)60108-2

[15] Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, *76*(4), 728-741. https://www.jstor.org/stable/1806070

[16]Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

[17] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. https://doi.org/10.1561/1500000011

[18] Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309-317. https://doi.org/10.1037/h0044319

[19] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425-478. https://doi.org/10.2307/30036540

[20]Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, *39*(2), 273-315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

[21] Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615-635. https://doi.org/10.1016/j.jeconom.2007.05.001

© 2026 The Author(s). Authors retain copyright in their articles.

[22] Küçük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys*, *53*(1), 1-37. https://doi.org/10.1145/3369026

[23] Camilleri, M. A. (2024). Factors affecting performance expectancy and intentions to use ChatGPT: Using SmartPLS to advance an information technology acceptance framework. *Technological Forecasting and Social Change*, *201*, 123247. https://doi.org/10.1016/j.techfore.2024.123247

[24] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 19-30). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1002

[25] Xie, L., Zhou, M., & Sun, M. (2012). Hierarchical multi-strategy Chinese microblog sentiment analysis and feature extraction [in Chinese]. *Journal of Chinese Information Processing*, *26*(1), 73-83.

[26] Wang, Z., Wang, S., Sun, L., Zhang, L., Wang, Y., & Hao, B. (2022). Weibo tendency analysis based on sentimental object recognition and sentimental rules [in Chinese]. *Journal of Beijing University of Aeronautics and Astronautics*, *48*(2), 301-310. https://doi.org/10.13700/j.bh.1001-5965.2020.0404

[27] Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

[28] Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39-60. https://doi.org/10.1016/0167-2681(80)90051-7

[29] Xia, L., Monroe, K. B., & Cox, J. L. (2004). The price is unfair! A conceptual framework of price fairness perceptions. *Journal of Marketing*, *68*(4), 1-15. https://doi.org/10.1509/jmkg.68.4.1.42733

[30] Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

[31] Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186-204. https://doi.org/10.1287/mnsc.46.2.186.11926

[32]Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157-178. https://doi.org/10.2307/41410412

[33] Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, *82*(6), 2295-2326. https://doi.org/10.3982/ECTA11757

[34] McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698-714. https://doi.org/10.1016/j.jeconom.2007.05.005

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

© 2026 The Author(s). Authors retain copyright in their articles.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.