The Digital Journey of Health Insurance Marketing: Narrative Ads and Product Page Design

Web Appendix

Appendix A: The Digital Customer Journey Interface in Our Research 2

Appendix B: Click-Through and Purchase Rates in the Observational Study 4

Appendix C: Experimental Stimuli and Manipulation Check 6

Appendix D: Empirical Diagnostics for the Initial Budget Cap 10

Appendix E: Multimodal Large Language Model–Based Narrativeness Measurement 11

Appendix F: Audio Analyses and Topic Entropy 22

Appendix G: Coding Ad Topic Entropy with the LDA Topic Model 23

Appendix H: Video Analysis and Facial Expression Coding 26

Appendix I: Robustness Check Using 1–10 Scale Measure of Ad Narrativeness 28

Appendix J: Empirical Details of the Causal Mediation Analysis 31

Appendix K: Hyperparameter Tuning and Final Specification of the Causal Forest Model 34

Appendix L: Heterogeneity of the Synergy Effect Across Audience Composition 40

References 42

Appendix A: The Digital Customer Journey Interface in Our Research

The customer journey we study spans four stages – Awareness, Interest, Engagement, and Purchase – clustered into two distinct touchpoint contexts: the in-platform Ad Page (Awareness and Interest on Douyin) and the off-platform Product Page (Engagement and Purchase on the insurer's site). The mandatory regulatory handoff between these two contexts is the structural break that motivates our cross-touchpoint congruence argument. Figure A1 presents this conceptual flow; Figure A2 illustrates the corresponding interface at each stage.

Figure A1. Conceptual Customer Journey in Social Media-Driven Health Insurance Marketing

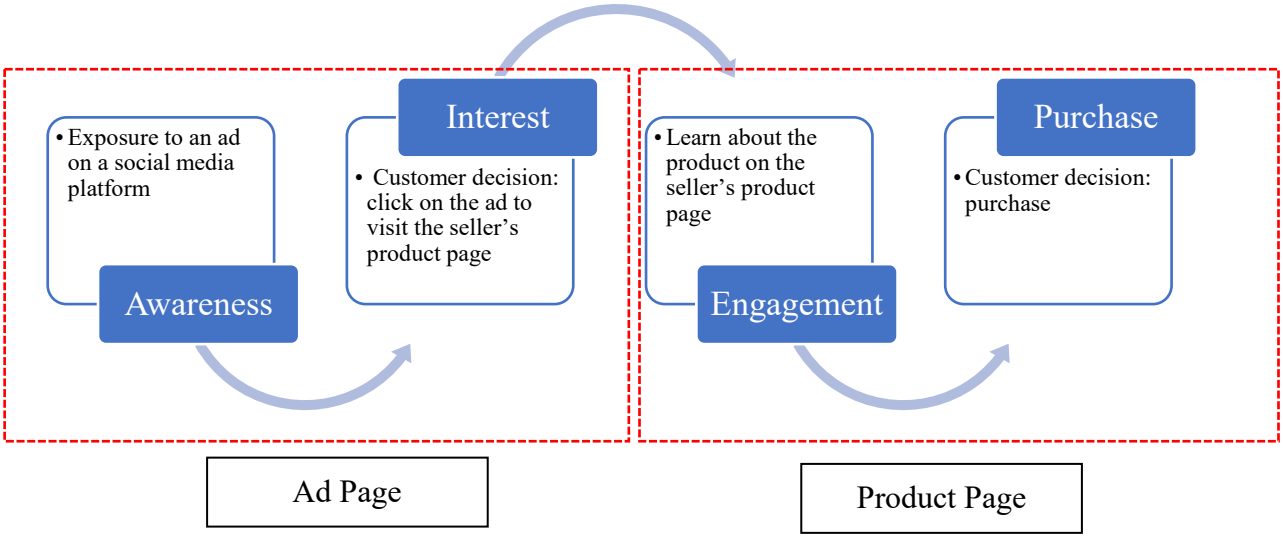
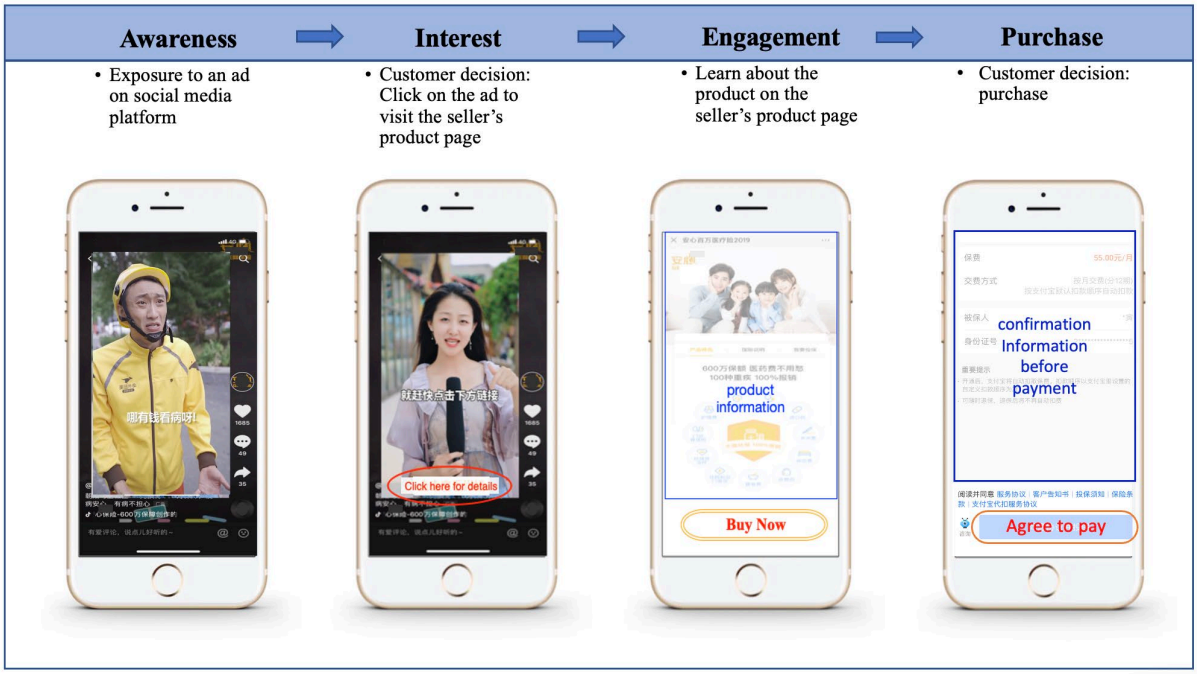


Figure A2. The Digital Customer Journey Interface in Health Insurance Marketing



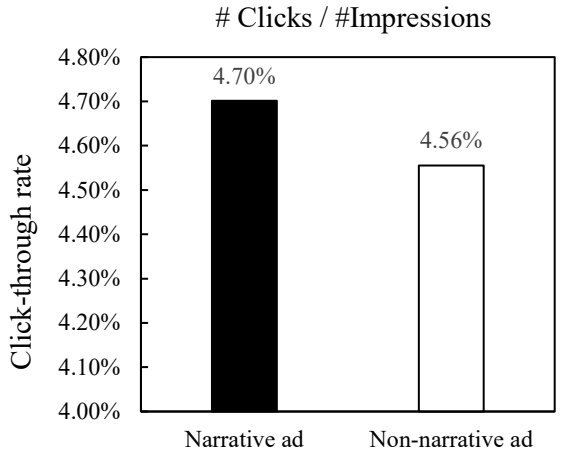
Appendix B: Click-Through and Purchase Rates in the Observational Study

Panel (a) of **Figure B** illustrates the difference in click-through rate (CTR) between narrative and non-narrative ads. Consistent with expectations, narrative advertisements achieved a significantly higher CTR ($M^{click}_{narrative}$ = 4.70% vs. $M^{click}_{non-narrative}$ = 4.56%, t (16931219) = 14.30, $p < .001$). The column difference directly reflects the positive effect of the ad type on initial engagement.

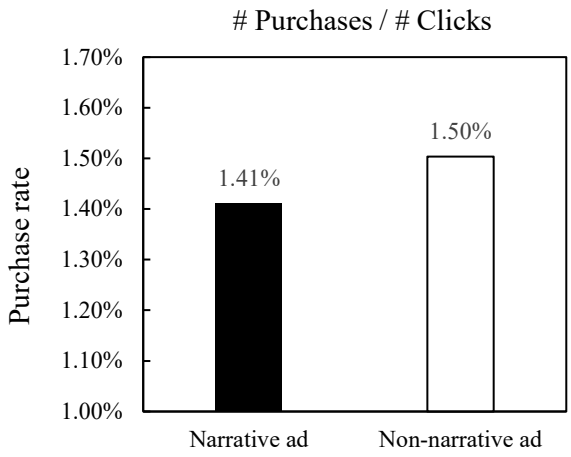
However, as shown in Panel (b), this initial engagement advantage does not translate to the immediate post-click conversion. The purchase rate (purchases/clicks) for narrative ads was lower than for non-narrative ads ($M^{purchase}_{narrative}$ = 1.41% vs. $M^{purchase}_{non-narrative}$ = 1.50%, t (782987) = 3.38, $p < .001$). This outcome suggests a conversion gap following the initial digital engagement.

We next compared the overall purchase rates (purchases/impressions) across the entire customer journey. Consistent with the click-conversion results, the narrative ads exhibited a lower overall purchase rate compared to the non-narrative ad ($M^{overall-purchase}_{narrative}$ = .066% vs. $M^{overall-purchase}_{non-narrative}$ =.068%, t (782987) = 1.6725, $p < .05$). These baseline results collectively indicate that while narrative advertising effectively drives upper-funnel engagement, it struggles with bottom-funnel conversion when executed in isolation using the fact-based product page.

Figure B. Click-through and Purchase by Ad Type in the Observational Study



(a) Influence of ad type on click-through rate (#Clicks / #Impressions)



(b) Influence of ad type on purchase conversion rate (#Purchases / # Clicks)

Appendix C: Experimental Stimuli and Manipulation Check

C.1. Experimental Design: Product Page Manipulations

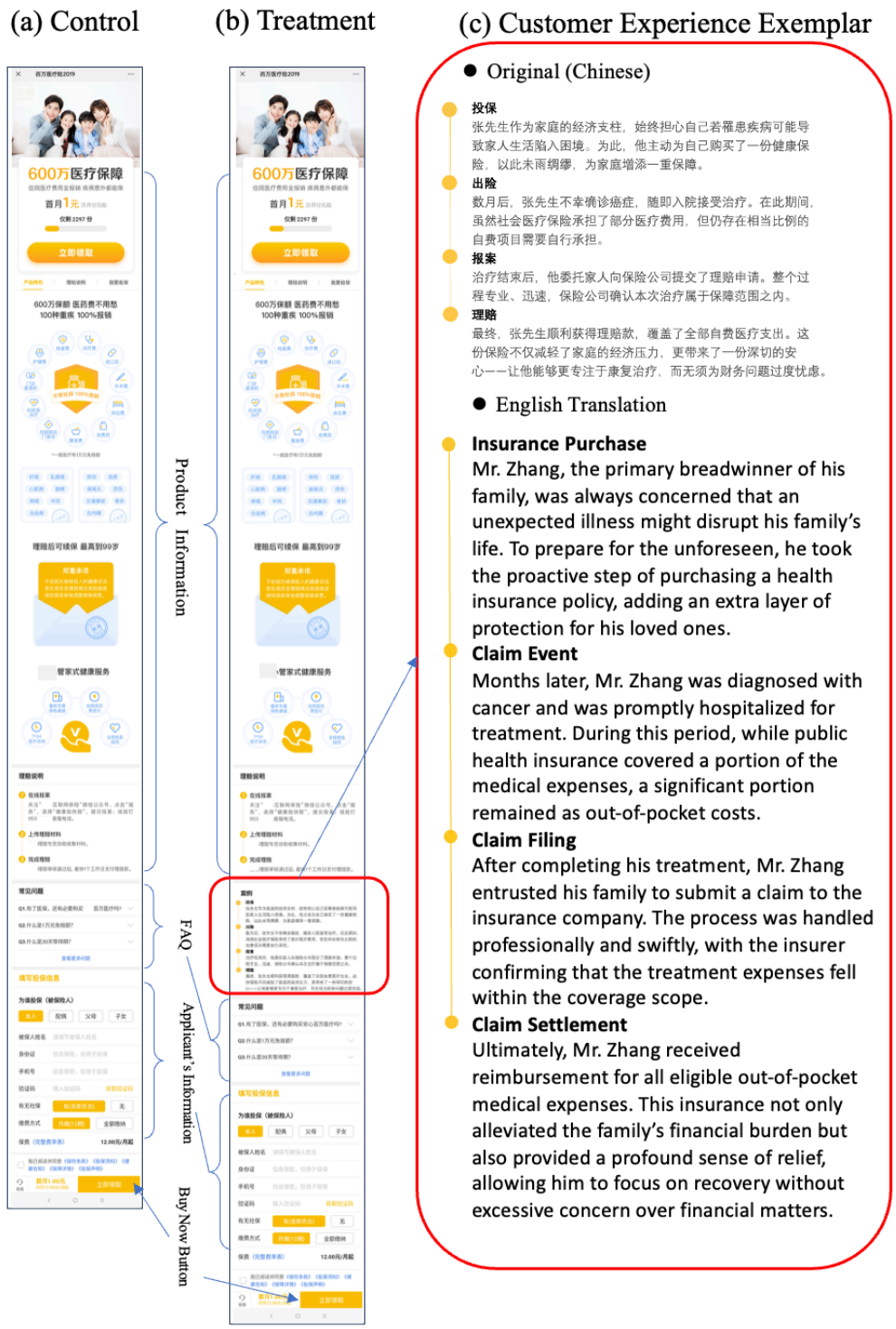
Figure C illustrates the two product page designs used in the randomized field experiment.

Panel (a): Control Condition (Factual Information Only). The standard product page presents purely factual information, including coverage details, benefits, pricing, and a conversion button.

Panel (b) Treatment Condition (Factual Information Plus Exemplar). This version retains all factual information from the control page and adds a customer experience exemplar aligned with the pre-click narrative advertisement to establish content congruence.

Panel (c) Close-up of the Customer Experience Exemplar. A detailed view of the narrative element embedded in the treatment page, recounting a specific customer's experience from policy purchase through benefit realization and claims.

Figure C. Illustration of the Two Product Page Versions in the Field Experiment



Notes: (a) shows the control page; (b) presents the treatment page featuring a customer experience exemplar; and (c) extracts and displays this exemplar in full (Chinese original and English translation).

C.2. Manipulation Check: Validation of Perceived Informativeness

To confirm that the experimental manipulation varied only along the dimension of narrativeness and not perceived informativeness, we conducted a manipulation check. Perceived informativeness was measured using a three-item scale adapted from Holzwarth et al. (2006), a measure also utilized in Kang et al. (2025).

We recruited 120 MBA students. The students were divided into two initial groups (n=60 each): the first group evaluated the control page, and the second group evaluated the treatment page. The first group subsequently evaluated the treatment page. This sequential design enabled both between-subjects (independent groups) and within-subjects (paired samples) comparisons, ensuring that any causal effect on insurance purchases could be cleanly attributed to the narrativeness manipulation.

Table C reports the results. The independent-samples t-test revealed no significant difference in perceived informativeness between the control ($M_{control}$ = 5.361) and treatment (M_{treat} = 5.444) pages, $t(118) = -.939$, $p = .348$, 95% CI [−.257, .091]. Similarly, the paired-samples t-test showed no significant difference ($M_{control}$ = 5.361, M_{treat}= 5.422; $t(59) = -1.401$, $p = .167$, 95% CI [−.149, .026]). Both tests confirm that the inclusion of the customer experience exemplar did not significantly alter perceived informativeness. Consequently, observed differences in subsequent purchase behavior can be robustly attributed to the narrativeness manipulation rather than confounding variations in information quantity or quality.

Table C. Comparison of Perceived Informativeness Between the Control and Treatment Pages

Test	Control	Treatment	Difference (SE)	t-value	p-value	95% CI of Difference
Independent-samples t-test	5.361	5.444	−.083 (.088)	−.939	.348	[−.257, .091]
Paired-samples t-test	5.361	5.422	−.061 (.044)	−1.401	.167	[−.149, .026]

Notes: The independent-samples *t*-test compares the informativeness ratings from two independent groups (n = 60 per group). The paired-sample *t*-test compares ratings from the same group of 60 participants who rated both pages.

Appendix D: Empirical Diagnostics for the Initial Budget Cap

Table D1. Balance Test on Audience Characteristics (Day1)

Audience Profile	Non-narrative Mean	Narrative Mean	Difference	t-stat	p-value
Average Age	44.23	43.63	.60	1.48	.139
Male Proportion	54.98%	54.64%	.34%	.25	.806
iPhone User Share	12.34%	11.12%	1.22%	.84	.401
Sample Size (N)	142	312			

Notes: This table reports two-sample t-tests comparing ad-day-level audience composition on Day 1 for narrative versus non-narrative ads. Average Age is the mean age of exposed viewers, Male Proportion is the percentage of exposed viewers who are male, and iPhone User Share is the percentage of exposed viewers accessing Douyin through Apple devices, all as reported by Douyin at the ad-day level. All p-values indicate no statistically significant differences in audience composition between narrative and non-narrative ads.

Figure D. Distribution of Budget Utilization (Day1)

Notes: Budget Utilization Rate is calculated as the ratio of actual advertising spend to the pre-allocated initial budget cap on Day 1.

Table D2. Balance Test on Budget Exhaustion Rates (Day 1)

Group	N	Mean (Full Exhaustion)	Std. Error	z-stat	p-value
Non-narrative Ads	142	57.04%	.042		
Narrative Ads	312	57.69%	.028		
Difference		-.65%	.050	-.13	.897

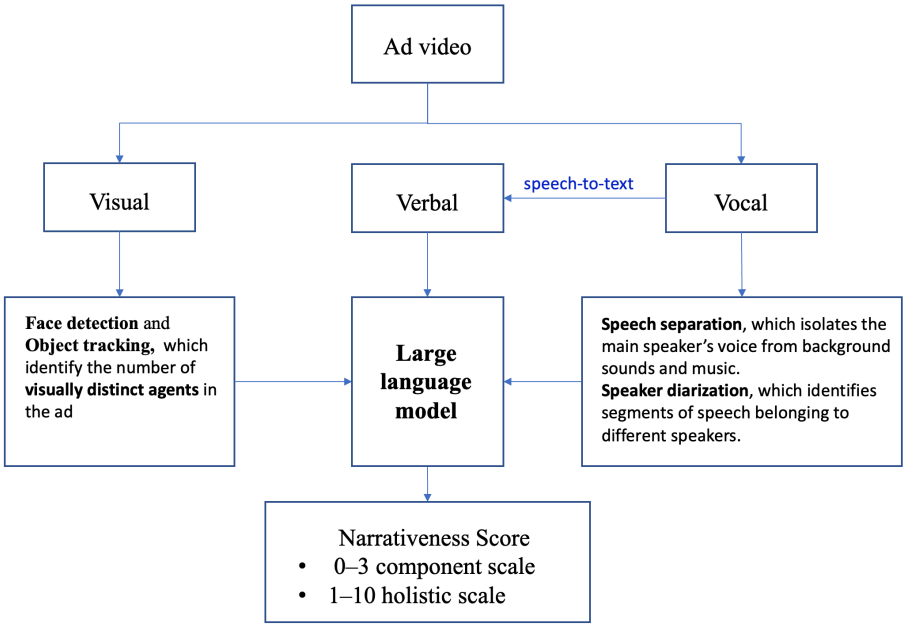
Notes: This table compares the probability of ads fully exhausting their daily budget. Full Exhaustion is a binary indicator equal to 1 if an ad's actual spend exactly reaches its initial budget cap on the first day of the campaign. The statistically insignificant difference (-.65%, $p = .897$) demonstrates that narrative and non-narrative ads are equally likely to be capped by administrative limits.

Appendix E: Multimodal Large Language Model-Based Narrativeness Measurement

To systematically analyze narrativeness embedded in video ads, we developed a multimodal data processing pipeline that integrates visual, vocal, and textual information extracted from raw video materials. This approach adheres to the structured four-step measurement framework proposed by Niu et al. (2025) for guiding Large Language Model (LLM) analysis of subjective constructs: (1) context definition, (2) model selection, (3) prompt engineering, and (4) measurement evaluation.

As depicted in Figure E1, the pipeline initiates feature extraction across three distinct modalities: Visual, Vocal, and Verbal. Visual Stream utilizes face detection and object tracking algorithms to identify and quantify salient visual agents and characters. Vocal Stream undergoes speech separation (to isolate the main speaker from background noise) and speaker diarization (to segment and identify distinct speakers). Finally, the LLM synthesizes visual and vocal cues integrated into Verbal Stream with transcribed ad text to evaluate narrativeness. This process yields narrativeness score on two metrics: a 0–3 component scale and a 1–10 holistic scale.

Figure E1. Framework for Extracting Narrativeness from Video Ads



E.1. Context Definition

Narrative advertisements embed product features within a story-like structure (Laurence 2018; Peracchio and Meyers-Levy 1997). This format is conceptually defined by two essential elements: a plot (a coherent sequence of events) and characters (identifiable agents who perform meaningful actions) (Bublitz et al. 2016; Escalas 2004). Building on this conceptualization, we operationalize narrativeness using LLM–based measures.

Our primary measure is the 0–3 component scale, adapted from Deighton et al. (1989). An ad receives 1 point for the presence of a plot, 1 point for the presence of characters, and 1 point if direct narration is absent (i.e., the ad does not rely on explicit speech or text directed at the audience, such as product claims or pack shots). This trichotomous scoring method captures the extent to which an ad relies on narrative features versus explicit expository communication.

As a complementary robustness measure, we also constructed a holistic 1–10 continuous scale. This score reflects the overall degree of narrativeness based on the integration and prominence of plot and character elements.

To establish a human benchmark for comparison, two trained coders independently evaluated the videos. An ad was labeled as narrative (1) only if both a plot and characters were present. The inter-rater reliability was high, with an overall agreement of 98% and a Cohen's kappa of .93, indicating near-perfect reliability.

E.2. Model Selection

We selected LLMs based on their capacity for deep semantic understanding and nuanced narrative reasoning. Our primary model is DeepSeek-V3, chosen due to its robust performance on Chinese-language benchmarks. The model's technical report documents superior results on various Chinese-language tasks (e.g., C-Eval, CLUWESC, and C-SimpleQA),[1] occasionally surpassing GPT-4o. This linguistic strength is critical, because all original ad materials are in Chinese. Furthermore, DeepSeek-V3's 128K token context window enables accurate semantic analysis of lengthy and multimodal inputs.

To assess cross-model robustness, we further tested ChatGPT-4o and ChatGPT-3.5. Scores generated by both models were highly correlated with the DeepSeek-V3 outputs, which provides strong confidence in the consistency of our measurement across LLM architectures.

[1] C-Eval: a comprehensive Chinese evaluation benchmark covering diverse domains; CLUWESC: Chinese Lexical Understanding Word Sense Comprehension benchmark; C-SimpleQA: Chinese Simple Question Answering benchmark.

E.3. Prompt Engineering

We developed two distinct sets of prompts to generate narrativeness scores for both scales.

E.3.1 Narrativeness: The 0–3 Component Scale

The prompt instructs the model to evaluate the presence of three criteria, each scored as 1 (present) or 0 (absent), yielding a total score from 0 to 3 :

- **Plot**: A coherent sequence of events unfolding over time, typically involving disruption, uncertainty, or resolution.
- **Character**: An identifiable agent who performs meaningful actions within the story world (e.g., making decisions, facing dilemmas, or experiencing change).
- **Un-narrated format**: The absence of dialogue directly addressing the audience (e.g., imperative statements, explicit product claims, or any speech acknowledging the viewer).

The exact prompt template is provided below:

Prompt Template: 0–3 Component Scale

You are a professional text analyst with interdisciplinary expertise in linguistics, psychology, and marketing. Your task is to analyze the following advertisement transcript and evaluate it based on the presence of three key narrative elements, as defined below. You are also provided with video analysis data, including the number of individuals detected in the ad. Please use this information to assist your evaluation, particularly for the character dimension.

Input Information

- Ad Transcript (speech-to-text): {text}
- Number of individuals detected in the video: {n}

Scoring Criteria

Evaluate the presence of the following three components. Each is scored as 1 (present) or 0 (absent). The total Narrativeness Score will range from 0 to 3.

1. **Narration**
   - Assign 1 point if the transcript is un-narrated—that is, if there is no language directly addressing the audience (e.g., no commands, claims, or explanations intended for the viewer).
   - Assign 0 points if the speaker talks directly to the audience (e.g., "Let me tell you why…", "You should consider…", "Our product helps you…").
   - Narration refers to any speech that explicitly acknowledges the audience's presence.
   - Justify your judgment by referencing specific language choices in the transcript.
2. **Character**
   - Assign 1 point if one or more protagonists are implied to be acting within a scene, as if unaware

of the audience.

- Use the video analysis result ({n} individuals detected) to help identify whether there are characters performing actions within the story world.
- Assign 0 points if no such characters are present or if all speech is directed at the viewer.
- Explain whether any speaker or implied agent demonstrates narrative agency by making decisions, facing dilemmas, or acting in context.

3. Plot

- Assign 1 point if the transcript reflects a sequence of events unfolding over time, especially involving tension, obstacles, or change.
- Assign 0 points if it only contains product descriptions, static states, or lacks temporal or causal development.
- A plot involves some movement or transformation—such as a challenge, conflict, or change that leads to a resolution.
- Identify any such event sequences and assess whether they constitute a narrative arc.

Instructions

1. Read the ad transcript and note the number of detected individuals.
2. Evaluate and score each criterion—Narration, Character, and Plot—providing a clear justification for each.
3. Compute the total Narrativeness Score (0–3 points).
4. Provide a brief summary of the transcript's overall Narrativeness and what the score implies.

Output Format

- Narrativeness Score: X point(s) (out of 3)

E.3.2 Narrativeness: The 1–10 Holistic Scale

For the 1–10 holistic scale, the model provides a single quantitative score reflecting the ad's overall degree of alignment with the narrative-ad definition, considering the seamless integration and prominence of plot and character. The corresponding prompt is provided below:

Prompt Template: 1–10 Holistic Scale

You are a professional text analyst with interdisciplinary expertise in linguistics, psychology, and marketing. Your task is to analyze the following advertisement transcript and assess how strongly it qualifies as a narrative ad, using a 1–10 scale. You are also provided with video analysis data, including the number of individuals detected in the ad. Please use this information to inform your judgment, especially for evaluating characters.

Input Information

- Ad Transcript list(speech-to-text): {text}
- Number of individuals detected in the video: {n}

Scoring Scale

Please evaluate the ad based on the two core narrative components. Each component should be scored from 1 to 10, where:

- 1 = Not present or very weak
- 10 = Strongly present and well-executed

You will report:

- A score for Plot (1–10)
- A score for Character (1–10)
- Then compute the Narrativeness Score as the average of the two (rounded to the nearest integer)

Narrative Criteria

1. Plot

- Evaluate whether the ad contains a coherent sequence of events that are either chronologically or causally linked.
- Consider whether these events are meaningfully connected and move toward a resolution or outcome.
- Mere listing of product features or facts does not constitute a narrative plot.

2. Character

- Assess whether the ad features identifiable agents who perform meaningful actions within a story.
- Characters must exhibit narrative agency—they should take actions, make decisions, face dilemmas, or otherwise drive the progression of events.
- Simply mentioning individuals (e.g., "I," "parents," "clients") is not sufficient unless they clearly act within the story.
- Use the number of individuals detected in the video ({n}) to inform whether such characters are likely present. However, scoring should be based on whether these individuals function as narrative agents in the transcript.

Instructions

1. Read the transcript carefully and consider the number of individuals detected.
2. Evaluate the presence and strength of Plot and Character based on the criteria above.
3. Assign a score (1–10) for each component.
4. Compute the Narrativeness Score as the average of the two scores (rounded to the nearest integer).

Output Format

- Narrativeness Score: X out of 10

E.4. Measurement Evaluation

To evaluate reliability and external validity, we benchmarked LLM-generated scores against human-coded binary classifications. We defined classification thresholds: ads were categorized as narrative when the 1–10 score was ≥ 6, and when the 0–3 component score was ≥ 2 . We calculated Precision, Recall, and F1 scores to assess the LLM's classification accuracy.

Table E1. Performance of LLM-Based Narrativeness Classification

Narrativeness Scale	Precision	Recall	F1 Score
0–3 scale	.945	1.000	.972
1–10 scale	.970	.920	.944

We further assessed discriminative performance using Receiver Operating Characteristic (ROC) curves. The Area Under the Curve (AUC) values were exceptionally high: .978 for the 0–3 component scale and .966 for the 1–10 holistic scale (Figure E2). These values demonstrate that the LLM-based scales possess excellent discriminative power. Combined with the favorable precision, recall, and F1 metrics, these results provide strong evidence supporting the validity and robustness of our multimodal, LLM-based narrativeness measurement.

Figure E2. ROC curves for LLM-Based Narrativeness Scores

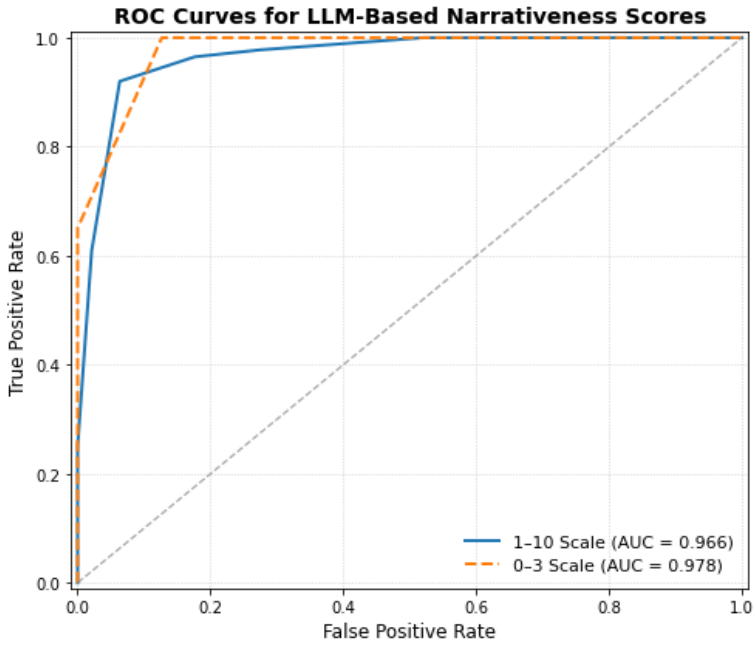


Figure E3 shows the frequency distributions of narrativeness scores assigned by the LLM to the 454 ads. Panel (a) reports the distribution for the 0–3 scale, whereas Panel (b) reports the distribution for the 1–10 scale.

Figure E3. Frequency Distributions for LLM-Based Narrativeness Scores

(a) Narrativeness (0-3) (b) Narrativeness (1-10)

Notes: Panel (a) shows the 0–3 scale distribution; Panel (b) shows the 1–10 scale distribution.

E.5. Illustrative Example

Table E2 presents an example advertisement transcript (original in Chinese, translated into English). The ad depicts a brief, story-like interaction involving three distinct characters: Old Wang, his coworker, and his son, Xiao Wang.

Using the 0–3 component scale, the LLM assigned a maximum score of 3 points:

- **Plot (Score: 1)**: The ad follows a coherent mini-narrative structure, including a problem (Old Wang fainting), rising tension (coworker's concern), and resolution (Xiao Wang offering insurance coverage).
- **Character (Score: 1)**: All three individuals act as identifiable agents within the story world, demonstrating agency through their decisions and interactions.
- **Narration (Score: 1)**: The dialogue is entirely scene-based, and does not directly address the audience (e.g., no imperative claims or pack shots).

This case demonstrates a fully narrative advertisement, in which the convergence of strong character agency, clear plot development, and an un-narrated format creates a compelling story. The alignment of the LLM-generated scores with human judgments provides compelling evidence of the model's capacity to accurately operationalize narrative structure.

Table E2. Example Ad Transcript with LLM Narrativeness Scores

Ad video	Ad Transcript
	[English translation (Original in Chinese)] Coworker: Old Wang, come on, I'll take you to the hospital! Let's go — your health matters! Old Wang: No, no, I'm fine! I just need to rest a bit. Coworker: Are you really okay? Old Wang: Yeah, it's nothing serious. Coworker: How is it nothing? You can't even stand up! I'm calling your son right now. (on the phone). Coworker: Yes, Xiao Wang, your dad just fainted — right here at the site. Come quickly! Bring plenty of money, it looks pretty serious. Xiao Wang (son): Let me talk to my dad. Coworker: Okay, here. Xiao Wang: Dad, please go to the hospital now. I've got 6 million in coverage ready for you! Old Wang: Six million?! Xiao Wang: I bought you the Million Medical Insurance last year — it covers both illness and accidents. Hospital and medical expenses are all reimbursable. Old Wang: All reimbursable? Xiao Wang: Yep — the first month was just 1 yuan, and only a dozen yuan after that. You can even pay monthly — it doesn't cost much! Old Wang: Okay, okay, I'll go, I'll go. Coworker: That's more like it! Let's go, Brother Wang — I'll take you to the hospital.

Narrativeness Score: 3 points (out of 3)

Narration Score: 1
Explanation: The ad transcript is un-narrated, meaning there is no direct address to the audience by any speaker. The dialogue takes place between characters (Old Wang, his coworker, and his son Xiao Wang), and the language is natural and scene-based. There is no instance where the speaker explicitly acknowledges the viewer (e.g., "You should consider…" or "Let me tell you…"), and all statements are made in the context of the fictional situation. Therefore, the narration criterion is satisfied with 1 point.

Character (based on 3 individuals) Score: 1
Explanation: The ad includes three distinct individuals (Old Wang, Coworker, and Xiao Wang), and all of them are implied to be acting within a scene, unaware of the audience. Their dialogue and actions (e.g., faking illness, calling the son, discussing insurance) suggest narrative roles and agency. For instance:
Old Wang is the protagonist who initially resists going to the hospital.
Coworker is concerned and takes initiative by calling the son.
Xiao Wang arrives and attempts to resolve the situation by offering insurance coverage.
Each character has a distinct voice and demonstrates narrative agency (e.g., making decisions, responding to a crisis, and showing concern). Therefore, the character criterion is satisfied with 1 point.

Plot Score: 1
Explanation: The ad reflects a working out of a story, which includes a problem (Old Wang fainting), rising tension (coworker's concern and insistence), and a resolution (Xiao Wang offering the insurance solution).
The expectation is that Old Wang is unwell and needs medical attention, but he initially resists it.
The opposition or doubt arises from Old Wang's denial of his condition and the coworker's frustration.
The resolution comes when Xiao Wang intervenes, offering the Million Medical Insurance as a solution. This product is introduced as a way to address the uncertainty of illness and accidents, and the dialogue leads to Old Wang finally accepting to go to the hospital.
This progression clearly follows the structure of a mini-narrative, with a setup, conflict, and resolution. Thus, the plot criterion is satisfied with 1 point.

Summary:
The ad exhibits a strong narrative structure with 3 full points. It tells a mini-story involving three characters who act within a realistic, emotionally engaging scene. The presence of a clear plot progression and the absence of direct audience address contribute to a compelling and immersive narrative, making the product (Million Medical Insurance) feel like a natural and

necessary resolution to the problem presented. This kind of storytelling is effective in advertising as it emotionally connects the viewer to the characters and subtly highlights the benefits of the product without being overtly salesy.

Narrativeness Score: 8 out of 10

Plot (Score: 9/10):
Explanation: The ad presents a clear and coherent sequence of events that are causally and chronologically linked, forming a mini-narrative. It begins with Old Wang fainting, which creates a problem or crisis. The coworker reacts by attempting to convince Old Wang to go to the hospital and calls his son. The son arrives, introduces the Million Medical Insurance, and uses it as the resolution to the health concern. This structure follows a classic narrative arc:
Setup: Old Wang fakes or genuinely appears to be unwell.
Conflict: His denial of the problem and the coworker's concern.
Climax/Resolution: Xiao Wang offers the insurance as a solution, and Old Wang finally agrees to go to the hospital.
While the narrative is short and somewhat formulaic, it effectively conveys a cause-effect progression and culminates in a product-oriented resolution. The product is not merely listed, but rather integrated into the resolution of the story, which enhances its narrative function. Hence, the Plot score is 9.

Character (based on 3 individuals) (Score: 7/10):
Explanation: The ad features three individuals (Old Wang, Coworker, and Xiao Wang), and each plays a distinct role in the narrative.
Old Wang is the central character whose health is the initial concern and who undergoes a change of mind at the end.
Coworker is a supporting character who initiates the action by offering help and calling the son.
Xiao Wang (the son) is a key agent who introduces the product and resolves the conflict by providing reassurance and the solution.
All three demonstrate narrative agency—they make decisions, respond to the situation, and interact with each other in a way that drives the story forward. However, the characters are somewhat stereotypical (e.g., the worried coworker, the responsible son), and their motivations and internal conflicts are not deeply explored. The lack of emotional depth and character complexity slightly limits the narrative strength. Therefore, the Character score is 7.

Summary:
The ad demonstrates strong narrative qualities, particularly in its plot development and character roles. The three individuals function as agents within a story world, and the events are structured in a way that builds tension and leads to a resolution. While the characters are not deeply nuanced, the narrative is effective in conveying the value of the product in a relatable and emotionally engaging context. The Narrativeness Score of 8 indicates that this is a well-crafted narrative ad, with clear storytelling and character agency, though there is room for improvement in terms of character depth and emotional resonance.

Appendix F: Audio Analyses and Topic Entropy

We convert the audio streams of all advertisements (in Mandarin Chinese) into text transcripts using a Recurrent Neural Network (RNN). As a state-of-the-art speech recognition algorithm, an RNN mimics the human brain's memory function, allowing it to make contextual predictions (Rumelhart et al. 1986). Its ability to process short-term features while accommodating long-term dependencies makes it particularly suitable for this task. Prior research has demonstrated that RNN-based speaker recognition improves with longer utterances (Yu and Deng 2016). Following transcription, we conducted content analysis to measure the topic entropy of each ad. Figure F illustrates the systematic approach.

Figure F. Audio Analysis



Appendix G: Coding Ad Topic Entropy with the LDA Topic Model

A Latent Dirichlet Allocation (LDA) model is a topic modeling technique that uncovers underlying topics within a collection of documents and infers the probability distribution of words within each topic. The LDA model operates under the assumption that a document is generated through a probabilistic process, where a topic is first sampled from the document's topic distribution, followed by the selection of a word from that topic's word distribution. To analyze the advertisement texts, we estimated a four-topic LDA model, selecting the number of topics based on multiple model fit measures. This decision was made through triangulation across different evaluation metrics to ensure robustness. We followed the approach of Choudhury et al. (2019), using entropy as a measure to capture variations in ad information. The topic entropy was calculated as:

$$entrop_j = \sum_{k=1}^{K} p_{jk} \times log_2 p_{jk} \qquad \text{(G1)}$$

where p_{jk} represents the proportion allocated to topic *k* for ad *j*, capturing the extent to which the content of ad *j* is focused on a few topics or distributed across multiple topics. Entropy serves as a measure of the breadth of informational content within a text. A low entropy value indicates that the ad message is concentrated around a few key topics, whereas a high entropy value suggests that the message covers a diverse range of topics.

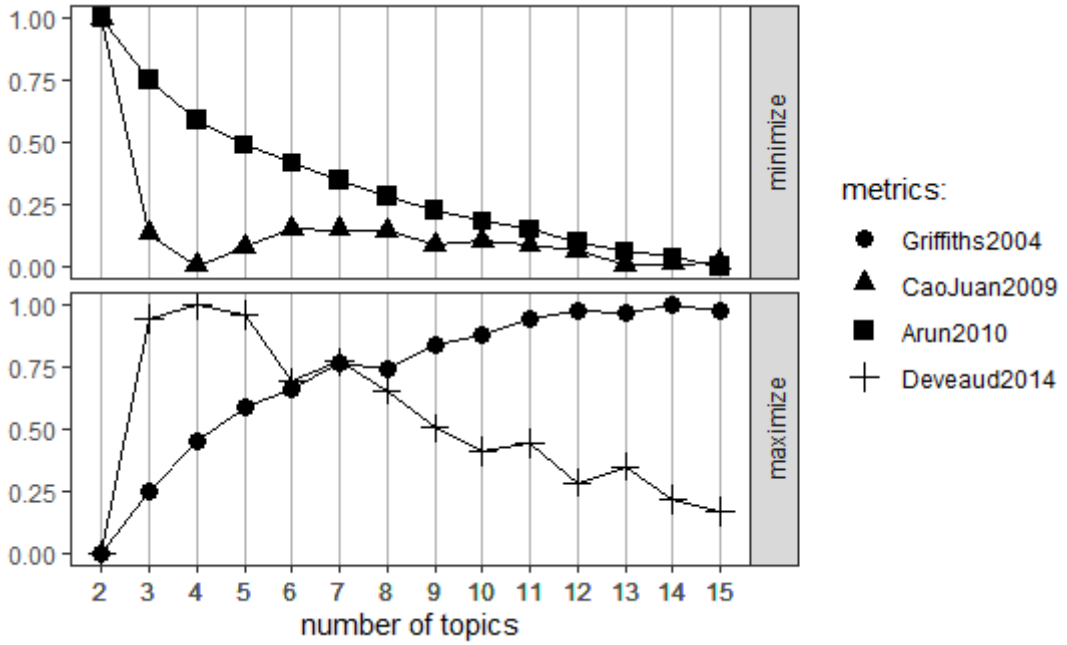
We followed the methodology of Choudhury et al. (2019) to determine the optimal number of topics (K) for the topic model, relying on key model fit indicators such as perplexity and log-likelihood. To facilitate this process, we used the ldatuning package in R, which enables a side-by-side comparison of four different model fit measures. These four measures, computed across

the entire document collection, are plotted against each other, with the x-axis representing the number of topics. Among these, the "CaoJuan2009" metric (represented by triangle points) estimates topic density by aiming to reduce the average cosine similarity between topics, ensuring greater distinctiveness (Cao et al. 2009). The Arun et al. (2010) metric (depicted by squared-off points) applies a Kullback-Leibler divergence measure, minimizing divergence to identify the most appropriate K for the model.

The bottom panel of the figure displays two metrics to maximize when selecting the optimal number of topics. The circular points represent the Griffiths and Steyvers (2004) indicator, which calculates the harmonic mean of the log-likelihood across different values of K, identifying the number of topics that best fits the data. The cross points denote the metric proposed by Deveaud et al. (2014), which maximizes the information divergence between all LDA topic pairs. This measure computes the Jensen-Shannon divergence for each pair of topics and selects the K that maximizes the average divergence across topic pairs, ensuring greater distinctiveness between extracted topics.

In each panel, fit measures are scaled between 0 and 1, with the optimal value indexed at 1. This side-by-side comparison allows for a more comprehensive evaluation, even though not all metrics may be equally relevant for a given problem. Based on the results presented in Figure G, we determined that the optimal number of topics (K) was 4.

Figure G. Fit Metrics for Topic Number Selection



Note: Figure G presents four different model fit measures across various topic numbers for our topic model estimation. The top panel displays two metrics that should be minimized to achieve the best fit, while the bottom panel presents two metrics that should be maximized. All fit measures are scaled between 0 and 1, with the optimal value indexed at 1, facilitating direct comparison across different topic numbers.

Appendix H: Video Analysis and Facial Expression Coding

Extracting visual features from video presents a significant computational challenge due to the sheer volume of data. With videos running at 24 frames per second over a total duration of 70,366 seconds, retaining all frames would be computationally prohibitive. To reduce this burden, we follow standard practice by preserving only the 12th frame per second, assuming it captures the dominant visual characteristics within that time frame. This approach yields approximately 70,366 images, each with a resolution of 360×640 pixels. The first step in our analysis involves detecting facial features within these images. Following recent marketing research (Lu et al. 2021), we utilize the dlib C++ library and its default face detector. This detector integrates the Histogram of Oriented Gradients (HOG) feature with a linear classifier, an image pyramid, and a sliding window detection mechanism. Trained on the iBUG 300-W face landmark database (Kazemi and Sullivan 2014; Sagonas et al 2013), this model achieves an accuracy rate of approximately 95% (Kazemi and Sullivan 2014).

Following recent applications (e.g., Choudhury et al. 2019; Hartmann et al. 2021), we employed a Convolutional Neural Network (CNN) algorithm to recognize actors' facial expressions in advertisements. A CNN transforms an input image into a multilayer hierarchical structure, where the first layer corresponds to the input image, and subsequent layers capture increasingly complex spatial features. The early layers detect shallow pixel groupings, such as edges and small facial components (e.g., parts of the nose or eye), while deeper layers identify more abstract facial structures by integrating distant pixel neighborhoods across the entire face. For each video, we extracted one frame per second and applied a pre-trained CNN, which

generated seven probability scores, each indicating the likelihood that the image depicted a particular facial expression: happiness, surprise, disgust, anger, sadness, fear, or neutrality. Figure H illustrates the systematic approach used for face recognition, detailing the outcomes and algorithms implemented at each step.

Figure H. Face Recognition and Facial Expressions Coding



Appendix I: Robustness Check Using 1–10 Scale Measure of Ad Narrativeness

To ensure that the causal findings regarding the effect of narrative congruity are not sensitive to the specific scaling or operationalization of the narrativeness measure, we performed a comprehensive robustness check. We re-estimated all primary models using a 1–10 holistic scale measure of ad narrativeness rather than the primary 0–3 component scale.

Tables I1–I3 report the corresponding results, including the two-stage control function model and the causal forest model estimations. The findings derived from the 1–10 scale are highly consistent in sign, magnitude, and statistical significance with those reported in the main analysis, confirming the robustness and stability of our conclusions across alternative measures.

Table I1. First-Stage Regression of AD Narrativeness (1-10) on Instrumental Variable

	Narrativeness(1-10 scale)
Instrument (Initial Budget Cap)	.853***
	(.083)
Age	-.064
	(.093)
Gender	-.759
	(4.884)
iPhone	-.515
	(.977)
Ad length	-.218
	(.182)
Entropy	-.832***
	(.197)
Locations	Yes
Emotions	Yes
Ad Agency	Yes
Day fixed effect	Yes
Number of Obs	16,962,383
R-squared	.910

Notes: * $p < .10$, ** $p < .05$, *** $p < .01$. Standard errors, clustered at the ad level, are reported in parentheses. Clustered robust standard errors by ad (accounting for within-ad correlation in the error terms) are reported in parentheses.

Table I2. Effects of Narrativeness(1-10) on Click-Through: Heckman First-Stage Results

	Click-Through
Instrument (Initial CTR)	.189**
	(.076)
Narrativeness(1-10 scale)	.002***
	(.001)
Age	.000
	(.002)
Gender	.068
	(.072)
iPhone	.007
	(.012)
Ad length	.002***
	(.001)
Entropy	.003***
	(.001)
$\hat{u}_cont$	-.003***
	(.001)
Locations	Yes
Emotions	Yes
Ad Agency	Yes
Day fixed effect	Yes
Number of Obs	16,962,383

Notes: * $p < .10$, ** $p < .05$, *** $p < .01$. $\hat{u}_cont$ is the residual from Equation (1) for the 1-10 scale measures of narrativeness. Clustered robust standard errors by ad (accounting for within-ad correlation in the error terms) are reported in parentheses.

Table I3. The Synergy Effects on Purchase: Heckman Second-Stage Results

	Purchase
Narrativenesss(1-10 scale)×Treat	.001***
	(.000)
Narrativenesss(1-10 scale)	-.001
	(.001)
Treat	-.011***
	(.003)
Age	-.002
	(.001)
Gender	.018
	(.077)
iPhone	.004***
	(.000)
Ad length	-.004***
	(.002)
Entropy	-.002
	(.002)
$\hat{u}_cont$×Treat	-.003**
	(.001)
$\hat{u}_cont$	.002*
	(.001)
IMR_cont	.107
	(.227)
Locations	Yes
Emotions	Yes
Ad Agency	Yes
Day fixed effect	Yes
Number of Obs	777,412
R-squared	.005

Notes: * $p < .10$, ** $p < .05$, *** $p < .01$. $\hat{u}_cont$ are the residuals from Equation (1) for the1-10 scale measure of narrativeness. IMR_cont are the inverse Mills ratios computed from Equation (2) based on the 1-10 scale narrativeness. Clustered robust standard errors by ad (accounting for within-ad correlation in the error terms) are reported in parentheses.

Appendix J: Empirical Details of the Causal Mediation Analysis

To examine the mechanism underlying the synergy effect, we test whether narrative congruence increases purchase conversion by sustaining consumer engagement. We operationalize engagement as the log-transformed count of clicks across 19 interactive elements on the product page. These actions reflect explicit user-initiated choices to deepen content interaction.

We estimate the Average Causal Mediation Effect (ACME) using the framework of Imai, Keele, and Yamamoto (2010) via Quasi-Bayesian Monte Carlo simulations (1,000 draws). Following the classification framework of Zhao, Lynch, and Chen (2010), the pattern of a significant indirect effect (ACME) and a non-significant direct effect (ADE) constitutes indirect-only mediation, providing strong evidence that the synergy effect operates through the proposed engagement pathway.

To quantify consumer engagement on the purchase page, we first derive a raw behavioral metric, *ClickCount*, calculated by summing the total number of clicks across 19 distinct interactive elements. To ensure the independence of this measure from our dependent variable, Purchase, we explicitly exclude the 'Buy Now' button from this count. The 19 interactive elements included in *ClickCount* are: three buttons leading to 'Product Features,' 'Claims Instructions,' and 'Apply'; four buttons that expand questions in the FAQ section; four buttons for selecting the insured party ('Self,' 'Spouse,' 'Parents,' or 'Children'); six links directing to the 'Rate Table,' 'Insurance Terms,' 'Purchase Notice,' 'Health Disclosure,' 'Coverage Details,' and 'Purchase

Statement'; and two buttons for 'Payment Method' and 'Social Insurance Status.' The visual layout and original interface design of these elements are provided in Appendix C.

Regarding the descriptive properties of *ClickCount*, the distribution exhibits substantial variation, with a mean of 9.67 and a standard deviation of 8.25. Observed values range from a minimum of 0 to a maximum of 89. We define *Engagement* as $\ln(1 + ClickCount)$ to normalize the distribution, stabilize variance, and minimize the influence of extreme values.

Table J1 presents the results. For the binary measure, the ACME is .0049 ($p < .001$), while the Average Direct Effect (ADE) is non-significant ($p = .468$). Similarly, for the continuous measure, the ACME is .0018 ($p < .001$), with the ADE remaining non-significant ($p = .472$). In both specifications, the mediated proportion exceeds 80%. These findings suggest that the synergy effect is primarily channeled through sustained engagement. The congruent ad-page design maintains the momentum generated by the narrative ad, motivating consumers to interact more actively with product information and thereby facilitating conversion.

Table J2 reports the estimates for both the mediator model and the outcome model across both operationalizations of narrativeness: the binary measure and the continuous measure.

Table J1. Causal Mediation Analysis of Synergy Effect on Purchase via Engagement

	Narrativeness Operationalization					
	(1) Binary			**(2) Continuous**		
Effect Type	**Estimate**	**[95% CI]**	**P-value**	**Estimate**	**[95% CI]**	**P-value**
ACME	.0049	[.0047, .0053]	< .001	.0018	[.0017, .0019]	< .001
ADE	.0012	[−.0017, .0046]	.468	.0004	[−.0007, .0016]	.472
Total Effect	.0061	[.0033, .0097]	< .001	.0022	[.0010, .0034]	< .001
Prop. Mediated	82.0%	[.5117, 1.5131]	< .001	80.1%	[.5215, 1.6978]	< .001

Notes: N = 777,412. All estimates, 95% confidence intervals (CI), and p-values are derived using the Quasi-Bayesian approximation following the framework of Imai et al. (2010). Model (1) uses a binary dummy for narrativeness. Model (2) uses a continuous narrativeness score. ACME and ADE represent Average Causal Mediation Effect and Average Direct Effect, respectively.

Table J2. Regression Estimates for Mediation Analysis

	Engagement		Purchase	
	（1）Binary	（2）Continuous	（1）Binary	（2）Continuous
Narrativenesss(binary)×Treat	.4738***		.0013	
	(.0122)		(.0016)	
Narrativenesss(binary)	.0021		-.0092***	
	(.0113)		(.0015)	
Narrativenesss(continuous)×Treat		.1692***		.0004
		(.0046)		(.0006)
Narrativenesss(continuous)		-.0053		-.0012**
		(.0037)		(.0005)
Treat	.0009	-.0417***	-.0065***	-.0065***
	(.0094)	(.0105)	(.0012)	(.0014)
Mediator(Engagement)			.0105***	.0105***
			(.0001)	(.0001)
Age	.0027	.0006	-.0022*	-.0018
	(.0093)	(.0093)	(.0012)	(.0012)
Gender	.0882	-.0038	.0017	.0199
	(.4992)	(.4996)	(.0654)	(.0655)
iPhone	-.0003	-.0004	.0044***	.0044***
	(.0031)	(.0031)	(.0004)	(.0004)
Ad length	-.0047	-.0081	-.0051***	-.0045***
	(.0065)	(.0065)	(.0008)	(.0009)
Entropy	.0121**	.0088*	-.0034***	-.0030***
	(.0051)	(.0051)	(.0007)	(.0007)
$\hat{u}_bin$ × Treat	.0364		-.0096*	
	(.0386)		(.0051)	
$\hat{u}_bin$	-.0311**		.0008	
	(.0153)		(.0020)	
IMR_bin	-.4248		-.2909**	
	(.8732)		(.1145)	
$\hat{u}_cont$ × Treat		-.0748***		-.0042**
		(.0133)		(.0017)
$\hat{u}_cont$		.0132**		.0024***
		(.0056)		(.0007)
IMR_cont		-.9847		.0865
		(.8143)		(.1067)
Locations	Yes	Yes	Yes	Yes
Emotions	Yes	Yes	Yes	Yes
Ad Agency	Yes	Yes	Yes	Yes
Day fixed effect	Yes	Yes	Yes	Yes
Number of Obs	777,412	777,412	777,412	777,412
R-squared	.005	.005	.011	.011

Notes: * p < .10, ** p <.05, *** p <.01. $\hat{u}_bin$ and $\hat{u}_cont$ are the residuals from Equation (1) for the binary and continuous measures of narrativeness, respectively. IMR_bin and IMR_cont are the inverse Mills ratios computed from Equation (2) based on the binary and continuous narrativeness. Standard errors are reported in parentheses.

Appendix K: Hyperparameter Tuning and Final Specification of the Causal Forest Model

K.1. Hyperparameter Tuning Strategy

Given the size of our dataset (exceeding 770,000 observations), exhaustive hyperparameter tuning on the full sample was computationally prohibitive. We therefore implemented a two-step cross-validation strategy.

First, we constructed a stratified subsample: we retained all positive purchase cases (where Purchase = 1) and randomly sampled 20,000 negative cases (where Purchase = 0). This preserved the rare-event structure while significantly reducing computational burden. On this subsample, we primarily concentrated tuning efforts on *min.node.size*, which directly controls the granularity of causal tree splits and thus governs the bias–variance trade-off.

Second, we explored alternative values of *mtry* (the number of covariates considered at each split) and *sample.fraction* (the proportion of the sample used to grow each tree). Consistent with prior evidence, we found that their influence on the estimated Conditional Average Treatment Effects (CATEs) was negligible compared with min.node.size.

Following extensive tuning, we determined that the optimal configuration closely aligned with the package default settings. Accordingly, we adopted the default hyperparameter specification for the full-sample estimation and increased the number of trees (num.trees = 2000) to ensure stable and robust inference.

K.2. Treatment Imbalance and Clustering

To address treatment imbalance between the control and treatment conditions, we applied **Inverse Probability of Treatment Weights (IPTW)**. The weight for each observation (Wi) was calculated as:

$$w_i = \begin{cases} \frac{.5}{p_1}, w_i = 1 \\ \frac{.5}{1 - p_1}, w_i = 0 \end{cases}$$

where p_1 denotes the treatment probability. This weighting scheme equalizes the effective sample sizes of the treatment and control groups, thereby improving estimation efficiency and preventing the Causal Forest from being dominated by the larger, non-weighted group.

To account for potential correlation among observations arising from exposure to the same advertisement, we specified ad-level clustering in the causal forest estimation by providing the ad identifier to the clusters argument. This ensures that the standard errors are robust to within-ad correlations, thereby providing conservative estimates while still preserving the unit-level heterogeneity in treatment effects.

K.3. Final Model Specification

The causal forest model was estimated using 2000 trees and the following configuration, primarily adopting package-default hyperparameters:

- min.node.size = 5
- sample.fraction = .5
- mtry = $min(\sqrt{p} + 20, p)$ (where p is the number of covariates)
- honesty = TRUE with honesty.fraction = .5
- alpha = .05
- imbalance.penalty = 0

K.4 Robustness of Main Results: Detailed Causal Forest Estimates

As a critical robustness check, we employ the Causal Forest approach (Wager and Athey 2018), a nonparametric, data-adaptive method that estimates treatment effects without strong functional-form assumptions. This approach is well suited to our setting for two reasons. First, it flexibly captures nonlinearities and high-order interactions among ad features and user characteristics; these patterns, often induced by algorithmic content distribution and heterogeneous user behavior, could evade conventional parametric models. Second, the Generalized Random Forest (GRF) framework incorporates orthogonalization and honest tree estimation (via sample splitting), yielding doubly robust and approximately unbiased treatment effect estimates even in high-dimensional settings (Athey and Wager 2019; Semenova and Chernozhukov 2021).

Together, these properties reduce model dependence and mitigate omitted-variable bias, strengthening the credibility of our inferences. The Causal Forest model not only validates the robustness of our main findings but also enables an exploration of treatment effect heterogeneity.

Following the potential outcomes framework (Rubin 1974), we define the treatment ($W = 1$) as exposure to the narrative ad paired with the exemplar product page, and the control ($W = 0$) as exposure to the alternative ad-product page combinations. For each consumer i, we compare two potential outcomes: Y_{i1} (outcome under treatment group) and Y_{i0} (outcome under control group).

The Conditional Average Treatment Effect (CATE) is: $\tau_i(x) = E[Y_{i1} - Y_{i0}|X_i = x]$, where X_i is the vector of observed characteristics. We estimate the CATE using the causal forest

approach (R package grf), which partitions data into subgroups with similar characteristics and estimates treatment effects within each subgroup. To mitigate bias, the method uses a two-sample procedure, where one subset defines the subgroups and a separate subset estimates the treatment effects. Variance is further reduced by repeating the procedure across random samples and averaging results (Chen et al. 2020). We estimated causal forests with 2,000 trees and applied inverse probability weights to account for treatment imbalance. Hyperparameter tuning details are reported in Appendix K.1-K.3 above.

Table K and Figure K1 present the results. The Average Treatment Effect (ATE) of the congruent combination on purchase conversion is .0051.

Table K. Summary of ATEs and Individual-Level CATEs

ATE	**N**	**ATE**	**SE**	**95% CI**	
$\hat{\tau}$	777,412	.0051	.003	[-.0008, .0110]	
CATE	**N**	**Mean**	**SD**	**Min**	**Max**
$\hat{\tau}_i$	777,412	.0065	.009	-.0277	.2101

Notes: CI refers to confidence interval. $\hat{\tau}$ refers to the estimated average treatment effect; $\hat{\tau}_i$ refers to the estimated conditional average treatment effect for consumer *i*.

Crucially, the results indicate substantial heterogeneity: 21% (163,524 out of 777,412) of observations exhibit negative CATEs, with estimates ranging from -.0277 to .2101. While the intervention is effective on average, its benefit is not universal. To validate this heterogeneity, we conducted an omnibus test using the Targeting Operator Characteristic (TOC) curve (Sverdrup et al., 2025). The estimated Area Under the Targeting Operator Characteristic (AUTOC) of .0073 ($t = 3.713$, $p < .001$) confirms that the causal forest model captures a significant predictive signal of treatment sensitivity rather than estimation noise.

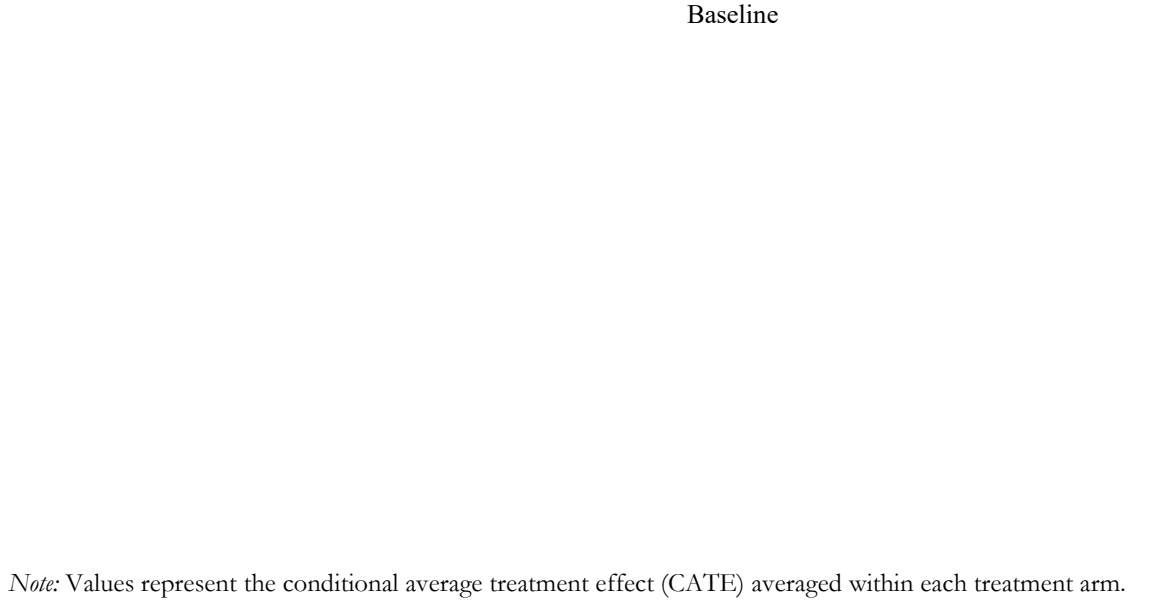
Figure K1. Distribution of CATEs on Purchase

Notes: ATE refers to the average treatment effect; SE refers to the standard error; CATEs refer to conditional average treatment effects.

To compare all experimental configurations, we employed a multi-arm causal forest, estimating the treatment effects of different configurations relative to a defined baseline. Figure K2 reports the estimated CATEs for the three treatment arms relative to the baseline (Non-narrative + Standard Page). Positive values indicate higher purchase likelihood than baseline. The relative CATEs are: Non-narrative + Exemplar = -.0057, Narrative + Standard = -.0070, and Narrative + Exemplar = .0015.

The congruence configuration yields the highest relative CATE, whereas both incongruent combinations produce negative relative effects. These findings reinforce that alignment between the ad format and the post-click exemplar presentation enhances purchase conversion, while misalignment detracts from it.

Figure K2. CATEs of Ad–Product Page Combinations Relative to Baseline



Note: Values represent the conditional average treatment effect (CATE) averaged within each treatment arm. Numbers indicate the relative CATE compared with baseline (Non-narrative + No Exemplar): Non-narrative + Exemplar = -.0057, Narrative + No Exemplar = -.0070, Narrative + Exemplar (narrative congruence) = .0015. Positive values indicate higher purchase likelihood than baseline. Narrative congruence yields the highest relative CATE, while non-congruence arms have negative relative CATEs.

Appendix L: Heterogeneity of the Synergy Effect Across Audience Composition

While the main text establishes the average synergy effect of pairing narrative ads with the exemplar product page, a natural question is whether this effect varies systematically with the composition of the audiences the ads reach. This appendix reports an exploratory heterogeneity analysis. We emphasize at the outset that the analysis is descriptive rather than theoretically motivated, and it is subject to a substantive data limitation discussed below.

L.1. Data Limitation and Interpretive Caveat

Our dataset contains ad-level audience aggregates for two of the three viewer-composition variables (age and gender) and an individual-level measure for the third (iPhone). *Age* and *Gender* are daily composition metrics computed across the viewers each ad reached on a given day (the average age of viewers and the proportion of male viewers). *iPhone* is an individual-level indicator recorded directly from the user's mobile device upon accessing the product page. The associations reported below for age and gender should therefore be interpreted as reflecting how the synergy effect differs across ads delivered to audiences with distinct demographic skews, not as causal statements about individual-level treatment-effect heterogeneity. We avoid phrasings such as 'the effect is stronger for women' and use formulations such as 'the effect is stronger for ads delivered to female-skewed audiences' throughout. The iPhone result, by contrast, can be interpreted at the individual level.

L.2. Estimation Approach

We quantify heterogeneity by estimating the Best Linear Projection (BLP) of the Conditional Average Treatment Effect (CATE) on observable ad-level characteristics (Athey and Wager 2021). Following Semenova and Chernozhukov (2021), we regress de-biased doubly robust scores derived from the causal forest (see Sections K.1–K.3 of Appendix K) on viewer-composition variables, controlling for ad-level features (topic entropy and the seven facial emotion scores). Robust standard errors are reported.

L.3. Results

Table L reports the BLP coefficients. We discuss the three viewer-composition dimensions in turn.

(a) **Gender composition**. The synergy is significantly weaker for ads with a higher share of male viewers ($\beta = -.1970$, $p < .05$), implying that narrative congruence is more effective for ads delivered to female-skewed audiences. This pattern is broadly consistent with prior work suggesting that women, on average, engage more readily with experiential and affective processing while men lean toward analytical processing (Stern 1992; Argo, Zhu, and Dahl 2008;

Van Laer et al. 2014). We note again that our data cannot distinguish whether the pattern reflects individual-level differences or other features that covary with audience gender composition.

(b) **Age composition**. We find no significant association between the average audience age and the magnitude of the synergy effect ($\beta = .0036$, n.s.). The effect of narrative congruence appears broadly stable across advertisements reaching audiences of different average ages.

(c) **Socioeconomic composition**. Using the individual-level iphone indicator as a proxy for higher socioeconomic status in the Chinese context (Ding, Gordon, and Knorr 2021), we find that the synergy is significantly attenuated for users accessing the product page through an iOS device ($\beta = -.0073$, $p < .01$). Because iphone is measured at the individual level in the purchase data, this association reflects individual-level heterogeneity rather than variation across ad-audience configurations: higher-SES users (proxied by iOS-device ownership) appear less susceptible to narrative congruence. One interpretation is that higher-socioeconomic-status audiences engage in more deliberate, analytic processing that reduces their susceptibility to narrative persuasion.

Collectively, the synergy appears stronger for ads delivered to female-skewed audiences, stable across ads with varying age compositions, and weaker for higher-SES (iPhone) users.

Table L. Sources of Heterogeneity in CATE on Purchase

	Doubly robust estimates of the CATE
Viewer demographics	
Age	.0036
	(.0022)
Gender	−.1970**
	(.0911)
iPhone	−.0073***
	(.0009)
Ad features	
Entropy	−.0184***
	(.0018)
Anger	.0799***
	(.0090)
Disgust	−.0069
	(.0071)
Fear	.0150
	(.0101)
Happiness	−.0265***
	(.0033)
Sadness	−.0320***
	(.0031)
Surprise	−.0557***
	(.0042)
Neutrality	−.0784***
	(.0069)
R-squared	.0021
Number of Observations	777,412

Notes: * $p < .10$, ** $p < .05$, *** $p < .01$. Robust standard errors are reported in parentheses. Coefficients for Age and Gender should be interpreted as associations between ad-level audience composition and the magnitude of the estimated CATE; they are subject to an ecological-fallacy caveat and should not be read as individual-level heterogeneity. The iPhone coefficient, measured at the individual level in the purchase data, reflects individual-level heterogeneity. The BLP specification is estimated on de-biased doubly robust scores and additionally controls for ad length, ad agency fixed effects, and day fixed effects; these controls were included in the original estimation and are disclosed here for transparency.

References

Argo JJ, Zhu R, Dahl DW (2008) Fact or fiction: An investigation of empathy and emotion in response to narrative advertisements. *J. Consum.Res.* 34(2):288–297.

Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN (2010) On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. Zaki MJ, Yu JX, Ravindran B, Pudi V, eds. *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, vol. 6118 (Springer, Berlin), 391–402.

Athey S, Wager S (2019) Estimating treatment effects with causal forests: An application. *Obs. Stud.* 5(2):37–51.

Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.

Bublitz MG, Escalas JE, Peracchio LA, Furchheim P, Grau SL, Hamby A, Scott A (2016) Transformative stories: A framework for crafting stories for social impact organizations. *J. Public Policy Mark.* 35(2):237–248.

Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive LDA model selection. *Neurocomputing* 72(7-9):1775--1781.

Chen Y, Lee JY, Sridhar S, Mittal V, McCallister K, Singal AG (2020) Improving cancer outreach effectiveness through targeting and economic assessments: Insights from a randomized field experiment. *J. Marketing* 84(3):1–27.

Choudhury P, Wang D, Carlson NA, Khanna T (2019) Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strateg. Manag. J.* 40(11):1705–1732.

Deighton J, Romer D, McQueen J (1989) Using drama to persuade. *J. Consum. Res.* 16(3):335–343.

Deveaud R, SanJuan E, Bellot P (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17(1):61–84.

Ding Y, Gordon BR, Knorr A (2021) Targeting with endogenous social networks: Theory and empirical evidence from mobile games. *J. Marketing Res.* 58(6):1079–1099.

Escalas JE (2004) Narrative processing: Building consumer connections to brands. *J. Consum. Psychol.* 14(1/2):168–180.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* 101(Supplement 1):5228–5235.

Hartmann J, Heitmann M, Schamp C, Netzer O (2021) The power of brand selfies. *J. Marketing Res.* 58(6):1159–1177.

Holzwarth M, Janiszewski C, Neumann MM (2006) The influence of avatars on online consumer shopping behavior. *J. Marketing* 70(4):19–36.

Imai K, Keele L, Yamamoto T (2010) Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* 25(1):51–71.

Kang SY, Kim J, Lakshmanan A (2025) Anatomical depiction: How showing a product's inner structure shapes product valuations. *J. Marketing* 89(1):56–76.

Kazemi V, Sullivan J (2014) One millisecond face alignment with an ensemble of regression trees. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* 1867–1874.

Laurence D (2018) Do ads that tell a story always perform better? The role of character identification and character type in storytelling ads. *Int. J. Res. Mark.* 35(2):289–304.

Lu S, Dai Y, Chen XY, Grewal R (2021) Do larger audiences generate greater revenues under pay what you want? Evidence from a live streaming platform. *Marketing Sci.* 40(5):964–984.

Niu Y, Wu J, Jiang S, Jiang Z (2025) The bullwhip effect in servitized manufacturers. *Manag. Sci.* 71(1):1–20.

Peracchio LA, Meyers-Levy J (1997) Evaluating persuasion-enhancing techniques from a resource-matching perspective. *J. Consum. Res.* 24(2):178–191.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66(5):688–701.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536.

Sagonas C, Tzimiropoulos G, Zafeiriou S, Pantic M (2013) 300 Faces in-the-wild challenge: The first facial landmark localization challenge. *Proc. IEEE International Conf. on Computer Vision Workshops* 397–403.

Semenova V, Chernozhukov V (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.* 24(2):264–289.

Stern BB (1992) Historical and personal perspectives on the role of emotions in advertising. *J. Consum.Res.* 19(3):234–247.

Sverdrup E, Petukhova M, Wager S (2025) Estimating treatment effect heterogeneity in psychiatry: A review and tutorial with causal forests. *Int. J. Methods Psychiatr. Res.* 34(2):e70015.

Van Laer T, De Ruyter K, Visconti LM, Wetzels M (2014) The extended transportation-imagery model: A meta-analysis of the antecedents and consequences of consumers' narrative transportation. *J. Consum.Res.* 40(5):797–817.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113(523):1228–1242.

Yu D, Deng L (2016) *Automatic Speech Recognition: A Deep Learning Approach*. Springer, London.

Zhao X, Lynch JG Jr, Chen Q (2010) Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *J. Consum. Res.* 37(2):197–206.