

Appendix

A. Summary Statistics

Table 1 Summary Statistics for 4,633 Ads Across 1,186 Quasi-Experiments

Variable	Mean	SD
Impressions	79,761	1,185,564
Clicks	545	8,979
CTR [%]	0.68	0.36
Ad runtime ^a [days]	8.65	10.62
	N	%
AI-generated image		
True	1,189	25.66
False	3,444	74.34
AI-generated caption		
True	337	7.27
False	4,296	92.73
Product category		
Technology and computing	1,054	22.75
Personal finance	742	16.02
Home and garden	641	13.84
Healthy living	339	7.32
Medical health	318	6.86
Other	1,539	33.22
Campaign objective		
Leads	2,424	52.32
Purchases	1,560	33.67
Page views	332	7.17
Other	317	6.84
Ad arbitrage advertiser ^b		
True	822	17.74
False	3,811	82.26

Notes. ^a indicates runtime prior to truncation, ^b calculated at the ad level.

B. Verbal Features

Table 2 Overview of Verbal Features

Control Variable	Description
LIWC-22^a	
<i>Caption</i>	
<i>Analytic</i>	Metric of logical, formal thinking
<i>Authentic</i>	Perceived honesty, genuineness
<i>BigWords</i>	Percentage of words seven letters or longer
<i>Clout</i>	Language of leadership, status
<i>Dic</i>	Percentage of words captured by LIWC
<i>Tone</i>	Degree of positive (negative) tone
<i>WC</i>	Total word count
<i>WPS</i>	Average words per sentence
<i>you</i>	Words referring to 2nd person (e.g., you, your, u, yourself)
<i>Description</i>	
<i>WC</i>	Total word count
Other	
<i>AI-generated caption</i>	Dummy variable indicating if ad caption is AI-generated

^a See Boyd et al. (2022).

C. Visual Features

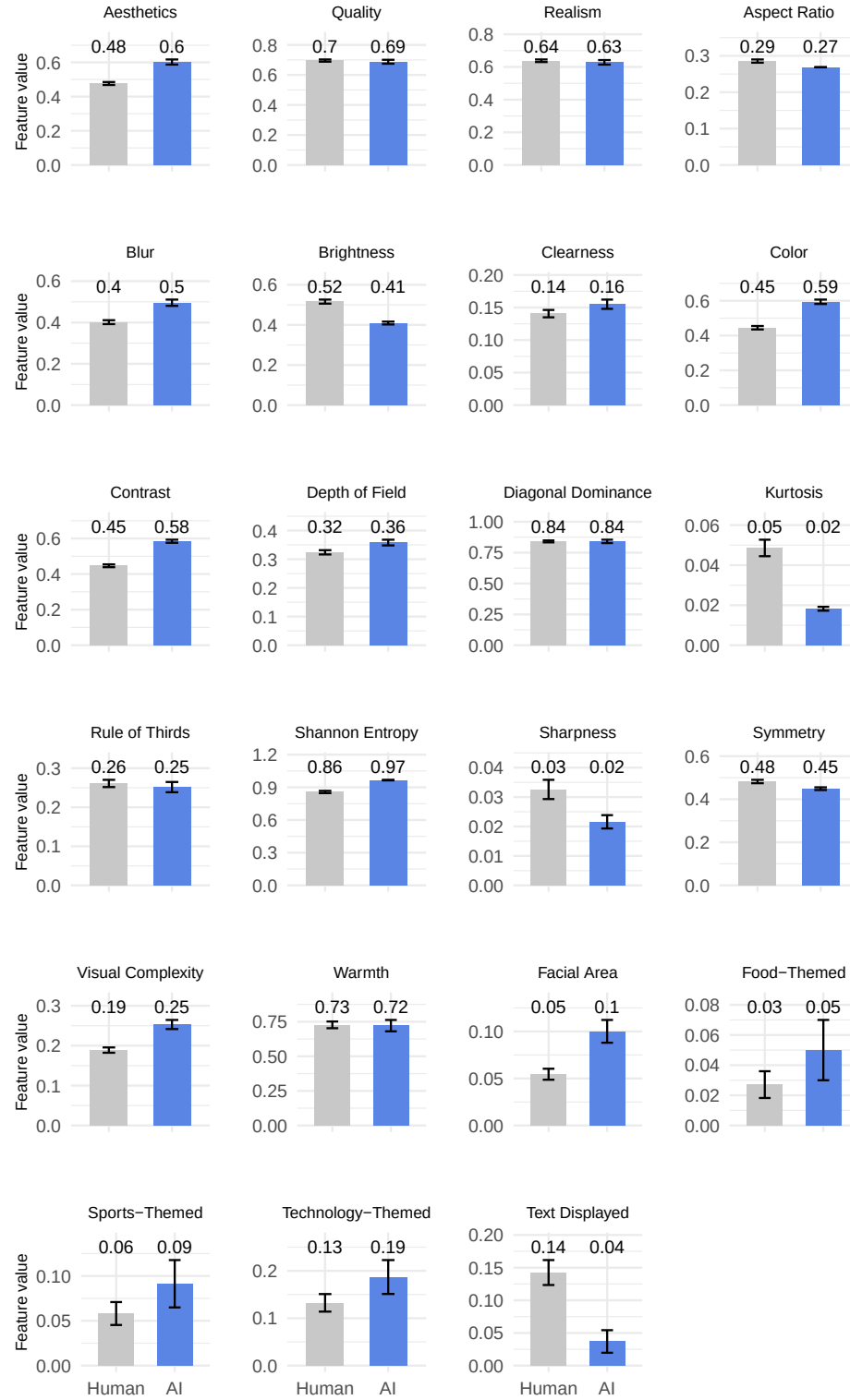
Table 3 Overview of Visual Features

Feature	Description	Source
Perceptual		
<i>Aesthetics</i>	Perceived image aesthetics (NIMA-predicted)	1
<i>Quality</i>	Perceived image quality (MTurk-rated)	2
<i>Realism</i>	Perceived image realism (MTurk-rated)	2
Structural		
<i>Aspect ratio</i>	Ratio of width to height of an image	3
<i>Blur</i>	Perceived blur annoyance	4
<i>Brightness</i>	Mean grayscale pixel intensity	5
<i>Clearness</i>	Mean horizontal and vertical Sobel operator gradient magnitude	1
<i>Color</i>	$\text{mean}(S_i) + \text{std}(S_i)$, where S_i is the S values in HSV color space for image i	5
<i>Contrast</i>	Standard deviation across all pixels in grayscale	5
<i>Depth of field</i>	Portion of an image that appears sharp in contrast to blurred areas in the background	6
<i>Diagonal dominance</i>	Shortest distance from saliency-based centroid to either diagonal of the image	6
<i>Kurtosis</i>	Kurtosis of the pixel-intensity distribution	5
<i>Rule of thirds</i>	Alignment of the main subject with the rule of thirds	6
<i>Shannon entropy</i>	Unpredictability of pixel intensities as proxies for image information amount	7
<i>Sharpness</i>	Variance of Laplacian operator (i.e., degree of change in image edges and details)	5
<i>Symmetry</i>	Normalized norm of difference between left and right image halves	5
<i>Visual complexity</i>	Count of distinct regions in an image	2
<i>Warmth</i>	Proportion of warm versus cold hues in HSV space	5
Content		
<i>Facial area</i>	Relative image area covered by faces' bounding boxes	2
<i>Food-themed</i>	Aggregates food-related COCO labels in a dummy variable	8
<i>Sports-themed</i>	Aggregates sport-related COCO labels in a dummy variable	8
<i>Technology-themed</i>	Aggregates tech-related COCO labels in a dummy variable	8
<i>Text displayed</i>	Indicates if text is displayed	8

Sources: 1 - Talebi and Milanfar (2018), 2 - Hartmann et al. (2025), Beichert et al. (2024), 3 - Dew et al. (2022), Raghubir and Greenleaf (2006), 4 - Crete et al. (2007), 5 - Zhang et al. (2022), 6 - Zhang and Luo (2023), 7 - Pieters et al. (2007), 8 - Lin et al. (2014). Web Appendix Figure 2 displays exemplary images for low and high feature values.

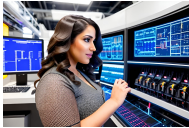
The variables *Food-themed*, *Sports-themed*, and *Technology-themed* are aggregated labels that are defined based on binary COCO labels (Lin et al. 2014) that are present in our data:

- *Food-themed* := *cake* OR *sandwich* OR *wine glass* OR *apple* OR *banana* OR *donut* OR *hot dog* OR *broccoli* OR *pizza* OR *orange* OR *carrot*
- *Sports-themed* := *bicycle* OR *frisbee* OR *kite* OR *skateboard* OR *skis* OR *snowboard* OR *sports ball* OR *surfboard* OR *tennis racket* OR *train* OR *truck*
- *Technology-themed* := *remote* OR *tv* OR *keyboard* OR *cell phone* OR *laptop*

Figure 1 Model-Free Means and Standard Errors for Visual Features of AI-Generated and Human-Made Ads

Note. For ease of comparability within the figure, the feature values in this plot are normalized ($x \in [0, 1]$).

Figure 2 Exemplary AI-Generated Images for Each Feature

Feature	Low	High	Feature	Low	High
Aesthetics			Rule of Thirds		
Quality			Shannon Entropy		
Realism			Sharpness		
Aspect ratio			Symmetry		
Blur			Visual Complexity		
Brightness			Warmth		
Clearness			Facial Area		
Color			Food-Themed ^a		
Contrast			Sports-Themed ^a		
Depth of Field			Technology-Themed ^a		
Diagonal dominance			Text Displayed ^a		
Kurtosis					

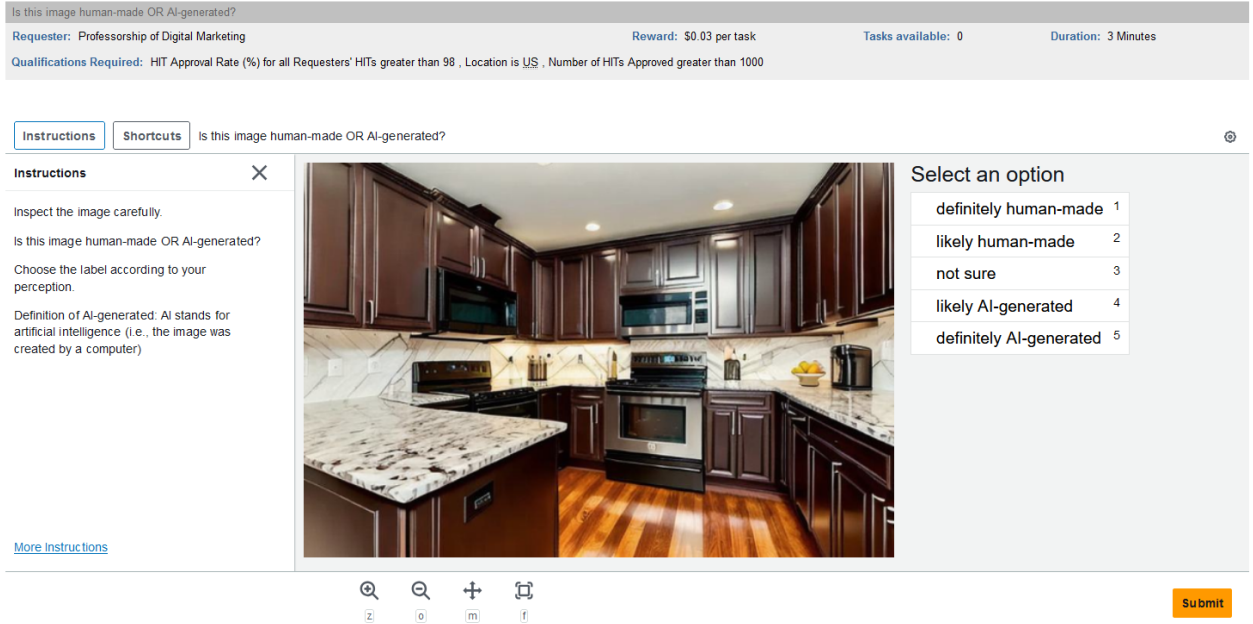
Note. The left image always shows a low feature value while the right image shows a high feature value (e.g., brightness for the left image is 30.37 and for the right image is 223.53 in mean grayscale pixel intensity). ^a indicates dummy variables, where low = 0 and high = 1.

D. MTurk Study on Perceived Artificiality

For all 1,751 images in our quasi-experiments (1,291 human-made images, 460 AI-generated images), we collect 5 ratings per image from 612 unique MTurkers in a within-subject design: “Is this image human-made OR AI-generated?” Each image is rated on a 5-point Likert scale anchored by ‘definitely human-made’ and ‘definitely AI-generated’ (adopted from Jakesch et al. 2023). This measure is intended to capture perceived artificiality based on visual cues rather than explicit beliefs about origin. On average, each MTurker rates 14.31 images and receives a reward of \$0.03 per rated image. Web Appendix Figure 3 shows the user interface as seen by the respondents. We use the median of the ratings per image to derive an image’s perceived artificiality (Web Appendix Table 10 displays consistent results when using the mean). Following Hartmann et al. (2023a), we normalize the looks-like-AI ratings $\in [0, 1]$. Our evaluators are selected based on the following criteria:

- Located in the United States
- > 1,000 tasks approved
- > 98% approval rate for all requesters’ tasks

Figure 3 Screenshot of the MTurk User Interface as Seen by the Respondents



E. Robustness Checks

E.1. Excluding Short-Running Ads

A concern one might have is that advertisers ‘played around’ with the novel GenAI Ad Maker and thereby launched AI-generated ads only for a short period of time to test their performance. To account for such behavior, we exclude short-running ads from our data and test various runtime cutoff dates between one and four days. Our results persist across these different runtime truncations, as demonstrated in Web Appendix Table 4.

Table 4 Robustness Analyses: Do Short-Running Ads Affect Our Findings?

Dependent variable:	CTR				
Model:	Reference (1)	RT >1d (2)	RT >2d (3)	RT >3d (4)	RT >4d (5)
<i>Variables</i>					
AI-generated image	0.0068 (0.0322)	-0.0031 (0.0318)	-0.0149 (0.0353)	-0.0061 (0.0357)	-0.0170 (0.0382)
looks-like-AI	-0.4593*** (0.1094)	-0.4571*** (0.1106)	-0.4154*** (0.1218)	-0.4290*** (0.1267)	-0.5629*** (0.1140)
AI-generated image × looks-like-AI	-0.2833** (0.0887)	-0.2911** (0.0888)	-0.2275** (0.0848)	-0.2277* (0.0889)	-0.2072* (0.0907)
<i>Controls</i>					
Caption	Yes	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>					
Quasi-Experiment	Yes	Yes	Yes	Yes	Yes
Calendar Date	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>					
Observations	29,592	28,788	26,894	26,260	24,336
Squared Correlation	0.19667	0.15282	0.28047	0.28782	0.20410
Pseudo R ²	0.98353	0.98411	0.98492	0.98511	0.98667
BIC	510,733.6	484,819.3	442,184.6	433,877.9	376,535.2

Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.

Notes. RT = runtime. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.2. Excluding Ads with Few Impressions

One potential concern is that, as generative AI was a relatively new technology, advertisers might have approached the GenAI Ad Maker with a ‘testing’ mindset. As a result, the ads created using this tool may not have been launched as seriously or strategically as the human-made ads, potentially limiting their comparability. To test this, we exclude ads from our quasi-experiments that have fewer than 1,000 impressions as well as ads that have fewer than 1,000 impressions OR fewer than 100 clicks (Hartmann et al. 2023b). Our findings are robust to these assessments, as detailed in Web Appendix Table 5, Models 2 and 3.

Table 5 Robustness Analyses: Do Ads with Very Few Impressions or Clicks Affect Our Findings?

Dependent Variable:	CTR		
	Reference	Excl. < 1,000 impressions	Excl. < 1,000 impressions or < 100 clicks
Model:	(1)	(2)	(3)
<i>Variables</i>			
AI-generated image	0.0068 (0.0322)	0.0077 (0.0313)	-0.0100 (0.0370)
looks-like-AI	-0.4593*** (0.1094)	-0.4599*** (0.1095)	-0.4792*** (0.1181)
AI-generated image × looks-like-AI	-0.2833** (0.0887)	-0.2708** (0.0885)	-0.3064** (0.1014)
<i>Controls</i>			
Caption	Yes	Yes	Yes
Description	Yes	Yes	Yes
Image	Yes	Yes	Yes
<i>Fixed effects</i>			
Quasi-Experiment	Yes	Yes	Yes
Calendar Date	Yes	Yes	Yes
<i>Fit statistics</i>			
Observations	29,592	26,733	16,110
Squared Correlation	0.19667	0.24436	0.38027
Pseudo R ²	0.98353	0.98412	0.98624
BIC	510,733.6	489,529.0	399,748.3

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.3. Excluding Ads with High CTRs

One might wonder if ads with higher-than-usual CTRs affect our findings. Such ads are generally observed for lower impression counts, which should be largely handled by the previous robustness check. Additionally, the binomial regression employed as part of our empirical strategy weights each ad-day observation's CTR by the number of impressions, giving low weight to ads with low impression counts. Nevertheless, we re-run our models excluding ads with CTRs more than one, two, or three standard deviations above the mean and find results consistent with those in the main manuscript (see Web Appendix Table 6).

Table 6 Robustness Analyses: Do Very High CTR Ads Affect Our Findings?

Dependent Variable:	CTR			
Model:	Reference (1)	CTR $< \mu + 1$ SD (2)	CTR $< \mu + 2$ SD (3)	CTR $< \mu + 3$ SD (4)
<i>Variables</i>				
AI-generated image	0.0068 (0.0322)	-0.0141 (0.0254)	-0.0004 (0.0262)	0.0081 (0.0290)
looks-like-AI	-0.4593*** (0.1094)	-0.5741*** (0.1147)	-0.5375*** (0.1131)	-0.5431*** (0.1100)
AI-generated image $\times$ looks-like-AI	-0.2833** (0.0887)	-0.3178** (0.1173)	-0.3087** (0.1192)	-0.2950** (0.1015)
<i>Controls</i>				
Caption	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes
<i>Fixed effects</i>				
Quasi-Experiment	Yes	Yes	Yes	Yes
Calendar Date	Yes	Yes	Yes	Yes
<i>Fit statistics</i>				
Observations	29,592	24,999	27,198	28,173
Squared Correlation	0.19667	0.29407	0.33847	0.41980
Pseudo R ²	0.98353	0.98691	0.98573	0.98513
BIC	510,733.6	377,659.2	416,025.2	447,676.0

Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.4. Focusing on Quasi-Experiments’ Early Horizons

One concern with the sibling-ad design is that the common comparison window may itself be selected by early ad performance. If advertisers stop ads based on early performance, then comparing sibling ads only up to the first stopping event could still reflect endogenous stopping rather than only differences in creative effectiveness. To empirically assess the potential impact of this concern, we examine ad performance over short horizons immediately after launch, before substantial stopping has occurred. The first two days of ad delivery should also help address concerns of divergent delivery as, to the extent that divergent delivery exists in our data, it should be minimal in the first two days of ad delivery before the observed CTRs for different ads in the campaign are learned.

Table 7 Robustness Analyses: Are Our Results Consistent Over the Early Horizons?

Dependent Variable:	CTR							
Model:	Reference (1)	First 1 day (2)	First 2 days (3)	First 3 days (4)	First 4 days (5)	First 5 days (6)	First 6 days (7)	First 7 days (8)
<i>Variables</i>								
AI-generated image	0.0068 (0.0322)	0.0941* (0.0392)	0.1396*** (0.0411)	0.0585 (0.0372)	0.0113 (0.0361)	0.0023 (0.0346)	-0.0016 (0.0335)	0.0088 (0.0319)
looks-like-AI	-0.4593*** (0.1094)	-0.1423 (0.1044)	-0.2413** (0.0919)	-0.1182 (0.1082)	-0.1380 (0.1039)	-0.1747 (0.1010)	-0.1844 (0.1037)	-0.2164* (0.1011)
AI-generated image $\times$ looks-like-AI	-0.2833** (0.0887)	-0.1124 (0.0912)	-0.4238** (0.1330)	-0.4556*** (0.1250)	-0.3991*** (0.1134)	-0.3920*** (0.0976)	-0.3788*** (0.0922)	-0.4030*** (0.0839)
Share of Active Quasi-Experiments		100%	81.3%	68.2%	59.2%	51.5%	46.9%	41.9%
<i>Controls</i>								
Caption	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>								
Quasi-Experiment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Calendar Date	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>								
Observations	29,592	4,507	7,574	9,927	11,981	13,710	15,198	16,426
Squared Correlation	0.19667	0.23838	0.24470	0.26399	0.27078	0.26803	0.26369	0.26102
Pseudo R ²	0.98353	0.97692	0.97599	0.97572	0.97708	0.97759	0.97799	0.97879
BIC	510,733.6	60,204.4	121,893.6	171,978.0	204,021.4	233,373.5	260,245.3	279,459.5

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses). A quasi-experiment j is counted as still active at day t if no ad in it has stopped before day t .

Specifically, we re-estimate our CTR model using only the first few days after ad launch. We start with the launch day only, then incrementally expand the window to the first two days, first three days, and so forth through the first seven days. We present the regression results in Web Appendix Table 7 and our coefficients of interest are the interaction terms. Across the first-two-day through first-seven-day windows, the AI-generated image $\times$ looks-like-AI interaction remains negative and statistically significant. The estimates are directionally consistent (somewhat larger in magnitude) with the reference regression using the full data, so the looks-like-AI moderation central to our finding is already present before substantial stopping occurs.

We also report in the same table the share of sibling sets that remain active at each horizon: all quasi-experiments are observable on the launch day, 81.3% remain observable through day 2, 68.2% through day 3, and 41.9% through day 7. Thus, the key pattern is already present in the early windows, when most sibling sets are still observable and before longer-run stopping can fully unfold. Of course, it does not prove that the sibling-ad design fully eliminates concerns about endogenous stopping, but this early-horizon analysis provides a useful bounding exercise and strengthens our confidence, as it does not support that stopping is the primary driver of our main results.

E.5. Accounting for Potential Delivery Imbalances

To account for potential daily imbalances in ad delivery within a quasi-experiment, we control for the quasi-experiment's Gini coefficient. The Gini coefficient is a common measure of inequality (i.e., imbalance) that quantifies the distribution of income (i.e., ad impressions) within a population (i.e., a quasi-experiment). It is defined on $[0, 1]$ where a value of 0 represents perfect balance (i.e., all ads in a quasi-experiment receive equally many impressions) and 1 represents perfect imbalance (i.e., one ad in a quasi-experiment receives all impressions and all others receive none). We test our results for robustness by controlling for the mean-centered Gini coefficient and its interaction with AI-generated image. Web Appendix Table 8 Model 2 shows that our findings are robust when controlling for delivery imbalances within quasi-experiments. Additionally, Web Appendix Table 8 Model 3 shows no evidence that AI-generated images differentially benefit from any imbalances in ad delivery measured by impressions versus human-made images.

Table 8 Robustness Analyses: Are Our Results Consistent When Accounting for Imbalances in Daily Impressions Within a Quasi-Experiment?

Dependent Variable:	CTR		
Model:	Reference (1)	Gini (2)	AI-generated image $\times$ Gini (3)
<i>Variables</i>			
AI-generated image	0.0068 (0.0322)	0.0068 (0.0322)	0.0032 (0.0318)
looks-like-AI	-0.4593*** (0.1094)	-0.4594*** (0.1092)	-0.4366*** (0.1053)
AI-generated image $\times$ looks-like-AI	-0.2833** (0.0887)	-0.2831** (0.0886)	-0.2699** (0.0844)
Gini coefficient (mean-centered)		-0.0299 (0.1562)	-0.1814 (0.2046)
AI-generated image $\times$ Gini coefficient (mean-centered)			-0.2720 (0.1665)
<i>Controls</i>			
Caption	Yes	Yes	Yes
Description	Yes	Yes	Yes
Image	Yes	Yes	Yes
<i>Fixed effects</i>			
Calendar Date	Yes	Yes	Yes
Quasi-Experiment	Yes	Yes	Yes
<i>Fit statistics</i>			
Observations	29,592	29,592	29,592
Squared Correlation	0.19667	0.19658	0.19760
Pseudo R^2	0.98353	0.98353	0.98357
BIC	510,733.6	510,735.7	509,555.6

*Signif. Codes: ***, 0.001, **, 0.01, *, 0.05.*

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Gini coefficient is defined $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.6. Excluding Pre-Go-Live Ads and Beta-Phase Participants

We investigate potential effects driven by the GenAI Ad Maker’s beta phase when selected advertisers were invited to use the tool. We exclude all ad observations prior to the GenAI Ad Maker’s go-live. In a separate assessment, we exclude all quasi-experiments by advertisers who participated in the beta phase. In both cases, our results remain consistent, as shown in Web Appendix Table 9, Models 2 to 4.

Table 9 Robustness Analyses: Do Pre-Go-Live Ads or Ads from Beta-Phase Advertisers Affect Our Findings?

Dependent Variable:	CTR			
	Reference	Excl. beta-phase advertisers	Excl. all pre-go-live ad observations	Excl. beta-phase advertisers or pre-go-live ad observations
Model:	(1)	(2)	(3)	(4)
<i>Variables</i>				
AI-generated image	0.0068 (0.0322)	0.0083 (0.0357)	-0.0111 (0.0345)	-0.0131 (0.0366)
looks-like-AI	-0.4593*** (0.1094)	-0.4442*** (0.1101)	-0.2829** (0.1089)	-0.2753* (0.1099)
AI-generated image $\times$ looks-like-AI	-0.2833** (0.0887)	-0.2625** (0.0949)	-0.3888*** (0.1012)	-0.3916*** (0.1053)
<i>Controls</i>				
Caption	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes
<i>Fixed effects</i>				
Quasi-Experiment	Yes	Yes	Yes	Yes
Calendar Date	Yes	Yes	Yes	Yes
<i>Fit statistics</i>				
Observations	29,592	27,482	19,363	18,401
Squared Correlation	0.19667	0.19465	0.18173	0.18112
Pseudo R ²	0.98353	0.98374	0.98167	0.98135
BIC	510,733.6	490,464.2	372,672.7	364,763.3

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0,1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.7. Using Mean-Based Instead of Median-Based Perceived Artificiality

To account for potential outliers in our MTurk ratings, we use the median of the five ratings received to determine a median looks-like-AI score per image. In this analysis, we check the robustness of our results to using the mean instead of the median. As detailed in Web Appendix Table 10 Model 2, our results remain consistent when using mean looks-like-AI.

Table 10 Robustness Analyses: Are Our Results Consistent When Using Mean Instead of Median Perceived Artificiality Scores?

Dependent Variable:	CTR	
Model:	Reference (1)	looks-like-AI (mean) (2)
<i>Variables</i>		
AI-generated image	0.0068 (0.0322)	0.0080 (0.0288)
looks-like-AI (default; median)	-0.4593*** (0.1094)	
AI-generated image $\times$ looks-like-AI (default; median)	-0.2833** (0.0887)	
looks-like-AI (mean)		-0.6860*** (0.1212)
AI-generated image $\times$ looks-like-AI (mean)		-0.3898** (0.1349)
<i>Controls</i>		
Caption	Yes	Yes
Description	Yes	Yes
Image	Yes	Yes
<i>Fixed effects</i>		
Quasi-Experiment	Yes	Yes
Calendar Date	Yes	Yes
<i>Fit statistics</i>		
Observations	29,592	29,592
Squared Correlation	0.19667	0.19620
Pseudo R ²	0.98353	0.98354
BIC	510,733.6	510,308.8

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.8. Excluding Human-Made Images That Are Detected as AI-Generated

Our data include a dummy variable for whether the image was generated by the GenAI Ad Maker. However, we cannot observe if advertisers generated images with third-party models and uploaded them manually. Our data would label such images as human-made. We argue that such spillover effects in the control group are a) unlikely due to the limited adoption of image generation models during the observation period and b) would, if present, make our control group (human-made images) more similar to our treatment group (AI-generated images), resulting in more conservative estimates. Irrespective of this, we would like to assess such advertiser behavior in this robustness check.

We employ SPAI (Karageorgiou et al. 2025), a state-of-the-art detector for AI-generated images based on masked spectral learning, which is also robust to (willful) image perturbations. For each image, SPAI gives an AI detection score $x \in [0, 1]$ where higher values of x indicate a higher likelihood of being AI-generated. We test if our results are robust to excluding images that feature elevated SPAI scores while being labeled as human-made. If an image is labeled as human-made and its SPAI score exceeds the threshold, its associated ads are removed from the respective analysis. The SPAI metrics code uses a default threshold of 0.5. Our findings are robust to these assessments, as detailed in Web Appendix Table 11, Models 2 through 4.

Table 11 Robustness Analyses: Are Our Results Consistent When Excluding Manually Uploaded but AI-Generated Imagery?

Dependent Variable:	CTR			
Model:	Reference (1)	SPAI threshold: 0.75 (2)	SPAI threshold: 0.5 (3)	SPAI threshold: 0.25 (4)
<i>Variables</i>				
AI-generated image	0.0068 (0.0322)	0.0704 (0.0414)	0.0527 (0.0461)	0.0510 (0.0477)
looks-like-AI	-0.4593*** (0.1094)	-0.4137** (0.1280)	-0.4362*** (0.1322)	-0.4467*** (0.1326)
AI-generated image $\times$ looks-like-AI	-0.2833** (0.0887)	-0.2608** (0.1006)	-0.2811** (0.1046)	-0.2808** (0.1065)
<i>Controls</i>				
Caption	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes
<i>Fixed effects</i>				
Calendar Date	Yes	Yes	Yes	Yes
Quasi-Experiment	Yes	Yes	Yes	Yes
<i>Fit statistics</i>				
Observations	29,592	20,307	18,969	18,036
Squared Correlation	0.19667	0.20420	0.20672	0.20601
Pseudo R ²	0.98353	0.98302	0.98338	0.98357
BIC	510,733.6	346,578.9	332,488.3	323,375.1

Signif. Codes: ***, 0.001, **, 0.01, *, 0.05.

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses).

E.9. Assessing Self-Selection in Human-Made Ads

It is possible that human-made ads that appear in a quasi-experiment are different from ads commonly used by the advertiser. On the one hand, human-made ads in quasi-experiments may be worse than other human-made ads because the advertiser may employ the GenAI Ad Maker in campaigns in which they struggle to generate human-made ads. On the other hand, it is possible that human-made ads in quasi-experiments are better than other human-made ads because of spillover or inspiration from the images suggested by AI to images uploaded by the advertisers. To test for such possible self-selection of human-made ads in our quasi-experiments, we assess whether the performance of human-made ads that are present in our quasi-experimental data differs from that of those we exclude. Specifically, we compare the CTR generated by human-made ads inside versus outside the quasi-experimental setting for the same advertiser using a paired t-test. We find no significant difference in CTRs between human-made ads that are part of the quasi-experiments and those that are not. See Web Appendix Table 12 for a more comprehensive analysis under full covariate controls.

Table 12 Robustness Analyses: Are Human-Made Ads Within the Quasi-Experimental Setting Different in Performance from Those That Were Not Included in Quasi-Experiments?

Dependent Variable:	CTR
Model:	(1)
<i>Variables</i>	
Ad in quasi-experiment (binary)	0.1590 (0.1631)
<i>Controls</i>	
Caption	Yes
Description	Yes
Image	Yes
<i>Fixed effects</i>	
Advertiser	Yes
Calendar Date	Yes
<i>Fit statistics</i>	
Observations	117,952
Squared Correlation	0.12985
Pseudo R ²	0.95889
BIC	7,444,992.4

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. Clustered standard errors at the advertiser level (shown in parentheses).

E.10. Alternative Sample Restrictions

Recall that we exclude some ads from the full data to construct the quasi-experiment dataset. One might wonder if our results are robust if we run the analysis using less strict exclusion criteria. To test this, we repeat our analyses using data without right truncation (to account for possible survival bias, we truncate all observations within a quasi-experiment once the first ad stops in the main analysis), data consisting of all campaigns with both AI- and human-generated ads, and the full dataset including 2,227,664 observations relative to the 29,631 observations in our most restrictive quasi-experimental data. Our findings remain robust across different specifications, as detailed in Web Appendix Table 13, Models 2 through 4.

Table 13 Robustness Analyses: Are the Results Sensitive to Alternative Sample Restrictions?

Dependent Variable:	CTR			
	Reference	Untruncated ads	Campaigns with AI & human ads	All observations
Model:	(1)	(2)	(3)	(4)
<i>Variables</i>				
AI-generated image	0.0068 (0.0322)	0.0605 (0.0372)	0.0665** (0.0238)	0.1250** (0.0408)
looks-like-AI	-0.4593*** (0.1094)	-0.2810* (0.1267)		
AI-generated image $\times$ looks-like-AI	-0.2833** (0.0887)	-0.2165* (0.0990)		
predicted looks-like-AI			-1.303*** (0.3330)	-1.059*** (0.2209)
AI-generated image $\times$ predicted looks-like-AI			-0.5307* (0.2314)	-1.151*** (0.2883)
<i>Controls</i>				
Caption	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes
<i>Fixed effects</i>				
Calendar Date	Yes	Yes	Yes	Yes
Quasi-Experiment	Yes	Yes	No	No
Campaign	No	No	Yes	No
Advertiser	No	No	No	Yes
<i>Fit statistics</i>				
Observations	29,592	40,024	40,044	2,218,322
Squared Correlation	0.19667	0.11862	0.06921	0.03938
Pseudo R ²	0.98353	0.98231	0.98006	0.95636
BIC	510,733.6	742,813.2	827,003.9	60,693,510.8

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. AI-generated image is effect-coded. Looks-like-AI is scaled $\in [0, 1]$ and mean-centered. Clustered standard errors at the quasi-experiment level (shown in parentheses) for Columns (1) and (2). Clustered standard errors at the campaign and advertiser level (shown in parentheses) for Columns (3) and (4), respectively. For each regression, the reported number of observations may be smaller than the input sample because observations are dropped when the dependent variable lacks within-group variation. Due to the size of our dataset, we were required to predict looks-like-AI for the remaining $N = 49,141$ images in our full data that are not part of the quasi-experimental setup for which we collected the $N = 1,751$ looks-like-AI ratings. The XGBRegressor model predicts looks-like-AI $\in [0, 1]$ with a hold-out MSE of 0.0662 and correlation to actual values of $\rho = 0.6843$, $p < 0.001$.

F. Disguising AI: Visual Features and Consumer Perception

Table 14 Visual Features Associated with Perceived Artificiality Versus AI-Generated Images

Dependent Variables: Model:	looks-like-AI (1)		AI-generated image (2)	
	Coef.	S.E.	Coef.	S.E.
<i>Variables</i>				
Perceptual				
Aesthetics	0.0247**	(0.0082)	0.1244***	(0.0095)
Quality	0.0019	(0.0075)	-0.0147	(0.0087)
Realism	-0.0208**	(0.0069)	-0.0030	(0.0082)
Structural				
Blur	0.0890	(0.0600)	-0.2312**	(0.0767)
Brightness	0.0112	(0.0093)	-0.0101	(0.0108)
Clearness	-0.0290*	(0.0141)	0.0711***	(0.0200)
Color	0.0357***	(0.0092)	0.0767***	(0.0118)
Contrast	0.0006	(0.0090)	0.1387***	(0.0111)
Kurtosis	0.0137	(0.0095)	0.0728***	(0.0145)
Sharpness	-0.0531	(0.0596)	0.1382	(0.0779)
Symmetry	0.0082	(0.0082)	0.0278**	(0.0103)
Warmth	-0.0188*	(0.0075)	-0.0229*	(0.0090)
Aspect ratio	-0.0056	(0.0074)	-0.0155	(0.0097)
Depth of field	0.0125	(0.0071)	0.0657***	(0.0097)
Diagonal dominance	0.0075	(0.0072)	-0.0201*	(0.0095)
Rule of thirds	-0.0052	(0.0072)	-0.0117	(0.0089)
Shannon entropy	-0.0146	(0.0122)	0.0746***	(0.0135)
Visual complexity	0.0116	(0.0086)	0.0280*	(0.0118)
Content				
Facial area	-0.0202*	(0.0083)	0.0790***	(0.0101)
Food-themed	0.0005	(0.0334)	0.1139*	(0.0500)
Sports-themed	-0.0017	(0.0261)	0.0714*	(0.0353)
Technology-themed	-0.0434*	(0.0216)	0.0120	(0.0259)
Text displayed	0.0257	(0.0240)	-0.0621*	(0.0272)
<i>Controls</i>				
AI-generated image	0.0925***	(0.0214)		
<i>Fixed effects</i>				
Advertiser	Yes		Yes	
<i>Fit statistics</i>				
Observations	1,751		1,751	
R ²	0.25767		0.57238	
Within R ²	0.09632		0.41721	

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

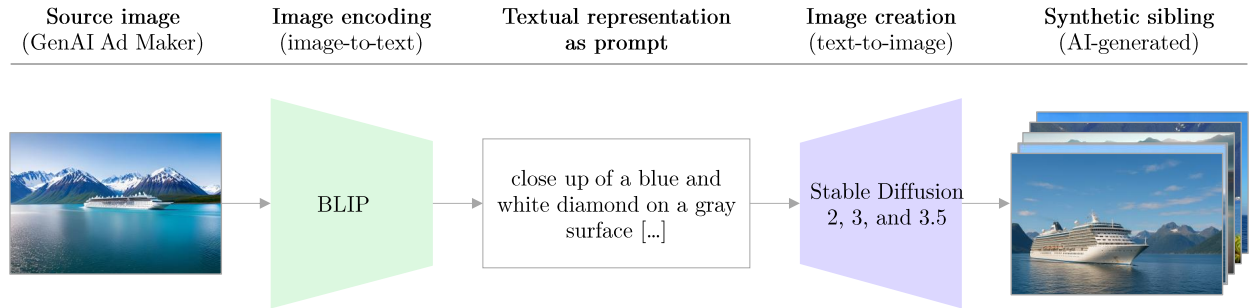
Notes. Robust standard errors in parentheses. Continuous independent controls are standardized (z-score). We obtain directionally consistent results when estimating a logistic regression for Model 2.

G. Regeneration of AI-Generated Images

To assess whether newer generative text-to-image models are able to close the gap in perceived artificiality to human-made images, we draw a random sample of 48 AI-generated images from our quasi-experiments. Our sample is stratified by advertiser industry. See Figure 5 subplot (a) in the main paper for looks-like-AI scores of the drawn sample.

We regenerate each of the 48 sampled images with later versions of Stable Diffusion, including Stable Diffusion 3.5, which was released in October 2024 and OpenAI’s GPT-4o’s image generation feature released in March 2025. For image regeneration, we follow the approach outlined by Hartmann et al. (2025), who first derive a textual description of a source image through an image-to-text model, that is, BLIP (for details see <https://huggingface.co/Salesforce/blip-image-captioning-large>, accessed November 11, 2024). In a second step, this textual description serves as a prompt for each generative text-to-image model, resulting in a regenerated image that closely resembles the source image. Web Appendix Figure 4 visualizes this process. For each of these regenerated images, we collect looks-like-AI ratings following the approach described in Web Appendix D. For Figure 5 in the main paper, we use the mean-based looks-like-AI rating for each image; we obtain directionally consistent results using median-based looks-like-AI.

Figure 4 Image-Regeneration Pipeline for Newer AI Models

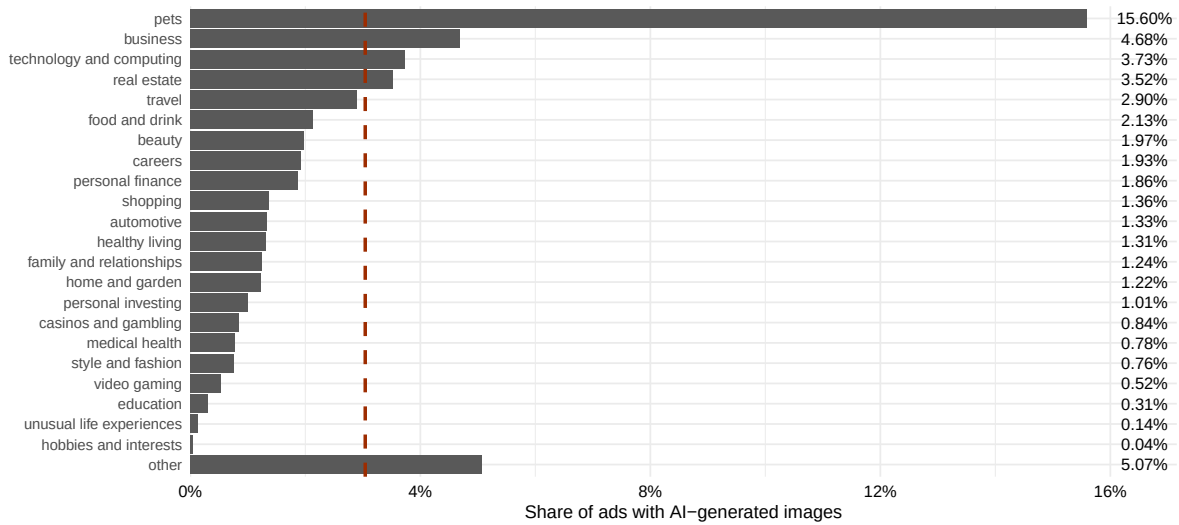


H. How Advertisers Use the GenAI Ad Maker

H.1. Adoption of the GenAI Ad Maker by Advertiser Industries

In our data, we observe which advertisers publish AI-generated ads and for what campaigns.¹ 6.25% of the advertisers used the GenAI Ad Maker at least once by the end of our observation period. Across all ads, the advertisers used the AI tool to generate 3.04% of the images after the GenAI Ad Maker went live. We observe heterogeneity in the adoption of the GenAI Ad Maker across advertiser industries. Figure 5 presents the tool adoption rates for AI-generated marketing imagery within the 34 days following the tool’s go-live. Advertisers in the categories *pets*, *business*, *technology and computing*, and *real estate* adopt the GenAI Ad Maker more frequently to generate ad visuals compared to the remaining industries.

Figure 5 Adoption of the GenAI Ad Maker’s Image Generation Feature Across Advertiser Industries



Note. The vertical dashed line at 3.04% indicates the mean adoption rate across all industries post go-live. Industries with fewer than 1,000 ads are clustered as *other*. Non-adopting industries are omitted.

H.2. Adoption of the GenAI Ad Maker by Advertiser and Campaign Characteristics

Looking beyond the industry verticals, to further understand which advertisers adopt the GenAI Ad Maker, we regress different adoption metrics on advertiser-level characteristics calculated from ads launched before the GenAI Ad Maker’s go-live, including the average number of ads per campaign, CTR, impressions, and spend in USD. Additionally, we include in the regression the campaign objective, industry, and whether an advertiser is an arbitrage advertiser². Table 15 Models 1 through 3 present the results of regression models examining which advertiser characteristics are associated with the adoption of GenAI Ad Maker. We use three different dependent variables to capture adoption: (1) whether an advertiser adopts the tool at least

¹ Our dataset does not include the clickstream of the advertiser interaction with the tool. We only observe full adoption of the GenAI Ad Maker’s generated output. This means we only observe if advertisers used AI-generated content as proposed by the GenAI Ad Maker without further modifications.

² Arbitrage advertisers operate on a volume-based business model where they use ads to drive traffic to content pages and then monetize that traffic by displaying (more profitable) ads to these visitors.

once, (2) the within-advertiser share of ads that are AI-generated, and (3) the within-advertiser share of campaigns that include at least one AI-generated ad.

Table 15 Relationship Between Advertisers' and Campaigns' Characteristics and GenAI Ad Maker Adoption

Dependent Variables:	Advertiser adopts	Share of AI ads	Share of campaigns with AI Ads
Model:	(1)	(2)	(3)
<i>Variables</i>			
Constant	-19.34*** (1.115)	-19.99*** (1.073)	-19.19*** (1.127)
Pre-go-live characteristics			
Ad count per campaign (log)	-0.0369 (0.1576)	0.0661 (0.2057)	0.1047 (0.1898)
Prior CTR	-1.718 (6.371)	-9.691 (10.19)	-15.79 (12.88)
Impressions (log)	0.1861 (0.1109)	0.0813 (0.1181)	0.0072 (0.1174)
Spend (log)	-0.1912 (0.1024)	-0.1407 (0.1137)	-0.0767 (0.1108)
Post-go-live characteristics (observation-specific)			
Campaign objective: Brand awareness	0.9202 (0.8407)	1.587 (0.8711)	1.434 (0.7952)
Campaign objective: Mobile app installs	1.320 (0.7873)	1.501* (0.7336)	1.401* (0.6932)
Campaign objective: Page views	0.4980 (0.3267)	0.8219 (0.4229)	0.6635 (0.3673)
Campaign objective: Purchases	0.7607** (0.2890)	0.9084* (0.3698)	0.7762* (0.3637)
<i>Controls</i>			
Arbitrage Advertiser	Yes	Yes	Yes
Industry	Yes	Yes	Yes
<i>Fit statistics</i>			
Observations	1,571	1,571	1,571
Squared Correlation	0.09988	0.16518	0.14076
Pseudo R ²	0.12883	0.33214	0.31392
BIC	1,230.7	782.64	864.77

*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.*

Notes. Robust standard errors are shown in parentheses. We exclude all advertisers that a) participated in the GenAI Ad Maker's beta phase ($N = 36$), b) have missing industry information ($N = 79$), and were inactive either c) prior ($N = 1,850$) or d) after the GenAI Ad Maker's go-live ($N = 3,538$). We obtain directionally consistent results when estimating OLS regressions for all three models.

We do not find a statistically significant relationship between advertisers' pre-go-live characteristics and their adoption of the GenAI Ad Maker's image generation feature (see Table 15 Models 1 through 3). Specifically, it does not seem that more active advertisers (higher spending) or more successful advertisers (higher CTR) are associated with higher AI adoption. Turning to campaign-level characteristics, we find that advertisers are more likely to adopt GenAI for conversion-oriented campaigns. Specifically, advertisers with a higher share of campaigns aimed at purchases or mobile app installs are significantly more likely to adopt the GenAI Ad Maker's image generation feature.

From a managerial perspective, the results imply that advertising platforms like Taboola can accelerate adoption by targeting AI tool promotion and support toward performance-oriented advertisers, who appear most receptive to experimenting with new generative AI technologies. These adoption patterns also highlight the importance of accounting for campaign objectives in our quasi-experimental design, since advertisers with different goals may systematically differ in their likelihood of adopting AI and possibly in how their audiences respond to AI-generated ads.

H.3. First-Stop Diagnostics for AI-Generated Versus Human-Made Ads

Our sibling-ad design compares ads that start on the same day and truncates the comparison window when the first ad in the quasi-experiment stops running. This design mitigates concerns about comparing ads over mechanically different exposure windows. However, the common endpoint may itself be selected by early ad performance. We therefore examine whether the ad that determines the end of the common comparison window is disproportionately likely to be AI-generated or human-made, conditional on prior relative performance.

H.3.1. Relative Performance Measurement Construction Let j index a quasi-experiment, i index an ad within j , and t index calendar date. Let τ_j denote the first calendar day on which any ad in quasi-experiment j stops running. For each ad-day, we construct cumulative clicks C_{ijt} and cumulative impressions N_{ijt} up to, but not including, date t . When $N_{ijt} > 0$, the raw lagged cumulative CTR is defined as $CTR_{ijt} = C_{ijt}/N_{ijt}$. Let $\overline{CTR}_{-ijt}$ denote the average cumulative CTR of observed sibling ads in the same quasi-experiment-date cell, excluding the focal ad. We use four measures of lagged relative performance³:

Normalized Relative CTR Difference. The normalized difference between the focal ad’s cumulative CTR and the sibling average is:

$$RelPerf_{ijt} = \frac{CTR_{ijt} - \overline{CTR}_{-ijt}}{CTR_{ijt}}.$$

Log Smoothed Relative CTR. We first define smoothed cumulative CTR as $CTR_{ijt}^s = (C_{ijt} + sp_0)/(N_{ijt} + s)$, where $p_0 = 0.5\%$ is the baseline CTR and $s = 200$ is the number of pseudo-impressions. Let $\overline{CTR}_{-ijt}^s$ denote the average smoothed cumulative CTR of observed sibling ads. The smoothed log-ratio measure is

$$RelPerf_{ijt} = \log \left(\frac{CTR_{ijt}^s}{\overline{CTR}_{-ijt}^s} \right).$$

Below-Sibling-Average Indicator. We also use a simple binary indicator for whether the focal-ad CTR is lower than the average of other ads in the same quasi-experiment or not:

$$RelPerf_{ijt} = \mathbf{1}\{CTR_{ijt} < \overline{CTR}_{-ijt}\}.$$

Within-Quasi-Experiment-Date Ranks. Finally, we use the focal ad’s within-quasi-experiment-date raw cumulative CTR rank (lower rank indicates higher CTR relative to siblings):

$$RelPerf_{ijt} = Rank\{CTR_{ijt}\}.$$

Table 16 First-Stop Diagnostics for AI-Generated Versus Human-Made Ads

Dependent Variable:	FirstStop			
Model:	Normalized CTR diff. (1)	Smoothed log ratio (2)	Below sibling avg. (3)	CTR rank (4)
<i>Variables</i>				
Relative performance	0.0002 (0.0017)	-0.0425* (0.0166)	0.0378* (0.0188)	0.0611* (0.0245)
AI-generated image	0.0156 (0.0172)	0.0158 (0.0138)	0.0183 (0.0198)	0.0201 (0.0214)
Relative performance $\times$ AI-generated image	-0.0004 (0.0017)	0.0176 (0.0149)	-0.0050 (0.0183)	-0.0090 (0.0252)
<i>Controls</i>				
Caption	Yes	Yes	Yes	Yes
Description	Yes	Yes	Yes	Yes
Image	Yes	Yes	Yes	Yes
<i>Fixed effects</i>				
Quasi-Experiment	Yes	Yes	Yes	Yes
<i>Fit statistics</i>				
Observations	2,732	4,040	2,837	2,837
R ²	0.49624	0.52527	0.48285	0.48370
Within R ²	0.02283	0.01872	0.02172	0.02333

Signif. Codes: ***: 0.001, **: 0.01, *: 0.05.

Notes. Clustered standard errors at the quasi-experiment level are shown in parentheses. AI-generated image is effect-coded. Column (1) measures normalized relative performance as the focal ad’s raw cumulative CTR minus the average raw cumulative CTR of its observed sibling ads, divided by the focal ad’s raw cumulative CTR. Column (2) uses the log of the focal ad’s smoothed cumulative CTR divided by the average smoothed cumulative CTR of its observed siblings. Column (3) indicates whether the focal ad’s raw cumulative CTR is below the observed sibling average. Column (4) uses the focal ad’s raw cumulative CTR rank within the same quasi-experiment-date cell. Larger values in Columns (1) and (2) indicate stronger focal-ad performance relative to siblings, while larger values in Columns (3) and (4) indicate weaker focal-ad performance. The unit of observation is an ad (evaluated on its quasi-experiment’s first-stop day), not an ad-day as in the main analysis. All columns start from the same set of sibling ads and the same first-stop timing used in the main analysis; the differences in sample size across columns reflect only what each relative performance measure requires to be well-defined. Forming a sibling comparison requires at least two active ads in a quasi-experiment-date cell, leaving 4,040 ads in Column (2). The raw-cumulative-CTR-based measures further require prior impressions, which exclude 1,203 ads stopped on their launch day, leaving 2,837 ads in Columns (3) and (4). The normalized CTR-difference measure is further undefined when the focal ad’s raw cumulative CTR equals zero, leaving 2,732 ads in Column (1). The smoothed measure (Column 2) is defined even without prior impressions because of its shrinkage term, so it retains the most observations.

H.3.2. Model and Results We estimate a first-stop diagnostic model at the ad level. The unit of observation is an ad within a quasi-experiment rather than an ad-day as in the main analysis, because first stopping is a one-time event defined at the ad level. We focus on observations on day τ_j , the first day when any ad in the quasi-experiment stops. The dependent variable $FirstStop_{ij}$ equals 1 if ad i is stopped on that first stopping day and 0 otherwise. If multiple sibling ads stop on the first stopping day, all such ads are coded as first-stopped.⁴ Specifically, we estimate the following linear probability model:

$$FirstStop_{ij} = \beta_0 + \beta_1 \phi_{ij} + \beta_2 RelPerf_{ij\tau_j} + \beta_3 \phi_{ij} \times RelPerf_{ij\tau_j} + \gamma X_{ij\tau_j} + \alpha_j + \epsilon_{ij},$$

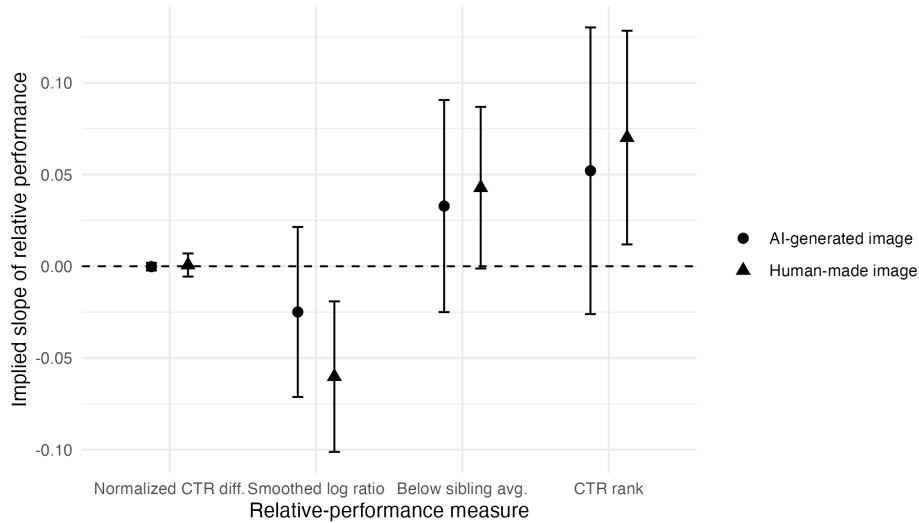
³ We only use relative CTR difference within a quasi-experiment rather than raw CTR difference because baseline CTR levels vary substantially across quasi-experiments, making the same absolute percentage-point difference difficult to compare across sibling sets.

⁴ As a robustness check, we also re-estimate the model after excluding quasi-experiments in which multiple ads stop on the first stopping day. This stricter restriction drops 38.9% of quasi-experiments, but the results remain consistent with those reported below.

where ϕ_{ij} is the effect-coded indicator for whether ad i in quasi-experiment j uses an AI-generated image ($\phi_{ij} = +1$ for AI-generated and $\phi_{ij} = -1$ for human-made content). $RelPerf_{ij\tau_j}$ is measured using cumulative performance before the first stopping day based on the four specifications described above. $X_{ij\tau_j}$ includes the same caption, description, and image controls as in the main analysis, including an indicator for whether the caption was AI-generated with the GenAI Ad Maker, verbal features of the caption and description, and visual features of the image. Web Appendices B and C list the full set of verbal and visual controls. We include quasi-experiment fixed effects α_j , and standard errors are clustered at the quasi-experiment level.

Table 16 reports the estimation results. In Columns (1) and (2), larger values of relative performance indicate higher focal-ad CTR relative to observed siblings; in Columns (3) and (4), larger values indicate weaker relative performance. For relative performance itself, the coefficient is statistically significant in Columns (2)-(4) and consistently indicates that lower-performing ads are more likely to be first stopped. The coefficient is not statistically significant in Column (1), consistent with the normalized CTR-difference measure being noisier than the other three measures. Importantly, across all four measures, the AI-generated image $\times$ relative performance interaction term is not statistically significant. We therefore do not detect statistically significant differences between AI-generated and human-made ads in how prior relative performance predicts first stopping.

Figure 6 Implied Relative-Performance Slopes by Ad Type in the First-Stop Diagnostics



Note. This figure plots the implied slopes of relative performance for AI-generated and human-made image ads from the first-stop diagnostics in Table 16. Points report the estimated slopes, and vertical bars show 95% confidence intervals based on standard errors clustered by quasi-experiment. In Columns (1) and (2), larger values indicate stronger focal-ad performance relative to observed siblings; in Columns (3) and (4), larger values indicate weaker focal-ad performance.

We also compute the implied slope of relative performance separately for AI-generated and human-made ads, recovered from the effect-coded specification (AI-generated image = +1, human-made = -1). We plot these estimates, along with 95% confidence intervals, in Figure 6. Consistent with the nonsignificant interaction terms in Table 16, the implied slopes for AI- and human-generated ads are close to each other, their

confidence intervals overlap in every specification, and none of the AI-human differences is statistically significant (p-values of 0.7925, 0.2382, 0.7849, and 0.7201 for the interaction term across the four measures).

Overall, the first-stop diagnostics suggest that relative performance does predict stopping decisions: lower-performing ads are more likely to be stopped first. At the same time, we are not able to detect in our data statistically significant differences between AI-generated and human-made image ads in how relative performance predicts stopping. This conclusion is unchanged when we exclude quasi-experiments in which multiple ads stop on the first stopping day, although this stricter sample restriction drops 38.9% of quasi-experiments. Taken together, the diagnostics do not eliminate the possibility of endogenous stopping. They do, however, help bound the concern. Lower-performing ads are more likely to be stopped, but across the four measures we construct, and given the estimated slopes and their confidence intervals, the data do not point to the pattern that would threaten the sibling-ad design, such that under-performing AI-generated or human-made ads are differentially likely to determine the truncation endpoint.

Web Appendix References

- Beichert M, Bayerl A, Goldenberg J, Lanz A (2024) Revenue generation through influencer marketing. *Journal of Marketing* 88(4):40–63.
- Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW (2022) The development and psychometric properties of LIWC-22 URL <https://www.liwc.app/>.
- Crete F, Dolmire T, Ladret P, Nicolas M (2007) The blur effect: perception and estimation with a new no-reference perceptual blur metric. *Human Vision and Electronic Imaging XII*, volume 6492, 196–206 (SPIE).
- Dew R, Ansari A, Toubia O (2022) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science* 41(2):401–425.
- Hartmann J, Bergner A, Hildebrand C (2023a) Mindminer: Uncovering linguistic markers of mind perception as a new lens to understand consumer-smart object relationships. *Journal of Consumer Psychology* 33(4):645–667.
- Hartmann J, Exner Y, Domdey S (2025) The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing* 42(1):13–31.
- Hartmann J, Netzer O, Zalta R (2023b) Diversity in advertising in times of racial unrest. *Available at SSRN 4466378* .
- Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120(11):e2208839120.
- Karageorgiou D, Papadopoulos S, Kompatsiaris I, Gavves E (2025) Any-resolution AI-generated image detection by spectral learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 18706–18717.
- Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P (2014) Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312* .

- Pieters R, Wedel M, Zhang J (2007) Optimal feature advertising design under competitive clutter. *Management Science* 53(11):1815–1828.
- Raghubir P, Greenleaf EA (2006) Ratios in proportion: What should the shape of the package be? *Journal of Marketing* 70(2):95–107.
- Talebi H, Milanfar P (2018) NIMA: Neural image assessment. *IEEE transactions on Image Processing* 27(8):3998–4011.
- Zhang M, Luo L (2023) Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. *Management Science* 69(1):25–50.
- Zhang S, Lee D, Singh PV, Srinivasan K (2022) What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Management Science* 68(8):5644–5666.