Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

How Persuasive Are LLMs in the Wild? Assessing Personalized Ads in Real-World Delivery

Asmaa El Fraihi[1,2,3], **Nardjes Amieur**[1,2,3], **Béatrice Roussillon**[4], **Oana Goga**[1,2,3]
[1]Institut National de Recherche en Informatique et en Automatique (Inria)
[2]Institut Polytechnique de Paris (IP Paris)
[3]Centre National de la Recherche Scientifique (CNRS)
[4]Université Grenoble Alpes
{asmaa.el-fraihi, nardjes.amieur, oana.goga}@inria.fr, beatrice.roussillon@univ-grenoble-alpes.fr

Abstract

Large language models (LLMs) have shown persuasive potential in controlled studies and surveys across commercial, political, and social domains; their effectiveness in real-world settings, however, remains largely underexplored. This work bridges that gap by evaluating LLM-generated personalized messages deployed through advertising campaigns on Meta platforms. We generate ads using three LLMs (GPT-4o, Gemini 1.5 Pro, and LLaMA 3.1), targeting four demographic groups through three distinct personalization strategies—adapting tone and language, introducing audience-relevant themes, and selectively emphasizing elements from the source material. We assess their performance across three dimensions: (1) user engagement through live testing on social media, (2) perceived appeal via user surveys, and (3) platform behavior through algorithmic delivery analysis. Our results show that personalized messages, deployed through Meta's ad delivery system, do not significantly improve user engagement compared to non-personalized alternatives, and for some demographic groups specific strategies reduced engagement further. We also find that survey-based assessments of ad appeal can diverge from observed behavioral outcomes, highlighting the limitations of relying solely on self-reported metrics to evaluate LLM-based personalization. Finally, we show that LLM-generated personalization cues can shift algorithmic ad delivery toward the intended audience by up to 8% without explicit targeting instructions, but that this influence is bounded by the platform's own relevance predictions. Together, these findings provide an empirical assessment of both the potential and the limits of LLM-based personalization within algorithmically mediated advertising systems.

1 Introduction

Recent advances in large language models (LLMs) have opened new possibilities for persuasive content generation across domains such as health communication, marketing, and political advocacy (Meguellati et al. 2024; Hackenburg and Margetts 2024; Breum et al. 2024; Karinshak et al. 2023; Altay et al. 2023). In particular, LLMs can adapt the tone, framing, and content of messages to match user attributes such as demographics or personality traits—a form of *automated personalization* that has been shown to produce messages comparably or more effective than human-written ones (Meguellati et al. 2024). By reducing the cost and expertise required to produce tailored content, LLMs make such personalization feasible at a scale that was previously impractical.

These capabilities have raised concerns about the risks of mass influence campaigns enabled by LLMs (Hendrycks, Mazeika, and Woodside 2023; Ostwal 2025; Ellis-Petersen 2024). Yet most evaluations of LLM-based personalization have been conducted in controlled environments such as user studies or simulated platforms, where participants rate messages in isolation from the systems that would ultimately deliver them. A critical gap therefore remains: *how do LLM-generated personalized messages perform when deployed at scale, in real time, and within the algorithmic logic of real-world delivery systems?* The ability to generate persuasive content is only one part of the equation. In practice, the reach and impact of any message depend not only on its content but also on how platforms select, prioritize, and distribute it—processes that may amplify, attenuate, or counteract the intended personalization.

This study addresses that gap by examining LLM-personalized content deployed as online ads on Meta platforms. Online advertising is a particularly relevant setting for this investigation, as these systems enable scalable message dissemination combined with precise audience selection based on inferred user traits, behaviors, and demographics. Platforms like Meta do not simply broadcast ads to a broad audience—they optimize delivery through algorithmic matching, determining who sees which message based on predicted engagement and relevance. LLM-generated ads thus operate within an algorithmic ecosystem that mediates their reach, potentially amplifying personalization cues for some audiences while suppressing them for others.

We conduct a multi-part empirical investigation centered on demographic personalization. We first generate a dataset of personalized ads using three LLMs (GPT-4o, Gemini 1.5 Pro, LLaMA 3.1 8B Instruct), targeting four demographic groups (women, men, young adults, and older adults) using news articles as source material. To address the high variability in unconstrained generation, we formalize three recurring personalization strategies observed in the generated content—linguistic, conceptual, and contextual—and use this typology to systematically generate and evaluate ads. Each ad is scored on *faithfulness* to the source ma-

terial, demographic *relevance*, and *persuasiveness*, and the highest-quality variants are selected for deployment.

We evaluate these ads through three complementary experiments: live testing on Meta platforms to measure how personalization strategies affect user engagement; user surveys to assess perceived appeal and compare subjective impressions with behavioral outcomes; and algorithmic delivery analysis to test whether message-level personalization cues can steer ad reach toward intended demographics without explicit targeting instructions.

Our contributions are as follows:

- We show that within live settings, *LLM-based personalization does not consistently improve observed engagement outcomes relative to non-personalized messages*. For some demographic groups, specific personalization strategies led to lower engagement than non-personalized alternatives. However, from the platform's perspective, personalized ads were generally more cost-efficient, indicating that the delivery algorithm favors them even when they do not produce higher engagement.
- By combining survey responses with field experiment outcomes, we show that *self-reported preferences do not reliably predict real-world performance*. For most demographics, participants preferred non-personalized content in surveys, yet that variant did not outperform others in the field. Among older adults, peer-directed judgments favored the only strategy that significantly reduced engagement. These findings suggest that survey-based evaluations, while informative about user perceptions, provide an incomplete picture of how personalized content performs in algorithmically mediated environments.
- We demonstrate that *LLM-based personalization cues can influence algorithmic ad delivery*, shifting demographic reach by up to 8% without explicit targeting instructions. However, this influence depends on how the platform already relates to the target demographic. When the platform strictly does not associate a piece of content with a given group, text-level personalization alone cannot redirect delivery toward that group, and in some cases reduces it further.

Together, these findings suggest that the real-world impact of LLM-based personalization in advertising is shaped less by the quality of the generated content than by the algorithmic infrastructure through which it is delivered. Evaluating such content in isolation from these delivery systems risks overestimating its practical influence. More broadly, our work highlights the need to study LLM-generated content not only as a product of generation but as an input to the algorithmic systems that determine its reach and effect.

2 Background and Related Work

2.1 Background

Our experiments rely on two key properties of Meta's ad delivery system, which we describe below.

Audience selection is not fully controlled by the advertiser. When an ad is eligible to be shown to a user, Meta's system evaluates it based on the advertiser's bid, the predicted likelihood that the user will engage, and an assessment of the ad's quality (Meta Business Help Centre 2026). While advertisers can specify the demographic attributes or interests of their intended audience, Meta has increasingly automated this process. Its Advantage+ audience system treats advertiser-defined attributes as suggestions rather than hard constraints, allowing the platform's own engagement predictions to determine which users ultimately see the ad (Meta Business Help Centre 2025a). As a result, the platform may deliver an ad to users outside the advertiser's intended audience if it predicts they are likely to engage, or withhold it from the intended audience if it predicts they are not.

Ad content influences who sees the ad. Because engagement predictions depend in part on the content of the ad—its text, imagery, and linked material—the ad itself acts as a signal in the delivery process (Ali et al. 2019; Imana, Korolova, and Heidemann 2021). Two ads with identical targeting parameters but different text may be shown to different users, because the platform's models assess their relevance to each user differently. This property is central to our algorithmic delivery experiment (Section 4.3), which tests whether textual personalization cues introduced by LLMs can shift who the platform delivers an ad to, even when no audience targeting criteria are specified.

2.2 Related Work

LLM-Based Personalization and Persuasion. A growing body of work has examined the ability of LLMs to generate persuasive content tailored to specific audiences. Meguellati et al. (2024) found that LLM-generated ads tailored to personality traits matched the effectiveness of human-written counterparts, while Simchon, Edwards, and Lewandowsky (2024) showed that personality-aligned political messages improved perceived persuasiveness, and Matz et al. (2024) demonstrated similar effects across multiple persuasion domains including marketing and advocacy. In interactive contexts, Salvi et al. (2024) found that LLM-driven personalization significantly increased persuasive impact in debates. Across these studies, a consistent finding is that LLMs can produce audience-adapted content at low cost and with minimal human oversight. However, this evidence has been gathered almost exclusively through surveys, controlled settings, or simulated platforms. This is a notable limitation, given that research in psychology and consumer behavior has consistently shown that stated preferences are imperfect predictors of actual behavior (Sheeran 2002; Chandon, Morwitz, and Reinartz 2005). Whether LLM-based personalization remains effective when messages are deployed through real-world channels—where reach and impact are mediated by algorithmic delivery systems—remains an open question that this study directly addresses.

Algorithmic Delivery Skews. As described in Section 2.1, ad delivery on Meta is shaped by the platform's own relevance predictions, not solely by advertiser intent. Prior auditing work has shown that this process can produce biased delivery outcomes along demographic lines—including

race, gender, and political leaning—even when advertisers do not explicitly target these attributes (Ali et al. 2019, 2021; Imana, Korolova, and Heidemann 2021; Lambrecht and Tucker 2019). Further work has demonstrated that the content of an ad itself can drive these skews: ad imagery featuring people of different races or genders shifts delivery along corresponding demographic lines (Kaplan et al. 2022), and similar effects have been documented in the delivery of climate-related ads across political and demographic groups (Sankaranarayanan, Sapiezynski, and O'Reilly 2025). These findings establish that content-level signals can serve as proxies for demographic attributes, influencing delivery without advertiser intent or awareness. With LLMs now capable of systematically generating text that varies along any number of such attributes, the question of whether text-level cues can steer algorithmic delivery becomes directly relevant. Our delivery experiment (Section 4.3) investigates this possibility.

3 Dataset

3.1 Generating Personalized Ads

We generate personalized ad texts using three language models: Gemini 1.5 Pro, GPT-4o, and LLaMA 3.1 8B Instruct. Each model is prompted to produce ads tailored toward one of four demographic groups—**men**, **women**, **young adults**, and **older adults**—drawing on 20 randomly sampled, non-political articles from *The Guardian* (The Guardian 2025) across three general-interest categories (Cities, Travel, and Culture). All three models generate ads for every article-demographic pair, yielding an initial set of 240 personalized ads. Using multiple LLMs broadens the pool of candidate ads and increases the likelihood of obtaining high-quality outputs for subsequent selection.

We focus on gender and age as personalization attributes for two reasons. First, Meta's ad delivery platform provides demographic breakdowns of campaign outcomes along these dimensions, enabling us to measure whether personalized ads reach their intended audiences in our subsequent experiments (Section 4.3). Second, by targeting demographically distinct groups—men versus women, young adults versus older adults—we maximize the contrast between conditions, making it easier to detect personalization effects if they exist.

This initial round of generation, however, revealed substantial variability in how models applied personalization. Some ads contained explicit tailoring—direct audience address, stylistic shifts, or demographic-specific framing—while others relied on subtle reframing or showed no clear personalization at all. Refining the prompts did not meaningfully reduce this inconsistency (see Appendix A). Rather than treating this variability as noise, we analyzed the outputs and identified three recurring strategies that models employed, which we formalized into a typology to guide a second, controlled round of generation:

- **Linguistic** personalization adapts the tone, style, or mode of address to match how the target audience communicates. An ad targeting young adults might use informal, high-energy language, while one targeting older adults might adopt a more measured and reflective tone.
- **Conceptual** personalization introduces themes associated with the target audience that go beyond what the source article contains. A travel article, for instance, might generate an ad around nostalgia and life reflection for older adults, or around self-discovery and adventure for young adults.
- **Contextual** personalization selectively emphasizes elements already present in the source article likely to resonate with a particular group. From the same hiking article, an ad targeting men might foreground physical challenge and endurance, while one targeting women might highlight scenic beauty and personal transformation.

Conceptual and contextual personalization differ in where the cue originates—external to the source article or within it. We maintain this separation because they represent distinct generation behaviors observed in the LLM outputs, and because they allow us to test whether the origin of the cue affects engagement and delivery differently.

Using this typology, we generate a controlled dataset that systematically varies strategy across demographics and articles. For each of the 20 articles, we produce one ad per strategy for each of the four demographic groups, resulting in 12 personalized ads per article (3 strategies × 4 groups). Each article also receives one non-personalized ad designed to appeal to a broad audience. This process is repeated with all three LLMs, yielding a total of 780 ads: 720 personalized and 60 non-personalized.

3.2 Quality Assessment

Before deploying ads in field experiments, we evaluate the generated outputs to assess whether the three models reliably produce persuasive, audience-aligned content across the three personalization strategies, and to select the highest-quality ads for real-world testing.

We adopt a model-as-a-judge approach (Liu et al. 2023), using GPT-4o to rate each ad along three dimensions: **faithfulness** to the source article, **relevance** to the target audience, and **persuasiveness**. Each dimension is scored on a 1–5 Likert scale (Batterton and Hale 2017) using structured prompts designed to standardize evaluation criteria (see Appendix A.2). While model-based evaluation lacks the nuance of human judgment, it ensures consistency across hundreds of ad variants and avoids annotator variability. Prior work has also shown that GPT-4o ratings approximate human judgments with reasonable fidelity (Bavaresco et al. 2024).

Table 1 summarizes the results. The three strategies exhibit distinct trade-offs. Contextual personalization scored highest on faithfulness, as it preserves the source article and selectively foregrounds relevant elements, but this conservatism resulted in lower relevance scores. Conceptual personalization, by contrast, scored highest on relevance and persuasiveness, but lowest on faithfulness—an expected consequence of introducing themes not present in the original text. Linguistic personalization fell between the two on faithfulness and relevance, but was rated the least persuasive,

Metric	Linguistic	Conceptual	Contextual
Faithfulness	3.31 (0.73)[a]	3.26 (0.76)[a]	3.77 (0.69)[b]
Relevance	3.48 (0.76)[a]	3.52 (0.71)[a]	3.06 (0.78)[b]
Persuasiveness	2.79 (0.37)[a]	2.98 (0.35)[b]	2.90 (0.43)[b]

Table 1: Mean ($\pm$ SD) scores for each evaluation metric by personalization strategy. Within each row, values with different superscript letters differ significantly (Tukey HSD, $p < 0.05$).

possibly because overt audience address or informal tone reduces perceived credibility, one of the sub-dimensions in our persuasiveness metric.

Ad Selection Based on these scores, we selected the top-performing ads from a single article for use in the field experiments. Selection followed a weighted scoring system that prioritized persuasiveness (60%), reflecting its central role in advertising effectiveness, with relevance and faithfulness each contributing 20% to ensure audience fit and source integrity. All selected ads were manually reviewed to confirm clarity, appropriate tone, and factual accuracy.

4 Experiments

Evaluating LLM-based personalization in real-world advertising requires disentangling several factors that jointly determine an ad's impact. When a personalized ad is deployed on a platform like Meta, the engagement it receives reflects not only the appeal of its content but also the platform's delivery optimization, which determines which users see the ad based on its own relevance predictions. To separate these effects, we design three complementary experiments. We first deploy ads on Meta through live testing, where all ad variants run under identical conditions, to measure whether personalization improves user engagement. We complement these results with a survey in which participants from the same demographic groups evaluate the same ads outside of any algorithmic context, isolating perceived appeal from platform-mediated outcomes. Finally, we turn from user response to platform behavior, testing whether LLM-generated personalization cues can shift algorithmic delivery toward intended demographics without explicit targeting, and whether the platform's own relevance predictions constrain this effect.

4.1 User Engagement on Social Media

We first assess whether LLM-based personalization improves user engagement when ads are deployed on Meta platforms. To compare engagement across ad variants, we use Meta's built-in A/B testing tool, which divides the target audience into non-overlapping groups, each exposed to a single ad variant, under identical budget, schedule, and bidding conditions (Meta Business Help Centre 2025b). An important caveat applies to interpreting results from this tool. As Boegershausen et al. (2025) have shown, Meta's A/B testing does not constitute a true randomized experiment. While the audience is split across conditions, the platform's delivery algorithm continues to optimize within each group, selectively showing ads to users it predicts are most likely to engage. Observed differences in engagement therefore reflect the joint effect of the ad's content and the platform's delivery decisions, and cannot be attributed to message content alone. We do not treat these results as causal estimates of personalization's effect on individual persuasion. Rather, they capture how personalized ads perform under realistic deployment conditions—the same conditions an advertiser using LLM-generated content would face.

Experiment Design. We conducted four separate tests, each targeting a distinct demographic group: women, men, young adults (ages 18–24), and older adults (ages 55+), all based in the United Kingdom. For each group, we created a campaign in Meta's Ad Manager with the optimization objective set to Engagement (any interaction such as likes, shares, comments, or saves). Within each campaign, five ad sets were configured with identical budgets (€20 per day), schedule (3 days), and bidding strategy (maximize results), each containing a single ad variant. The five variants were:

1. A **no-text** version containing only an image, serving as a baseline to isolate the effect of textual content on engagement.
2. A **non-personalized** version using the best-performing broadly appealing ad.
3. A **linguistic** variant adapting language and tone to the target audience.
4. A **conceptual** variant introducing audience-relevant themes.
5. A **contextual** variant emphasizing source material elements aligned with the audience.

Meta's A/B testing tool assigned each variant to a separate, non-overlapping audience segment, ensuring that each user saw only one version of the ad.

Measurement. We extract outcome data for each test using Meta's Graph API (Meta for Developers 2025b), collecting demographic breakdowns of reach (unique users shown each ad) and engagement (total interactions). We fit a logistic regression model predicting engagement (1 = engaged, 0 = did not engage) with user gender, age group, and ad variant as predictors, along with two-way interactions between demographics and variant (age $\times$ variant and gender $\times$ variant) to capture differential effects across groups.

We also extract Meta's internal *win rate* metric, which simulates thousands of delivery scenarios to estimate the probability that a given ad outperforms others in terms of cost-per-action (CPA) (Meta Business Help Centre 2025d). CPA is computed as total campaign cost divided by the number of engagement outcomes—a lower CPA indicates more engagement per unit of cost. Because the CPA and win rate reflect the platform's assessment of an ad's consistency and algorithmic favorability across delivery scenarios, they capture a dimension of performance that engagement rates alone do not.

Hypotheses. We test two expectations. First, that personalized ad variants yield higher engagement rates than non-personalized alternatives within the platform's delivery optimization (**H1**). Second, that personalization effects are not uniform—different strategies may be more or less effective depending on the demographic group (**H2**).

Results. Table 2 summarizes the results. We organize findings around engagement rate, which captures user response, and platform-side metrics (CPA and win rate), which capture how the delivery system values each variant.

Engagement Rate. Personalized ad variants did not consistently outperform the non-personalized baseline (full model results in Appendix A.4). None of the personalization strategies showed significant main effects relative to the controls (Linguistic: β = 0.11, p = 0.33; Conceptual: β = 0.05, p = 0.65; Contextual: β = 0.10, p = 0.37). However, the effects of personalization varied across demographics. Linguistic personalization significantly reduced engagement among older adults (Older Adults $\times$ Linguistic: β = -0.62, p = 0.005), while showing a marginally positive effect among women (Women $\times$ Linguistic: β = 0.29, p = 0.06). This suggests that, under the platform's delivery optimization, certain personalization strategies may be associated with lower engagement than ads with no text at all.

CPA and Win Rate. Platform-side metrics tell a different story. Across most demographic groups, personalized ads achieved lower CPA and higher win rates than both the no-text and non-personalized variants. The most effective strategy varied by group: linguistic personalization produced the lowest CPA for young adults and women, while conceptual personalization performed best for men and older adults. This indicates that the platform's delivery algorithm responds favorably to personalized content—even when users themselves do not engage more—suggesting that personalization aligns with the platform's internal relevance signals.

Takeaways. Within the platform's delivery optimization, LLM-generated broadly appealing content performed as well as personalized variants, suggesting that LLMs produce effective ad content without requiring demographic tailoring. At the same time, the platform consistently favored personalized ads in terms of cost efficiency, creating a disconnect: personalization is rewarded by the delivery system even when it does not produce higher user response. This raises a question that the following experiments address: if personalization does not improve user engagement, what is its actual effect within algorithmically mediated advertising?

4.2 Survey-Based Evaluation of Ad Appeal

The previous experiment captures how personalized ads perform under realistic deployment conditions, but observed outcomes are shaped in part by the platform's delivery optimization. To assess how users perceive these ads independently of algorithmic mediation, we conducted a survey in which participants from the same demographic groups evaluated the same ad variants used in the live test.

Design. We designed four surveys, one per demographic group: men, women, young adults (ages 18–24), and older adults (ages 55+). Each survey recruited 200 participants through Prolific ($N = 800$ total), pre-screened for UK residency and English fluency to match the population of our field experiment. To prevent overlap between samples, participants who completed one survey were excluded from all others. Gender-specific surveys included participants across age groups; age-specific surveys did not restrict by gender.

Each participant was shown four ad variants tailored to their demographic group: one non-personalized, one linguistic, one conceptual, and one contextual. They answered three forced-choice questions, selecting the ad they found (1) most personally appealing, (2) most appealing to people like them, and (3) most likely to engage with if encountered online. Participants could also indicate no preference.

Hypothesis. Based on the live deployment results, where personalization was not consistently associated with improved engagement for most groups, we expect that participants will report similar or higher appeal for the non-personalized variant compared to personalized alternatives (**H3**).

Results. We compute the percentage of respondents selecting each variant per question and assess significance using binomial tests. Results are summarized in Table 3 (see Appendix A.3 for full results). For men, women, and young adults, participants selected the non-personalized variant as most appealing across all three questions. In the live deployment, however, none of the variants—personalized or not—differed significantly in observed engagement for these groups. The survey thus reveals a stated preference for broadly appealing content that does not correspond to a measurable performance difference on the platform.

The older adult group presents a more nuanced picture. In the survey, participants rated conceptual personalization as the most personally appealing and the most likely to prompt their engagement. In the live test, conceptual personalization achieved the best platform-side metrics for this group (lowest CPA and highest win rate), though without a statistically significant improvement in engagement. However, when asked what their peers would find most appealing, older adults selected linguistic personalization—the strategy that produced the only significant negative effect on engagement in our regression for this group. Older adults' personal preferences thus aligned with the platform's assessment of ad quality, but their peer-directed judgments pointed toward the least effective strategy.

Takeaways. These results illustrate two ways in which survey measures can diverge from observed engagement outcomes under real deployment conditions. For most demographics, participants expressed a preference for non-personalized content that did not translate into a performance difference in the field. Among older adults, personal preferences aligned with platform-side metrics, but peer-directed judgments pointed toward the only strategy that significantly reduced engagement. In both cases, self-reported preferences captured nuances in how users perceive personalized content, but these nuances did not reliably predict real-world impact. This is particularly relevant given

Demographic	Ad Variation	Reach	Engagement (%)	CPA (€)	Win Rate
Women	No text	656	26	0.12	0.28
	No Personalization	607	21	0.14	<0.05
	Linguistic	**616**	**27**	**0.12**	**0.30**
	Conceptual	601	26	0.12	0.15
	Contextual	637	25	0.12	0.24
Men	No text	805	21	0.12	0.23
	No Personalization	736	22	0.12	0.10
	Linguistic	733	20	0.14	<0.05
	Conceptual	**730**	**24**	**0.11**	**0.36**
	Contextual	816	21	0.12	0.28
Young	No text	482	24	0.14	<0.05
	No Personalization	515	20	0.13	<0.05
	Linguistic	**578**	**26**	**0.10**	**0.88**
	Conceptual	567	22	0.13	<0.05
	Contextual	556	21	0.13	<0.05
Old	No text	710	13	0.17	0.07
	No Personalization	580	19	0.15	0.28
	Linguistic	718	14	0.16	0.07
	Conceptual	**534**	**20**	**0.14**	**0.46**
	Contextual	543	19	0.16	0.12

Table 2: Ad performance on Meta platforms by demographic group and ad variant. Reach is the number of unique users shown each variant. Engagement is the proportion of reached users who interacted with the ad. CPA is the average cost per engagement outcome in euros. Win rate is the estimated probability that a variant outperforms others across Meta's simulated delivery scenarios.

that much of the existing literature on LLM-based persuasion draws its conclusions from survey instruments alone. Our findings suggest that such evaluations, while informative about user perceptions, provide an incomplete picture of how personalized content performs once it enters an algorithmically mediated environment.

4.3 Algorithmic Delivery Analysis

Our first two experiments show that LLM-based personalization has limited impact on observed engagement outcomes under the platform's delivery optimization and that survey preferences do not reliably predict deployment performance. We now ask a different question: even if personalized text does not consistently produce higher engagement outcomes, does it influence which users the platform delivers the ad to?

Design. Using the same ads selected for the live test, we run delivery experiments for four demographic targets: women, men, young adults (18–24), and older adults (55+). For each target, we test all three personalization strategies separately. Each test consists of a campaign with 30 identical ad sets, all containing the same ad text paired with a generic image and a link to the source article. Running 30 simultaneous ad sets per condition allows us to account for variability in Meta's delivery system and obtain more stable estimates of demographic reach. Critically, all audience targeting parameters are left unrestricted—ads are eligible to be shown to users of any gender and age group in the United Kingdom, with no predefined interests or behavioral filters. All ad sets share an identical budget (€10) and run simultaneously for 12 hours, with the optimization objective set to maximize engagement. Under these conditions, the platform autonomously determines which users see the ad based entirely on its own relevance predictions. Campaigns were implemented using the Facebook Business SDK (Meta for Developers 2025a) and Meta Ads Manager (Meta Business Help Centre 2025c), with performance data collected through the Graph API (Meta for Developers 2025b).

Baseline. To measure how much the textual content of an ad influences delivery, we include a control condition: an image-only ad with the same link and settings but no text (see Figure 1 in the Appendix). Any difference in demographic reach between the text-bearing ads and this baseline can be attributed to the platform's response to the textual content of the ad.

Measurement. For each condition, we compute the proportion of users in the target demographic reached across 30 ad sets. For gender experiments, we report the median proportion of male users reached; for age experiments, the median proportion in the target age bracket. We compare each personalization condition to the no-text baseline using z-tests for proportions.

Results. *Gender.* Table 4 reports the median proportion of male users reached. The baseline already exhibits a gender skew: 62.7% of impressions were served to men. This likely reflects the platform's association of the linked content—a Guardian article—with male audiences, consistent with documented gender gaps in news media consumption (Kassova 2023), combined with lower average advertising costs for

Demographic	Preferred Ad	Appeal (%)	Peer Appeal (%)	Engagement (%)
Men	No Personalization	41.8**	36.8**	39.8**
Women	No Personalization	41.8**	37.3**	37.3**
Young	No Personalization	36.6**	32.7**	36.1**
Old	Conceptual	28.5*	-	28*
	Linguistic	-	37.5**	-

* p < 0.05, ** p < 0.01, *** p<0.001

Table 3: Survey results by demographic group. For each question, the table shows the most frequently selected ad variant and the percentage of respondents who chose it. Significance is assessed using binomial tests. Dashes indicate that the variant was not selected significantly more than others in the category.

	Ad Variation	Total Reach	Men (%)
Baseline	No text	3748	62.65
Men	Linguistic	4430	**64.19***
	Conceptual	5289	62.42
	Contextual	5635	61.86
Women	Linguistic	4583	**54.37*****
	Conceptual	5562	**58.93***
	Contextual	5316	**58.72****

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Demographic reach for gender-targeted ad variants. Each row reports the total reach and the median percentage of male users across 30 ad set trials. All comparisons are against the no-text baseline using z-tests for proportions.

male users on the platform (Lebesgue 2024).

Against this skewed baseline, linguistic personalization—the most explicit form of tailoring—produced clear shifts in both directions. Ads linguistically tailored for men increased male reach ($z = 2.20, p = 0.027$), while those tailored for women significantly reduced it ($z = -6.80, p < 0.001$), effectively shifting delivery by over 8 percentage points toward female users. Conceptual and contextual strategies showed no significant effect for male-targeted ads, but both shifted delivery toward women when the ads were women-targeted (Conceptual: $z = -2.04, p = 0.041$; Contextual: $z = -2.88, p = 0.003$), reducing the male share by approximately 4 percentage points.

Unlike the ads examined in prior delivery audits, where the content itself is strongly tied to a demographic, such as job listings in male-dominated fields, or imagery associated with specific groups (Ali et al. 2019; Imana, Korolova, and Heidemann 2021; Kaplan et al. 2022), the ads in our experiments promote a general-interest travel article with no inherent demographic association. The only difference between variants is the ad text, generated by an LLM through a single prompt. Yet these text-level cues alone were sufficient to shift the platform's delivery decisions. This suggests that the proxy targeting mechanisms identified in prior work extend to LLM-generated content, where the barrier to producing such cues is negligible, and the process is trivially scalable.

Age. Tables 5 and 6 report results for age-based personalization. The baseline reach among young adults (18–24) was 41%, indicating that the platform already associates the content with younger audiences. Among this group, only contextual personalization produced a significant effect, increasing youth reach by over 8 percentage points ($z = 6.42, p < 0.001$). Linguistic and conceptual strategies had no significant impact. This contrasts with the gender results, where linguistic personalization was most effective at shifting delivery, suggesting that the type of cue the platform responds to varies across the tested dimensions.

The pattern for older adults (55+) is strikingly different. The baseline reach for this group was just 14%, indicating that the platform does not associate the content with older users. None of the personalization strategies improved delivery to this group. Linguistic and conceptual personalization further *reduced* reach by approximately 6 percentage points ($z = -8.64$ and $z = -8.22$, respectively; $p < 0.001$), while contextual personalization had no significant effect. This means that explicitly tailoring an ad for older adults—using direct address, age-relevant language, or thematic framing—led the platform to show it to *fewer* older users, not more.

This finding has important implications. It suggests that when the platform's relevance model does not associate a piece of content with a demographic group, adding personalization cues to the text does not override that assessment—it may instead introduce signals that the platform interprets as reducing relevance for the broader audience, further narrowing delivery. In the context of LLM-based personalization, this means that generating targeted text is not sufficient to reach the intended audience. The platform acts as a gatekeeper whose relevance predictions can override, and even counteract, the advertiser's intent.

Takeaways. Our findings show that LLM-generated text, produced through a simple prompting step, is sufficient to shift ad delivery toward a target demographic without any explicit targeting parameters, demonstrating that proxy targeting through ad content is trivially scalable. However, this capacity is not uniform across demographic attributes, and personalization can be counterproductive for groups the platform already deprioritizes. LLM-based personalization deployed without explicit targeting may therefore correct delivery imbalances for some groups while deepening them for others, with implications for equitable access to information in algorithmically mediated environments.

	Ad Variation	Total Reach	Young (%)
Baseline	No text	3748	40.94
Young	Linguistic	4732	41.10
	Conceptual	4923	45.66
	Contextual	4627	**49.07*****

$^{*}\ p < 0.05, ^{**}\ p < 0.01, ^{***}\ p < 0.001$

Table 5: Demographic reach for young adult–targeted ad variants (ages 18–24). Each row reports the total reach and the median percentage of users aged 18–24 across 30 ad set trials. All comparisons are against the no-text baseline using z-tests for proportions.

	Ad Variation	Total Reach	Old (%)
Baseline	No text	3748	14.11
Old	Linguistic	4826	**8.09*****
	Conceptual	4931	**7.89*****
	Contextual	4846	14.74

$^{*}\ p < 0.05, ^{**}\ p < 0.01, ^{***}\ p < 0.001$

Table 6: Demographic reach for older adult–targeted ad variants (ages 55+). Each row reports the total reach and the median percentage of users aged 55 and older across 30 ad set trials. All comparisons are against the no-text baseline using z-tests for proportions.

5 Limitations

Our study has several limitations that should be considered when interpreting the results. Our field experiments and delivery analysis were conducted on a single platform—Meta. Other platforms employ different optimization algorithms and delivery mechanisms, and our findings may not generalize to systems such as Google Ads or X. Ad delivery systems are also subject to frequent updates, which may affect the reproducibility of specific results over time. Our personalization targets are limited to binary gender and two age brackets, a constraint imposed by Meta's demographic reporting infrastructure. Richer attributes, such as interests or personality traits, may enable more effective tailoring but would require alternative measurement approaches. Moreover, the prompts for personalized and non-personalized conditions reflect different generation objectives—broad appeal versus demographic tailoring—which necessarily involve different instructions. While this means the conditions are not identical in prompt structure, they reflect the realistic distinction between personalized and non-personalized ad generation as it would be practiced. More generally, LLM outputs are sensitive to prompt design, and it is difficult to establish whether a given prompt elicits the best possible output from a model. Different prompt formulations, more detailed instructions, or expert-crafted few-shot examples may yield more effective personalized ads than those tested here—a limitation shared by any study that evaluates LLM-generated content. Finally, our field experiments use a single source and single-attribute personalization for consistency and interpretability; different content or multi-attribute targeting may produce different dynamics.

6 Conclusion

This study evaluates LLM-based personalization not in isolation, but within the algorithmic infrastructure that determines an ad's reach and impact. Across three experiments, a consistent picture emerges: personalized messages are not consistently associated with improved engagement outcomes over broadly appealing alternatives, a pattern observed in both live deployment and survey settings. Where personalization does have an effect, it is primarily on the platform rather than on observed user response—generated ads achieve better cost efficiency, and their textual cues can shift algorithmic audience selection by up to 8% without any explicit targeting. These results suggest a reframing of the concern around LLM-enabled influence at scale. The risk may be less that LLMs produce messages that are more persuasive to individuals, and more that they make it trivial to generate content that interacts with delivery algorithms in ways that shape who is exposed to what. At the same time, our findings show that LLMs are capable of producing broadly appealing content that performs well without any demographic tailoring—a capacity that, combined with the low cost and effort of generation, may prove more consequential for the advertising landscape than personalization itself.

As LLMs become embedded in advertising workflows, the question is not only whether their outputs are persuasive, but how those outputs interact with the opaque delivery systems that mediate their reach. Our work takes a first step toward answering that question empirically, and we hope it encourages further research at the intersection of LLM-generated content and algorithmic delivery.

Ethical Statement

This study was approved by the institutional ethics review board (ERB) under a project-wide protocol. All LLM-generated ads used in the field experiments were manually reviewed to exclude misleading, harmful, or factually inaccurate content, and the source material was limited to a neutral topic with positive connotations (Travel, Cities, and Culture) to minimize the risk of exposing users to objectionable content.

We acknowledge a transparency limitation in the field experiments: Meta's Ad Manager did not support labeling ads as AI-generated at the time we conducted the experiments, and embedding such disclosures in the ad text would have introduced a confound by altering both user response and algorithmic delivery. Users exposed to ads on the platform were therefore not informed that the content was generated by an LLM. We believe this is justified given the benign nature of the advertised content and the study's contribution to understanding the real-world implications of LLM-based personalization—findings that are themselves relevant to informed public discourse on AI-generated advertising. Survey participants were fully debriefed after completing the study and informed that the ads they evaluated were LLM-generated and demographically tailored. They were

also provided with resources about large language models (see Appendix A.3).

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This research is supported by the European Research Council (ERC) grant no. 101041223.

References

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–30.

Ali, M.; Sapiezynski, P.; Korolova, A.; Mislove, A.; and Rieke, A. 2021. Ad delivery algorithms: The hidden arbiters of political messaging. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 13–21.

Altay, S.; Hacquin, A.-S.; Chevallier, C.; and Mercier, H. 2023. Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, 29(1): 52.

Batterton, K. A.; and Hale, K. N. 2017. The Likert scale what it is and how to use it. *Phalanx*, 50(2): 32–39.

Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Boegershausen, J.; Cornil, Y.; Yi, S.; and Hardisty, D. J. 2025. On the persistent mischaracterization of Google and Facebook A/B tests: How to conduct and report online platform studies. *International Journal of Research in Marketing*, 42: 886–903.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 152–163.

Chandon, P.; Morwitz, V. G.; and Reinartz, W. J. 2005. Do intentions really predict behavior? Self-generated validity effects in survey research. *Journal of marketing*, 69(2): 1–14.

Ellis-Petersen, H. 2024. Revealed: Meta approved political ads in India that incited violence. *The Guardian*. https://www.theguardian.com/world/article/2024/may/20/revealed-meta-approved-political-ads-in-india-that-incited-violence.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hackenburg, K.; and Margetts, H. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Imana, B.; Korolova, A.; and Heidemann, J. 2021. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the web conference 2021*, 3767–3778.

Kaplan, L.; Gerzon, N.; Mislove, A.; and Sapiezynski, P. 2022. Measurement and analysis of implied identity in ad delivery optimization. In *Proceedings of the 22nd ACM Internet Measurement Conference*, 195–209.

Karinshak, E.; Liu, S. X.; Park, J. S.; and Hancock, J. T. 2023. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–29.

Kassova, L. 2023. The Gender Consumption Gap for News and How Publishers Can Address It. https://pressgazette.co.uk/comment-analysis/gender-consumption-gap-news-publishers/.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science*, 65(7): 2966–2981.

Lebesgue. 2024. Analyzing Facebook CPM by Gender. https://lebesgue.io/facebook-ads/facebook-cpm-by-gender. Accessed: 2025-01-01.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Matz, S.; Teeny, J.; Vaid, S. S.; Peters, H.; Harari, G.; and Cerf, M. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1): 4692.

Meguellati, E.; Han, L.; Bernstein, A.; Sadiq, S.; and Demartini, G. 2024. How Good are LLMs in Generating Personalized Advertisements? In *Companion Proceedings of the ACM on Web Conference 2024*, 826–829.

Meta Business Help Centre. 2025a. About Advantage+ Audience. https://en-gb.facebook.com/business/help/273363992030035. Accessed: 2025-01-01.

Meta Business Help Centre. 2025b. About Experiments. https://en-gb.facebook.com/business/help/1915029282150425. Accessed: 2025-01-01.

Meta Business Help Centre. 2025c. How to Create Ad Campaigns in Meta Ads Manager. https://en-gb.facebook.com/business/help/621956575422138. Accessed: 2025-01-01.

Meta Business Help Centre. 2025d. How Winning Campaigns Are Determined in A/B Tests. https://www.facebook.com/business/help/166313650471318. Accessed: 2026-03-18.

Meta Business Help Centre. 2026. About Ad Auctions. https://en-gb.facebook.com/business/help/430291176997542. Accessed: 2026-03-01.

Meta for Developers. 2025a. Ads Buying — Meta Business SDK — Documentation. https://developers.facebook.com/docs/business-sdk/common-scenarios/ads-buying/. Accessed: 2025-01-01.

Meta for Developers. 2025b. Graph API — Documentation. https://developers.facebook.com/docs/graph-api/. Accessed: 2025-01-01.

Ostwal, T. 2025. Meta and X Approve AI Ads Referencing Nazi War Crimes Ahead of German Elections, Research Finds. https://www.adweek.com/media/meta-and-x-approve-ai-ads-referencing-nazi-war-crimes-ahead-of-german-elections-research-finds/.

Salvi, F.; Ribeiro, M. H.; Gallotti, R.; and West, R. 2024. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380*.

Sankaranarayanan, A.; Sapiezynski, P.; and O'Reilly, U.-M. 2025. Facebook algorithm's active role in climate advertisement delivery. *Nature Climate Change*, 15(7): 719–724.

Sheeran, P. 2002. Intention—behavior relations: a conceptual and empirical review. *European review of social psychology*, 12(1): 1–36.

Simchon, A.; Edwards, M.; and Lewandowsky, S. 2024. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS nexus*, 3(2): pgae035.

The Guardian. 2025. Latest news, sport and opinion from the Guardian. https://www.theguardian.com/europe.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see Ethics Statement
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Introduction, Background and Related Work.
   (d) Do you clarify what are possible artifacts are in the data used, given population-specific distributions? Yes. Since we are conducting field experiments, the observations reflect real-world behavior within the studied platform. These discussions appear throughout the whole paper.
   (e) Did you describe the limitations of your work? Yes, see Section 5.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Takeaways and Conclusion sections.
   (g) Did you discuss any potential misuse of your work? Yes, see Conclusion.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethics Statement and Appendix A.3.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, see Ethics Statement.
2. Additionally, if your study involves hypothesis testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, we cite The Guardian as the source of article content used to generate ads.
   (b) Did you mention the license of the assets? Yes, the use falls under fair use for research purposes; The Guardian articles are publicly available.

(c) Did you include any new assets in the supplemental material or as a URL? No, we provide the materials in the Appendix. Specifically, the prompts that were used (Appendix A and A.2) and the final advertised copies (Appendix A.4).

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, for participants recruited through crowdsourcing, consent is obtained through a Prolific-IRB-compliant process. For users involved in the ad delivery experiments, see Section 6 for our discussion on Ethics within this research.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, the use of generated content for field experiments was manually reviewed to avoid harmful, misleading, or offensive material (see Ethics Statement).

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, all the questions and details related to the survey are provided in Section 4.2. Additional materials are in Appendix A.3.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see Ethics Statement

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, see Appendix A.3

(d) Did you discuss how data is stored, shared, and deidentified? NA, aside from participants' answers, we do not collect any other type of information from participants.

A Appendix

A.1 Generating Ads Prompts

Simple Prompt The following is the prompt used to generate the initial ads dataset. The prompt is simple and is similar to a lot of prompts used in prior research in this area (Matz et al. 2024; Hackenburg and Margetts 2024):

```
Write 5 persuasive Facebook Ad Text Bodies (Three
sentences maximum) for the following article:
{article}.
Each ad is tailored for one of the following
audiences:
  1. Neutral: a broad, general audience.
  2. Men: a male demographic.
  3. Women: a female demographic.
  4. Young adults 18--24 years old: a younger
demographic no older than 24 years old.
  5. Old adults 55+ years old: a senior demographic
no younger than 55 years old
```

Detailed Prompt We added details about how targeting can be achieved and instructions to clarify the objectives.

```
Write five persuasive Facebook ad texts to promote
the provided media content from the Guardian:
{article}
Each ad variation is designed to resonate with one
of the following demographics:
  1. Neutral: Appeal to a broad, general audience.
  2. Men: An ad targeted to a male demographic.
  3. Women: An ad targeted to a female demographic.
  4. Young adults 18--24 years old: An ad that
resonates with a younger demographic no older than
24 years old.
  5. Old Adults 55+ years old: An ad that resonates
with an older demographic no younger than 55 years
old.

Instructions:
  * Ad Text: The tailored ads for women, men, young
adults and old adults should be distinct from the
neutral ad text, and clearly reflect the unique
interests and preferences of each demographic.
  * Targeting Explanation: For each ad variation,
provide a concise explanation that clearly
articulates how and why the ad is tailored to
its target demographic.
  * Content Relevance: Ensure all generated ad
text is coherent, factually accurate, and directly
relevant to the provided news article.
  * Clear Call to Action: Add a clear call to
action ("Read more," "Learn more") within each ad
text.
  * Character Count: Aim for a character count
suitable for Facebook ads, ideally under 280
characters. However, prioritize clarity and
persuasiveness over strict adherence to this
limit.
```

One-Shot To encourage stronger personalization, we appended a worked example to the zero-shot prompt, showing demographically tailored ads for a sample article. This example was designed to illustrate the expected style and level of differentiation across demographic groups.

```
[Detailed Zero-shot prompt]

Here's an example of personalized ads crafted for
the following article: [Provided article]

Personalized Ad texts:
  *Men*: A gripping tale of friendship, ambition,
and the fight for survival. Lisa Ko's Memory
Piece masterfully blends personal struggles with a
stark dystopian future, challenging perceptions of
success and freedom.

  *Women*: Discover the power of female friendship
and resilience in Lisa Ko's Memory Piece. This
moving novel explores the lives of three women
across time, highlighting their triumphs and
struggles against a backdrop of societal change.

  *Old adults*: Reflect on the passage of time
and the enduring power of memory in Lisa Ko's
```

```
captivating Memory Piece. This thought-provoking
novel explores themes of friendship, societal
change, and the importance of preserving our
history. Discover more!

  *Young Adults*: Get ready for a wild ride
through time! Memory Piece is a mind-bending
novel that tackles big issues -- from the art
world to a dystopian future -- with sharp wit and
unforgettable characters. Click to learn more!

  *Neutral*: Explore the intersection of memory,
art, and dystopia in Lisa Ko's Memory Piece.
This poignant novel follows three friends across
decades, weaving together personal narratives with
a chilling vision of the future. Read more!
```

Controlled Personalized Ad Generation Prompt The following prompt was used to generate the final dataset of ads, with the personalization strategy specified:

```
Task Description:
Write 3 persuasive Facebook ad texts to promote
the provided media content from the Guardian:
{article}.
The ad is tailored specifically to an audience of:
{attribute}

Instructions:
  *Ad Generation*:
    The first ad will use *Language-based
personalization*: the ad uses specific language
or tone that align with the target audience, such
as explicitly naming the group or using a tone
that fits the target audience.

    The second ad will use *Concept-based
personalization*: the ad introduces new themes
or ideas that reflect the characteristics or
interests of the target audience, beyond what is
in the original article.

    The third ad will use *Context-based
personalization*: the ad adapts arguments or
elements directly from the original article to
appeal to the target audience.

  *Clear Call to Action:* Add a clear call to
action ("Read more," "Learn more") within each ad
text.

  *Character Count:* Aim for a character count
suitable for Facebook ads, ideally under 280
characters. However, prioritize clarity and
persuasiveness over strict adherence to this
limit.
```

Non-Personalized Ad Generation The following prompt was used to generate the broadly appealing, non-personalized ads:

```
Task Description:
Write a persuasive Facebook ad text to promote
the provided media content from the Guardian:
{article}

Instructions:
  *Ad Generation*: The ad is persuasive to a broad
and general audience and does not target any
specific group or demographic attribute.

  *Clear Call to Action*: Add a clear call to
action ("Read more," "Learn more") within the ad
text.

  *Character Count*: Aim for a character count
suitable for Facebook ads, ideally under 280
characters. However, prioritize clarity and
persuasiveness over strict adherence to this
limit.
```

A.2 Evaluation Prompts

We detail below the prompts used for each of the three evaluation metrics: faithfulness, relevance, and persuasiveness. Each prompt instructs GPT-4o to score an ad along specified dimensions and return a weighted composite score.

Faithfulness

```
You are an expert evaluator tasked with assessing
the faithfulness of a generated ad text to the
provided context. Your evaluation will be based on
the following dimensions:

  1. **Content Accuracy (CA):** How accurately
does the text reflect the details provided in the
context? (Score: 1 to 5)
  2. **Message Alignment (MA):** How well does the
text align with the intended message or theme in
the context? (Score: 1 to 5)
  3. **Contextual Completeness (CC):** Does the
text include **all** key elements specified in the
context? (Score: 1 to 5)

**Scoring Guidelines:**
  - 1 = Poor (completely inconsistent with the
context)
  - 2 = Below Average (largely inconsistent with
the context)
  - 3 = Average (partially consistent with the
context)
  - 4 = Good (mostly consistent with the context)
  - 5 = Excellent (fully consistent with the
context)

**Context:** {context}
*Generated Ad Text:** {ad_text}

**Task:**
  1. Assign a score (1 to 5) for each dimension.
  2. Calculate the Faithfulness Score (FS) using
the formula:
    FS = (CA * 0.50) + (CC * 0.25) + (MA * 0.25)

**Output Format:**
Format your response as a JSON object with the
following structure:
  {{ "faithfulness" : <FS> }}
```

Relevance

```
You are an expert evaluator tasked with assessing
the relevance of a generated text as an
advertisement. Your evaluation will be based on
the following dimensions:

  1. **Attention Capture (AC):** How effectively
does the text grab the reader's attention? (Score:
1 to 5)
  2. **Highlighted Information (HI):** How clearly
and effectively does the text communicate the core
```

```
idea or message? (Score: 1 to 5)
  3. **Target Audience Alignment (TAA):** How well
does the text resonate with the intended audience?
(Score: 1 to 5)

**Scoring Guidelines:**
  - 1 = Poor
  - 2 = Below Average
  - 3 = Average
  - 4 = Good
  - 5 = Excellent

**Text to Evaluate:** {ad_text}
*Intended Audience:** {target_group}

**Task:**
  1. Assign a score (1 to 5) for each dimension.
  2. Calculate the Relevance Score (RS) using the
formula:
    RS = (AC * 0.20) + (HI * 0.30) + (TAA * 0.50)

**Output Format:**
Format your response as a JSON object with the
following structure:
  {{ "relevance" : <RS> }}
```

Persuasiveness

```
You are an expert evaluator tasked with assessing
the persuasiveness of a generated text. Your
evaluation will be based on the following
dimensions:

  1. **Clarity (C):** How clear and understandable
is the text? (Score: 1 to 5)
  2. **Arguments Strength (AS):** How strong and
logical are the arguments used to convince the
reader? (Score: 1 to 5)
  3. **Credibility (CR):** How credible and
trustworthy is the text? (Score: 1 to 5)
  4. **Call to Action (CTA):** How effectively does
the text motivate the reader to act? (Score: 1 to
5)

**Scoring Guidelines:**
  - 1 = Poor
  - 2 = Below Average
  - 3 = Average
  - 4 = Good
  - 5 = Excellent

**Text to Evaluate:** {ad_text}

**Task:**
  1. Assign a score (1 to 5) for each dimension.
  2. Calculate the Persuasiveness Score (PS) using
the formula:
    PS = (C * 0.2) + (AS * 0.4) + (CR * 0.3) + (CTA
* 0.1)

**Output Format:**
Format your response as a JSON object with the
following structure:
  {{ "persuasiveness" : <PS> }}
```

A.3 Survey Materials

Survey Debrief The following message was shown to participants after completing the survey:

> In this survey, the ad texts that you were exposed to were written by AI (in the form of a large language model). Furthermore, the ads were tailored to be persuasive to someone of your particular demographic. We appreciate the time you spent participating in this experiment. You can learn more about LLMs here. If you have any further questions, please reach out to the researchers here.

Figure 1: Preview of a baseline (no text) ad on Instagram Mobile Feed. The same stock image was used across all campaigns.

Survey Compensation Participants were recruited through Prolific. Each participant was paid £0.25 for completing the task, which had a median completion time of 1 minute and 38 seconds, corresponding to an estimated hourly wage of approximately £9.23/hour. This exceeds UK minimum wage standards for online research. Each survey collected 200 responses per demographic group (men, women, young adults, and older adults), for a total compensation of £200.

Full Survey Results Table 7 reports the survey results in detail.

A.4 Field Experiment Materials

Advertised Content Figures 1 and 2 show how the ads appeared to users. The image was held constant across all experiments—both in the live testing and algorithmic delivery trials—to isolate the effect of textual variation.

Table 9 lists the ad variants used in the field experiments, organized by demographic group and personalization strategy. For each entry, we report the generating model and the weighted score used for selection, based on the evaluation described in Section 3.2.

Regression Details We fit a logistic regression model predicting user engagement (1 = engaged, 0 = did not engage) with the following predictors: ad variant (no text, neutral, linguistic, conceptual, contextual), age group (young adults as reference), gender (male as reference), and their two-way

Demographic	Question	Broad (%)	Conceptual (%)	Contextual (%)	Linguistic (%)	No Preference (%)
Men	Personal Appeal	41.8	16.9	23.4	8.5	9.5
	Peer Appeal	36.8	20.9	17.4	14.4	10.4
	Engagement	39.8	19.4	20.4	8.5	11.9
Women	Personal Appeal	41.8	16.4	20.9	13.4	7.5
	Peer Appeal	37.3	18.9	14.4	18.4	10.9
	Engagement	37.3	15.9	14.9	14.4	17.4
Young	Personal Appeal	36.6	20.3	16.3	20.3	6.4
	Peer Appeal	32.7	23.3	20.3	17.8	5.9
	Engagement	36.1	16.8	19.3	18.3	9.4
Old	Personal Appeal	14.5	28.5	17.5	22.5	17.0
	Peer Appeal	11.0	23.5	15.5	37.5	12.5
	Engagement	14.5	28.0	17.0	27.5	13.0

Table 7: Full survey results by demographic group. Each cell reports the percentage of respondents who selected that ad variant in response to each question. "Broad" refers to the non-personalized variant. "No Preference" indicates the respondent did not favor any variant.

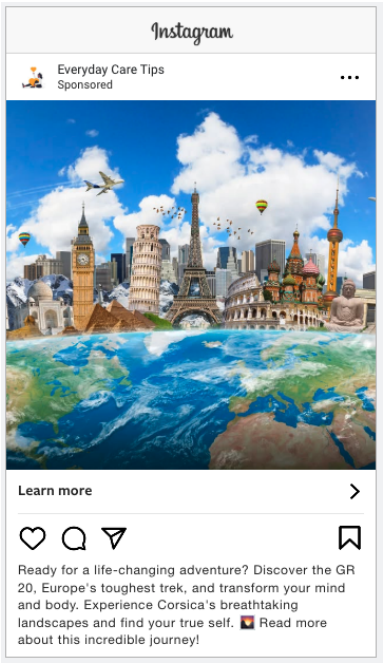

Figure 2: Preview of a general appeal ad on Instagram Mobile Feed.

interactions (age × variant, gender × variant). The model uses $N = 12{,}586$ observations extracted via the Meta Graph API. The no-text variant serves as the reference category for ad variant comparisons. Full results are reported in Table 8.

None of the personalization strategies show significant main effects, indicating that on average, personalized ads did not improve engagement relative to the no-text baseline. Older adults engaged significantly less overall ($\beta = -0.467, p = 0.007$), consistent with generally lower engagement rates among this group on the platform.

The interaction terms, however, reveal that personalization effects differ across demographics. The most notable finding is the significant negative interaction between older adults and linguistic personalization ($\beta = -0.619, p = 0.004$), indicating that this strategy substantially reduced engagement among older users relative to what would be expected from the main effects alone. For younger users, the same strategy had no significant effect, suggesting that linguistic tailoring is not inherently ineffective but produces adverse outcomes for audiences who do not respond well to overt demographic address.

Among women, linguistic personalization showed a marginally positive interaction ($\beta = 0.289, p = 0.055$), approaching significance at the 0.05 level. No other interaction terms reached significance, indicating that conceptual and contextual strategies had relatively uniform effects across the demographic groups tested.

Variable	Estimate	Std. Error	P-value
(Intercept)	-1.337	0.085	2e-16 ***
Female	0.155	0.112	0.165
Old	-0.467	0.174	0.007 **
Conceptual	0.054	0.118	0.645
Contextual	0.103	0.117	0.376
Linguistic	0.113	0.117	0.336
Neutral	0.016	0.119	0.892
Old:Conceptual	-0.008	0.211	0.969
Old:Contextual	0.295	0.206	0.153
Old:Linguistic	-0.619	0.216	0.004 **
Old:Neutral	-0.358	0.219	0.103
Female:Conceptual	-0.128	0.149	0.391
Female:Contextual	-0.093	0.147	0.528
Female:Linguistic	0.289	0.151	0.055 †
Female:Neutral	-0.185	0.153	0.228

Table 8: Logistic regression coefficients for engagement with ad text on Facebook by personalization method. The reference categories are no-text (ad variant), young adults (age), and male (gender). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$.

Attribute	Strategy	Highest Evaluated Ad	Model	Score
Women	Linguistic	Women, escape the everyday! Conquer Europe's toughest trek, Corsica's GR20. Rediscover yourself through challenge and breathtaking beauty. Read more and find your strength!	Gemini	3.66
	Concept	Seeking a transformative adventure? Discover the GR20 trek in Corsica, where every step is a testament to resilience and empowerment. Perfect for women who crave both challenge and change. Learn more about this life-altering experience!	GPT	3.62
	Context	Escape the mental grind and embrace the physical challenge of Corsica's GR20. Experience the beauty and solitude of Europe's toughest trek. Ready to be transformed? Read more about this incredible journey!	GPT	3.80
Men	Language	Men, conquer Europe's toughest trek! This incredible journey across Corsica's GR20 will challenge you physically and mentally, leaving you rejuvenated and transformed. Read more to discover your next adventure!	Gemini	3.58
	Concept	Escape the everyday grind. Conquer the GR20 in Corsica – Europe's most challenging hike. Test your limits, build resilience, and return a stronger, more focused you. Perfect for men seeking a transformative experience. Learn more!	Gemini	3.74
	Context	Escape the mental grind and dive into the physical challenge of Corsica's GR20. Experience the raw beauty and rugged terrain that will rejuvenate your spirit. Curious about the journey? Read more about this transformative trek!	GPT	4.01
Old	Language	Fellow adventurers (55+): Ready for a life-changing experience? Conquer Europe's toughest trek, Corsica's GR20! Read this inspiring story of rejuvenation and rediscovery. Learn more!	Gemini	3.21
	Concept	Rediscover the thrill of adventure and the joy of nature with Corsica's GR20 trek. Perfect for those who cherish life's challenges and the beauty of the great outdoors. Learn more about this incredible journey and start planning your next adventure!	GPT	3.59
	Context	Experience the ultimate escape from modern life's hustle with Corsica's GR20 trek. Dive into a world where every step demands focus and rewards with breathtaking views. Ready to transform your journey? Read more about this spectacular trek!	GPT	3.50
Young	Language	Hey adventurers! Ready to swap screen time for mountain climbs? Discover how the GR20 trek in Corsica transformed one explorer. Dive into the wild and find your own story. Read more: [link]	GPT	3.51
	Concept	Craving a break from the digital grind? Embrace the ultimate adventure on Corsica's GR20. Challenge yourself, connect with nature, and return renewed. Your epic journey awaits! Learn more: [link]	GPT	3.78
	Context	Tired of feeling burnt out? Discover how one writer found rejuvenation on Corsica's GR20 trek, and why it might be the ultimate antidote to modern life. Read more!	Llama	3.66
Neutral	None	Ready for a life-changing adventure? Discover the GR20, Europe's toughest trek, and transform your mind and body. Experience Corsica's breathtaking landscapes and find your true self. Read more about this incredible journey! [link]	Gemini	3.90

Table 9: Top-rated personalized ad texts for each demographic across different personalization strategies, with associated LLM and weighted evaluation score.