Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise

Zenan Chen[1] and Jason Chan[1]

[1]Carlson School of Management, University of Minnesota

Abstract

Since the launch of ChatGPT in Dec 2022, Large Language Models (LLMs) are rapidly adopted by businesses to assist users in a wide range of open-ended tasks, including creative work. While the versatility of LLM has unlocked new ways of human-AI collaboration, it remains uncertain how LLMs should be used to enhance business outcomes. To examine the effects of human-LLM collaboration on business outcomes, we conducted an experiment where we tasked expert and non-expert users to write an ad copy with and without the assistance of LLMs. Here, we investigate and compare two ways of working with LLMs: (1) using LLMs as "ghostwriters," which assume the main role of content generation task and (2) using LLMs as "sounding boards," to provide feedback on human-created content. We measure the quality of the ads using the number of clicks generated by the created ads on major social media platforms. Our results show that different collaboration modalities can result in very different outcomes for different user types. Using LLMs as sounding boards enhances the quality of the resultant ad copies for non-experts. However, using LLMs as ghostwriters did not provide significant benefits and is in fact detrimental to expert users. We rely on textual analyses to understand the mechanisms and learned that using LLMs as ghostwriters produces an anchoring effect which leads to lower-quality ads. On the other hand, using LLMs as sounding boards helped non-experts achieve ad content with low semantic divergence to content produced by experts, thereby closing the gap between the two types of users.

1 Introduction

Large Language Models (LLMs), a subcategory of generative artificial intelligence (AI), represent significant advances in the domain of text generation. Because of their ability to produce remarkably coherent natural responses to user prompts by predicting subsequent wording within textual

sequences, LLM applications have instruction-following capabilities (Ouyang et al. 2022) that provide human users with unprecedented support for a wide array of open-ended and non-routine creative tasks that were beyond the capabilities of other AI counterparts (Bubeck et al. 2023). Not surprisingly, the investment in generative AI startups witnessed considerable growth in the first quarter of 2023, with total funding of $10.7 billion.[1] To harness the power of LLMs, companies have begun to integrate LLMs into their core operations. For instance, Coca-Cola has formed a partnership with the consulting firm Bain & Company to integrate ChatGPT to assist with marketing operations that involve creative work.[2] In addition, companies like Jasper and Copy.ai are providing LLM-based content generators to assist with creative tasks (e.g., advertisement copywriting), with the belief that the use of these AI tools can help to complete tasks significantly faster (Eloundou et al. 2023). Given the rapid and widespread adoption of LLMs, a pressing managerial concern pertains to gaining a better understanding of how best to harness the power of LLMs for enhancing business outcomes.

While the past work provides valuable insights into human-AI collaboration, it falls short of illuminating how best to utilize LLMs in work domains that were underserved by previous generations of AI. Past research on human-AI collaboration has focused mainly on rule-based systems or machine-learning models that are designed for specific, standardized, and codifiable tasks (e.g., classification) (Fügener et al. 2021, Bauer and Gill 2023). Although insightful, the codifiable tasks studied previously were quite different from the tasks that LLMs are capable of assisting with. Specifically, LLMs can facilitate users with a wide range of open tasks (e.g., question answering, customer support, summarization), and creative tasks (e.g., writing poetry, songs, and marketing messages). To this end, there is a need to systematically research whether and how human-LLM collaboration would enhance work performance, especially in the context of creative tasks. Initial academic inquiries into this question pave the way for future research that explores the integration of LLMs into businesses and its impact on the workforce of the future.

Answers to the impact of human-LLM collaboration on work performance are shrouded in uncertainty due to the presence of competing theoretical perspectives. On the one hand, the quality of LLM-generated content should arguably reflect the combined creativity across a broad and di-

[1] https://pitchbook.com/news/articles/Amazon-Bedrock-generative-ai-q1-2023-vc-deals
[2] https://www.bain.com/vector-digital/partnerships-alliance-ecosystem/openai-alliance/

verse set of individuals, as they are trained with a massive amount of data from a wide range of domains (Bubeck et al. 2023). Thus, the creative capabilities of LLMs would arguably surpass that of a single individual (or a small team). However, due to their autoregressive nature, LLMs may be limited to generating expected outputs as they are trained to adhere to existing data distributions (Bunescu and Uduehi 2019).

Hence, a first-order question pertains to whether the use of generative AI would lead to better business outcomes for creative tasks, and for what situations. With this main guiding question, it follows that we need to understand the optimum way of utilizing LLMs. Given that LLM-based applications have only been recently introduced to the public, practitioners and businesses lack guidelines on how best to use them. One of the most basic ways of using LLMs is to employ them as "ghostwriters" and simply let them take on the bulk of the work. Here, users could instruct these conversational AI to generate an initial set of content to kickstart the creative process. Based on this generated initial content, users can then iteratively instruct LLMs to re-generate the content to improve certain aspects until a satisfactory output is produced. Such an approach could be an effective way for LLMs to overcome challenges such as "writer's block."

However, it is known that algorithms and machine learning techniques can impose anchoring effects on humans in decision-making contexts (e.g., Jacowitz and Kahneman 1995, Fügener et al. 2021, Bauer and Gill 2023). Consequently, having LLMs take the lead in the creative process may not be optimal. In situations where creativity is a crucial component of the output quality, the anchoring effect can be especially salient as it limits the creativity of the final output derived (Berg 2014).

An alternative approach entails employing LLM as a "sounding board," in which the LLMs are prompted to offer feedback and assessments of content created by human users. Here, the AI does not assume the main role in the creative process but instead serves to evaluate and critique users' creative work. Past literature has shown that feedback from an external agent can enhance performance and creativity (Zhou 2003). Nonetheless, the quality of the LLM-generated feedback and the extent to which users can effectively integrate such feedback into their work remain uncertain. In particular, users may not fully exploit the benefits of AI because of difficulties in evaluating the correctness of AI advice (Jussupow et al. 2021).

Finally, to set the effects of collaboration modality in reality, one needs to also consider the

users who work with the LLMs. Specifically, the use of information technology is documented to produce differential impacts between skilled and unskilled workers (e.g., Akerman et al. 2015). Thus, understanding whether the various collaboration modalities could have differential impacts on workers of varying skill levels would provide more precise and holistic guidelines for utilizing LLMs.

To address these gaps in our understanding, we designed a randomized experiment wherein participants were tasked with crafting advertisement copies for a common consumer product, under one of three conditions: a ghostwriter modality group, a sounding board modality group, and a control group. We developed a customized LLM-based (GPT-4) interface that operationalizes the abovementioned collaboration modalities to be used in our experiment. The ghostwriter group is only allowed to use the LLM interface to generate ad copies, whereas the sounding board group receives feedback from our LLM interface about their ad copies. The control group did not receive any AI assistance.

We adopt a commonly used definition, ad effectiveness, as measured through ad clicks, to define the quality of an ad (Naik et al. 1998, Agarwal and Mukhopadhyay 2016, Kim et al. 2019). To obtain an objective real-world assessment of the effectiveness of the created ads, these ads were deployed on actual social media ad campaigns, to which the resultant number of clicks generated by each ad is captured and used as a dependent variable.

Our results provide several intriguing insights. We find that modality and user expertise jointly determine the performance benefits of LLM usage. In particular, the use of LLMs as sounding boards helped non-experts to create significantly higher-quality ads compared to the non-experts who did not have access to the LLM interface. This improvement is however not observed for experts. Interestingly, we learn that the ghostwriter modality causes experts to perform worse compared to experts in the control group. Such a negative effect is not detected among non-experts.

A set of follow-up textual analyses were used to explore the possible underlying mechanisms. We find that the use of LLMs as a sounding board led non-experts to write ads that are more similar to that of experts. However, the ghostwriter modality results in anchoring effects, wherein users in this modality are less likely to create ads that deviate from initial LLM-generated outputs which ultimately yield outputs of lower variety compared to users in the sounding board condition. We further learned that while the use of LLM improves the execution process of the ad creation process

(through the sounding board modality), it does not enhance the creativity aspects of the created ad.

We contribute to the growing body of literature on human-AI collaboration (e.g., Fügener et al. 2022, Ge et al. 2021, Bauer et al. 2023), particularly towards the use of LLMs. This line of research on LLMs mostly focuses on evaluating the effects of their adoption by comparing differences between the usage and non-usage scenarios (Noy and Zhang 2023, Brynjolfsson et al. 2023, Dell'Acqua et al. 2023). We add depth to this literature by assessing whether and how the usage modality of LLMs can influence work outcomes. This is especially important in the context of LLM technology, as this powerful AI is flexible enough to be used in different ways (Dell'Acqua et al. 2023), to which the optimum usage modality under different conditions needs to be better understood in order for businesses to derive value from its deployment.

By shedding light on the differential effects of these modalities on experts and non-experts, we also contribute to a stream of literature on the heterogenous effects of AI adoption on workers (Akerman et al. 2015, Brynjolfsson et al. 2023, Lysyakov and Viswanathan 2022). Not only do our results provide practical implications on optimum collaboration modalities for different worker types, but they also shed light on the disruptive potential of LLMs on the labor market in creative domains.

2 Background

2.1 Large Language Models

Generative AI, particularly Large Language Models (LLMs), has recently gained significant attention. LLMs are an application of neural network models that process data sequentially (Bubeck et al. 2023). The training process primarily entails forecasting subsequent words in a sequence, drawing from their prior context. This pre-training process utilizes extensive text corpora to learn patterns from statistical word co-occurrences, enabling LLMs to create grammatically and semantically coherent new texts. More recent LLMs include an additional fine-tuning step to further "align" the LLM outputs with human preferences. For instance, a given prompt may yield numerous potential responses from an LLM, some of which might be factually inaccurate, biased, or less desirable. In the training process, human evaluators rank these responses to train a reward

function that prioritizes certain outputs over others. Besides improving the overall output quality, this type of refinement also allows the model to be more appropriately adapted to its specific use (Ouyang et al. 2022).

LLMs are found to simulate human-like thought processes in generating natural language output in economic (Horton 2023) and marketing settings (Brand et al. 2023). Moreover, the state-of-the-art LLMs possess attributes of General Intelligence, as signified by their remarkable performance in a wide range of tasks (e.g., coding, medicine, law, psychology, etc.) without task-specific training and instruction (Bubeck et al. 2023).

Current studies on human-LLM collaboration can be categorized into two main themes. First, a set of studies investigate the effect of LLM adoption in various tasks and contexts. They find that LLMs can effectively be used to enhance tasks such as writing press releases, short reports, analysis plans, emails (Noy and Zhang 2023), and in customer support settings that involve a stable product involving technical questions (Brynjolfsson et al. 2023). However, the use of LLM is found to be less effective for certain consulting tasks (Dell'Acqua et al. 2023). While these studies provide useful insights into the types of tasks that can be more effectively assisted by LLMs, there is still a lack of understanding on how best users should collaborate with LLMs. To this end, the second stream of research examines different ways of utilizing LLMs. Dell'Acqua et al. (2023) found that users exhibit different choices of using LLMs, such as integrating AI and human capabilities at a sub-task level versus strategically delegating sub-tasks to AI.

Although it is useful to know that users deploy different ways of collaborating with LLMs, it is of managerial importance to know the business implications of utilizing LLMs in different ways, and who would benefit most from these different collaboration modalities. If differences in performance ensue, it would also be prudent to understand the underlying mechanisms for such results.

Our work differs from past work in that we do not look at how users choose to work with LLMs, but instead focus on the downstream question of how different collaboration modalities can affect performance outcomes. In other words, while Dell'Acqua et al. (2023) focuses on the upstream aspect of LLM usage, our work completes the picture by showcasing the downstream consequences of utilizing LLMs under different collaboration modalities.

2.2 Creative Tasks

Creativity is considered one of the hallmarks of human intelligence, and it is required for a set of tasks that are critical in business operations. Creative work tends to involve two elements. The first element involves coming up with new ideas or inspirations and exploring new perspectives, i.e., the creative part of the creation process.

The second element of the creative process involves the use of skills and strategies to allow these creative ideas/perspectives to materialize, i.e., the instrumental aspects needed in idea execution. To give an example, tasks that involve creative writing (e.g., writing novels, poetry, lyrics, stories) require the writer(s) to have a good command of a language and writing skills in order to convey their ideas. As such, an outstanding novel should have an interesting plot (creative aspect) that is conveyed effectively through good writing (execution aspect). Similarly, for ad copy creation, a successful ad should aim to have an innovative angle of gaining the attention and piquing the interest of the audience (creative aspect) and effective means of communicating this pitch through a good choice of words (execution aspect).

An often cited creative task in business is advertising, which bears strong corporate interest due to its substantial market size, amounting to approximately 180 billion USD in 2022[3]. Well-designed advertisements (ads) bring about a range of important positive business outcomes, including increased ad recall, liking of the brands/products, and purchase intentions (e.g., Smith et al. 2008, Pieters et al. 2002). Consumers are constantly exposed to a plethora of ads in various forms and mediums. Thus, ads need to excel not only in their intrinsic quality but also in distinguishing themselves from a myriad of competing messages, as they vie for the finite attention resources of consumers. Therefore, crafting effective ads requires a heightened level of creativity, to successfully "break through the competitive clutter" (Pieters et al. 2002). Owing to the impressive ability of LLMs to produce coherent and meaningful text, new LLM-based services such as ChatGPT, Copy.ai, and Jasper are increasingly employed among marketers to help with the ad creation process (Davenport and Mittal 2022).

[3] `https://ark-invest.com/articles/analyst-research/digital-advertising-market/`

2.3 Anchoring Effect

While AI has the potential to enhance the performance of creative tasks (Di Fede et al. 2022, Stevenson et al. 2022), prior works have found that the usage of AI can negatively influence human outputs in prediction and decision-making contexts. One of the mechanisms present in Human-AI collaboration is the anchoring effect. These cognitive biases are characterized by the inclination of an individual's attitudes, behaviors, and beliefs to be influenced by a given reference point or anchor (Tversky and Kahneman 1974). The detrimental impact of anchoring bias on human performance happens when individuals do not adjust their beliefs from the initial anchor value towards a more subjectively plausible solution (Jacowitz and Kahneman 1995). Users can be anchored by the outputs of algorithms, which in turn influences their behavior (Bauer and Gill 2023). Similarly, anchoring effects may also influence the creative process. The novelty and usefulness of the final ideas often depend on the initial ideas generated at the beginning of creative tasks (Berg 2014), to which anchoring effects can adversely affect the creative outputs when humans rely on AI for initial ideas.

Because LLMs are trained to predict the most probable token given the preceding input tokens, they may generate homogenous content when given similar prompts (Jentzsch and Kersting 2023, Padmakumar and He 2023). As such, an over-reliance on LLMs in the creative process can potentially lead to the production of content that is expected and less distinctive. A potential consequence is that using LLMs as the main driver of the creative process may not produce desirable outcomes, especially when users are anchored to the outputs generated by LLMs. Within the advertising domain, ad content created under the ghostwriting modality may fall short in originality, and be unable to compete effectively for consumers' attention, resulting in a poorer advertising outcome (Smith et al. 2008, 2007, Pieters et al. 2002).

3 Experimental Setup

To investigate our research questions, we conducted an experiment using 355 participants recruited from Prolific, a widely accepted participant recruitment platform (Palan and Schitter 2018). To allow for greater result generalizability, participants from the United States and the United King-

dom were solicited. Standard eligibility criteria were used to pre-screen participants, which include maintaining a minimum task acceptance rate of 95% and possessing at least a college degree (Palan and Schitter 2018, Noy and Zhang 2023). Our study features a 3 (two human-LLM collaboration modalities and control group) $\times$ 2 (expert and non-expert) between-subject design. Power analysis showed that our sample size allows a small to medium effect size of Cohen's $d = 0.25$ to be detected.

In the experiment, participants were asked to complete an ad copywriting task. Specifically, they were tasked with creating advertisements for an iPhone protective case. The rationale behind the choice of this product is that it is a common consumer product that does not require industry-specific knowledge to perform the task, making most research participants solicited online adequate candidates. At the same time, the product has a good number of unique features that allow for sufficient flexibility for participants to exercise their creativity. To create the ad, participants are provided with detailed product descriptions and images (Figure A1a in appendix), along with a sample image of how advertisements will look on Facebook feeds (Figure A1b in appendix).

To obtain a balanced sample of experts and non-experts, we first utilize Prolific's screening tool to identify potential experts based on the industry they have worked. Since participants from marketing & sales employment sectors are more likely to bear marketing-related knowledge, roles, and responsibilities, we recruited half of our participants from marketing & sales employment sectors, and the remaining participants were recruited from other employment sectors. From this initial pool of participants, we further rely on a set of questionnaires to distinguish marketing experts from non-experts on a more granular level.

Expertise refers to experience and knowledge in a certain domain (Ericsson et al. 2018), to which experts in marketing have greater marketing-related knowledge than an average person (Arnett and Wittmann 2014). Thus, individuals with more marketing knowledge, skills, and experience would have higher expertise compared to individuals who had less exposure to marketing roles/tasks. Furthermore, considering the possibility that one may have past marketing experience but has since moved to other occupations, our measure of expertise needs to account for both past marketing experience and current job responsibility. With this in mind, we asked participants to report (1) whether their current job involves marketing-related responsibility (Yes or No), and (2) their prior experiences of crafting advertisement copies (No experience, Less than a year, 1-2 years,

3-5 years, 5 or more years), prior to them performing the experimental tasks.

Using these responses, we define experts in our study as individuals who answered "Yes" to the first question, or those having at least 1 year of experience in crafting advertisement copies. Through this categorization, individuals with current marketing responsibility and/or past experience are deemed as experts *vis-a-vis* individuals who have little to no experience in marketing at all.

In addition, we have collected human-rated perceived quality of their advertisement copies. Specifically, we measured an array of items that are shown to be important factors of ad effectiveness, namely perceived informativeness (Lee et al. 2018), perceived positive affect (product interests) (Watson et al. 1988), and purchase intention (Teixeira et al. 2014).

Utilizing this dependent variable, we conducted a manipulation check. We found that experts in the control group produce ads with higher perceived quality than non-experts in the control group, wherein the average item ratings of expert-written ads (*M=6.084, SD=2.027*) are about 25% higher than non-experts-written ads (*M=4.854, SD=2.027*), $p < 0.01$.

3.1 Conditions

Participants were randomly assigned to three conditions: (1) utilizing AI as a ghostwriter for writing the advertisement text; (2) utilizing AI as a sounding board to solicit AI-generated feedback; and (3) no AI technology, which serves as a control group. We choose the state-of-the-art GPT-4 as the foundational LLM used in our LLM interfaces. The model parameters are set to be identical to the default parameters (i.e., temperature = 1.0 as of April 2023) used on OpenAI's consumer-facing web interface, as this is likely to be the most widely used setting.

To operationalize the different usage modalities, we developed two separate LLM interfaces and manipulated the LLM's behavior via carefully engineered system prompts, which were not visible to the participants. In the sounding board condition, the LLM is prompted to solicit ad copy drafts from participants and would subsequently offer feedback for their ad copies. We designed this LLM to deny any user requests that directly ask it to produce ad copies. In the ghostwriter condition, the LLM is prompted to compose ad copies in response to user instructions. Specific prompts used in the design of the LLMs are located in Appendix A.2. We tested the two interfaces extensively to ensure that user requests do not produce unintended outputs. In a *post hoc* analysis, we

reviewed all chat histories and did not find any unintended use of the LLM (e.g., "jailbreaking"). [4] In the No-AI control condition, participants do not have access to the LLM interface entirely.

For both the ghostwriter and sounding board conditions, the participants are allowed to use the LLM as much (or as little) as they want. Participants are instructed to submit the advertisement copy once they are satisfied with the output. Users in the ghostwriter condition are allowed to edit the ads manually prior to submission. These design considerations mirror real-world situations wherein users have access to AI technologies, and are allowed flexibility on the extent of utilization. To further preserve the integrity of the experiment, the use of external AI technologies is prevented to which copying and pasting functionality is confined exclusively within the experiment platform. We also capture records of all messages exchanged between participants and their assigned interfaces, providing data for subsequent textual analyses.

3.2 Familiarizing with the LLM Interface

To minimize differences in experience with LLM usage, participants were to complete a training task closely resembling the main task, albeit without the submission of advertisement copy prior to the main copywriting task. The training task familiarizes all participants with the LLM interface. Participants are required to send at least one message to their respective LLM. The LLM interface deployed in the training task is similar to the participant's assigned condition.

To account for any potential extraneous effects due to the training procedure (i.e., subject fatigue), participants in the control group also interacted with an LLM interface in the training session. For this purpose, control participants are randomly matched to either the ghostwriter or sounding board LLM. In this manner, we ensure that participants across all conditions receive the same pre-experiment stimuli preceding the main writing task, effectively mitigating any potential carry-over effects.

3.3 Incentives

A two-part incentive structure was implemented to encourage participants to exert their best efforts in writing high-quality ad copies. Participants received a base payment of $5, with additional rewards up to $3 based on ad performance rankings within their condition, measured by their ad

[4] We include additional manipulation checks in Appendix A.5.

clicks.[5] This incentive structure is prominently displayed throughout the task, which reminds the participants of the competing nature of the industry. In our instructions, we further highlight that generic advertisement copies are unlikely to fare well with more carefully constructed and creative advertisement copies by other participants (e.g., Pieters et al. 2002).

3.4 Measurements

We collected a range of demographic information (gender, age, and highest academic degree earned) and experience from the participants at the beginning of the study, allowing us to control for potential influence between these factors and the outcomes of the tasks. Responses were measured using a five-point Likert scale. In addition, we inquired whether their job responsibilities include marketing tasks and their previous job exposure to marketing.

Following each task, we asked participants to estimate the percentile rank of their advertisement among all the participant-created advertisements, using a slider ranging from 0% (worst) to 100% (best). This approach has been used in the literature as a measure of evaluating an individual's self-assessment of their performance (Brandts et al. 2015). We also conducted an exit survey with our treated participants, asking them to provide comments about their experience with the LLM.

3.5 Advertising Procedure

To obtain an objective measure of ad performance, ad copies produced by study participants were deployed as advertisements on social media platforms. Since the ad copy (i.e., the ad message) is the only component that varies between different ads, Facebook and Instagram are particularly suitable advertising platforms, given that their newsfeed advertisement format allocates considerable space for text content. A dedicated advertising campaign was allocated for each participant, resulting in 355 campaigns. Each campaign contains a single ad with the ad copy crafted by the participant along with the same set of product images.

It is plausible that the same viewer may see multiple ads crafted by different experiment participants because of the retargeting algorithm, thereby obfuscating the contribution of each ad. To eliminate this possibility, we randomly assigned each ad to a unique US county, so that each

[5] See Appendix A.3 for more details.

Table 1: Summary Statistics

		No-AI (N=117)		Sounding board (N=122)		Ghostwriter (N=116)	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Age (segment)		3.0	1.2	3.1	1.4	2.8	1.4
Education		2.2	0.6	2.2	0.6	2.3	0.6
Ad Clicks		16.2	6.6	16.4	6.2	13.0	5.0
Duration (mins)		6.1	4.2	12.7	8.7	6.0	4.6
# of Messages				2.8	1.5	2.6	1.7
Estimated Rank		46.8	23.6	57.1	22.9	56.0	20.6
		N	Pct.	N	Pct.	N	Pct.
Gender	Female	49	41.9	54	44.3	47	40.5
	Male	67	57.3	65	53.3	68	58.6
	Prefer not to say	1	0.9	3	2.5	1	0.9
Expertise	Non-Expert	57	48.7	67	54.9	67	57.8
	Expert	60	51.3	55	45.1	49	42.2
Emp. Sector	Marketing & Sales	61	52.1	62	50.8	53	45.7
	Other	56	47.9	60	49.2	63	54.3

viewer would be exposed to exactly one ad from the experiment. The county selection process and randomization check are described in detail in Appendix A.4.

The advertisements from all three conditions were launched over six days including weekdays and weekends to account for day-specific effects. The same number of ads from each study condition was launched on each day of the experiment to account for potential differential day-of-week effects. Each ad was allocated the same amount of advertising budget, such that they get the same amount of exposure.

3.6 Randomization & Manipulation Checks

Table 1 provides the summary statistics derived from the experimental data. One-way ANOVA tests for the age, gender and education levels revealed no significant variations ($p > 0.1$) across the treatment groups for these demographic features, indicating that randomization is successful.

Our main manipulation lies in the modalities of human-LLM collaboration. We manually reviewed all the chat histories, we found that our manipulation was successful in that participants interacted with the LLM in the intended fashion.

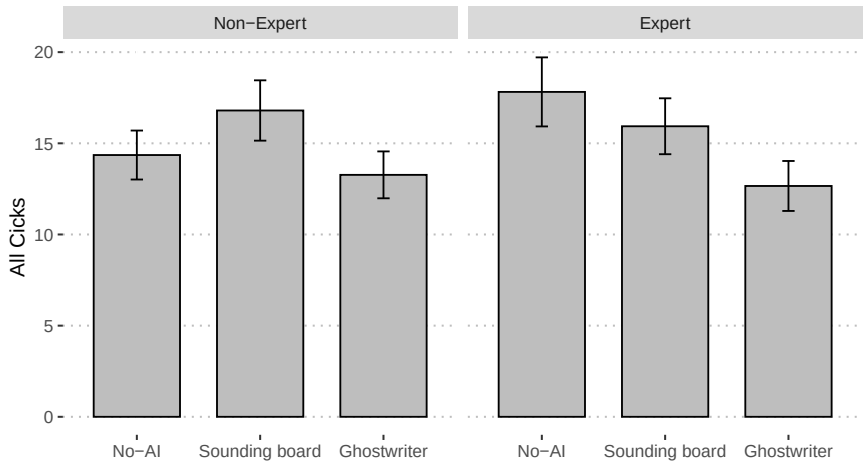
4 Results

4.1 Effects on Advertisement Clicks

The advertisement quality is defined as the effectiveness of the ad in fulfilling its intended purpose (Agarwal and Mukhopadhyay 2016, Naik et al. 1998). Thus, the effectiveness of an online ad is ultimately measured by how well it fulfills its purpose of bringing users to the marketer's product homepage (Agarwal and Mukhopadhyay 2016, Kim et al. 2019, Naik et al. 1998). Following the best practice in the literature, we measure the quality of the ad copy using the number of consumer responses to the ad, i.e., ad clicks.

The model-free evidence in Figure 1 shows the average number of ad clicks across conditions and user expertise. Linear regressions that control for age, gender, and education produced consistent results (Appendix A.7). Results from the Poisson regressions (see Appendix A.8) are consistent with the results from linear regressions.

Figure 1: Mean Ad Clicks by Conditions and User Expertise (95% CI)



A two-way interaction effect was observed, in which (1) non-experts benefit from AI utilization but only in the sounding board condition, while (2) experts do not benefit from AI use, regardless of the modality, with a decrease in performance observed in the ghostwriter modality.

Table 2: Heterogenous Effects of AI Usage on Estimated Rank

	DV: Estimated Rank (%)			
	All		Sounding board	Ghostwriter
	(1)	(2)	(3)	(4)
(Intercept)	34.735***	33.078***	34.184***	30.108***
	(3.806)	(4.905)	(5.215)	(5.988)
Condition [Sounding board]	10.587***	12.298**	12.360**	
	(3.029)	(4.516)	(4.561)	
Condition [Ghostwriter]	9.331**	10.761**		10.620*
	(2.948)	(4.106)		(4.122)
Expertise [Expert]	−0.882	1.246	1.373	0.559
	(2.759)	(4.629)	(4.767)	(4.744)
Condition [Sounding board] × Expertise [Expert]		−3.481	−3.754	
		(6.080)	(6.111)	
Condition [Ghostwriter] × Expertise [Expert]		−2.934		−3.048
		(5.928)		(5.939)
Num.Obs.	355	355	239	233
R2	0.072	0.074	0.072	0.086
R2 Adj.	0.045	0.041	0.031	0.049

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Robust standard errors in parentheses. All models controlled for age, gender, and education.

4.2 Effects on Self-assessment of Performance

Regressions in Table 2 examine the effects of AI usage on the estimated rank of their advertisement performance. The results show that users in both modalities significantly deemed their advertisement performance to be higher than the control group (10.6% for the sounding board group and 9.3% for the ghostwriter group), suggesting that the users of LLM have more optimistic self-evaluations. However, we note that this higher self-confidence is not always translated to better advertisement performance as seen in our main result.

4.3 Mechanism

Our analysis so far has found significant differential effects of human-LLM collaboration modalities on ad quality. We next conduct a series of analyses to examine the mechanisms leading to these results.

4.3.1 Effect of Collaboration Modality on Semantic Divergence

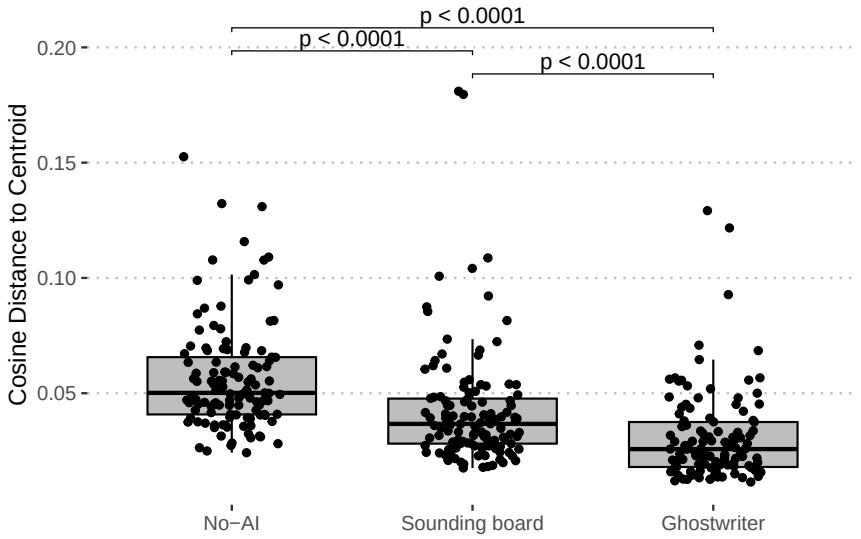
Given that similar prompts to LLMs can result in homogenous outputs (Padmakumar and He 2023), a key potential explanation for the inferior performance of advertisement copies generated through the ghostwriter modality is the presence of a potential anchoring effect, wherein the initial draft produced by the LLM may serve as an anchor that reduces one's capacity to think deeper and out of the box to improve their ad messages (Berg 2014), making the final ad copy semantically closer to LLM-generated first drafts (i.e., the anchor). Semantic divergence captures the average distance an ad is from the rest of the ads within the same condition. In other words, it is a variance measure of the ad content within its condition, which can reveal the presence of divergent thinking and originality (Orwig et al. 2021, Olson et al. 2021). With a lower semantic divergence, users within ghostwriting conditions tend to generate advertisement copies that are more similar to one another in the same condition. Ads that read in a generic or expected fashion are unlikely to fare well, as they do not "pop" to the audience.

Should an anchoring effect be present in the ghostwriter group, the final advertisement copies from the ghostwriter group would exhibit a lower semantic divergence compared to other groups. To test the presence of the anchoring effect, we first conduct a group-level analysis of the textual semantic distances. Text semantics similarity can be obtained by calculating the cosine distances between text embeddings, which are high-dimensional representations of the semantic meaning. A minimal cosine distance between two text embeddings reflects high semantic similarities. To obtain these embeddings, we employ the "text-embedding-ada-002", a text embedding model that is shown to have state-of-the-art ability for capturing semantic meanings and accessing similarities (Muennighoff et al. 2023). Group-level semantic divergence is then measured by comparing the cosine distances between each text embedding and the group centroid. We detail our methods in Appendix A.9.

Figure 2 visualizes the semantic divergences of advertisement copies from the three study conditions. Here, the control group has the largest semantic divergence for its ads (M=0.057, SD=0.024), indicating that users without any LLM inputs tend to produce ad copies that differ from one another within the group the most. In contrast, the ghostwriter group has the lowest semantic divergence (M=0.031, SD=0.020) across all three groups, suggesting that anchoring effects

are highly salient among participants in the ghostwriter group.

Figure 2: Effect of LLM on Semantic Divergence



We note that the sounding board group has a lower semantic divergence (*M=0.043, SD=0.026*) than the control group, but significantly higher than the ghostwriter group (Figure 2). The successful use of creativity in advertising contexts requires not only originality but also relevance and appropriateness (Smith et al. 2007), as irrelevant or inappropriate advertisements are counterproductive. Considering that non-experts in the sounding board group are creating more effective ads compared to the non-experts in the control group, a lower semantic divergence for this group of participants may simply be due to non-experts reflecting on the feedback to create more relevant and appropriate ads compared to their original ads.

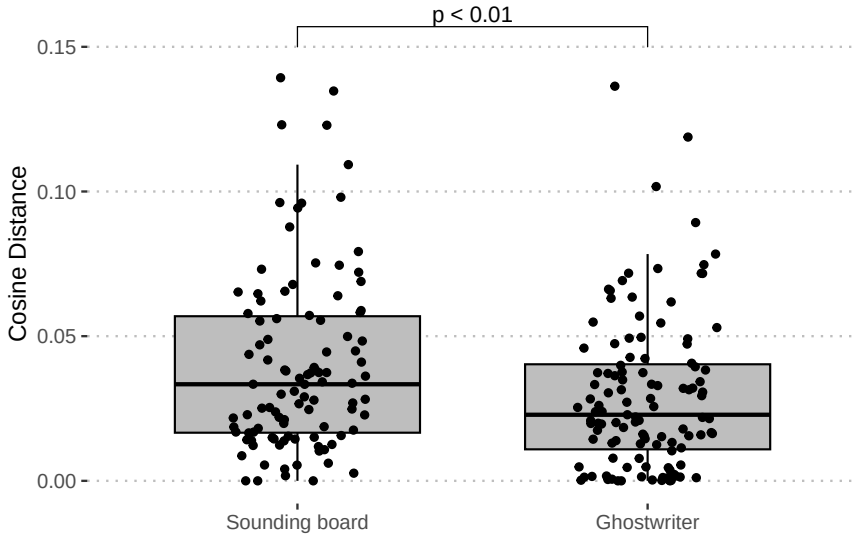
4.3.2 Effect of Collaboration Modality on Extent of Draft Revision

The anchoring effect can also be observed in the ad creation process. We further performed a within-subject comparison of the text embeddings to understand how much participants in the two LLM treatment groups (i.e., ghostwriter and sounding board) have revised their advertisement copies throughout the creation process. Out of 238 total participants in these two treatment groups, 25 participants did not include a first draft in the chat history[6], leaving us with 213 participants.

[6]For instance, some participants did not ask the LLM to provide further advice (sounding board group) or did not use LLM (ghostwriter group).

For each participant, the extent of revision can be measured by the semantic distance between the first draft and the final copy, wherein a lower semantic distance means that the participant made fewer revisions to their drafts. Figure 3 shows that users in the sounding board modality revised their copies to a greater extent (*M=0.040, SD=0.031*) relative to users in the ghostwriter modality (*M=0.029, SD=0.026*), $p < 0.01$.

Figure 3: Semantic Distance between First Draft and Final Copy



It is possible that using LLM as a ghostwriter may produce a more complete and high-quality first draft, reducing the need for further revisions. To assess this possibility, we obtained human-rated perceived quality of the first drafts with new participants on Prolific (see Appendix A.11). Table 3 presents the regression results controlling for the subjective perceived quality. In model 2, as expected, when the first draft has a higher initial perceived quality, the participants generally make fewer revisions ($\beta = -0.006, p < 0.01$). Even after controlling for the perceived quality of the initial draft, the collaboration modality remains a significant factor, wherein the participants in the ghostwriter group still revised less than the participants in the sounding board group ($\beta = -0.009, p < 0.05$). This underscores the significance of collaboration modality on the anchoring effect and provides another evidence for the presence of the anchoring effect in the ghostwriter group.

Table 3: Effects of Modality on Revision

	Degree of Revision (Cosine Distance)		
	(1)	(2)	(3)
(Intercept)	0.037***	0.070***	0.075***
	(0.009)	(0.014)	(0.014)
Condition [Ghostwriter]	−0.009*	−0.009*	−0.013*
	(0.004)	(0.004)	(0.005)
First Draft Perceived Quality		−0.006**	−0.007***
		(0.002)	(0.002)
Expertise [Expert]			0.001
			(0.006)
Condition [Ghostwriter] × Expertise [Expert]			0.010
			(0.008)
Num.Obs.	213	213	213
R2	0.041	0.102	0.119
R2 Adj.	0.009	0.066	0.076

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Robust standard errors in parentheses. All models controlled for age, gender, and education. Here, the sample only contains the two treatment conditions, wherein the dummy variable for the reference group (Sounding board) is omitted. The dummy variable for the Ghostwriter group thus encodes the effect of Ghostwriter relative to Sounding board on the dependent variable.

4.3.3 Advertisement Copy Difference between Experts and Non-experts

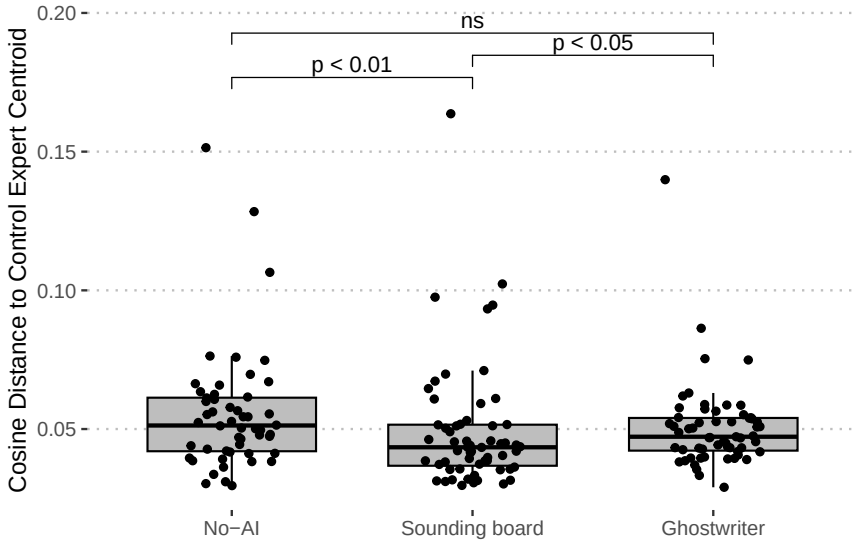
A potential benefit of LLMs is that they are trained on massive unlabeled web pages (OpenAI 2024), which contain marketing or advertisement copywriting techniques. These techniques, frequently employed by experienced marketers, may not be familiar to novices who lack relevant domain knowledge. Using either collaboration modality may be more helpful for non-experts, by getting them to incorporate them in ad copies.

Experts in the treatment groups are influenced by LLMs and are not considered in this analysis. In contrast, Here, we consider experts in the control group as the reference point. For each study group, we calculate the semantic distance between non-experts and the reference experts. We document the comparison process in Appendix A.10.

As non-expert users rely on LLM-generated ad copies in the ghostwriter group or act on LLM-generated feedback in the sounding board group, they would be producing ad copies that are more similar (compared to the non-expert users in the control group) to those that are produced by reference experts. Should this hold, the usage of LLM reduces the semantic difference between

experts in the control group and non-experts.

Figure 4: Semantic Distances between Control Group Expert Centroid and Non-experts



Comparisons of the cosine distances (Figure 4) reveal that non-experts in the sounding board group (*M=0.049, SD=0.022*) produced advertisement copies that are semantically more akin ($p < 0.01$) to those written by reference experts, compared to the outputs generated by non-experts in the control group (*M=0.055, SD=0.022*) and ghostwriter group (*M=0.050, SD=0.016*). This result suggests that using LLM as a sounding board helps non-experts to write more like experts.

4.4 Textual Characteristics

To gain further insights into the influence of collaboration modality on the created ad, we analyzed various textual characteristics of the produced ads. The analysis of textual characteristics can give us further insights into the main results. Specifically, we examined 1) sentiment (looking at both Polarity and Subjectivity), 2) readability (measured by the Gunning Fog Index, where a higher index indicates lower readability), 3) ad length (the number of words/tokens), 4) extent of emoji usage (the number of emojis used) and 5) extent of hashtags usage (the number of hashtags used).

We regress ad clicks on these independent variables to understand the relationship between these ad characteristics and ad effectiveness. Since the distribution of the number of tokens is skewed, we take its log value in our regression.

In Table 4, we see that ads with more subjective sentiments—those expressing more personal

Table 4: Predictors of Ad Performance

	Ad Clicks
	(1)
(Intercept)	2.440***
	(0.116)
Subjectivity	−0.209*
	(0.097)
Polarity	0.114
	(0.094)
Gunning Fog	−0.020***
	(0.005)
Emoji	−0.033***
	(0.004)
log(token)	0.147***
	(0.026)
Hashtag	0.033
	(0.021)
Num.Obs.	355
Log.Lik.	−1168.908
F	14.820

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes. Standard errors in parentheses.

opinions, emotions or judgments—tended to receive fewer clicks compared to their objective counterparts which presented facts or neutral information. This negative association suggests that consumers prefer ads that convey information in a straightforward and factual manner, potentially perceiving subjective ads as less credible or less informative (Darley and Smith 1993).

The coefficient for $log(token)$ suggests that longer ads tend to be more effective ($p < 0.001$). We performed a follow-up analysis to understand whether the positive effect of ad length will dissipate beyond a certain length, given the intuition that additional information becomes redundant or is not processed beyond a set length. To investigate this possibility, we split the ads into two groups by the median ad length (70 tokens) and ran two separate regressions for ads above and below the median length. Interpreting the results in Table 5, the ad length only positively predicts ad effectiveness when the ad length is below the median token length. When the ad length is sufficiently long (longer than 70 tokens), ad length does not have a significant effect on ad effectiveness. Intuitively, an ad that is too short may not provide enough information to the consumer, but an overly long ad does not provide additional benefits toward ad clicks. We discuss the implication of this insight later.

Table 5: Effect of Ad Length on Ad Effectiveness

	Ad Clicks	
	Tokens below median	Tokens above median
(Intercept)	2.240***	2.789***
	(0.209)	(0.267)
log(token)	0.125*	−0.011
	(0.056)	(0.057)
Num.Obs.	172	183
Log.Lik.	−564.162	−647.634
F	5.039	0.040

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Notes. Standard errors in parentheses.

Additionally, Table 4 suggests that less readable ads (i.e., higher Gunning Fog Index) and ads that contain more emojis tend to perform worse. It is intuitive to see that ads need to be easy to understand in order to elicit user willingness to click. We conjecture that the presence of emojis may make the ad seem informal and unprofessional, leading to a lowered willingness to click on ads with such elements. Finally, ad sentiment (both subjectivity and polarity) does not seem to predict ad effectiveness.

We dug deeper to evaluate the link between different study conditions and the prevalence of these various ad attributes, by performing a comparison of means of these ad attributes across study conditions. From the summary statistics table (Table 6), we observe that users in the ghost-writer group used emojis and hashtags more frequently ($p < 0.05$, Table A10) compared to users from the other study groups. Given that the use of emojis is negatively related to ad performance, the prevalent use of emojis by users in the ghostwriting group could explain why ads produced in this condition are sub-par.

This set of findings led us to conduct a follow-up analysis to explore how these ad characteristics are introduced and evolved during users' interaction with the LLM. On average, treated users interacted with their LLM 2.6 times during the study, meaning that most users created 2 to 3 ad copies during the study. With that, we examine how collaboration modality affects the ad creation process by examining the change in textual characteristics across three versions of the ad copy: (1) the first draft, (2) the last draft (before the submission), and (3) the final copy submitted. We compute these text characteristics for the three ad copy versions (when available) separately and compare them across the two treatment conditions. Note that since there are no intermediate ad

Table 6: Summary Statistics For Text Features By Condition

	CONDITION	No AI	Sounding board	Ghostwriter
Subjectivity	Mean	0.616	0.595	0.643
	Std	0.215	0.151	0.114
	Min	0.000	0.000	0.361
	Max	1.000	1.000	1.000
Polarity	Mean	0.313	0.270	0.309
	Std	0.230	0.153	0.136
	Min	-0.525	-0.058	-0.082
	Max	1.000	0.891	0.750
Gunning Fog	Mean	9.193	9.636	9.745
	Std	2.935	3.524	2.857
	Min	4.000	2.700	2.267
	Max	17.651	33.924	16.200
Emoji	Mean	0.171	0.025	6.060
	Std	1.061	0.272	4.889
	Min	0.000	0.000	0.000
	Max	7.000	3.000	19.000
Tokens	Mean	56.615	83.516	103.379
	Std	38.022	41.287	56.583
	Min	20.000	21.000	19.000
	Max	255.000	252.000	265.000
Hashtag	Mean	0.000	0.000	0.353
	Std	0.000	0.000	1.105
	Min	0.000	0.000	0.000
	Max	0.000	0.000	7.000

Table 7: Characteristics of First Draft Ad across Conditions

	Sounding board (N=102)		Ghostwriter (N=115)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Gunning Fog	9.011	2.882	9.943	2.132	0.931**	0.348
Emoji	0.029	0.297	9.200	4.395	9.171***	0.411
Tokens	61.431	33.880	136.687	50.982	75.256***	5.818
Hashtag	0.010	0.099	0.557	1.326	0.547***	0.124

Notes. Not all users have first drafts, as some submit their initial ad as the final ad. Users without first drafts are excluded from this analysis.

Table 8: Characteristics of Last Draft Ad across Conditions

	Sounding board (N=74)		Ghostwriter (N=74)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Gunning Fog	9.533	2.961	9.623	2.943	0.091	0.485
Emoji	0.041	0.349	6.514	5.032	6.473***	0.586
Tokens	85.270	40.883	98.095	53.806	12.824	7.856
Hashtag	0.000	0.000	0.500	1.397	0.500**	0.162

Notes. Not all users have more than one draft. They are excluded from this analysis as their only draft is considered as the first draft.

drafts in the control condition, we focus this analysis mainly on the two treatment conditions. The summary statistics of the comparison are reported in Table 7, 8, and 9.

We found that, in the initial draft, there were significant differences across the readability (gunning fog index), emoji usage, ad length, and hashtag usage across the two LLM treatment groups. Specifically, first drafts generated by the LLM in ghostwriter modality are less readable (higher gunning fog index, $p < 0.01$), include more emojis ($p < 0.001$), are longer in length ($p < 0.001$), and use more hashtags ($p < 0.001$), compared to the first drafts written by users in sounding board modality.

Table 9: Characteristics of Final Copy Ad across Conditions

	Sounding board (N=122)		Ghostwriter (N=116)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Gunning Fog	9.636	3.524	9.745	2.857	0.109	0.415
Emoji	0.025	0.272	6.060	4.889	6.036***	0.455
Tokens	83.516	41.287	103.379	56.583	19.863**	6.448
Hashtag	0.000	0.000	0.353	1.105	0.353***	0.103

Notes. This table contains full sample as all users are required to submit their final ad copy.

Specifically, we learn that the initial ad copy produced in the ghostwriting condition contains, on average, nine emojis. That is, emojis are an artifact introduced by the ghostwriter modality of the LLM and are not added by users in the ads. A likely reason why LLMs tend to generate ad copies with more emojis is that they are trained on a variety of publicly accessible web pages, including ads for various products on the Internet. Given that past research has shown that usage of emojis could enhance ad effectiveness (Das et al. 2019), it is speculated that online ads would include the use of emojis widely, which subsequently is learned and incorporated by LLMs.

A deeper examination of the existing work on the effectiveness of emojis (e.g., Das et al. 2019) revealed that the study was conducted at least six years prior to our study. However, a newer study published in 2023 found that usage of emojis in messages decreases message credibility and source trustworthiness (Koch et al. 2023) which can reduce the willingness to click on such ads. This change in effectiveness is akin to the "banner blindness" effect (Cho and Cheon 2004), wherein users learn banner ads are linked to scam sites, which they subconsciously ignore in subsequent encounters. Similarly, consumers could learn that content with emojis is typically linked to low-quality sites, and would pay less attention to such content. Given the LLM is trained on past data and is not optimized with newer data, it fails to recognize that the use of emojis may not resonate well with consumers more recently.

Next, we compare various draft versions to understand how ads evolve in the ad creation process. Here, for each variable V, we derive the individual-specific differences between first draft and the last draft ($\Delta V_{LastFirst} = V_{LastDraft} - V_{FirstDraft}$), last draft and the final ad copy ($\Delta V_{FinalLast} = V_{FinalCopy} - V_{LastDraft}$), for each participant. We then perform a series of one-tailed t-tests to test whether the differences across the drafts are significantly different from zero. The results are reported in Table A6 and A7, respectively.

Table A6 shows that relative to users in the ghostwriter group, users in the sounding board group lengthened their ad copy considerably by 26.2 words on average ($p < 0.001$), as they moved from the initial draft to the last draft. Note that we have earlier found ad length is positively associated with ad effectiveness for ads fewer than 70 words. We see that the average length for the first drafts created by users in the sounding board is only 61.43 words (see Table 7), to which further increases in the ad length would be beneficial. Moving from the first to the last draft, sounding board users increase the length of their ad, likely through the addition of elements suggested by the

LLM. This observation is consistent with the responses we received from the exit survey conducted at the end of the experiment, included in the Appendix A.14.

In contrast, users in the ghostwriter modality substantially reduced the length of their ads by 30.27 words ($p < 0.001$) in their last draft, suggesting that the initial ad produced by the LLM in the ghostwriter condition might be somewhat verbose. Also, we see that users in the ghostwriter condition reduced the number of emojis by 2.46 ($p < 0.001$) on average. In the process of shortening the ad, the users may have found an excessive usage of emojis and took some of them out in their revision. The difference in hashtag usage was not statistically significant ($p > 0.1$).

Finally, Table A7 shows that these ad characteristics remain consistent across the last ad draft and the final copy in the sounding board group. However, users in the ghostwriter group exhibit some interesting patterns. On average, users in the ghostwriter group had manually reduced ad length by 5.18 words ($p < 0.01$), the number of emoji by 0.85 ($p < 0.001$), and hashtags by 0.16 ($p < 0.05$). Despite the effort to reduce the number of emojis used, users in the ghostwriting condition did not remove the emojis completely. This illustrates how the anchoring effect of the ghostwriting modality manifests. The ghostwriting modality introduces undesirable ad elements early on in the ad creation process. While users do remove some of these, they are inclined to keep some of them as they are anchored to the initial ad elements presented by the LLM.

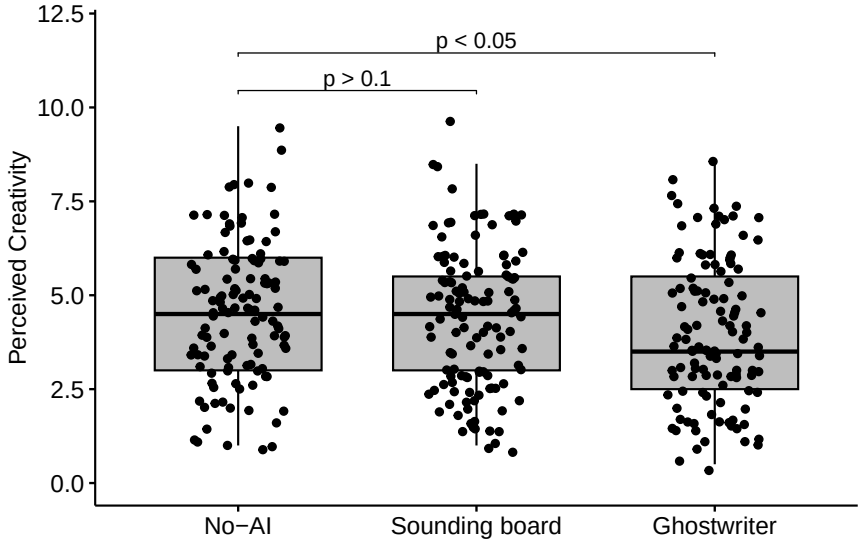
Next, we further examined potential heterogeneities in the revision patterns across user expertise levels. Specifically, we compared the change between the first draft and the final copy across experts and non-experts for the two treatment groups. The results are reported in Table A8 for the sounding board group and Table A9 for the ghostwriter group, respectively.

We found that experts in the ghostwriter group shortened the ad copy more than the non-experts ($p < 0.05$). Experts also tend to improve the readability (or decrease the usage of more complex language) more than the non-experts ($p < 0.05$).

4.5 Effects on Perceived Creativity of Ad Copies

Finally, we obtained a subjective rating of the creativity of the ad copy across the study conditions and assessed how each LLM modality would influence the creative component of an ad. Specifically, workers on Prolific (excluding prior participants) were tasked to provide labels for the perceived creativity and quality of the ads produced in our study on a scale of 1 to 10. Each ad copy is

Figure 5: Perceived creativity ratings of the ad copies



rated by at least two independent workers to which we derive an average rating. If the difference between the two ratings is more than 5, then we engage a third label from another worker, to which the two closest ratings are used to calculate the average rating.

Our results (Figure 5) revealed that using LLMs as a sounding board does not improve the perceived creativity of the ads (*M=4.369, SD=1.863*) relative to that of the control group (*M=4.521, SD=1.823*). This result is in line with the findings from Boussioux et al. (2024) who found that the use of LLM does not improve user creativity. While using LLM as a sounding board did not improve creativity, it could help users in the execution part of the ad creation process. This is evidenced by our result that showed that non-experts in the sounding board condition were able to write better and produce ads that bear semantic similarities to that of experts (section 4.3.3).

In contrast, we learned that ad copies in the ghostwriter modality exhibit the lowest perceived creativity (*M=3.974, SD=1.913*) and were in fact lower than ad copies created in the control group. Not only does this corroborate our mechanism analysis which shows that ad copies in the ghostwriter group have lower semantic divergence, but it also shows that using LLM as a ghostwriter decreases the creativity of the ad produced. We performed a subsequent analysis and found that the perceived creativity of the ads is highly correlated with the perceived quality, $r(352) = 0.670, p < .001$.

5 Discussion

Through a randomized online experiment, we study how collaboration modality between users and LLM could affect the performance of creative tasks. We contribute to the nascent and rapidly growing literature on human-LLM collaboration by considering the effect of collaboration modality and user type. Our results demonstrate that the modality plays an instrumental role in reaping the benefits of employing LLM. Specifically, the sounding board modality helps non-experts improve their ad quality by making their writing more similar to that of experts. However, the use of the ghostwriter modality produces worse outcomes, especially among experienced individuals.

Our results mirror a jagged frontier phenomenon, to which the use of LLMs is beneficial under certain situations but not so in other circumstances (Dell'Acqua et al. 2023). This heterogeneity is likely to arise from the differential impact of user bias across expertise levels and collaboration modalities. Additional analyses were used to shed light on these variations. First, we observed that an anchoring effect in the ghostwriter modality is present for both experts and non-experts. We argue that this effect has a greater negative impact on the experts relative to non-experts because the baseline ad performance of non-expert users was already low, to begin with. In contrast, experts in the ghostwriting condition produced ads that performed poorly relative to the experts in the control group. Consequently, the use of the ghostwriting modality is more detrimental to experts, as the resultant anchoring effects limit the use of their individual skills and creativity.

The lack of performance improvement among experts when using LLMs as a sounding board could be due to a "ceiling effect," wherein the performance of general-purpose LLMs is already near the upper limit of ad performance in our experiment. This result aligns with findings from Gaube et al. (2023), which suggested that feedback from the AI does not provide additional benefits for human experts who already have a high baseline performance. Moreover, general-purpose LLMs are trained using a variety of web content (OpenAI 2024) which include advertising content and non-marketing content. That is, they are not trained exclusively with ad content of excellent quality. Thus, while general-purpose LLMs are capable of providing outputs with decent quality, their advice is unlikely to surpass the knowledge and experience of experts. Hence, while non-experts are able to improve their ad-writing skills from the feedback offered by the LLM tool, experts are unlikely to surpass the global level of ad-writing skills gleaned from the LLM, as they

already possess these marketing knowledge and skill sets.

Finally, through our tests comparing the creativity levels of the created ads, three important insights were revealed. First, the ad copy creation is indeed a creative task as ad content creativity is clearly part of the ad evaluation process. Second, while the use of LLM does not directly improve the creative aspect of ad creation, it does improve the execution part of the creative process. Finally, the wrong choice of usage modality hurt the creative aspect of the ad creation process through anchoring effects. In sum, our sub-findings reveal the nuanced effects that LLM usage has on the creative and execution aspects of the ad creation process.

Our findings offer valuable insights for practitioners. Our results underscore the need to align the choice of AI collaboration modality and user expertise levels. Here, using LLMs as sounding boards can effectively help inexperienced employees without specialized skills to achieve performance levels akin to that of experts. Companies may employ LLMs as sounding boards to support employees who are new to the job or are facing unfamiliar operational problems. As seen in our study results, the use of LLMs as sounding boards can help novices accelerate their learning processes. However, management should be careful about the usage of LLM as a ghostwriter, as such a collaboration modality can lead to unintended consequences. In particular, managers should provide LLM usage guidelines to educate employees about the potential effects of anchoring when using LLM as a ghostwriter to drive the bulk of the creative process.

In addition, while a general-purpose LLM is unlikely to directly improve the quality of work from experts, companies may wish to explore the development of LLMs that are specially trained with domain-specific data, that may provide new insights or knowledge to these skilled users.

Our study opens up several avenues for future research. First, it would be interesting to explore possible ways to overcome the anchoring effect that is present with the use of the ghostwriter modality. For instance, a solution might involve providing LLM-specific training to raise workers' awareness of the anchoring effects. Second, as one of the first studies on the topic, we focus on the interaction between user expertise and collaboration modality, to develop a thorough understanding of the nuanced interaction. Future research could examine higher-order effects, such as the interactions with other LLM characteristics (e.g., temperature) and human characteristics (e.g., expertise, LLM prompt engineering techniques). While the current literature found no direct evidence of temperature on the creativity of LLM themselves (Stevenson et al. 2022), future work

may want to examine whether different LLM temperature parameters bear interaction effects with different user types. Third, our research focused on creative tasks in the context of LLM; future studies could extend the scope to generative AIs in other creative domains (e.g., art, music, and video generation), to assess the generalizability of usage modality in those settings.

In conclusion, our study emphasizes the importance of understanding the nuanced effects of using LLMs for creative tasks along with their heterogeneous effects on different users. As the adoption of LLMs continues to grow, it is crucial for individuals and organizations to be aware of the potential benefits and pitfalls of these technologies. By leveraging the strengths of LLMs while mitigating their limitations, companies can better harness the power of generative AI to enhance human creativity and improve business outcomes in various creative domains.

References

Agarwal, A. and Mukhopadhyay, T. (2016). The Impact of Competing Ads on Click Performance in Sponsored Search. *Information Systems Research*, 27(3):538–557.

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The Skill Complementarity of Broadband Internet *. *The Quarterly Journal of Economics*, 130(4):1781–1824.

Arnett, D. B. and Wittmann, C. M. (2014). Improving marketing success: The role of tacit knowledge exchange between sales and marketing. *Journal of Business Research*, 67(3):324–331.

Bauer, K. and Gill, A. (2023). Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies. *Information Systems Research*.

Bauer, K., von Zahn, M., and Hinz, O. (2023). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*.

Berg, J. M. (2014). The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes*, 125(1):1–17.

Boussioux, L., N. Lane, J., Zhang, M., Jacimovic, V., and Lakhani, K. R. (2024). Generative AI and Creative Problem Solving.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using GPT for Market Research.

Brandts, J., Groenert, V., and Rott, C. (2015). The Impact of Advice on Women's and Men's Selection into Competition. *Management Science*, 61(5):1018–1035.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at Work.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Bunescu, R. C. and Uduehi, O. O. (2019). Learning to Surprise: A Composer-Audience Architecture. In *International Conference on Innovative Computing and Cloud Computing*.

Burtch, G., He, Q., Hong, Y., and Lee, D. (2021). How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit. *Management Science*.

Cho, C.-H. and Cheon, H. J. (2004). Why Do People Avoid Advertising on the Internet? *Journal of Advertising*, 33(4):89–97.

Darley, W. K. and Smith, R. E. (1993). Advertising Claim Objectivity: Antecedents and Effects. *Journal of Marketing*, 57(4):100–113.

Das, G., Wiener, H. J. D., and Kareklas, I. (2019). To emoji or not to emoji? Examining the influence of emoji on consumer reactions to advertising. *Journal of Business Research*, 96:147–156.

Davenport, T. H. and Mittal, N. (2022). How Generative AI Is Changing Creative Work. *Harvard Business Review*.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.

Di Fede, G., Rocchesso, D., Dow, S. P., and Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*, C&C '22, pages 623–627, New York, NY, USA. Association for Computing Machinery.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.

Ericsson, K. A., Hoffman, R. R., and Kozbelt, A. (2018). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with Ai. *MIS Quarterly*, 45(3):1527–1556.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, 33(2):678–696.

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T. K., Hudecek, M. F. C., Ackery, A. D., Grover, S. C.,

Coughlin, J. F., Frey, D., Kitamura, F. C., Ghassemi, M., and Colak, E. (2023). Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports*, 13(1):1383.

Ge, R., Zheng, Z. E., Tian, X., and Liao, L. (2021). Human–Robot Interaction: When Investors Adjust the Usage of Robo-Advisors in Peer-to-Peer Lending. *Information Systems Research*, 32(3):774–785.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?

Jacowitz, K. E. and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166.

Jentzsch, S. and Kersting, K. (2023). ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models.

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, 32(3):713–735.

Kim, T., Barasz, K., and John, L. K. (2019). Why Am I Seeing This Ad? The Effect of Ad Transparency on Ad Effectiveness. *Journal of Consumer Research*, 45(5):906–932.

Koch, T., Denner, N., Crispin, M., and Hohagen, T. (2023). Funny but not Credible? Why Using (Many) Emojis Decreases Message Credibility and Source Trustworthiness. *Social Media + Society*, 9(3):20563051231194584.

Lee, D., Hosanagar, K., and Nair, H. S. (2018). Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science*.

Lysyakov, M. and Viswanathan, S. (2022). Threatened by AI: Analyzing Users' Responses to the Introduction of AI in a Crowd-Sourcing Platform. *Information Systems Research*.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). MTEB: Massive Text Embedding Benchmark.

Naik, P. A., Mantrala, M. K., and Sawyer, A. G. (1998). Planning Media Schedules in the Presence of Dynamic Advertising Quality. *Marketing Science*, 17(3):214–235.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., and Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118.

OpenAI (2024). GPT-4 Technical Report.

Orwig, W., Diez, I., Vannini, P., Beaty, R., and Sepulcre, J. (2021). Creative Connections: Computational Semantic Distance Captures Individual Creativity and Resting-State Functional Connectivity. *Journal of cognitive neuroscience*, 33(3):499–509.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.

Padmakumar, V. and He, H. (2023). Does Writing with Language Models Reduce Content Diversity?

Palan, S. and Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Pieters, R., Warlop, L., and Wedel, M. (2002). Breaking Through the Clutter: Benefits of Advertisement Originality and Familiarity for Brand Attention and Memory. *Management Science*, 48(6):765–781.

Smith, R. E., Chen, J., and Yang, X. (2008). The Impact of Advertising Creativity on the Hierarchy of Effects. *Journal of Advertising*, 37(4):47–62.

Smith, R. E., MacKenzie, S. B., Yang, X., Buchholz, L. M., and Darley, W. K. (2007). Modeling the Determinants and Effects of Creativity in Advertising. *Marketing Science*, 26(6):819–833.

Stevenson, C., Smal, I., Baas, M., Grasman, R., and van der Maas, H. (2022). Putting GPT-3's Creativity to the (Alternative Uses) Test.

Teixeira, T., Picard, R., and el Kaliouby, R. (2014). Why, When, and How Much to Entertain Consumers in Advertisements? A Web-Based Facial Tracking Field Study. *Marketing Science*.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

Zhou, J. (2003). When the presence of creative coworkers is related to creativity: Role of supervisor close monitoring, developmental feedback, and creative personality. *Journal of Applied Psychology*, 88:413–422.

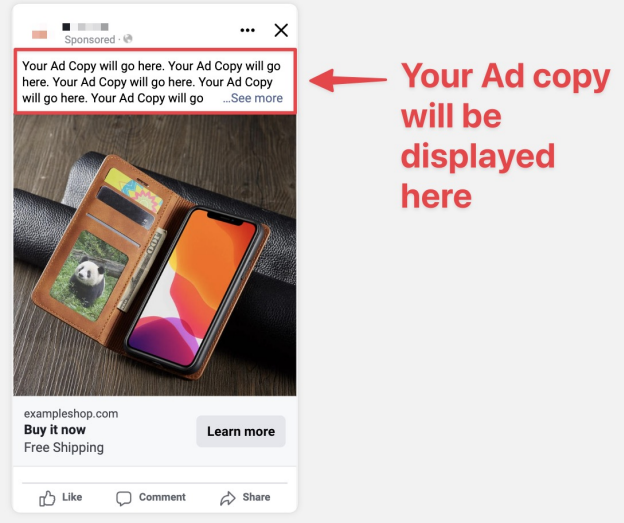
A Appendix

A.1 Product Details

The product description shown below is provided on the ad writing task webpage, along with two product images (Figure A1a) and how their ads will be displayed (Figure A1b).

- **Product Name:** iPhone Flip CardHolder Wallet Leather Case
- **Product Features & Highlights:**
  - **Material:** Crafted with the finest Retro PU Leather and Soft TPU for unparalleled strength and durability. Keeping it clean is effortless, just use a damp rag to wipe off dust and dirt.
  - **Prefect Protection:** Experience the perfect protection, thanks to our innovative design featuring a built-in Kickstand and a luxurious Card Holder Pocket. Our case offers 360 degrees of protection, with raised edges for ensuring maximum protection for your camera and screen. Say goodbye to annoying fingerprint marks and scratches, thanks to the scratch-resistant performance and drop protection of our case. Plus, the soft and anti-skid lining on the interior ensures super cushioning rebound, protecting your phone from any abrasions.
  - **Kick Stand:** Our built-in kickstand mode allows you to watch videos or chat with friends hands-free, making it perfect for multitasking.
  - **Compatibility:** Our case boasts precise access to all ports, controls, and sensors. They are tailored to a wide range of iPhone models and sizes, ranging from iPhone 7 to the latest iPhone 14, 14 Plus, 14 Pro, and 14 Pro Max.
- **Sale price:** $12.99 (50% off) + Free Shipping
- **Regular price:** $25.99

(a) Product images shown to participants and included in ad campaigns



(b) Preview of the ad shown to participants

Figure A1: All images shown to participants

A.2 Prompt Engineering for Human-LLM Collaboration Modalities

To implement the ghostwriter modality, the following system prompt was used when calling GPT-4 API:

> You are a marketing assistant who is responsible for writing a social media ad for our company. The ad will be shown on Instagram and Facebook feeds, where users may have short attention spans. Now you are given the following product:
>
> "<Product Description Inserted Here>"

Your first message to the human was: "Welcome! As your AI assistant, I will help you in creating an ad copy. I've already read the product description. Note that my role is not to provide feedback but to follow your instructions and guidelines in creating the ad copy. Please provide your first set of instructions and I will create a first draft of the ad copy. I cannot determine whether the ad copy is ready. Once you feel comfortable about the quality of the ad copy, you can go ahead and submit it."

You will work with a human to write the ad copy. You will follow the directions and feedback from the human. Use your knowledge of marketing and explain it in the feedback. The human may ask high-level questions regarding the product or marketing in general. When they ask questions, do not simply repeat the product information. Instead, you should provide more thought using your marketing expertise.

To implement the sounding board modality, the following system prompt was used when calling GPT-4 API:

You are a marketing assistant who is responsible for writing a social media ad for our company. The ad will be shown on Instagram and Facebook feeds, where users may have short attention spans. Now you are given the following product:

"<Product Description Inserted Here>"

This is for your information only. Do not repeat it.

You are also assigned a human counterpart who will be working with you. You and the human counterpart will form a team of two, and your job is to provide directions and feedback.

Remember, your job is to provide feedback and to direct. You should never write the ad copy. When asked to give an example, do not directly show how to write the ad or provide any examples. Do not provide a draft under any circumstances. Do

NOT give the headline or the body of the ad. Do not answer any requests that are unrelated to the focal task. You should never give concrete examples of how to write the ad. You do not repeat yourself.

You have used your knowledge of writing ad copies in the context of social media and came up with a few distinct and creative directions of how the ad copy should be written. You have asked the human to come up with the draft.

Your first message to the human was:

"Welcome! As your AI assistant, I can provide feedback based on the ad copy you sent to me. I've already read the product description. Note that my role is not to create the ad copy for you but to assist you by giving you my perspectives. Please draft an ad copy and I'll provide you with feedback to help you refine it further. I cannot determine whether the ad copy is ready. Once you feel comfortable about the quality of the ad copy, you can go ahead and submit it."

If no draft has been provided by the human, politely ask them to write it. If the draft has been provided, you should evaluate the draft and provide constructive feedback without giving concrete examples of the writing. Use your knowledge of marketing and explain it in the feedback. You will provide concrete feedback based on the marketing literature. You need to be consistent. When providing feedback, be sure to check if previous feedback has been incorporated. It is OK that humans may disagree with you on your feedback. You can provide additional feedback based on the revised draft. Remember, you should not provide examples of the writing. The human may ask high-level questions regarding the product or marketing in general. When they ask questions, do not simply repeat the product information. Instead, you should provide more thought using your marketing expertise.

A.3 Incentive Structure

The participants received $5 immediately upon the completion of the task. This ensured that all participants were remunerated for their time and efforts, regardless of the quality of their submissions. To motivate participants to produce their best advertisements, additional financial incentives were offered based on the relative performance of each participant's ad copy with respect to those submitted by other participants. Participants were informed that they would be rewarded with an additional $3, should the clickthrough rate of their advertisements be among the top 20% of their cohort (i.e., within their experiment condition, although the various conditions were not revealed to participants). Participants whose ad performance lies in the top 20-50% range within their experiment condition would receive an additional $2. The additional rewards were paid within a week after the task completion.

The incentive structure was displayed in a prominent position on the consent page prior to subject enrollment and on each subsequent page of the main copywriting task to be salient.

A.4 County Selection and Randomization

To avoid biases introduced through the geolocation selection process, we use data from the American Community Survey (ACS) 2020 to ensure that populations covered by counties assigned to three conditions are comparable. We exclude counties with disproportionately large populations (greater than 1,000,000) as well as those with very small populations (less than 10,000). Selecting counties with mid-sized populations helped us generate enough impressions, while evenly distributing the ads across counties with comparable population counts. We then randomly assigned these counties to one of the ad campaigns that were set up for each participant.

We further conducted one-way Analysis of Variance (ANOVA) tests that demonstrated no statistically significant difference between the counties in terms of population ($p = 0.18$) and household income ($p = 0.33$). This allowed us to proceed with confidence that our geographical allocation of ads did not introduce any unintended biases into the study. Summary statistics of assigned counties for each condition are presented in Table A1.

Table A1: Summary Statistics of Counties by Condition

	Population				Household Income			
	Mean	**Std. Dev**	**Min**	**Max**	**Mean**	**Std. Dev**	**Min**	**Max**
Sounding board	273,791.57	217,494.38	100,011	977,203	85,982.56	19,460.05	60,163	139,828
Ghostwriter	236,100.21	163,776.05	99,956	856,553	83,007.32	16,367.26	60,910	135,842
No-AI	234,400.16	173,791.66	99,727	995,567	85,870.95	16,473.84	61,523	135,597

A.5 Manipulation Checks

We performed two manipulation checks. First, we randomly inspected the chat history in the two treatment conditions and inspect for signs of failed manipulation. An example of a failed manipulation would be users successfully tricking the LLM into directly providing generated advertisement copy instead of feedback in the sounding board group. We did not find signs of this behavior in the chat histories.

Second, we asked the participants to report their experience with the ChatGPT after the main task. We asked whether "ChatGPT helped me craft the ad copies based on my input" (ghostwriter), and whether "ChatGPT gave feedback on my writing of the ad copies" (sounding board) using two survey items measured on a 5-point Likert agreement scale. Using two Welch's t-tests, we find that participants in the sounding board group agree with the sounding board items more than the ghostwriter items ($p < 0.001$), and vice versa.

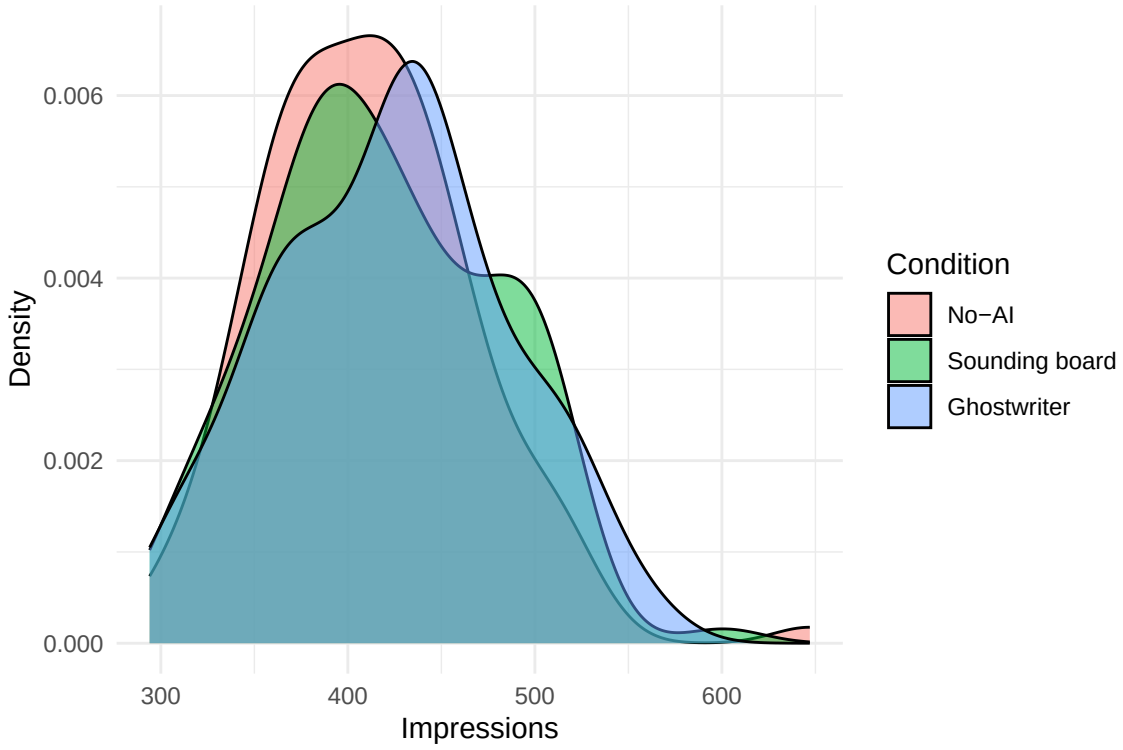
A.6 Ad Impressions

As a robustness check, we plotted Figure A2 to compare the numbers of impressions across treatment conditions, and we further find that ads across all conditions share a somewhat similar distribution. One-way ANOVA analysis also shows that the impressions are not statistically significantly different across treatment conditions ($p > 0.1$). Therefore, it is unlikely that the number of impressions will influence the result.

A.7 Additional Linear Regression Results

We first assess if experts differ from non-experts in the quality of the ads produced. Model 1 in Table A2 shows that ads written by experts attract significantly more clicks than non-experts in the control group ($\beta = 4.14, p < 0.01$), indicating that our manipulation of expertise through the

Figure A2: Impressions density



selection of participants using domain expertise is appropriate. Interestingly, as shown in Model 2, the performance gap between experts and non-experts becomes non-significant when LLM is involved, suggesting that LLM helps close the performance gap. To understand the overall effect of LLM usage (regardless of the modality), we created a dummy variable (AI) indicating whether LLM is involved. As per Model 3, employing AI results in a marginally significant decrease in clicks ($\beta = -1.40, p = 0.057$). To isolate the differential effects of human-LLM collaboration modalities, we assigned a dummy variable to each treatment group, using the No-AI group as the reference in Model 4. Consistent with the model-free evidence, the ghostwriter condition produces a significant negative effect, wherein its usage produces three fewer ad clicks compared to the control condition.

To investigate the potential interaction between collaboration modality and user expertise, We re-estimated treatment effects for each expertise level in Model 5 and Model 6. Estimates in Model 5 reveal that the ghostwriter condition produces a significant decrease in clicks for experts compared to the control group ($\beta = -5.07, p < 0.001$). In contrast, Model 6 demonstrates a significant increase of 2.4 advertisement clicks for non-experts employing LLM as a sounding board compared

Table A2: Effect of AI Usage on Ad Clicks

	DV: Ad Clicks					
	No-AI	With AI	All Sample		Non-Expert	Expert
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	16.410***	15.638***	17.356***	15.696***	12.699***	5.244**
	(1.895)	(0.789)	(1.048)	(1.154)	(1.318)	(1.776)
Expertise [Expert]	3.916**	−0.913				
	(1.246)	(0.810)				
Condition [AI]			−1.400+			
			(0.714)			
Condition [Sounding board]				0.140	2.296*	−2.148+
				(0.828)	(1.013)	(1.292)
Condition [Ghostwriter]				−3.014***	−1.004	−5.171***
				(0.762)	(0.915)	(1.209)
Num.Obs.	117	238	355	355	191	164
R2	0.102	0.048	0.046	0.089	0.118	0.157
R2 Adj.	0.044	0.015	0.024	0.065	0.074	0.113

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Robust standard errors in parentheses. All models controlled for age, gender, and education.

to the control group ($p < 0.05$). Taken together, the sounding board modality benefits non-experts, whereas the ghostwriter modality hurts the performance of experts.

Here, we run a batch of additional linear regressions which include interaction terms (Table A3). Model 1 and 2 include the full sample. Model 3 and 4 contrast the sounding board group and the ghostwriter group with the control group, respectively. These regression results are consistent with those of Table A2.

A.8 Poisson Regression Results

We replicated the linear regression models using Poisson regressions in Table A4 and A5. The results are consistent.

A.9 Comparing Text Similarity and Semantic Divergence

A common method to gauge the semantic similarity of text is by analyzing the cosine distances between text embeddings, which are high-dimensional vector representations encapsulating the semantic information of text (e.g. Brynjolfsson et al. 2023, Burtch et al. 2021). Cosine distance reflects the angle between the two high-dimensional vectors. A minimal cosine distance (approaching zero) between two vectors (embeddings) indicates high semantic similarity between the respec-

Table A3: Heterogenous Effects of AI Usage on Ad Clicks

	DV: Ad Clicks			
	All		Sounding board	Ghostwriter
	(1)	(2)	(3)	(4)
(Intercept)	15.589***	13.496***	13.425***	14.972***
	(1.150)	(1.231)	(1.338)	(1.618)
Condition [Sounding board]	0.177	2.327*	2.244*	
	(0.824)	(1.024)	(1.008)	
Condition [Ghostwriter]	−2.962***	−1.046		−1.018
	(0.760)	(0.917)		(0.928)
Expertise [Expert]	0.619	3.375**	3.657**	3.406**
	(0.686)	(1.190)	(1.210)	(1.199)
Condition [Sounding board] × Expertise [Expert]		−4.364**	−4.220*	
		(1.629)	(1.634)	
Condition [Ghostwriter] × Expertise [Expert]		−3.951**		−4.020**
		(1.510)		(1.513)
Num.Obs.	355	355	239	233
R2	0.091	0.116	0.074	0.131
R2 Adj.	0.064	0.085	0.033	0.096

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Robust standard errors in parentheses. All models controlled for age, gender, and education.

Table A4: Effect of AI Usage on Ad Clicks (Poisson Regressions)

	DV: Ad Clicks					
	No-AI	With AI	All Sample		Non-Expert	Expert
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	2.793***	2.747***	2.860***	2.753***	2.557***	1.612***
	(0.121)	(0.252)	(0.253)	(0.254)	(0.257)	(0.452)
Expertise [Expert]	0.248***	−0.063				
	(0.052)	(0.038)				
Condition [AI]			−0.091**			
			(0.029)			
Condition [Sounding board]				0.008	0.147**	−0.127**
				(0.032)	(0.047)	(0.046)
Condition [Ghostwriter]				−0.208***	−0.073	−0.340***
				(0.035)	(0.049)	(0.051)
Num.Obs.	117	238	355	355	191	164
Log.Lik.	−397.349	−782.618	−1194.732	−1174.984	−604.640	−546.007
F	4.501	3.389	5.046	8.682	5.445	8.819

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. All models controlled for age, gender, and education.

Table A5: Heterogenous Effects of AI Usage on Ad Clicks (Poisson Regressions)

	DV: Ad Clicks			
	All		Sounding board	Ghostwriter
	(1)	(2)	(3)	(4)
(Intercept)	2.745***	2.611***	2.607***	2.702***
	(0.254)	(0.256)	(0.256)	(0.100)
Condition [Sounding board]	0.011	0.149**	0.145**	
	(0.032)	(0.046)	(0.046)	
Condition [Ghostwriter]	−0.205***	−0.076		−0.074
	(0.035)	(0.049)		(0.049)
Expertise [Expert]	0.041	0.210***	0.228***	0.212***
	(0.030)	(0.048)	(0.049)	(0.049)
Condition [Sounding board] × Expertise [Expert]		−0.270***	−0.262***	
		(0.065)	(0.065)	
Condition [Ghostwriter] × Expertise [Expert]		−0.253***		−0.258***
		(0.070)		(0.070)
Num.Obs.	355	355	239	233
Log.Lik.	−1174.083	−1163.698	−804.750	−756.229
F	7.993	8.490	4.392	8.423

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. All models controlled for age, gender, and education.

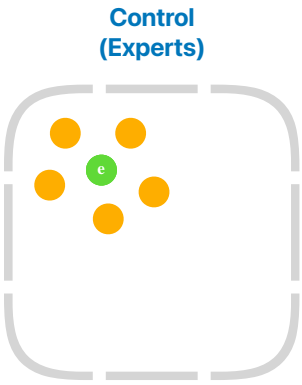
tive texts. We first construct the embeddings of the final submitted advertisement copies using the "text-embedding-ada-002", a text embeddings model that is shown to have state-of-the-art ability for capturing semantic meanings and accessing similarities (Muennighoff et al. 2023).

The group-level semantic divergence can be measured by calculating the cosine distances between each of the advertisement copy embeddings in a group and the group centroid (Hartigan and Wong 1979). If ads in a study group have a greater mean cosine distance than other groups, then they are more diverse and unique in terms of their content, as they tend to be different from one another in the same group. Following Burtch et al. (2021), we use the non-parametric Mann-Whitney U test to compare the cosine distances at the group level, as these values do not exhibit a normal distribution.

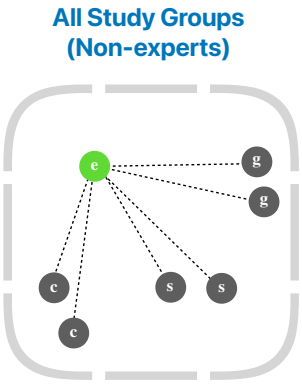
A.10 Comparision of Non-experts and Reference Experts

To make this comparison, we first calculate the centroid $\mathbf{e}$ of advertisement copy embeddings from reference experts in the control group (Figure A3a). Following this, we compute the cosine distances between each non-expert participant's advertisement copy embeddings from each experi-

Figure A3: Calculation of Experts and Non-expert Distances



(a) Control group expert embeddings and calculated centroid $\mathbf{e}$



(b) Distances between non-expert embeddings $(\mathbf{c}, \mathbf{s}, \mathbf{g})$ and control group expert centroid $\mathbf{e}$

mental group (i.e., Control $\mathbf{c}$, Sounding board $\mathbf{s}$, and Ghostwriter $\mathbf{g}$) and the aforementioned centroid $\mathbf{e}$ (Figure A3b). The nonexpert-expert distances of the control group (i.e., distances between $\mathbf{c}$ and $\mathbf{e}$) serve as a baseline. We then compare the nonexpert-expert distances (the lengths of the dashed lines in Figure A3b) of the sounding board and ghostwriter group with the baseline.

A.11 Subjective Quality Rating of First Drafts

We conducted a follow-up study to obtain subjective measures of the quality of the initial ad copy made by participants for the ghostwriter and sounding board group. We recruited a set of participants on Prolific (excluding prior participants in the main study) for this task. The first draft of each ad copy is rated by two independent participants on its perceived quality. The specific items (10-point scale) used in the survey are as follows:

- Quality: How would you rate the overall quality of the ad?
- Effectiveness: If this ad is displayed on social media (such as Facebook and Instagram) as a post, how would you rate the effectiveness of the ad in terms of attracting clicks?

In this labeling task, we utilized an incentive-compatible payment schedule, wherein the participants are awarded a bonus payment that is proportional to the degree of agreement among the two raters. This is done so that participants are incentivized to rate as accurately as possible. Using the subjective ratings, we derive a composite score of the perceived quality by taking the average of these two items.

A.12 Representive Advertisement Copies

We first split the ad copies into two pools by ad effectiveness (top 15% and bottom 15% in terms of ad clicks). From each pool, we select representative ad copies with embeddings closest to the group centroids within the pool, grouped by treatment condition and user expertise level.

A.12.1 Top 15% Ad Copies

No AI (Control), Expert

Tired of carrying around a bulky wallet and a phone? With our iPhone Flip Cardholder Wallet Leather Case, you can keep your phone and cards all in one convenient place.

Our case is made from high-quality leather and designed with 360 degrees of protection for your phone. Plus, our built-in kickstand mode allows you to watch videos or chat with your friends hands-free, making it perfect for multitasking.

For a limited time, we are offering 50% off our regular price. Click the link below to learn more and order today!

No AI (Control), Non-Expert

Get this high quality, durable PU leather IPhone Flip CardHolder Wallet Leather Case for half price, 12.99 ($25.99 original price). It comes complete with anti-smudge, enhanced 360 degree screen protection, multitasking conveniences with a kickstand to watch videos and chat with friends, and has broad adaptibility with IPhones 7 to the most current IPhone 14, 14 Plus, 14 Pro, as well as 14 Pro Max. Order now and as a bonus receive free shipping.

Sounding board, Expert

Got a new iPhone? Bet you want to keep it protected! :-)

The iPhone Flip Cardholder Wallet Leather Case protects your phone and makes it easy to enjoy your favorite shows by holding your screen in an easy-to-see manner.

The leather is luxurious and great looking and also provides 360 degrees of protection. It has a built-in kickstand to make it easy to watch videos and a cardholder pocket to make a wallet unnecessary!

For the next week, you can purchase it for just $12.99 with free shipping - 50% off of the regular price of $25.99. Buy it now, it is sure to sell out!

Sounding board, Non-Expert

iPhone flip card holder wallet leather case, made with the finest Retro PU Leather and Soft TPU for effortless durability, 360-degree protection, and scratch resistance. Compatible with a wide variety of iPhone models from 7 to 14, 14 Plus, 14 Pro, and 14 Pro Max. Has perfect protection with a built-in kickstand to allow hands-free usage. The innovative card holder pocket can hold various credit cards for ease of access on the go. Currently $12.99 + free shipping! Limited-time offer at 50% off! Buy now!

Ghostwriter, Expert

Protect, organize, and elevate your iPhone with our Flip CardHolder Wallet Leather Case! 📱🛡️

🔥 Premium Retro PU Leather & Soft TPU material for a sleek look

🌟 360° protection secures your phone from scratches & drops

🎬 Built-in kickstand for hands-free entertainment

💳 Convenient card holder pocket for essentials

📲 Compatible with iPhone 7 to iPhone 14 Pro Max

Upgrade now and SAVE 50% + FREE SHIPPING! Limited time offer! 💰🚚

Experience the benefits TODAY! Click to shop now! 👉🛒

#iPhoneProtection #FunctionalStyle #GetYoursNow

Ghostwriter, Non-Expert

📱Upgrade Your iPhone Experience with Our Flip CardHolder Wallet Leather Case!🔥

✨Introducing the perfect accessory for your iPhone - the Flip CardHolder Wallet Leather Case, crafted with the finest Retro PU Leather and Soft TPU for exceptional strength and durability.

💎Protect Your Phone in Style: Enjoy 360° protection with raised edges guarding your camera and screen. The soft, anti-skid lining offers superb cushioning to keep your phone safe from scratches and damage.

💼Card Holder Convenience: Keep your essentials close with the luxurious Card Holder Pocket built into our sleek case.

💻Hands-Free Kickstand: Watch videos or chat with friends easily using the built-in kickstand for hands-free multitasking.

📲Perfect Compatibility: Designed for a wide range of iPhone models (iPhone 7 to iPhone 14, 14 Plus, 14 Pro, and 14 Pro Max), our case ensures precise access to all ports, controls, and sensors.

🚀Sale Alert: Get this amazing deal at just $12.99 (50% off) + FREE Shipping! Don't miss out on upgrading your iPhone experience. Offer ends soon! 🛒

👉Click the link to shop now and protect your iPhone in style!

A.12.2 Bottom 15% Ad Copies

No AI (Control), Expert

HALF PRICE + FREE SHIPPING! Say 'goodbye' to bulky wallets and scratched phone screens — and 'hello' to the ultimate in convenience and protection. The iPhone Flip CardHolder Wallet Leather Case is fully compatible with all iPhones, featuring built-in kickstand, raised edges and crafted with quality leather for long-lived durability.

No AI (Control), Non-Expert

Protect and elevate your iPhone with our Flip Card Wallet. This retro leather case features a built-in kickstand for hands-free use, scratch resistance, drop protection, and easy access to all of your phone's ports and controls. On sale now for just $12.99 (50% off) + free shipping. Get yours now!

Sounding board, Expert

IPhone Flip CardHolder Wallet Leather Case $12.99 + free shipping (regular $25.99) LIMITED TIME OFFER This case is made of premium leather and durable construction and comes in several different colors. Super easy to keep clean, wiping with just a damp cloth. Protect your IPhone from scratches, drops and fingerprints via the innovative design. Features include: *360-degree protection *raised edges *soft anti-skid lining Holds all your cards and IDs too! Also includes a kickstand for hands-free use. Fits IPhone models from 7 up to 14, including 14 Plus, 14 Pro and 14 Pro Max.

Sounding board, Non-Expert

The Iphone is crafted with the finest Retro PU Leather and Soft TPU for unparalleled strength and durability. Keeping it clean is effortless, just use a damp rag to wipe off dust and dirt. The phone is designed with innovation featuring a built-in Kickstand and a luxurious Card Holder Pocket. Our case offers 360 degrees of protection, with raised edges for ensuring maximum protection for your camera and screen. Say goodbye to annoying fingerprint marks and scratches, thanks to the scratch-resistant performance and drop protection of our case. Plus, the soft and anti-skid lining on the interior ensures super cushioning rebound, protecting your phone from any abrasions. It has a built-in kickstand mode allows you to watch videos or chat with friends hands-free, making it perfect for multitasking. The case boasts precise access to all ports, controls, and sensors. They are tailored to a wide range of iPhone models and sizes, ranging from iPhone 7 to the latest iPhone 14, 14 Plus, 14 Pro, and 14 Pro Max.

You can purchase the phone for the prices below: Sale price: $12.99 (50% off) + Free Shipping Regular price: $25.99

Ghostwriter, Expert

🔥50% OFF iPhone Flip CardHolder Wallet Leather Case - Limited Time Offer!🔥

Upgrade your iPhone game with the stylish and functional iPhone Flip CardHolder Wallet Leather Case! Crafted with premium Retro PU Leather and Soft TPU, this case offers unparalleled strength, durability, and protection. 📱

Key Features: ✓ Luxurious Card Holder Pocket ✓ Built-in Kickstand for hands-free video watching

✓ 360-degree protection for camera and screen

✓ Compatible with iPhone 7 up to the latest iPhone 14 models

✓ Easy to clean and scratch-resistant

Don't miss out on this amazing deal! 💸 Get yours now for ONLY $12.99 (regular price $25.99) + FREE Shipping! 🚚

Stocks are limited, so grab yours while you can! Shop Now 👉 [LINK]

Ghostwriter, Non-Expert

🔥50% OFF iPhone Flip CardHolder Wallet Leather Case!🔥

Upgrade your iPhone experience with our stunning iPhone Flip CardHolder Wallet Leather Case - now just $12.99 (regular price $25.99) + FREE shipping! 😲

Experience elegance and protection like never before: ✅ Finest Retro PU Leather & Soft TPU for unmatched durability

✅ 360° protection with raised edges to secure your camera and screen

✅ Scratch-resistant, anti-fingerprint and drop protection

✅ Built-in kickstand for hands-free viewing or chatting

✅ Luxurious card holder pocket for added convenience

✅ Perfect fit for iPhone 7 up to the latest iPhone 14 models

Don't miss out on this incredible deal! Upgrade and protect your iPhone in style NOW! 💼📱💸

Click the link in our bio to shop now! Limited time offer! ⏳

A.13 Textual Characteristics

Additional comparsions across the first drafts and final drafts are included in this section. Table A6 compares the difference in Ad characteristics between the final copy and first draft. Table A7 compares the difference in Ad characteristics between the final copy and last draft.

Table A6: Difference in Ad Characteristics between Last and First Draft

	Sounding board (N=74)		Ghostwriter (N=74)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
$\Delta GunningFog$	0.552	2.631	-0.167	3.226	-0.719	0.484
$\Delta Emoji$	0.000	0.000	-2.459	4.579	-2.459***	0.532
$\Delta Tokens$	26.243	35.989	-30.270	52.940	-56.514***	7.442
$\Delta Hashtag$	-0.014	0.116	-0.054	0.594	-0.041	0.070

Table A7: Difference in Ad Characteristics between Final Copy and Last Draft

	Sounding board (N=74)		Ghostwriter (N=74)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
$\Delta GunningFog$	0.148	2.213	-0.195	1.450	-0.343	0.308
$\Delta Emoji$	0.000	0.000	-0.851	2.131	-0.851***	0.248
$\Delta Tokens$	1.176	8.473	-5.176	16.662	-6.351**	2.173
$\Delta Hashtag$	0.000	0.000	-0.162	0.759	-0.162+	0.088

Table A8 and A9 breaks down the comparisions between final copy and first draft across experts and non-experts within the sounding board group and the ghostwriter group, respectively.

Table A8: Difference in Ad Characteristics between Final Copy and First Draft in Sounding Board Modality

	Non-Expert (N=55)		Expert (N=47)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
$\Delta GunningFog$	0.386	3.111	0.658	3.621	0.272	0.674
$\Delta Tokens$	24.218	39.484	25.298	22.716	1.080	6.271
$\Delta Hashtag$	0.000	0.000	-0.021	0.146	-0.021	0.021

Notes. The difference in emoji usage is not calculated as there is only one user in sounding board modality using emojis, whose emoji usage did not change across first draft and final copy.

Table A9: Difference in Ad Characteristics between Final Copy and First Draft in Ghostwriter Modality

	Non-Expert (N=67)		Expert (N=48)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
$\Delta GunningFog$	0.406	2.545	-1.004	3.115	-1.410*	0.547
$\Delta Emoji$	-3.015	4.937	-3.271	5.135	-0.256	0.956
$\Delta Tokens$	-23.836	50.494	-45.521	52.892	-21.685*	9.815
$\Delta Hashtag$	-0.090	0.621	-0.354	1.021	-0.265	0.166

Table A10 provides the p-values for the t-tests shown in Table 6. Note that the missing value is because the No-AI and Sounding board groups both have completely zero hashtag usage. Therefore, the p-value is infinite.

Table A10: P-Values For Text Features Comparision Across Features

	Sounding Board v.s. No-AI	Ghostwriter v.s. No-AI	Sounding Board v.s. Ghostwriter
Polarity	0.087+	0.842	0.043*
Subjectivity	0.382	0.23	0.006**
Gunning Fog	0.293	0.147	0.794
Tokens	0.0***	0.0***	0.002**
Emoji	0.142	0.0***	0.0***
Hashtag	nan	0.001***	0.0***

A.14 Qualitative Analysis of Optional Open-ended User Feedback

At the end of the task, an optional open-ended question was used to understand participants' experience with the generative AI used. A few themes emerged from our reading of these comments. First, several users in the ghostwriter modality indicated that the use of AI allows them to spend less cognitive effort. The following are a few excerpts demonstrating this.

> Subject 10 (Expert): It was a quick way to write ads, without having to think too hard about it.
>
> Subject 25 (Non-expert): I didn't really have to write anything for the ads, and I was able to ask it to improve upon the initial work it gave me so the end product would be acceptable.
>
> Subject 63 (Expert): I feel like it saved me mental effort at the end of a mentally tiring day.

This set of feedback reveals that users are relying on the LLM tool quite a bit, which highlights the possibility that users in the ghostwriter condition might be experiencing an anchoring effect. A user in this condition mentioned this explicitly, stating that his writing is heavily influenced by the AI and he would have come up with something that is quite different if not for the LLM tool. This piece of qualitative evidence adds to our theoretical mechanism of anchoring effect.

> Subject 94 (Non-expert): This was a really interesting task. I think the AI influenced my writing a lot. There are certainly sentences that it created which I would not have done. I feel that if I was writing this without AI then my writing would probably have been MORE radical and unconventional.

However, a small subset of users with the ghostwriting modality noted that the AI could help them in the brainstorming phase and be more efficient, but recognized the shortcomings produced.

> Subject 54 (Expert): Yes, it was very useful though there still needed to be a fair bit of human input. I didn't have too much familiarity with the advertising platforms and I didn't feel I could fully trust the AI's judgements every single time. I did have to make some executive decisions.
>
> Subject 123 (Expert): AI can be useful to make tasks more efficient but not without human review quite yet.
>
> Subject 97 (Non-expert): The AI is mainly useful for brainstorming and to get background information. After the stage of brainstorming, it become less useful. However, it's still great to proofread.

We went ahead to check the ad performance of these users who were aware of the limitations of the LLM tool. This group of users consists of both experts and non-experts. We learned that these users produced ads that performed relatively well and were ranked higher than the average subjects.

Responses from participants in the sounding board modality indicated a learning effect, to which the AI-assisted users in the ad creation task by providing suggestions and tips to enhance the effectiveness of their ads. It is interesting to note that users found the feedback to be useful and similar to feedback that they would get from human experts.

Subject 5 (Expert): It was useful in giving tips I might not have thought of otherwise, it allowed me to add more detail without being verbose

Subject 35 (Non-expert): It seems to do a good job beta-reading ad copy. The feedback I received was consistent with what I get from human editors or friends when I do short marketing tasks.

Subject 8 (Non-expert): It helps me hone and tailor the ad to sound more enticing and enhance features.

This result further corroborated our analysis that non-experts in the sounding board modality learn to write more similarly to experts.