The Effect of Education in Prompt Engineering: Evidence from Journalists

Amirsiavosh Bashardoust[1*], Yuanjun Feng[1*], Dominique Geissler[2*], Stefan Feuerriegel[2], Yash Raj Shrestha[1]

[1]University of Lausanne, Switzerland
[2]LMU Munich, Munich Center for Machine Learning (MCML), Germany
{amirsiavosh.bashardoust, yuanjun.feng, yashraj.shrestha}@unil.ch, {d.geissler, feuerriegel}@lmu.de,

Abstract

Large language models (LLMs) are increasingly used to create content for social media, specifically in the context of journalism. In this paper, we analyze whether training in prompt engineering can improve the interactions of users with LLMs. For this, we conducted an experiment where we asked journalists to write short texts before and after training in prompt engineering. We then analyzed the effect of training on three dimensions: (1) the user experience of journalists when interacting with LLMs, (2) the domain expert perception, and (3) the non-expert reader perception, such as clarity, engagement, and other text quality dimensions. Our results show: (1) Our training improved the perceived expertise of journalists but also decreased the perceived helpfulness of LLM use. (2) The effect on expert perception varied by the difficulty of the task. (3) There is a mixed impact of training on reader perception across different text quality dimensions.

1 Introduction

Generative AI such as large language models (LLMs) are increasingly used to create content for social media (Statista 2024; Feng et al. 2023). Especially journalists can leverage the power of LLMs for various tasks such as writing texts, drafting social media posts, and creating multi-media content (Schmidt 2024; Felten, Raj, and Seamans 2023). For example, the *Los Angeles Times* already uses LLMs to automatically generate reports on seismic activity in real-time (Los Angeles Times 2019). However, oftentimes, employees use LLMs without training or when there are no prevalent policies in place in their organization. As a result, prompt engineering training is important – especially in a journalistic context where the impact of quality of web content creation is large (Feuerriegel et al. 2023; Beauchene et al. 2023).

The quality of LLM outputs depends – to a large extent – on the quality of prompts, meaning the user input that the LLM receives (Atreja et al. 2024). For example, poorly structured prompts tend to result in vague, incorrect, or irrelevant output (Lin 2024). Output quality can be assessed through different lenses, i.e., through domain experts and non-expert readers (McGuire, de Cremer, and van de Cruys 2024). Moreover, the quality of prompts also shapes the overall experience of users. Novice users have difficulties designing effective prompts and struggle to adjust prompts effectively to improve their outcomes (Zamfirescu-Pereira et al. 2023). This can lead to frustration and dissatisfaction among users and therefore reduce the perceived helpfulness of LLMs (Kim et al. 2024).

Effective prompt writing, commonly referred to as *prompt engineering*, is widely regarded as an essential skill when using LLMs (Yang et al. 2024). Common strategies include asking the LLM to use chain-of-thought reasoning (Wei et al. 2022) or to adopt the role of a specific persona (White et al. 2023). As we hypothesize later, effective prompt engineering can not only improve the output quality of LLMs but also the user experience. This raises the question of whether users can be trained in prompt engineering training.

In this paper, we analyze whether prompt engineering training can improve the user experience and output quality of journalists interacting with LLMs. For this, we conducted an experiment where $N = 29$ participants receive in-person training in prompt engineering.[1] We performed our experiment with professional journalists whom we asked to write short social media posts (similar to those that are published on social media platforms such as Twitter/X and LinkedIn) about scientific research articles – a task that exemplifies the growing intersection of scientific communication and social media engagement. The posts were written with the support of ChatGPT-3.5 both before and after training (see Figure 1 for an overview of our study flow). We analyze the effect of training on three dimensions:

RQ1: *How does prompt engineering training influence the perceived expertise and perceived helpfulness of journalists when interacting with LLMs?*

We measure the perceived expertise of journalists with the LLM and the perceived helpfulness of the LLM for the given task both before and after the journalists receive the training. We expect that, after training, users may find LLMs more helpful as they better understand how to generate the desired output through prompts. In addition, the training should also

*These authors contributed equally.

[1]Code and data for this project are available on Github: https://github.com/vosh-96/The-Effect-of-Education-in-Prompt-Engineering-Evidence-from-Journalists.

make users rate their expertise in using LLMs higher (see Section 4).

RQ2: *How does prompt engineering training influence the domain expert perception of texts written by journalists with LLMs?*

In RQ2, we focus on the perception of domain experts as one measure of output quality. We expect that creating a social media post about a scientific article can potentially introduce inaccuracies in the text, either due to the more concise and engaging format of the social media post (as compared to the original scientific article) or due to the use of an LLM because LLMs often tend to hallucinate (Bang et al. 2023). Hence, we conduct an external, post-hoc evaluation of output quality, where we ask domain experts to rate the accuracy of the texts that were written during the experiment.

RQ3: *How does prompt engineering training influence the non-expert reader perception of texts written by journalists with LLMs?*

The generated social media posts may be perceived differently by domain experts (who may focus primarily on accuracy) and the general non-expert audience (who may prefer social media posts that are interesting and engaging) (McGuire, de Cremer, and van de Cruys 2024). We thus examine how non-expert readers perceive the written social media text as another dimension of output quality. Hence, we conduct another external, post-hoc evaluation where we let non-expert readers evaluate the texts from the experiment along several non-expert reader-specific dimensions for assessing scientific journalism such as clarity and engagement.

2 Background

We review literature on prompt engineering that provides strategies and behavioral evidence. An overview over literature aimed at improving AI literacy in general can be found in Appendix A.

The quality of the LLM output highly depends on the quality of the user input (i.e., the prompt) (Atreja et al. 2024). Some general guidelines on how to effectively prompt LLMs have been developed, which include: (1) guiding the model to solutions, (2) adding relevant context, (3) being explicit in the instructions, (4) asking for lots of options, (5) giving examples of good answers, and (6) breaking complex tasks into subtasks (Lin 2024; OpenAI 2024; Meskó 2023; White et al. 2023).

Beyond simple guidelines, several, more advanced strategies for *prompt engineering* have been identified (Liu et al. 2023). One strategy is to ask the model to use *chain-of-thought reasoning*. Chain-of-thought describes a series of intermediate natural language reasoning steps that lead to the final output (Wei et al. 2022). This allows LLMs to decompose multi-step problems into solvable, intermediate steps, which, in turn, can positively affect the accuracy and relevance of the output. Furthermore, chain-of-thought prompting enables users to understand the model's reasoning process, allowing them to verify the output against their own domain knowledge and can boost the perceived helpfulness of LLMs (Wei et al. 2022).

Another common strategy is the use of *personas*. Here, the LLM is asked to adopt the standpoint of a particular persona (White et al. 2023), such as, for example, a journalist who needs to write an article. This strategy is particularly useful when users seek to write for a specific audience or from a specific standpoint (White et al. 2023). As a result, using personas creates outputs that are often more personalized for different target audiences and can thus improve readers' perceptions. While there is evidence that these strategies influence the output quality of LLMs (Liu et al. 2023; Wei et al. 2022; White et al. 2023), little is known about the impact of *training* users in prompt engineering.

Recent evidence shows that prompt engineering affects both output quality as well as user experience and especially novice users often struggle with prompt engineering (Zamfirescu-Pereira et al. 2023; Jahani et al. 2024; Dang et al. 2022). One reason for this is that they have an incomplete understanding of the capabilities of LLMs and they further tend to create prompts that mimic human-to-human instructions, for example, by relying more on instructions rather than giving examples (Zamfirescu-Pereira et al. 2023). In addition, novice users tend to over-generalize from single observations and, hence, do not make systematic progress when engineering prompts (Zamfirescu-Pereira et al. 2023). Other works have found that novice users interact with LLMs as they would with a human interlocutor, such as, for example, by using socially desirable phrases like "hello" and "thank you" or explaining inner thoughts and motives (Knoth et al. 2024). This raises the need to understand how novice users can be trained in effective prompt engineering and how this affects LLM output quality and user experience, which is the focus of our study.

Previous work used audience-specific writing tasks and assessment methods to evaluate prompt engineering skills. Knoth et al. (2024) tested students' prompt engineering by asking them to plan a scientific project using an LLM. In addition, Knoth et al. (2024) collected both the user prompts and the LLM output to assess prompt engineering skills and asked the students additional questions to evaluate their user experience. We later adopt this audience-specific design approach by designing tasks that are specific to science journalists, asking them to share their entire dialogues with ChatGPT with us, as well as asking follow-up questions.

Research gap Writing effective prompts is essential for leveraging the full potential of LLMs, but this can be particularly challenging for novice users. To the best of our knowledge, there is no work that analyzes the effect of training users in prompt engineering on user experience as well as output quality. In this study, we are the first to train professional journalists in effective LLM prompting and assess the impact on user's perceived expertise and perceived helpfulness, as well as output quality (e.g., perception of domain experts in terms of accuracy but also non-expert perceptions such as clarity and engagement).

3 Experiment

We conducted an experiment with a group of journalists specializing in science communication from Switzerland to measure the effect of prompt engineering training on user

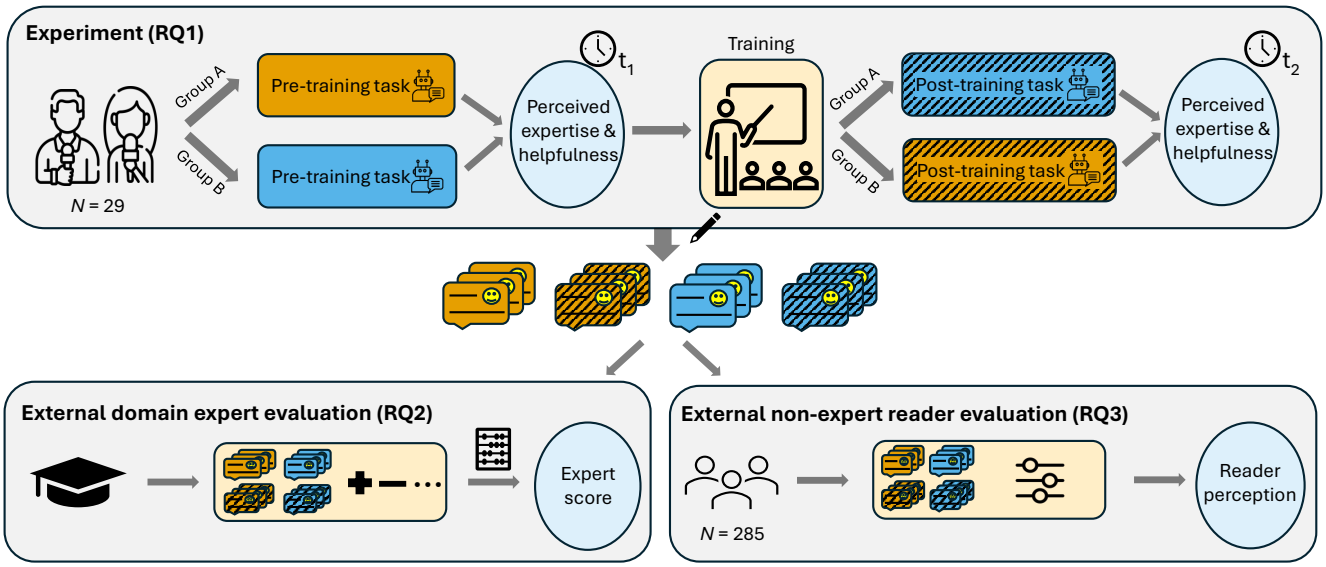

Figure 1: Overview of our study process.

experience and output quality. The experimental protocol was pre-registered at Open Science Framework (OSF)[2]. The journalists (N=29) participated in two writing tasks, namely, a pre-training task and a post-training task that took place before and after training, respectively. In both tasks, the journalists created a short social media post about a scientific article using an LLM (specifically, ChatGPT-3.5). We used a counter-balanced measures design to examine the effect of prompt engineering training. Since treatment order can influence participants' behavior due to fatigue or other external factors (Gaito 1961), we randomly varied the order of the articles in the writing task. We also asked journalists to rate their experience with LLMs with respect to *Perceived Expertise* and *Perceived Helpfulness*.

The written social media posts subsequently went through two separate post-hoc external evaluations: (i) an evaluation of accuracy by domain experts and (ii) an evaluation of reader perception (e.g., clarity and engagement) by a group of non-expert readers.

3.1 Participants

We invited 37 science journalists to our experiment. They were non-incentivized but motivated to learn about prompt engineering for science communication. Participating and submitting responses to the two writing tasks (*pre-training task* and *post-training task*), as well as demographic information, were optional. In the end, 29 participants provided valid responses. One participant did not provide demographic information, but we still used the responses from the writing task to estimate the learning effect. In general, the number of participants aligns with previous experiments that involved similar tasks for professionals (Li et al. 2024; Benk et al. 2022; Huang et al. 2020; Senoner et al. 2024).

Among the participants, 16 are women, 11 are men, and one is non-binary. The average age is 45.82 years, with ages ranging from 27 to 66 years. Except for one journalist with a Bachelor's degree, the rest of the participants held at least a Master's degree (or equivalent), and 14 even had a doctorate degree (or equivalent). In addition, 16 participants had over 10 years of work experience. We used participants' self-rated perceived expertise before training as a proxy for their prior expertise level. We observed that no participant selected the highest rating ("7"), which suggests that none of them considered themselves experts in prompt engineering before training.

3.2 Tasks

In our experiment, participants (journalists) were asked to perform a science communication task. Participants should transform an extended abstract of a scientific article into an engaging Twitter/X post with the help of ChatGPT-3.5.[3] We imposed a character limit of 280 characters for each post, similar to Twitter/X. For the writing tasks, we chose two scientific articles, which we refer to as ARTICLE PSY and ARTICLE GUNVIOLENCE:

- ARTICLE PSY (Ross et al. 2024) was published in *Journal of the American Medical Association (JAMA)*. It studies the effectiveness of psychiatric hospitalization in reducing subsequent suicidal behaviors in patients with suicidal ideation or suicide attempts and develops a machine learning model to individualize treatment recommendations (Ross et al. 2024).
- ARTICLE GUNVIOLENCE (Semenza et al. 2024) was published in *JAMA Open* journal. It examines the association of exposure to gun violence with suicidal ideation

[2]https://osf.io/sp5fx/overview?view_only=b5072c31187b471284cab5111d199531

[3]Participants used ChatGPT-3.5, accessed through the ChatGPT website interface on 25/04/2024.

and behaviors among black adults in the United States who have been subject to interpersonal gun violence (Semenza et al. 2024).

We provide the extended abstracts of the two scientific articles in Appendix I. In our counterbalanced design, participants conducted the writing tasks for both articles but with random article display orders; that is, half of the participants received ARTICLE PSY before training and ARTICLE GUNVIOLENCE after, while the rest received the articles in reverse order. We chose to use two articles in our experiment because we had to balance the cognitive effort required by the training and the time constraints of the training session. Including more articles would likely have overwhelmed participants and compromised the quality of their engagement with the writing tasks.

This task is common for science journalists in newsmaking (Tenenboim-Weinblatt and Baden 2018), and our participants are familiar with it. We intentionally selected the two articles from above as we expected several challenges that are common in science communication (Jensen and Gerber 2020; Guenther and Joubert 2017). This includes differentiating between correlation and causality (e.g., ARTICLE GUNVIOLENCE offers associative but not causal findings), understanding the scope and boundary conditions of findings (e.g., ARTICLE GUNVIOLENCE is limited to people identified as Black or African American in the US), interpreting data (e.g., statistical uncertainty), and translating jargon and technical language. Later, in RQ2, we also grade the quality of the social media posts along the previous dimensions.

3.3 Training in Prompt Engineering

Our prompt engineering training for journalists was designed as a comprehensive, 2-hour interactive training. All journalists attended the same session. This intensive session combined theoretical input with hands-on exercises, aiming to familiarise journalists with the capabilities of LLMs and train them in writing effective prompts. The training is conducted as an in-person workshop because that way we have better control over confounding variables compared to other formats. All participants receive the same training at the same speed in the same physical learning environment without external interruptions. This helps us to better disentangle the effects of training.

In developing this workshop, we drew upon best practices from both academic research (Schulhoff et al. 2024; Huang and Chang 2022; Mosbach et al. 2023; Saravia 2022) and OpenAI's published guidelines for using ChatGPT (OpenAI 2024). For reasons of reproducibility, we make the slides of our course publicly available together with our code on our repository. We emphasize that several parts of our course involved interactive elements.

The primary objective of our training was to equip journalists with a robust understanding of prompt engineering principles and to develop their practical skills in crafting and refining prompts. As outlined in Table 5 in Appendix B, the training was structured into two main parts. The first part laid the groundwork by introducing foundational concepts, beginning with an explanation of prompts and the concept of context length. We then progressed to explore various LLM applications, including text summarization, question answering, text classification, role-playing, and reasoning. This section concluded with an introduction to three key prompting strategies: zero shot, few shot, and chain-of-thought techniques.

- *Zero shot*: The user provides a task description to the model without any further examples, and the model will generate the response based on the task description.
- *Few shot*: Besides the task description, the user also provides a few example outputs in the desired format to the model. The model will use the task description and the examples to generate the response.
- *Chain-of-thought*: The user will ask the model to break down the task description into a series of intermediate steps and imitate human-like reasoning to solve them.

The second part covered prompt refinement methods. Here, we focused on the iterative nature of prompting. We broke the process into three main subprocesses: *getting started*, *refining the focus*, and *exploring in depth*. The getting started phase is mainly the first attempt to give the prompt context and objective. This was followed by iterating the prompt to improve the final result. We ended the training by reviewing the risks and ethical considerations of using LLMs in real life.

3.4 Procedure

We conducted our experiment on April 25, 2024, at the University of Lausanne. We started with onboarding, where we explained the goal of the experiment, briefly introduced prompt engineering, and collected informed consent. We then asked participants to fill in their demographic information, including age, gender, level of education, and years of experience, while the survey with demographic information was made optional.

Next, participants were randomly assigned to two conditions, based on which we distributed the *pre-training task*. Each group received either ARTICLE PSY or ARTICLE GUNVIOLENCE for writing social media posts. After completing the task, we asked them to rate (i) the perceived expertise with ChatGPT and (ii) the perceived helpfulness of ChatGPT in the task (both on a 7-point Likert scale).

Subsequently, we proceeded with the prompt engineering training.

After training, we asked participants to perform the *post-training task*, where the opposite task is given to each group. This allows us to isolate the learning effect later. Again, we collected the participants' (i) perceived expertise with ChatGPT and (ii) perceived helpfulness of ChatGPT in the task.

Lastly, we invited participants to share their thoughts and experiences about using LLMs and prompt engineering for the task in a qualitative form, and all participants were further invited to voluntary discussions. We debriefed the participants and appreciated their participation.

3.5 Ethical Considerations

We respect the privacy and autonomy of all participants involved in this study. Data were collected anonymously, with all personally identifiable information being removed. The study followed ethical research standards (Rivers and Lewis 2014), and the experimental design received approval from the Ethics Commission at the University of Lausanne. We debriefed participants after the study and allowed them to ask questions and provide feedback.

4 Effect of Training on Perceived Expertise and Perceived Helpfulness (RQ1)

Our prompt engineering training aims to equip journalists with the skills to effectively utilize LLMs for their professional tasks. In RQ1, we examine how our training affects journalists' *Perceived Expertise*, *Perceived Helpfulness*, and their interactions with LLMs.

4.1 Method

User Perceptions We asked participants (N=29) to rate their (i) *Perceived Expertise* and (ii) *Perceived Helpfulness* with LLMs before and after training using a 7-point Likert scale from 1 ("extremely unhelpful/no expertise") to 7 ("extremely helpful/expert"). We estimated the effect of our prompt engineering training, demographics, and article display order on participants' experience based on Equation 1.

$$\begin{aligned} Y_i = \beta_0 + u_{0i} + \beta_1 \textit{TrainingStatus}_i + \beta_2 \textit{ArticleOrder}_i + \\ \beta_3 \textit{Age}_i + \beta_4 \textit{Gender}_i + \beta_5 \textit{Education}_i + \\ \beta_6 \textit{WorkExperience}_i + \epsilon_i \end{aligned} \quad (1)$$

where Y_i denotes the experience with LLMs (i.e., *Perceived Expertise* or *Perceived Helpfulness*) for the i-th participant, with intercept β_0, participant-level random effects u_{0i}, coefficients β_1 to β_6, and an error term ϵ_i. The other variables are as follows: *TrainingStatus* represents the participant's training status, encoded as 0 (before training) and 1 (after training). *ArticleOrder* indicates how articles are displayed in the two tasks. *Age*, *Gender*, *Education*, and *WorkExperience* correspond to the participants' demographic variables.

User Interactions To complement and explain the observed shifts in user perceptions, we conducted an open discussion after training to explore participants' attitudes toward their interactions with LLMs. Additionally, one of the authors manually annotated participants' chat dialogues with ChatGPT to identify prompting strategies introduced during training. Based on these annotations, we created binary variables indicating whether each participant employed a given prompting technique: in-context prompting (*InContext*), chain-of-thought (*CoT*), iterative refinement (*IterativeRefine*), and few-shot prompting (*FewShot*). Appendix H details our annotation process.

To further investigate whether these prompting techniques influenced participants' perceived expertise or helpfulness, we fit a mixed-effects linear regression model (see Eq. (2)) using the binary technique indicators as predictors for *Perceived Expertise* and *Perceived Helpfulness*.

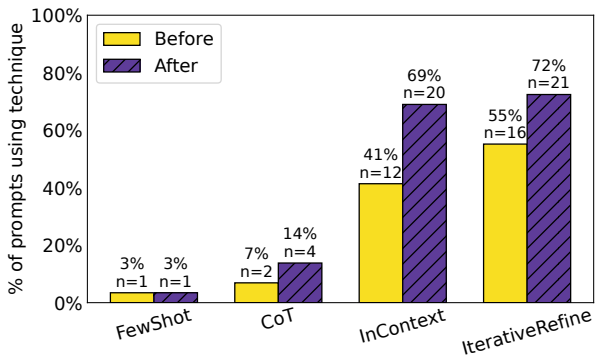

Figure 2: Percentage and Frequency of prompting technique usage before and after training.

4.2 Results

Perceived Expertise with LLMs On a 1-7 Likert scale, participants reported increased perceived expertise after training, with mean ratings improving from 3.38 to 3.72 ($\hat{g}_{\text{Hedges}}$=-0.39) and the median rising from 3 to 4.

Regression results from Equation 1 (see Table 1) show that *TrainingStatus* significantly improved perceived expertise (β=0.357, p<0.05) and that older participants rated their perceived expertise to be lower (β=$-$0.083, p<0.05). The former indicates that the training effectively enhanced participants' self-assessed ability to use LLMs, which was not necessarily the case for older participants.

Participants' prompting technique usage in Figure 2 reinforce these findings. In the post-training task, participants used techniques taught in the training more often, especially in in-context prompting and iterative refinement. According to the regression results in Table 9 in Appendix H, we observed that the usage of chain-of-thought, in-context prompting, and few-shot prompting positively influenced participants' perceived expertise (p>0.05). Meanwhile, iterative refinement was negatively associated with perceived expertise (p>0.05).

Perceived Helpfulness of LLMs Despite increased expertise, participants' average perceived helpfulness decreased slightly post-training, from 4.93 to 4.76 ($\hat{g}_{\text{Hedges}}$=0.17), while the median remained stable at 5. Regression analysis revealed a negative effect of *TrainingStatus* on perceived helpfulness (β=$-$0.286, $p > 0.05$). This suggests that prompt engineering training may lower participants' perceived helpfulness of LLMs.

In the discussion, participants expressed frustration regarding stylistic limitations and ChatGPT's cultural adaptability. They described outputs as "not trustable" or similar to a "secondary school dissertation", which highlights incompatibilities, particularly within the Swiss context. Positive remarks emphasized the speed and efficiency of LLMs. Such heightened awareness of the limitations likely contributed to reduced perceived helpfulness. Detailed insights from post-training discussions are provided in Appendix C.

According to the regression results in Supplementary Table 9, the usage of in-context prompting and few-shot

prompting had a positive effect on perceived helpfulness (p>0.05), while iterative refinement and chain-of-thought exhibited a negative association (p>0.05).

	Perceived Expertise	*Perceived Helpfulness*
ArticleOrder	0.063	−0.241
	(0.551)	(0.437)
TrainingStatus	0.357*	−0.286
	(0.164)	(0.153)
Age	−0.083*	−0.037
	(0.033)	(0.026)
Gender	0.800	0.726
	(0.624)	(0.495)
Education	0.026	−0.212
	(0.478)	(0.379)
WorkExperience	0.217	−0.043
	(0.215)	(0.170)
Intercept	6.228***	6.965***
	(1.556)	(1.235)

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 1: Mixed-effect linear regression results on perceived expertise and perceived helpfulness.

4.3 Interpretation of Findings

The results reveal a two-sided impact of prompt engineering training on journalists' experiences with LLMs. On the one hand, participants' perceived expertise in using LLMs significantly improved following the training, as reflected in both self-reported scores and the adoption of more advanced prompting strategies.

On the other hand, though not significant, perceived helpfulness tended to decline after the training. We hypothesize that, as participants interacted more with the LLM and became more proficient, they also became more aware of the models' limitations, particularly in terms of stylistic quality, cultural adaptability, and trustworthiness of outputs. These limitations were not part of the training curriculum, where we discussed data bias and potential inequalities in LLM outputs, privacy, as well as user dependencies on this technology. Hence, we attribute the decrease in perceived helpfulness to the users' repeated interactions with the LLM during the training. The lower perceived helpfulness of LLMs might negatively impact users' interest and willingness to use LLMs in the future. Future training, especially for professionals with strong domain-specific standards and expectations, should directly address these concerns.

5 Effect of Training on Domain Expert Perception (RQ2)

Next, we assess the domain expert's perception of the participants' posts[4]. This is particularly important due to the challenging nature of the task, which involves distinguishing correlational and causal language and interpreting field-specific jargon. In addition, LLMs tend to hallucinate and produce wrong output, which needs to be accounted for.

5.1 Method

To assess the **domain expert perception** of the LLM output, we asked four experts – in this case, academics with experience in both machine learning and medical research at renowned European universities – to conduct a blind evaluation. To ensure reproducibility, we developed a robust scoring system that awards or penalizes each post by one point based on specific criteria (see Table 2). Our scoring system distinguishes different dimensions: (1) We refer to *factual errors* when there is a misuse of causal language, terminology, or incorrect information. We penalize a post with a minus point for each factual error. (2) We refer to *representation errors* as those that convey a misleading or incorrect narrative compared to the original materials. Here, we again penalize each representation error with one minus point (e.g., these are mostly due to hallucination of the LLM, such as using hashtags that are unrelated to the research). (3) We reward posts for having depth (e.g., mentioning limitations of the study). Finally, we calculate an *overall score* by summing over the points awarded for depth minus the negative points for factual errors and representation errors. Appendix E gives examples of participant-generated posts and the corresponding feedback from the domain experts.

Given the non-normal distribution of the collected data, we employ the Wilcoxon-Mann-Whitney test (Wilcox 2011) for trimmed means to conduct between-subject comparisons of each article in terms of the above accuracy scores before and after training. To analyze the relationship between prompting techniques and expert evaluations, we ran the linear regression model (see Eq. (2) in the supplements and Appendix H) to obtain confidence intervals (CIs).

5.2 Results

On average, experts agreed on the ratings 71.8% of the time, which shows consistency in the rating. Overall, ARTICLE PSY showed non-significant improvement in the mean overall score from -1.31 (95% CI = $[-1.89, -0.72]$) before training to -0.69 (95% CI = $[-1.28, -0.10]$) post-training, while the mean overall score decreased non-significantly from 0.13 (CI = $[-0.32, 0.59]$) to -0.62 (CI = $[-1.16, -0.08]$) after training for ARTICLE GUNVIOLENCE. A comparison of the overall score is in Appendix F. Here, we provide a detailed comparison of factual errors, representation errors, and depth (see Figure 3 and Table 3).

Factual errors (Figure 3, left): For ARTICLE PSY, the mean number of factual errors per post reduced from 0.76 (95% CI = [0.50, 1.03]) before training to 0.54 (95% CI = [0.29, 0.78]) after training. Conversely, it increased for ARTICLE GUNVIOLENCE from 0.36 (95% CI = [0.15, 0.58]) before training to 0.48 (95% CI = [0.27, 0.69]) after training. However, caution is warranted as the results are not statistically significant at common significant thresholds. Regression results in Table 10 suggest that the usage of few-shot

[4]To ensure a fair evaluation, we evaluated users' submitted posts rather than LLM outputs directly as we observed that some users prompted the LLM to provide several alternatives and picked the final submission out of them.

Factual errors (subtract 1 point)	**Representation errors** (subtract 1 point)	**Depth** (receive 1 point)
ARTICLE PSY		
Misuse of causal language	Lack of effect size mentioned	Specify subjects being veterans
Misuse of terminologies	Insufficient consideration of heterogeneity effects on study outcomes	Highlight limitations
Incorrect report of numbers	Write the post as they are the researchers	Mention of time-frame of outcome or of population
Incorrect report of concepts	Unrelated hashtags	Complete mention of outcomes
	Incorrect mention of the source of heterogeneity	Mention of the study's baseline as current practice
ARTICLE GUNVIOLENCE		
Misuse of causal language	Absence of reported effect sizes and quantitative results	Specifying geographical context (e.g., US, American)
Incorrect report of study's subject	Incorrect policy implications	Highlight limitations
Incorrect numbers in the posts	Write the post as they are the researchers	Mention of both outcomes
Incorrect messages	Mix different results	Mention detailed inclusion criteria

Table 2: Scoring criteria for evaluating the posts through domain experts. The base score is zero.

prompting and chain-of-thought is associated with fewer factual errors ($p>0.05$).

Representation errors (Figure 3, center): The mean number of representation errors decreased for ARTICLE PSY from 1.01 (95% CI = [0.75, 1.28]) to 0.79 (95% CI = [0.48, 1.10]). For ARTICLE GUNVIOLENCE, it increased from 0.88 (95% CI = [0.72, 1.04]) before training to 1.05 (95% CI = [0.88, 1.21]) after training. All prompting techniques except iterative refinement were associated with decreased representation errors ($p>0.05$) (see Table 10).

Depth (Figure 3, right): The mean score for depth per post increased for ARTICLE PSY from 0.47 (95% CI = [0.12, 0.81]) to 0.63 (95% CI = [0.33, 0.93]) after training. In contrast, it decreased for ARTICLE GUNVIOLENCE from 1.38 (95% CI = [1.12, 1.64]) to 0.91 (95% CI = [0.55, 1.26]) after training. From the regression results Table 10, all prompting techniques, except few-shot prompting, resulted in less depth ($p>0.05$).

We conducted a post-hoc power analysis to determine the minimum sample size to capture the observed effect size. With the minimum effect size of $d = -0.387$ and assuming a power of 0.8 and a two-tailed significance level of $\alpha = 0.05$, the minimum sample size was estimated to be $N = 222$.[5] Evidently, the required sample size would exceed the number of journalists who receive training annually in Switzerland by several orders of magnitude (which is known to pose challenges in human-computer interaction studies involving specialized user groups (Jonassen et al. 2025)). This is why we combine quantitative methods with qualitative insights.

5.3 Interpretation of Findings

Our results show that ARTICLE PSY improved on both errors (fewer factual and representation errors) and depth after training, while ARTICLE GUNVIOLENCE saw decreased

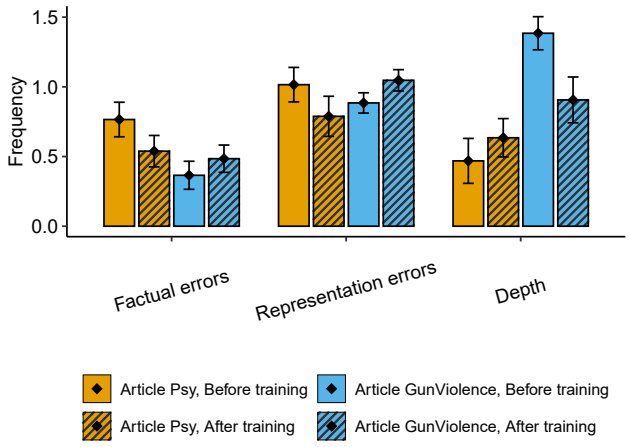

Figure 3: The average number of factual errors, misrepresentation errors, and bonus points for depth. All scores were evaluated by the domain expert. Whiskers refer to standard errors.

depth and higher error counts. This suggests a possible synergy between the depth of posts and their error rates: when participants included more detailed information (ARTICLE PSY), the number of errors declined, which boosted the overall expert score. Conversely, a drop in depth (ARTICLE GUNVIOLENCE) coincided with increased factual and representation errors. A possible explanation of our divergent findings might be that the training examples may have been more aligned with the clinical, data-driven content of ARTICLE PSY, which made it easier for participants to transfer their learned skills. Another factor could be that iterative prompting, encouraged by the training, triggered stricter guardrails in the LLM for the sensitive sociopolitical content of ARTICLE GUNVIOLENCE,leading to increasingly vague or error-prone outputs. Taken together, this highlights that prompt engineering training might not uniformly benefit all

[5]We used $d = \frac{2r_{rb}}{\sqrt{1-r_{rb}^2}}$ to convert our minimum observed r_{rb} to Cohen's d.

	Mean (95% CI)		Median		Effect size	p-value
	Before training	After training	Before training	After training		
Overall Score						
ARTICLE PSY	-1.31 (-1.89, -0.72)	-0.69 (-1.28, -0.10)	-1.50	-0.50	-0.46	< 0.05
ARTICLE GUNVIOLENCE	0.13 (-0.32, 0.59)	-0.62 (-1.16, -0.08)	0.25	-0.62	0.48	< 0.05
Factual Errors						
ARTICLE PSY	0.76 (0.50, 1.03)	0.54 (0.29, 0.78)	0.75	0.50	0.25	0.25
ARTICLE GUNVIOLENCE	0.36 (0.15, 0.58)	0.48 (0.27, 0.69)	0.25	0.50	-0.19	0.39
Representation Errors						
ARTICLE PSY	1.01 (0.75, 1.28)	0.79 (0.48, 1.10)	1.00	0.75	0.32	0.15
ARTICLE GUNVIOLENCE	0.88 (0.72, 1.04)	1.05 (0.88, 1.21)	0.75	1.00	-0.29	0.17
Depth						
ARTICLE PSY	0.47 (0.12, 0.81)	0.63 (0.33, 0.93)	0.12	0.75	-0.24	0.27
ARTICLE GUNVIOLENCE	1.38 (1.12, 1.64)	0.91 (0.55, 1.26)	1.25	1.00	0.44	0.05

Note: We performed Wilcoxon signed-rank test for each group. We reported rank-biserial correlation as the effect size.

Table 3: Statistics testing on before and after training on expert grades.

content types and that its effectiveness may depend heavily on the nature and sensitivity of the topic.

6 Effect of Training on Reader Perception (RQ3)

Next, we assess the **reader perceptions** of the Twitter/X posts across a general audience that was initially targeted by the journalists (i.e., average news readers). To do so, we conducted a pre-registered, post-hoc evaluation using non-experts recruited from prolific.com to evaluate the posts created by our journalists both before and after training.

6.1 Method

An overview of the post-hoc evaluation is shown in Supplementary Figure 5. We examine the effect of prompt engineering training on ten different dimensions of text quality and reader response, as evaluated by non-expert readers. These dimensions are: (1) readability, (2) clarity, (3) informativeness, (4) perceived trustworthiness, (5) appropriateness for the target audience, (6) depth, (7) appropriateness for the mode of communication, (8) engagement, (9) intention to re-share, and (10) intention to seek further information. A full overview is in Table 7.

The above dimensions have been widely employed in the literature to assess the quality of text from a reader's perspective (van der Lee et al. 2019; Celikyilmaz, Clark, and Gao 2020). Our hypothesis is that prompt engineering training may positively influence these aspects of text quality and reader response. The first eight dimensions focus on intrinsic text qualities, while the last two measure behavioral intentions such as information seeking and information sharing, which are both important factors for news consumption and social media more broadly (Pröllochs and Feuerriegel 2023). Together, these dimensions provide a comprehensive evaluation of both content characteristics and potential reader actions. We used Yuen's test (Wilcox 2011) for trimmed means to compare the dimensions before/after training.

Procedure First, we obtained informed consent and participants completed an attention check. We note that we retained all participants in our primary analysis, regardless of their performance on the attention check. However, we conducted a robustness check excluding those who failed the attention check, which yielded consistent results. We randomly assigned each participant to a pair of social media posts from the same journalist so that we could later rule out between-journalist variability in terms of style. For each post, participants responded to the above-mentioned questions along 10 dimensions (i.e., eight pertaining to intrinsic text quality and two pertaining to behavioral intentions). All responses were recorded on a 7-point Likert scale. To mitigate order effects, the sequence of X/Twitter posts was randomized for each participant in the post-hoc evaluation.

Participants We collected responses from 318 participants through the Prolific platform. To maintain data integrity and minimize confounding factors, we implemented strict inclusion criteria. Specifically, we excluded participants who failed to provide their unique Prolific ID in the survey or those who attempted to complete the survey multiple times. This exclusion process was crucial to ensure that each non-expert reader was exposed to posts from only one journalist, thereby eliminating potential bias in our data. After filtering, we had 285 responses; on average, each pair of X/Twitter posts created by the journalists received 19.65 evaluations. Participants' ages ranged from 19 to 77 years old, with a mean of 36.13 years. Out of the 285 non-expert readers, 124 are men, 155 are women, 4 identified as non-binary, and 2 preferred not to answer.

6.2 Results

The effects of training on ARTICLE PSY and ARTICLE GUNVIOLENCE varied across different metrics. The results are shown in both Figure 4 and Table 4. For ARTICLE PSY, we found improvements in *Appropriateness for the target audience* (increased from 4.89 to 5.23), and *Depth* (increased from 4.10 to 4.70). For ARTICLE GUNVIOLENCE, improvements were seen in *Engagement* (increased from

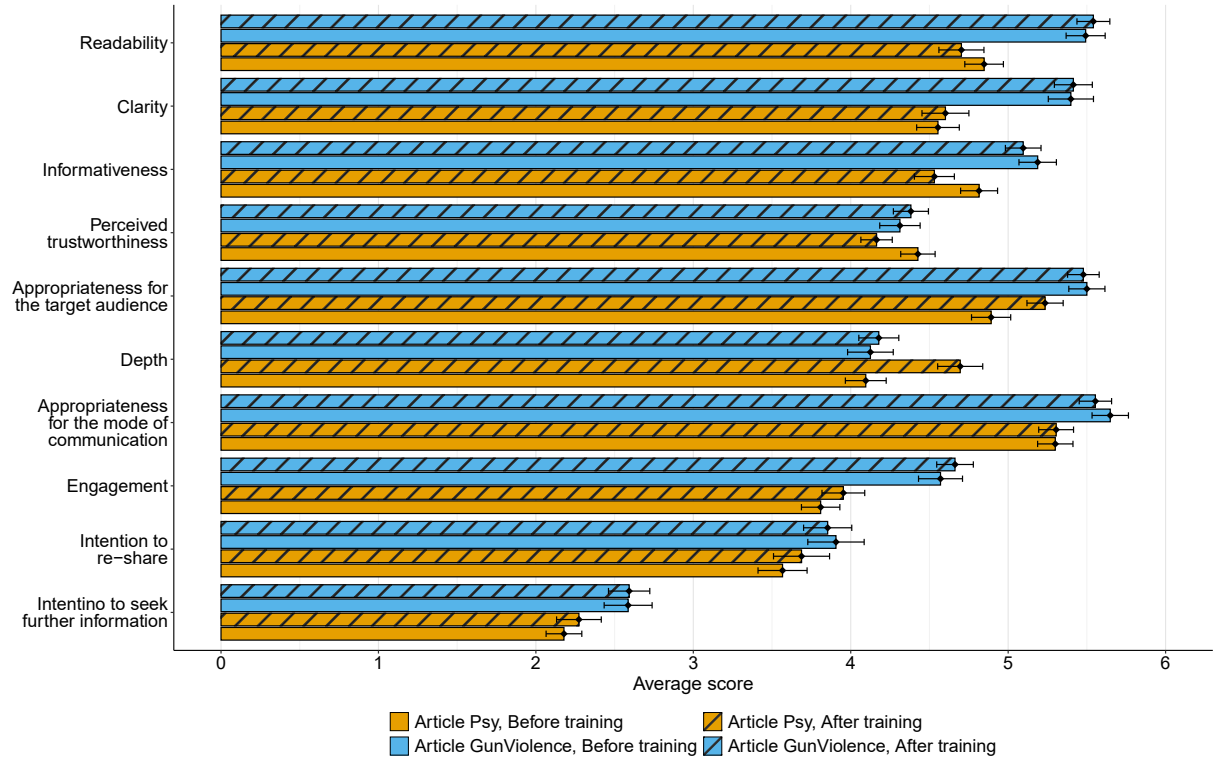

Figure 4: External evaluation of reader perception (assessed by non-expert readers). The bars represent the score averaged over all non-expert readers and posts from the corresponding article. Whiskers represent standard errors.

4.57 to 4.66). Interestingly, some metrics decreased slightly after training, such as *Informativeness* for both articles. For example, for ARTICLE PSY, the *Informativeness* dropped from 4.82 to 4.53, while, for ARTICLE GUNVIOLENCE, it decreased from 5.19 to 5.10. The *Intention to seek further information* remained relatively stable for both articles, with ARTICLE PSY showing a minimal increase from 2.18 to 2.27, and ARTICLE GUNVIOLENCE maintaining the same mean of 2.59 with slightly adjusted confidence intervals. However, for none of the variables, the difference is statistically significant at common thresholds.

6.3 Interpretation of Findings

Our findings indicate that prompt engineering training had a varied impact on the perceived quality of X/Twitter posts by non-expert readers. Specifically, there are slight improvements in dimensions like *Style Appropriateness* and *Depth*, suggesting that journalists may have been able to craft more tailored and detailed content after receiving training. This aligns with the idea that prompt engineering strategies such as chain-of-thought reasoning and persona adoption help to improve the relevance of output text for the target audience (Wei et al. 2022; White et al. 2023).

Further, the small changes in *Engagement* and the slight decline in *Informativeness* may suggest that prompt engineering training can improve technical aspects of writing but does not necessarily translate into more compelling or informative posts from a reader's perspective. Moreover, the *behavioral intentions* to seek further information or to share the posts are fairly similar. This could imply that the training did not significantly impact the overall appeal or effectiveness of the posts in driving reader action. This could be due to the complex nature of social media engagement, where clarity and engagement might be more critical than the technical aspects that prompt engineering tends to enhance.

However, we are careful with our interpretation as the above comparisons were not statistically significant. This may indicate that LLMs can generate content that is of a similar style as professional journalists in the eyes of non-expert readers. Hence, our findings suggest that it may be difficult for non-expert readers to identify meaningful differences between content that is LLM-generated or human-generated. Similar observations were made in studies comparing how people perceive misinformation generated by LLMs vs. humans, finding that non-experts cannot distinguish the veracity of both (Bashardoust, Feuerriegel, and Shrestha 2024).

The above analysis underscores that prompt engineering training should not only focus on the efficiency gains for LLM users but should also account for the downstream impact on the perception of potential audiences. Hence, a good prompt engineering strategy is one that not only improves the precision and depth of content but also addresses how to maintain – or even enhance – information-seeking and reader engagement. This is particularly important in the context of social media, where behavioral intentions to reshare content are crucial for social media posts to go viral. Our

	Mean (95% CI)		Median		Effect size	p-value
	Before training	After training	Before training	After training		
Readability						
ARTICLE PSY	4.84 (4.60, 5.09)	4.70 (4.42, 4.99)	5.00	5.00	0.00	1.00
ARTICLE GUNVIOLENCE	5.49 (5.25, 5.74)	5.54 (5.34, 5.75)	6.00	6.00	0.01	0.91
Clarity						
ARTICLE PSY	4.56 (4.29, 4.82)	4.60 (4.31, 4.90)	5.00	5.00	−0.10	0.35
ARTICLE GUNVIOLENCE	5.40 (5.11, 5.68)	5.41 (5.18, 5.65)	6.00	6.00	0.06	0.59
Informativeness						
ARTICLE PSY	4.81 (4.58, 5.04)	4.53 (4.28, 4.78)	5.00	5.00	0.11	0.31
ARTICLE GUNVIOLENCE	5.19 (4.95, 5.42)	5.10 (4.87, 5.32)	5.00	5.00	0.08	0.44
Engagement						
ARTICLE PSY	3.81 (3.57, 4.05)	3.95 (3.68, 4.22)	4.00	4.00	−0.12	0.29
ARTICLE GUNVIOLENCE	4.57 (4.29, 4.85)	4.66 (4.43, 4.89)	5.00	5.00	−0.08	0.46
Style appropriateness						
ARTICLE PSY	4.89 (4.65, 5.13)	5.23 (5.01, 5.46)	5.00	6.00	−0.23	0.05
ARTICLE GUNVIOLENCE	5.50 (5.27, 5.73)	5.48 (5.28, 5.68)	6.00	6.00	0.08	0.45
Language appropriateness						
ARTICLE PSY	5.30 (5.08, 5.52)	5.30 (5.09, 5.52)	6.00	6.00	−0.03	0.82
ARTICLE GUNVIOLENCE	5.65 (5.42, 5.88)	5.55 (5.35, 5.76)	6.00	6.00	0.01	0.91
Depth						
ARTICLE PSY	4.10 (3.84, 4.36)	4.70 (4.41, 4.98)	5.00	5.00	−0.29	< 0.01
ARTICLE GUNVIOLENCE	4.13 (3.83, 4.41)	4.18 (3.93, 4.43)	4.00	4.00	−0.01	0.91
Perceived trustworthiness						
ARTICLE PSY	4.43 (4.21, 4.64)	4.16 (3.97, 4.36)	4.00	4.00	0.15	0.22
ARTICLE GUNVIOLENCE	4.31 (4.06, 4.57)	4.38 (4.16, 4.60)	4.00	4.00	−0.05	0.67
Intention to reshare						
ARTICLE PSY	3.57 (3.26, 3.87)	3.69 (3.34, 4.03)	4.00	4.00	−0.08	0.47
ARTICLE GUNVIOLENCE	3.90 (3.56, 4.26)	3.85 (3.55, 4.16)	4.00	4.00	0.04	0.74
Intention to seek further information						
ARTICLE PSY	2.18 (1.95, 2.40)	2.27 (1.99, 2.55)	2.00	2.00	−0.13	0.26
ARTICLE GUNVIOLENCE	2.59 (2.28, 2.89)	2.59 (2.33, 2.86)	2.00	2.00	0.001	0.99

We performed Wilcoxon signed-rank test for each group. We reported rank-biserial correlation as the effect size.

Table 4: External evaluation metrics before and after training.

findings align with the broader literature, suggesting that, while AI literacy and prompt engineering are crucial for leveraging LLMs effectively, they must be adapted to the specific needs of the task and audience to be truly impactful (Zamfirescu-Pereira et al. 2023; Ng et al. 2021).

7 Discussion

In this paper, we study the effects of prompt engineering training on perceived user expertise and perceived helpfulness (RQ1), domain expert perception (RQ2), and non-expert reader perception (RQ3) among professional journalists. Our findings reveal that, while our training significantly improved journalists' perceived expertise with LLMs, including an increased use of advanced prompting techniques, it decreased the perceived helpfulness of LLMs. The effect of training on the expert score was mixed, where an increase in errors coincided with a decrease in depth. Lastly, we find a nuanced impact of training on reader perception across different text quality dimensions.

Practical implications: Our findings offer several practical takeaways for the design of prompt engineering training, particularly in professional contexts such as journalism, where accuracy and audience perception are crucial. *First,* prompt engineering trainings should align with task-specific content and domain sensitivities. The mixed effects we find for the two articles – particularly the decline in expert score for ARTICLE GUNVIOLENCE – suggest that prompt engineering strategies may not generalize equally well across domains. This implies that future training should be adapted to the characteristics of the subject, especially when dealing with sensitive topics. For example, training should include domain-specific case studies and emphasize the potential limitations of prompting in sensitive content areas.

Second, prompt engineering trainings should address the trade-off between user expertise and critical perception. Our training successfully increased journalists' perceived expertise and lead to the adoption of more advanced prompting techniques. However, this gain in strategic engagement was accompanied by a decline in perceived helpfulness of the model, as participants became more aware of its shortcomings during repeated interactions with the LLM. This highlights a broader tension: as users become more skilled, they may also become more critical. Training designers should

anticipate this by incorporating modules on model limitations, failure modes, and appropriate fallback strategies. This not only helps set realistic expectations but may also help users to deal with unsatisfying outcomes. This helps to sustain user motivation, particularly among professionals who rely on LLMs for high-quality output.

Third, effective prompt engineering should optimize for both internal quality and external audience perception. Our non-expert readers' evaluation showed that improvements in technical user capabilities did not necessarily translate into stronger behavioral engagement or informativeness. This suggests that prompt engineering should not only focus on optimizing responses from the LLM but also consider how outputs will be received by target audiences (e.g., in medicine, a prompt may be more effective if the output can be verified by clinicians (Spitzer et al. 2025)).

Limitations and future research: While our study provides valuable insights into the impact of prompt engineering training, several limitations present opportunities for future research. First, the relatively small sample size of 29 participants, though consistent with prior experiments involving professionals (Li et al. 2024; Huang et al. 2020; Senoner et al. 2024), may limit the generalizability of our findings. Future research could expand the sample size and include participants from diverse backgrounds, including other high-stake professions such as law, policy, and education, to see whether similar patterns in perceived expertise, perceived helpfulness, expert scores, and reader perception emerge. Nevertheless, our analysis is based on domain experts (journalists), which corroborates the ecological validity of our findings but which introduces natural limits to the maximum available sample size. Second, future research should also explore the long-term effects of prompt engineering training, particularly as LLMs continue to evolve. Third, we only use one LLM, namely ChatGPT-3.5, which may not reflect the capabilities of other LLMs. Yet, ChatGPT-3.5 was widely accessible to the public at the time of the experiment and its performance was state-of-the-art. Fourth, our findings are mixed and some post-hoc analyses lacked statistically significant results, potentially due to the exploratory nature of our study. This suggests that prompt engineering may not yield pronounced and consistent improvements in output as perceived by the readership.

Further, our work underscores the role of teaching good prompting to equip users with the skills to enhance their interactions with LLMs. Unlike studies that compare different prompting strategies, our research focuses on the human aspect of crafting prompts. This highlights an important area for further exploration: how personalized and context-specific prompting strategies can be taught to consistently improve task outcomes across different scenarios. Understanding the nuances of effective prompting is essential for professionals aiming to leverage LLMs for more accurate and relevant results in their work (Feuerriegel et al. 2025).

Acknowledgments

Funding by the Swiss National Science Foundation (Grant: 215542) and the German Research Foundation (Grant: 543018872) is acknowledged.

References

Atreja, S.; Ashkinaze, J.; Lingyao, L.; Mendelsohn, J.; and Hemphill, L. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *arXiv:2406.11980*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; Do, Q. V.; Xu, Y.; and Fung, P. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *AACL*.

Bashardoust, A.; Feuerriegel, S.; and Shrestha, Y. R. 2024. Comparing the willingness to share for human-generated vs. AI-generated fake news. *CSCW*.

Beauchene, V.; de Bellefonds, N.; Duranton, S.; and Mills, S. 2023. AI at work: What people are saying. *BCG Global*.

Benk, M.; Weibel, R. P.; Feuerriegel, S.; and Ferrario, A. 2022. "Is it my turn?": Assessing teamwork and taskwork in collaborative immersive analytics. *CSCW*.

Celikyilmaz, A.; Clark, E.; and Gao, J. 2020. Evaluation of text generation: A survey. *arXiv:2006.14799*.

Dang, H.; Mecke, L.; Lehmann, F.; Goller, S.; and Buschek, D. 2022. How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. *CHI Workshops*.

Felten, E. W.; Raj, M.; and Seamans, R. 2023. Occupational heterogeneity in exposure to generative AI. *SSRN Electronic Journal*.

Feng, Y.; Poralla, P.; Dash, S.; Li, K.; Desai, V.; and Qiu, M. 2023. The impact of ChatGPT on streaming media: A crowdsourced and data-driven analysis using Twitter and Reddit. *International Conference on Big Data Security on Cloud*.

Feuerriegel, S.; Hartmann, J.; Janiesch, C.; and Zschech, P. 2023. Generative AI. *Business and Information Systems Engineering*, 66: 111–126.

Feuerriegel, S.; Maarouf, A.; Bär, D.; Geissler, D.; Schweisthal, J.; Pröllochs, N.; Robertson, C. E.; Rathje, S.; Hartmann, J.; Mohammad, S. M.; Netzer, O.; Siegel, A. A.; Plank, B.; and van Bavel, J. J. 2025. Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 4.

Gaito, J. 1961. Repeated measurements designs and counterbalancing. *Psychological Bulletin*, 58(1): 46–54.

Guenther, L.; and Joubert, M. 2017. Science communication as a field of research: Identifying trends, challenges and gaps by analysing research papers. *Journal of Science Communication*, 16(2): A02.

Huang, C. A.; Koops, H. V.; Newton-Rex, E.; Dinculescu, M.; and Cai, C. J. 2020. AI song contest: Human-AI co-creation in songwriting. *arXiv:2010.05388*.

Huang, J.; and Chang, K. C.-C. 2022. Towards reasoning in large language models: A survey. *arXiv:2212.10403*.

Jahani, E.; Manning, B.; Zhang, J.; TuYe, H.-Y.; Alsobay, M. A. M.; Nicolaides, C.; Suri, S.; and Holtz, D. 2024. As generative models improve, people adapt their prompts. *OSF Preprint*.

Jensen, E. A.; and Gerber, A. 2020. Evidence-based science communication. *Frontiers in Communication*, 4.

Jonassen, Z.; Lawrence, K.; Wiesenfeld, B. M.; Feuerriegel, S.; and Mann, D. 2025. A Qualitative Analysis of Remote Patient Monitoring: How a Paradox Mindset Can Support Balancing Emotional Tensions in the Design of Healthcare Technologies. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Kandlhofer, M.; Steinbauer, G.; Hirschmugl-Gaisch, S.; and Huber, P. 2016. Artificial intelligence and computer science in education: From kindergarten to university. *IEEE Frontiers in Education Conference*.

Kim, Y.; Lee, J.; Kim, S.; Park, J.; and Kim, J. 2024. Understanding users' dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge level. *International Conference on Intelligent User Interfaces*.

Knoth, N.; Tolzin, A.; Janson, A.; and Leimeister, J. M. 2024. AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6: 100225.

Laupichler, M. C.; Aster, A.; Schirch, J.; and Raupach, T. 2022. Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 3: 100101.

Li, J.; Cao, H.; Lin, L.; Hou, Y.; Zhu, R.; and El Ali, A. 2024. User experience design professionals' perceptions of generative artificial intelligence. *CHI*.

Lin, Z. 2024. How to write effective prompts for large language models. *Nature Human Behavior*, 8: 611–615.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Long, D.; and Magerko, B. 2020. What is AI literacy? Competencies and design considerations. *CHI*.

Los Angeles Times. 2019. What is the Quakebot and how does it work?

Markus, A.; Pfister, J.; Carolus, A.; Hotho, A.; and Wienrich, C. 2024. Effects of AI understanding-training on AI literacy, usage, self-determined interactions, and anthropomorphization with voice assistants. *Computers and Education Open*, 6: 100176.

McGuire, J.; de Cremer, D.; and van de Cruys, T. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports*, 14.

Meskó, B. 2023. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, 25: e50638.

Moore, R. C.; and Hancock, J. T. 2022. A digital media literacy intervention for older adults improves resilience to fake news. *Scientific Reports*, 12(1): 6008.

Mosbach, M.; Pimentel, T.; Ravfogel, S.; Klakow, D.; and Elazar, Y. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv:2305.16938*.

Ng, D. T. K.; Leung, J. K. L.; Chu, S. K. W.; and Qiao, M. S. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2: 100041.

OpenAI. 2024. Prompt engineering.

Pinski, M.; and Benlian, A. 2023. AI literacy - Towards measuring human competency in artificial intelligence. *Hawaii International Conference on System Sciences*.

Pinski, M.; and Benlian, A. 2024. AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans*, 2(1).

Pröllochs, N.; and Feuerriegel, S. 2023. Mechanisms of true and false rumor sharing in social media: Collective intelligence or herd behavior? *CSCW*.

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3(38).

Ross, E. L.; Bossarte, R. M.; Dobscha, S. K.; Gildea, S. M.; Hwang, I.; Kennedy, C. J.; Liu, H.; Luedtke, A.; Marx, B. P.; Nock, M. K.; Petukhova, M. V.; Sampson, N. A.; Zainal, N. H.; Sverdrup, E.; Wager, S.; and Kessler, R. C. 2024. Estimated average treatment effect of psychiatric hospitalization in patients with suicidal behaviors: A precision treatment analysis. *JAMA Psychiatry*, 81(2): 135–143.

Saravia, E. 2022. Prompt Engineering Guide.

Schmidt, H. 2024. Generative KI im Journalismus: Texte erstellen und Informationen suchen stehen im Vordergrund. *Frankfurter Allgemeine Zeitung*.

Schulhoff, S.; Ilie, M.; Balepur, N.; Kahadze, K.; Liu, A. A.; Si, C.; Li, Y.; Gupta, A.; Han, H.; Schulhoff, S.; Dulepet, P. S.; Vidyadhara, S.; Ki, D.; Agrawal, S.; Pham, C.; Kroiz, G.; Li, F.; Tao, H.; Srivastava, A.; Da Costa, H.; Gupta, S.; Rogers, M. L.; Goncearenco, I.; Sarli, G.; Galynker, I.; Peskoff, D.; Carpuat, M.; White, J.; Anadkat, S.; Hoyle, A.; and Resnik, P. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv:2406.06608*.

Semenza, D. C.; Daruwala, S.; Brooks Stephens, J. R.; and Anestis, M. D. 2024. Gun violence exposure and suicide among black adults. *JAMA Network Open*, 7(2): e2354953.

Senoner, J.; Schallmoser, S.; Kratzwald, B.; Feuerriegel, S.; and Netland, T. 2024. Explainable AI improves task performance in human-AI collaboration. *Scientific Reports*, 14.

Spitzer, P.; Hendriks, D.; Rudolph, J.; Schlaeger, S.; Ricke, J.; Kühl, N.; Hoppe, B. F.; and Feurriegel, S. 2025. The effect of medical explanations from large language models on diagnostic decisions in radiology. *medRxiv 2025.03.04.25323357*.

Statista. 2024. Social media and artificial intelligence (AI) - statistics & facts.

Tenenboim-Weinblatt, K.; and Baden, C. 2018. Journalistic transformation: How source texts are turned into news stories. *Journalism*, 19(4): 481–499.

Theophilou, E.; Koyutürk, C.; Yavari, M.; Bursic, S.; Donabauer, G.; Telari, A.; Testa, A.; Boiano, R.; Hernandez-Leo, D.; Ruskov, M.; Taibi, D.; Gabbiadini, A.; and Ognibene, D. 2023. Learning to prompt in the classroom to understand AI limits: A pilot study. *AIxIA – Advances in Artificial Intelligence*, 14318: 481–496.

UNESCO. 2022. K-12 AI curricula: A mapping of government-endorsed AI curricula.

van der Lee, C.; Gatt, A.; van Miltenburg, E.; Wubben, S.; and Krahmer, E. 2019. Best practices for the human evaluation of automatically generated text. *ACL*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv:2302.11382*.

Wilcox, R. R. 2011. *Introduction to robust estimation and hypothesis testing*. Academic Press.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large language models as optimizers. *ICLR*.

Yang, W. 2022. Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation. *Computers and Education: Artificial Intelligence*, 3: 100061.

Zamfirescu-Pereira, J. D.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023. Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *CHI*.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) no

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA)

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA)

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA)

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA)

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA)

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA)

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) no

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA)

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no)

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no)

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no)

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no)

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA)

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no)

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no)

4.10. This paper states the number of algorithm runs used

to compute each reported result (yes/no)

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no)

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no)

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA)

Appendix

A Improving AI Literacy

AI literacy refers to "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (Long and Magerko 2020). Given the increasing integration of AI into professional environments, AI literacy plays a crucial role in preparing workers to effectively interact with AI technologies (Pinski and Benlian 2023). As a result, there is a growing demand for AI literacy training programs that equip individuals with the necessary skills to make effective use of AI technology (Ng et al. 2021). To achieve this goal, UNESCO developed a curriculum that defines essential learning domains for AI literacy training. These include AI fundamentals, ethical considerations, societal impacts, and AI application and development (UNESCO 2022). Our work advances this effort by extending the understanding of how to effectively enhance users' practical skills in prompt engineering, resulting in improved output quality when interacting with LLMs.

Previous work found that AI literacy trainings need to adjust their structure and content for different audiences (Yang 2022; Laupichler et al. 2022; Pinski and Benlian 2024) and has developed various trainings for AI literacy. For example, Kandlhofer et al. (2016) designed AI literacy trainings for different age groups from kindergarten to university and found that the trainings improved AI literacy. Another work (Markus et al. 2024) developed online trainings, where participants were taught how AI technologies function to improve their understanding of intelligent voice assistants. The work found that the trainings (i) led to an improved understanding of the potential and risks of intelligent voice assistants, (ii) promoted effective interactions with the voice assistants, and (iii) elicited user interest in exploring the technology. Moore and Hancock (2022) designed a 1-hour interactive module where participants were trained in recognizing online misinformation. The authors found that the training improves truth discernment, especially for older adults. Theophilou et al. (2023) studied how prompt engineering training can help users understand the opportunities and limits of LLMs. The authors found that training reduced negative sentiments toward LLMs and that students had a better understanding of the limitations of LLMs. Yet, unlike our work, this research focuses on understanding AI limitations and does not consider the output quality of LLMs. Still, empirical evidence is missing on how prompt engineering training can affect task-specific output quality, particularly in professional environments.

B Prompt Engineering Training Content

The training consisted of two parts: an introduction to prompt engineering and practical tips. Table 5 contains the outline of the prompt engineering training during the two hours session with journalists.

Topic	Key concepts
Part 1: Introduction to prompt engineering (55 minutes)	
Fundamentals of prompts	Core elements of prompts
Prompt applications	Text summarization, question and answer, text classification, and role playing
Prompt techniques	Zero shot, few shot, and chain-of-thought
Part 2: Practical aspects (55 minutes)	
Prompt optimization	Iterative process for increasing the quality of prompts
Risks and challenges	Data bias, inequality, data privacy

Table 5: Outline of the prompt engineering training.

C Post-Training Discussions

Here are additional findings from the oral interviews that we conducted after the experiment. We focus on some notable impressions from the discussions with the journalists.

Question: What is your experience with ChatGPT?
Responses:

- *"It is good for sorting information, but the style is very low (secondary school dissertation). The style and originality are not high-level."*
- *"Speed up really much, but the outcome is not up to expectation. We could not convert to the final 100% outcome that we would like."*
- *"When it does not listen to the commands, it makes the journalist frustrated and reduces creativity. It burns the energy and makes us ask, do I really want to work with it? At this stage, they are not trustable."*
- *"Concerns: These tools are developed in the US, and they are not necessarily adapted to Europe, especially Swiss culture."*

Findings: Overall, the participants expressed dissatisfaction with the style of the LLM outputs and indicated that the LLM did not achieve the results as desired. The LLM was primarily used to generate initial drafts. Additionally,

one participant raised a concern that relying on such technology might reduce cognitive exercise, potentially weakening mental acuity.

D Post-hoc Evaluation Overview

Figure 5 illustrates the overview of our post-hoc evaluation of non-expert reader perceptions.

E Example of Expert Evaluation

Table 6 illustrates example of Twitter/X post with accuracy assessment from domain expert.

ARTICLE PSY
Post: *"New study shows psychiatric hospitalization reduces immediate suicide attempt risk but varies by patient history. Precision treatment could cut suicide attempts by 16% and reduce hospitalizations by 13%. #MentalHealthAwareness #SuicidePrevention"* **Accuracy assessment:** The use of "reduce" suggests causal language (*factual error*). The post further fails to address the study's heterogeneity in its sample (representation error). However, it successfully covers multiple research outcomes (*depth*).
ARTICLE GUNVIOLENCE
Post: *"Hey, young changemakers! Did you know gun violence hits Black youth hard, leading to higher rates of suicidal ideation & attempts? Let's unite to demand safer communities for our generation. #YouthForChange #MentalHealthAwareness"* **Accuracy assessment:** The post implies causal language and incorrectly reports the subjects as "youth" (2x *factual error*). The post does not mention any quantitative result or effect size (*representation error*). The post mentions both outcomes (*depth*).

Table 6: Examples of written Twitter/X posts with additional feedback about the accuracy from the domain expert.

F Results for Expert Overall Score

The comparison of the expert overall score is in Figure 6. We calculate an overall score by summing over the points awarded for depth minus the negative points for factual errors and representation errors.

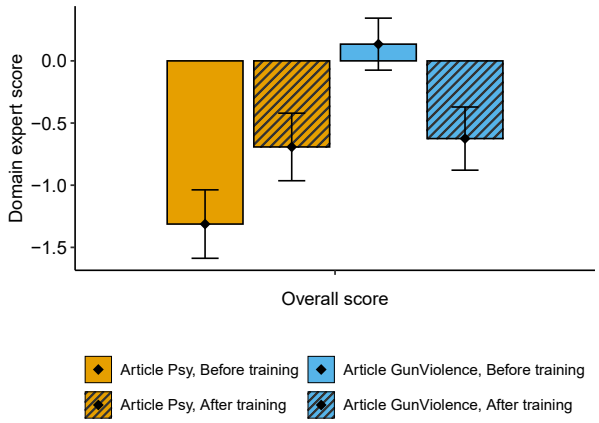

Figure 6: The overall score measuring accuracy (as assessed by the domain expert). Whiskers refer to standard errors.

G Summary of Evaluation Variables

Table 7 summarize the variables used to evaluate the reader perception.

Variable name	Question
Intrinsic text quality	
Readability	*The post is easily readable.*
Clarity	*While reading, I immediately understood the post.*
Informativeness	*The post is informative.*
Perceived trustworthiness	*I trust the information in the post.*
Appropriateness for the target audience	*The language used in the post is appropriate for X/Twitter.*
Depth	*The post is superficial and lacks detail.*
Appropriateness for the mode of communication	*The writing style of the post is appropriate for X/Twitter.*
Engagement	*The post is engaging.*
Behavioral intentions	
Intention to re-share	*Would you share this post on your social media?*
Intention to seek further information	*Would you seek more information on this topic?*

Table 7: Summary of variables used for assessing the reader perception and the corresponding survey questions. All variables were collected on a 7-point Likert scale.

H Prompting Techniques

To analyze how participants applied prompting techniques during their interactions with ChatGPT, one of the authors manually annotated each chat session. The annotation followed a deductive coding scheme grounded in prior research on LLM-prompting strategies. We identified whether participants employed any of these prompting techniques: in-context prompting (*InContext*), chain-of-thought (*CoT*), iterative refinement (*IterativeRefine*), and few-shot prompting (*FewShot*). Each technique was coded as a binary variable (1

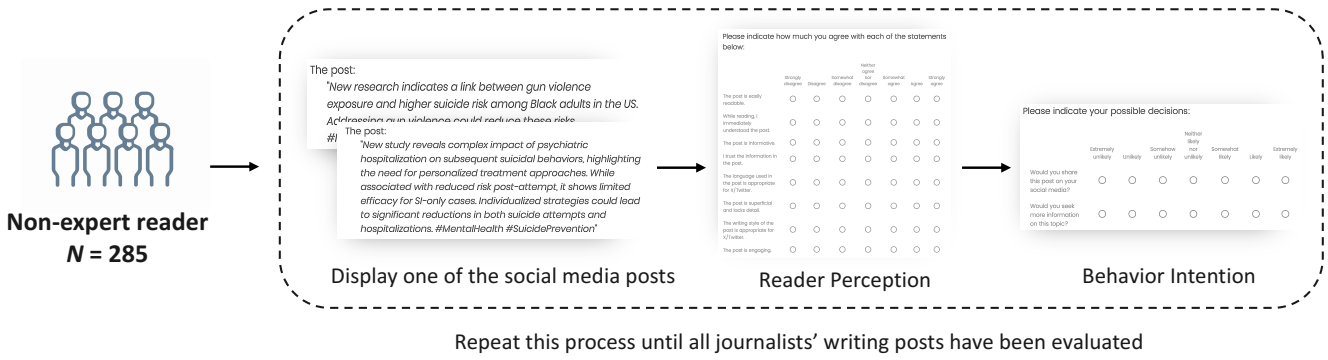

Figure 5: Procedure for external, post-hoc evaluation in terms of non-expert reader perceptions (e.g., clarity, engagement).

Technique	Annotation criteria
ContextPrompt	Add background or contextual details to enhance relevance or accuracy.
CoT	Prompt the model to reason step by step or include intermediate steps.
IterativeRefine	Revise prompts across multiple turns to improve responses.
FewShot	Include a few examples to demonstrate the desired task or output.

Table 8: Annotation criteria for prompting techniques used in participants' ChatGPT dialogues.

= used at least once; 0 = not used). Table 8 provides definitions and criteria used for annotation.

	Perceived Expertise	*Perceived Helpfulness*
ArticleOrder	0.253	0.009
	(0.548)	(0.454)
InContext	0.330	0.482
	(0.335)	(0.308)
CoT	0.559	−0.226
	(0.608)	(0.546)
FewShot	0.592	0.153
	(1.450)	(1.196)
IterativeRefine	−0.263	−0.497
	(0.395)	(0.361)
Intercept	3.340***	4.862***
	(0.501)	(0.430)

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 9: Mixed-effects linear regression results for perceived expertise and helpfulness.

We then modeled the effect of the prompting techniques on the participant outcomes and the expert evaluation outcomes. The regression equation is

$$Y_i = \beta_0 + u_{0i} + \beta_1 \textit{ArticleOrder}_i + \beta_2 \textit{InContext}_i + +\beta_3 \textit{CoT}_i \\ \beta_4 \textit{FewShot}_i + \beta_5 \textit{IterativeRefine}_i + \varepsilon_i, \tag{2}$$

where Y_i denotes *Perceived Expertise*, *Perceived Helpfulness* or three expert evaluation dimensions (*Factual errors*, *Representation errors*, and *Depth*) for the i-th participant. In-context prompting (*InContext*), Chain-of-Thought (*CoT*), iterative refinement (*IterativeRefine*), and few-shot prompting (*FewShot*) are binary variables indicating whether the participant used each prompting technique. β_1 to β_5 are coefficients. *ArticleOrder* controls the counterbalanced order of article presentation. The participant-level random intercept u_{0i} accounts for individual variability, and ε_i represents the error term.

	Factual	*Representation*	*Depth*
ArticleOrder	−0.148	−0.236	0.290
	(0.133)	(0.124)	(0.186)
InContext	−0.017	0.118	−0.071
	(0.131)	(0.128)	(0.191)
CoT	0.363	0.172	−0.167
	(0.204)	(0.194)	(0.290)
FewShot	0.606	0.059	0.050
	(0.341)	(0.318)	(0.475)
IterativeRefine	−0.045	−0.060	−0.339
	(0.148)	(0.143)	(0.213)
Intercept	0.581***	1.008***	0.985***
	(0.142)	(0.134)	(0.200)

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 10: Mixed-effects regression results for expert-evaluated factual accuracy, representation, and depth.

I Scientific Articles Used in the Experiment

Here are the two article summaries for the writing tasks during the experiment.

ARTICLE PSY

The following abstract is quoted from (Ross et al. 2024):

Estimated Average Treatment Effect of Psychiatric Hospitalization in Patients With Suicidal Behaviors. A Precision Treatment Analysis

Importance: Psychiatric hospitalization is the standard of care for patients presenting to an emergency department (ED) or urgent care (UC) with high suicide risk. However, the effect of hospitalization in reducing subsequent suicidal behaviors is poorly understood and likely heterogeneous.
Objectives: To estimate the association of psychiatric hospitalization with subsequent suicidal behaviors using observational data and develop a preliminary predictive analytics individualized treatment rule accounting for heterogeneity in this association across patients.
Design, Setting, and Participants: A machine learning analysis of retrospective data was conducted. All veterans presenting with suicidal ideation (SI) or suicide attempt (SA) from January 1, 2010, to December 31, 2015, were included. Data were analyzed from September 1, 2022, to March 10, 2023. Subgroups were defined by primary psychiatric diagnosis (nonaffective psychosis, bipolar disorder, major depressive disorder, and other) and suicidality (SI only, SA in past 2-7 days, and SA in past day). Models were trained in 70.0% of the training samples and tested in the remaining 30.0%.
Exposures: Psychiatric hospitalization vs nonhospitalization.
Main Outcomes and Measures: Fatal and nonfatal SAs within 12 months of ED/UC visits were identified in administrative records and the National Death Index. Baseline covariates were drawn from electronic health records and geospatial databases.
Results: Of 196 610 visits (90.3% men; median [IQR] age, 53 [41–59] years), 71.5% resulted in hospitalization. The 12-month SA risk was 11.9% with hospitalization and 12.0% with nonhospitalization (difference, −0.1%; 95% CI, −0.4% to 0.2%). In patients with SI only or SA in the past 2 to 7 days, most hospitalization was not associated with subsequent SAs. For patients with SA in the past day, hospitalization was associated with risk reductions ranging from −6.9% to −9.6% across diagnoses. Accounting for heterogeneity, hospitalization was associated with a reduced risk of subsequent SAs in 28.1% of the patients and increased risk in 24.0%. An individualized treatment rule based on these associations may reduce SAs by 16.0% and hospitalizations by 13.0% compared with current rates.
Conclusions and Relevance: The findings of this study suggest that psychiatric hospitalization is associated with reduced average SA risk in the immediate aftermath of an SA but not after other recent SAs or SI only. Substantial heterogeneity exists in these associations across patients. An individualized treatment rule accounting for this heterogeneity could both reduce SAs and avert hospitalizations.

Gun Violence Exposure and Suicide Among Black Adults

Importance: Black individuals are disproportionately exposed to gun violence in the US. Suicide rates among Black US individuals have increased in recent years.
Objective: To evaluate whether gun violence exposures (GVEs) are associated with suicidal ideation and behaviors among Black adults.
Design, Setting, and Participants: This cross-sectional study used survey data collected from a nationally representative sample of self-identified Black or African American (hereafter, Black) adults in the US from April 12, 2023, through May 4, 2023.
Exposures: Ever being shot, being threatened with a gun, knowing someone who has been shot, and witnessing or hearing about a shooting.
Main Outcomes and Measures: Outcome variables were derived from the Self-Injurious Thoughts and Behaviors Interview, including suicidal ideation, suicide attempt preparation, and suicide attempt. A subsample of those exhibiting suicidal ideation was used to assess for suicidal behaviors.
Results: The study sample included 3015 Black adults (1646 [55%] female; mean [SD] age, 46.34 [0.44] years [range, 18–94 years]). Most respondents were exposed to at least 1 type of gun violence (1693 [56%]), and 300 (12%) were exposed to at least 3 types of gun violence. Being threatened with a gun (odds ratio [OR], 1.44; 95% CI, 1.01–2.05) or knowing someone who has been shot (OR, 1.44; 95% CI, 1.05–1.97) was associated with reporting lifetime suicidal ideation. Being shot was associated with reporting ever planning a suicide (OR, 3.73; 95% CI, 1.10–12.64). Being threatened (OR, 2.41; 95% CI, 2.41–5.09) or knowing someone who has been shot (OR, 2.86; 95% CI, 1.42–5.74) was associated with reporting lifetime suicide attempts. Cumulative GVE was associated with reporting lifetime suicidal ideation (1 type: OR, 1.69 [95% CI, 1.19–2.39]; 2 types: OR, 1.69 [95% CI, 1.17–2.44]; ≥3 types: OR, 2.27 [95% CI, 1.48–3.48]), suicide attempt preparation (≥3 types; OR, 2.37; 95% CI, 2.37–5.63), and attempting suicide (2 types: OR, 4.78 [95% CI, 1.80–12.71]; ≥3 types: OR, 4.01 [95% CI, 1.41–11.44]).
Conclusions and Relevance: In this cross-sectional study, GVE among Black adults in the US was significantly associated with lifetime suicidal ideation and behavior. Public health efforts to substantially reduce interpersonal gun violence may yield additional benefits by decreasing suicide among Black individuals in the US.

ARTICLE GUNVIOLENCE

The following abstract is quoted from (Semenza et al. 2024):