Study Information

We are organizing a workshop on prompt engineering (intervention) for a group of science communication journalists from Switzerland. They will be given two tasks on scientific communication and are expected to use ChatGPT. The aim is to study the effect of prompt engineering training on text quality.

Hypothesis

H1: We hypothesize that participants, after receiving the intervention (course on prompt engineering), would deliver higher-quality output (with regard to the dimensions, which are listed under dependent variables).
H2: We hypothesize that the perceived expertise of using ChatGPT after the intervention would increase.

Study Design

Study type

We propose a counterbalanced repeated measures design to investigate the effect of a course on prompt engineering on the quality of created outcomes for a scientific communication task. Participants will face two conditions: half will face task A before the course and task B after the course, and the rest will face task B beforehand and task A afterward.

Blinding

There is no blinding in this study.

Study design

This study will be conducted at a workshop for science communication journalists in Switzerland. Before the course, they will be asked to complete a scientific communication task using ChatGPT. After the course, they will be asked to do the task one more time with different materials. For each task, they should transform the abstract of a scientific paper into an X/Twitter post with a recommended maximum length of 280 characters. Before beginning the first task, we will collect variables like age, educational background, and years of experience as a scientific journalist. After each task, participants will be asked to evaluate their expertise with ChatGPT and their perceived helpfulness of ChatGPT. Besides the collected variables, we will run a post-hoc analysis of the X/Twitter posts they created. The first post-hoc analysis will evaluate the accuracy of the posts using experts. The second post-hoc analysis will be through a sub-study. We will ask raters (who will be hired from an online platform) to rate the secondary variables for two posts from Task A and Task B. We will collect ten ratings for each of the posts from Task A and Task B. Then, we will test if there is an effect before and after the intervention.

Sub-study

In order to investigate the user's perception of created X/Twitter posts, we will run a between-group design sub-study in which each social media post will be rated by ten different persons (raters). Each rater evaluates the texts from tasks A and B of one journalist; hence, the rater only evaluates two texts in total. All the variables in the sub-study are on a 7-level Likert scale. For each X/Twitter post per person, these secondary variables will be averaged on raters' scores. Raters at the beginning of the survey will face an attention test; if they pass the attention test, they will be shown one X/Twitter from task A and one from task B, and then these posts will be rated on all secondary variables.

Randomization

Principle study
We will randomly assign each participant to each of the two conditions.

Sub-study
We will randomly assign one post from Task A and one post from Task B to raters.

Sampling Plan

Existing data
The preregistration is done prior to collecting the data.

Data collection procedure

Principle study
We will collect the data during the workshop. The first round will be conducted before the intervention, and the second round will be conducted after the intervention. The data collection will be conducted through Qualtrics, and we ask the participants to share their logs from ChatGPT with us as well.

Sub-study
We will use an online platform to hire raters to evaluate the created X/Twitter posts based on secondary variables. The data collection platform will be through Qualtrics as well.

Sample size

Principle study
Based on the registration to the workshop, we estimate that there will be a participation size of around 40 journalists.

Sub-study
The sample size for the sub-study is at least 10 times the number of participants.

Variables

Manipulated variables

We manipulate the conditions (sequence of tasks) of the participants.

Measured variables

Measured variables at the workshop

- Independent variables:
    - Age
    - Educational background
    - Years of experience as a science communication journalist
    - Expertise with ChatGPT (7-level Likert scale)
        - Question: How do you rate your expertise in using LLM such as ChatGPT?
    - Helpfulness of ChatGPT within the task (7-level Likert scale)

Measured variables after the workshop

- Variable on expert's evaluation of social media post
    - Accuracy (How accurate the X/Twitter post reflects the abstracts)
- Variables on crowd evaluation of social media post (evaluated through sub-study all on 7-level Likert scale)
    - Clarity
    - Readability
    - Engagement
    - Appropriateness of language for target audience (use of jargon)
    - Appropriate language for the mode of communication (overall writing style)
    - Perceived trustworthiness
    - Informativeness
    - Depth
    - Behavioral intentions (intention to seek further information; intention to reshare)

Analysis Plan

Statistical models

We will use subject-level linear regression and control for the type of task and the participants' conditions.

Additional analysis:

We will check if there is a difference between the secondary variables and the conditions. This will be achieved using the Chi-squared test. Another analysis will test if their perceived

experiences and ChatGPT’s helpfulness change after the intervention. This will be achieved using the Chi-squared test as well.

Transformations

Inference criteria

Data exclusion

Principle study
If a subject fails to complete a task, we will exclude that from our data.

Sub-study data exclusion:
If raters fail the attention test, they will be excluded from the data.

Missing data

Exploratory analysis
Through an oral discussion after the class, we will ask about participants’ perceptions of the course and use ChatGPT on the task.