PDF Download
3701571.3701586.pdf
08 January 2026
Total Citations: 3
Total Downloads: 995

Published: 01 December 2024

Citation in BibTeX format

MUM '24: International Conference on Mobile and Ubiquitous Multimedia
December 1 - 4, 2024
Stockholm, Sweden

RESEARCH-ARTICLE

An LLM-driven Transcription Task for Mobile Text Entry Studies

A. KOMNINOS, University of Patras, Rio, Achaia, Greece

ANNA MARIA FEIT, Saarland University, Saarbrucken, Saarland, Germany

LUIS A. LEIVA, University of Luxembourg, Esch-sur-Alzette, Luxembourg

FLORIAN LEHMANN, University of Bayreuth, Bayreuth, Bayern, Germany

IOULIA SIMOU, University of Patras, Rio, Achaia, Greece

DIMOSTHENIS MINAS, University of Patras, Rio, Achaia, Greece

View all

Open Access Support provided by:

Saarland University

University of Patras

University of Bayreuth

University of Luxembourg

MUM '24: Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (December 2024)
https://doi.org/10.1145/3701571.3701586
ISBN: 9798400712838

An LLM-driven Transcription Task for Mobile Text Entry Studies

Andreas Komninos
University of Patras
Rio, Greece
akomninos@ceid.upatras.gr

Anna Maria Feit
University of Saarland
Saarbrücken, Germany
feit@cs.uni-saarland.de

Luis A. Leiva
University of Luxembourg
Luxembourg, Luxembourg
luis.leiva@uni.lu

Florian Lehmann
University of Bayreuth
Bayreuth, Germany
Florian.Lehmann@uni-bayreuth.de

Ioulia Simou
University of Patras
Rio, Greece
simou@ceid.upatras.gr

Dimosthenis Minas
University of Patras
Rio, Greece
d.minas@ac.upatras.gr

Aggelos Fotopoulos
University of Patras
Rio, Greece
fotopoulos@ceid.upatras.gr

Michalis Xenos
University of Patras
Rio, Greece
xenos@ceid.upatras.gr

Abstract

We explore a novel transcription task in mobile text entry research, presenting stimuli within LLM-generated conversational contexts to improve participant engagement and phrase memorability. We conducted two studies: an eye-tracking study examining participants' attention when presented with conversational contexts alongside stimuli, and an experiment comparing LLM-generated and human-generated prompt-response pairs in transcription tasks, involving both high and low memorability stimuli. Key findings reveal that presenting conversational contexts improves recall for low memorability phrases and results in fewer uncorrected errors during transcription. No significant effects were observed on other basic text entry metrics, or participant subjective appraisals of engagement with the novel task, suggesting it can be used safely as an alternative to the traditional transcription task. We discuss the potential of LLMs in improving text entry evaluation methods, including generating diverse linguistic styles, emotionally loaded contexts, and even simulating entire evaluation processes. Our study highlights the need for systematic approaches to generate and evaluate LLM outputs for research purposes, and for proposing new metrics and evaluation methods.

CCS Concepts

• **Human-centered computing → Laboratory experiments**; **Ubiquitous and mobile computing design and evaluation methods**; **Empirical studies in ubiquitous and mobile computing**.

MUM '24, December 01–04, 2024, Stockholm, Sweden
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1283-8/24/12
https://doi.org/10.1145/3701571.3701586

Keywords

Mobile text entry; Evaluation Methodologies; Transcription tasks; Large Language Models

ACM Reference Format:
Andreas Komninos, Anna Maria Feit, Luis A. Leiva, Florian Lehmann, Ioulia Simou, Dimosthenis Minas, Aggelos Fotopoulos, and Michalis Xenos. 2024. An LLM-driven Transcription Task for Mobile Text Entry Studies. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '24), December 01–04, 2024, Stockholm, Sweden.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3701571.3701586

1 Introduction

Evaluation methods for mobile text entry research typiclly involve transcription tasks, in which participants are presented with phrases (stimuli), and are then asked to copy these using an input method, as quickly and as accurately as possible. In some studies, the stimuli remain visible to participants for the duration of the task, while in others, the phrases disappear as soon as the participant begins writing, in order to generate a more ecologically valid setup (since, in real life, text composition rarely involves copying a reference text visible to the user). In both cases, it is desirable to use stimuli that are *memorable*, in order to minimise the amount of time looking at the stimulus while composing text (therefore avoiding dilution of extracted metrics such as words-per-minute), and also to avoid circumstances where the participant has forgotten what they need to type (e.g. after committing and correcting an error), and are unable to complete the task successfully [20, 22]. For this purpose, the text entry community typically employs one of two standard phrase sets (MacKenzie [22] or MobileEmail [36]), which have been validated for memorability [19].

There are, however, several issues with the experimental setup as described above. Firstly, the transcription task has been criticised as having *low external validity* [37]. Secondly, there is doubt as to whether the stimuli included in the MobileEmail set are representative of general text entry contexts, since they are mostly derived from a business setting, and use language that is related to professional settings [13, 19]. Further, the task of copying data presented to participants entirely out of context can be tedious and tiring, leading to increased fatigue and disengagement from the task after a while, and subsequently, worse performance [21, 29].

To address these problems, we hypothesised that presenting stimuli to participants as part of a messaging conversation, instead of stand-alone decontextualised phrases, we might be able to improve engagement with the task. Providing conversational *contexts* to accompany the stimuli, might assist the participant to remember the phrases to be transcribed, leading to fewer recall issues and thus allowing the use of less memorable phrases in text entry experiments. This would allow researchers to use phrase sets derived from various sources that are more representative of today's use of language, and communication contexts (e.g. social networks, websites, books etc.), without having to worry about validating their memorability. Naturally, creating such conversations manually could be very inefficient. We thus wondered if could leverage Large Language Models to automatically create these conversations, therefore automating the process in a reliable and useful manner.

In this paper, we present two studies aimed at addressing the previous questions. In the first study, we investigate participants' attention when a conversational context (prompt) is presented together with a stimulus (response), through an eye-tracking study. Next, we demonstrate a method to evaluate the quality of LLM-generated conversation prompts. Finally, in the second study, we perform a second experiment with 53 participants to compare the effects of using LLM and human-generated prompt-response pairs in transcription tasks, including both validated high and low memorability stimuli as the response phrases.

2 Related Work

Text entry research has relied heavily on controlled laboratory studies for accurate quantification of speed and errors. These studies typically employed transcription tasks where users copied presented text as quickly and accurately as possible, making this the *de-facto* evaluation method [32]. Transcription protocols allow for error correction through researcher-specified combinations of backspace, cursor movement, or auto-correction features, and focus on the computation of metrics such as words per minute and various error rates. Some researchers also explored free text composition tasks (e.g. based on a contextual prompt, such as an instruction, incoming message, situation description or images [24, 37]). A challenge in such task remains that because there is no reference text to compare user entry against, it is impossible to calculate certain error rate metrics, as it is not possible to determine user intent.

Lab studies with transcription tasks suffer from the main drawback of lacking external validity. As noted by Vertanen and Kristensson [37], users real behaviour is to imagine and compose text, rather than memorise and copy text that is presented to them by some unknown authority. In real life, the text entry process may also be frequently interrupted due to external events and thus a lab study does not capture behaviours and problems caused by attention shifts and other cognitive processes involved in interruptions. To observe text entry behaviours in more realistic settings, researchers have experimented with in-the-wild studies that aim for data collection outside the laboratory setting. Some of these studies merely transfer the transcription task outside the lab (e.g. [28, 32]). Others have experimented with embedding these tasks in mobile games, to increase compliance [15] and avoid boredome or fatigue. Further studies have employed passive sensing, i.e. using dedicated keyboards and data collection frameworks to collect typing behaviour data from users' free text entry using their own applications for every-day contexts [6, 12, 18, 25, 34]. An important consideration in such efforts is to take privacy as a critical aspect of the software development and data capture methodology, since participants' text can include sensitive information such as passwords, credit card numbers, and of course, private messages. To this end, some researchers resort to recording only non-sensitive data, such as inter-key intervals, non-character keys, word counts and other privacy-preserving metrics [4, 6].

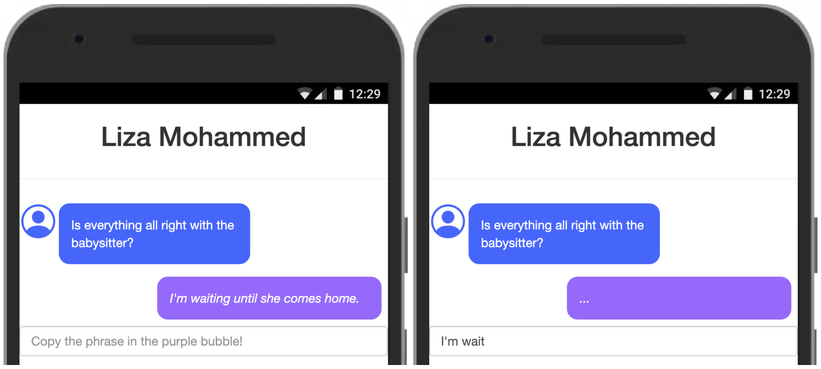
These studies demonstrated that lab behaviour during text entry, may often be quite different from real-world settings. However, it is unlikely that the traditional lab evaluation, using the transcription task, will be entirely replaced by other laboratory methodologies or in-the-wild studies, as the former allows the capture of more precise and detailed metrics that help to understand user behaviour, and further allow for data capture using additional sensing equipment such as eye trackers or EEG headsets, which cannot be used in the wild [16, 30]. It therefore remains important to consider whether there might be improvements that we can make on the transcription task, in order to improve its external validity and reduce some of its inherent shortcomings.

Increasing realism in the *appearance* and *content* of an interface during an experimental task, may help to remove unwanted behaviours and their interactions with treatment conditions, thereby improving the power of an experiment to discover effects [21]. Researchers have also found that embedding contextual information in a basic skills assessments may positively affect performance (e.g. in applying numerical literacy [31]). Considering mobile text entry as a basic digital literacy skill, our main research question is to examine if, and how presenting a transcription stimulus (phrase) from the validated phrase set in [36], in an ecologically valid manner (i.e. as a response to an incoming text message, see Fig. 1a), may affect participants performance and engagement with the task.

3 Study 1 - Visual Attention during Transcription

3.1 Objective

A significant concern at the start of our work, was to ensure that participants would approach the task in an ecologically valid manner, i.e. reading the incoming messages (prompt phrases) before attempting to memorise the intended response phrases and then transcribing them. We were mindful of the danger that participants, knowing the focus of the experiment is the transcription, might not pay any attention to the prompt phrases, and instead focus only on the stimulus phrase, possibly foregoing any positive effects of having this additional task context presented to them. Therefore in our first study, our primary objective was to validate the UI design of tasks and its ability to produce the ecologically valid experience we desired. A second objective was to examine whether presenting response phrases all at once, or using a gradual word-by-word presentation might also have any impact on participants - this would a text equivalent to previous studies where stimulus phrases are read out to participants, instead of presenting them in writing [1].



Is everything all right with the babysitter?

I'm waiting until she comes home.

Copy the phrase in the purple bubble!

Is everything all right with the babysitter?

...

I'm wait

(a) Prompt-response pairs presented to participants prior to entry.

(b) Response phrase disappears when participants begin entry.

Figure 1: Demonstration of the interface in the LLM-prompt condition. Participants need to transcribe the phrase in the purple bubble, which disappears as soon as they start writing (replaced with “...”).

3.2 Methods

The study design is a 2x2 experiment with one factor being context (baseline: only showing the phrase to be transcribed; LLM-prompts: showing LLM-generated prompt-response pairs) and the second factor being a phrase presentation effect (no effect and gradual appearance). Therefore, conditions are 1) baseline, with presentation effect (B-E), 2) baseline, no presentation effect (B), 3) LLM-prompts, with presentation effect(L-E), and; 4) LLM-prompts, no presentation effect (L). We used highly memorable phrases from the MobileEmail set (`mem_count_cer0>7`, words≥ 5, total 121 phrases) and their associated LLM-generated prompts (see next section for details).

3.2.1 Experiment environment. Participants took the study on a desktop computer environment in a controlled laboratory setting, using a browser simulating a mobile device with a custom HTML5 - JavaScript interface that presented the phrase pairs as a text conversation between two persons. Apart from keystroke events, we captured participants’ gaze using a Tobii T120 Eye-tracker integrated in a 17-inch TFT monitor. We opted to use a desktop setting for this study, as our aim was mostly to examine how participants behave when the prompt-response pairs are presented to them. Using a desktop eye-tracker with a large monitor helps to circumvent calibration and accuracy problems that are inherent to using wearable eye trackers to observe use on handheld devices.

Using the experiment application interface, initially, participants entered their demographic details, and a random order for the experiment conditions was computed for them. After entering their details, participants could begin the first block of phrases. Completing a block returned them to the demographics screen, from where they could begin the next block. Each condition was associated with 10 phrases selected randomly without duplication from the pool of available phrases.

In all conditions we simulate a messaging conversation where the prompt (incoming message) appears in a blue bubble, while the response to be transcribed appears in a purple bubble (Fig. 1a). The baseline condition presents an empty prompt bubble, and in the purple bubble, a response phrase from the MobileEmail set, asking participants to transcribe it. The response phrase is presented either in its entirety at once (no effect condition), or gradually on a word-per-word basis to the participant (gradual condition). The input field is locked until the prompt phrase has been fully displayed. The response phrase disappears once the participant begins transcription and appears again once the participant has submitted their input (Fig. 1b). The LLM-generated condition presents prompt-response phrase pairs using both bubbles. The prompt phrase appears first, with a pop-out effect to draw participants’ attention to it. We then apply a suitable delay based on the length of the prompt phrase, to allow for adequate time to read the prompt. Past literature indicates a mobile reading speed of 170-190 WPM [5, 35], so we used a conservative estimate of 150WPM translated to 80ms/character (given that some participants may not be proficient English speakers). We also add a further 3000ms to the calculated delay, to allow participants time to begin reading the prompt, as some might not start right away, and also to revisit it before the response phrase appears. After this delay, the response phrase appears as per the baseline condition. After completing a phrase block, the participants are returned to the demographics page, where they can begin the next block until they complete the study.

3.2.2 Prompt-response generation. We used the *Llama3-8B Instruct* model (with 6-bit quantization to fit the memory of our available GPU - RTX3070 with 8GB VRAM) as the LLM to generate the prompts to accompany the response phrases selected from the MobileEmail set. The quality of the output was not a priority for this study, but after several iterations of refinement and following guidance by White et al. [39], we concluded to use the prompt in Listing 1 to instruct the LLM in terms of its output, using a few-shot technique to prime the LLM to conform to providing JSON output (in order to facilitate further processing):

3.2.3 Participants. Participants were selected via convenience sampling from the student population of a local university and included 13 participants (1 female), mean age 25.07 years old ($\sigma = 4.83$), with a self-reported proficiency in English equivalent to CEFR C2-Proficiency (n=8), C1-Advanced (n=3), B2-Upper Intermediate (n=1), or non-native without English certification (n=1).

```
    Write a mobile messaging conversation between
        two persons.
    The first person opens the conversation with
        either a question or a statement.
    The second person responds with the phrase [
        RESPONSE PHRASE].
    The first person's phrase must have high
        contextual relevance to the answer.
    The dialogue must consist only of one opening
        phrase and one response.
    Your response must be a single JSON object
        with two fields: 'opening', 'response'.
    You must not include any commentary in your
        response, ONLY than the JSON object.

    EXAMPLE PHRASE: "this is a classroom test"
    EXAMPLE OUTPUT:
    {
        "opening":"what is this document?",
        "response":"this is a classroom test",
    }
```

Listing 1: Prompt used for Study 1 (P1).

3.3 Study 1 Results

3.3.1 Keyboard metrics. From the collected keystroke events we computed the typing speed (WPM), minimum string distance (MSD) between the response phrase and submitted text, and backspaces per phrase during entry (BSP). In terms of performance, we did not find any statistically significant differences in MSD (Friedman $\chi^2 = 2.592, p = .459$), BSP (Friedman $\chi^2 = 4.890, p = .18$) or WPM (ANOVA $F = 0.019, p = .996$), see Table 1.

3.3.2 Eye-tracking metrics. We measured participants' gaze behaviour on the blue bubble area containing the prompt phrase (where appropriate) and the purple bubble area containing the response phrase. Metrics include the total fixation duration, total number of fixations, and fixation duration on prompts as a percentage of overall fixations in both the prompt and response areas.

The total *number of fixations* on the prompt bubble is statistically significant across conditions (ANOVA $F = 26.3092, p < .001$). Pairwise t-tests with post-hoc Bonferroni correction demonstrate statistical significance between conditions (B-E, L-E): $t = -6.349, p <$ .001; (B-E, L): $t = -6.747, p < .001$; (B, L-E): $t = -6.3205, p < .001$, and; (B, L): $t - 6.7138, p < .001$. For the response bubble, again the difference is statistically significant (ANOVA $F = 6.5008, p <$ .001). Pairwise t-tests with post-hoc Bonferroni correction demonstrate statistical significance between conditions (B-E, L-E): $t = -5.1826, p < .001$; (B-E), L): $t = -6.6458, p < .001$, and; (B, L-E): $t = -4.5346, p = .001$.

Accordingly, total *fixation duration* difference on the prompt bubble is statistically significant across conditions (ANOVA $F = 25.1127, p < .001$). Pairwise t-tests with post-hoc Bonferroni correction demonstrate statistical significance between conditions (B-E, L-E): $t = -6.3141, p < .001$; (B-E, L): $t = -6.422, p < .001$; (B, L-E): $t = -6.139, p < .001$ and; (B, L): $t = -6.2484, p < .001$. For the response bubble, again the difference is statistically significant (ANOVA $F = 25.1127, p < .001$). Pairwise t-tests with post-hoc Bonferroni correction demonstrate statistical significance between conditions (B-E, L-E): $t = -6.3141, p < .001$; (B-E, L): $t = -6.422, p < .001$; (B, L-E): $t = -6.139, p < .001$, and; (B, L): $t = -6.2484, p < .001$.

When considering the duration of fixations on the prompt area as a percentage of the duration of fixations in both areas, we note that, as expected, it makes for a marginal amount of attention in the baseline conditions (with effect: $\bar{x} = 2.946\%, \sigma = 1.845\%$; no effect: $\bar{x} = 3.388\%, \sigma = 2.288\%$), whereas in the LLM prompt conditions, the percentage is significantly higher (with effect: $\bar{x} = 27.526\%, \sigma = 10.885\%$; no effect: $\bar{x} = 26.293\%, \sigma = 10.715\%$).

3.4 Study 1 discussion

Our results show that participants are not paying attention to the empty prompt bubble, as would be expected in the baseline conditions, but they are spending some time to observe it in the LLM-prompt conditions (Fig. 3a). This dismissed our worry that participants might ignore the prompt phrase altogether and simply wait to read the response phrase that needs to be memorised. Nevertheless, the attention spent on the prompt is considerably lower than the time spent observing the response phrase. As for the response area, the presentation effect does not seem to have an impact on the duration of fixations on the response phrase within the same basic condition (baseline, LLM prompts). Interestingly, participants spend less time fixating on the response phrase when it is presented without an accompanying prompt. Overall, we find supporting evidence that presenting phrases to be transcribed together with a prompt does attract participants' attention to the prompt, and also increases the time spent examining the response phrase, therefore potentially improving their chances to memorise the phrase correctly.

4 Study 2 - Memorability and performance with human vs. LLM-generated prompts

4.1 Objective

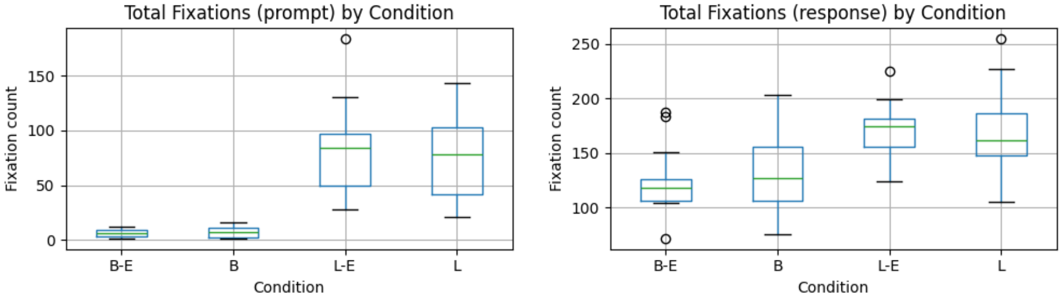
In this study, we sought to investigate how participants in a transcription experiment might be affected by the presentation of high and low memorability phrases, accompanied by conversation context (prompt) phrases that are generated by LLMs. We contrast this with prompt phrases hand-crafted by humans for the same responses, and against a baseline which presented only the response phrase to participants. This study concerns a 2x3 experiment design with two main factors: Context (Baseline - no prompts [B], LLM-generated prompts [L], Human-generated prompts [H]), and Response Phrase Memorability (High [HM], Low [LM]). We thus have six conditions labelled as: 1) B-HM; 2) B-LM; 3) L-HM; 4) L-LM; 5) H-HM and; 6) H-LM. The aim was to be able to compare LLM-generated prompts, which can be constructed massively and with minimal effort, vs. prompts generated by humans, a time-consuming manual effort.

4.2 Experiment environment

We modified the experiment application environment used in Study 1, with minor presentation and usability tweaks, in order to provide

	MSD		BSP		WPM	
	Mean	SD	Mean	SD	Mean	SD
Baseline - w. effect (B-E)	0.962	0.984	14.231	8.477	45.091	14.414
Baseline (B)	1.531	1.424	18.154	10.754	46.432	14.754
LLM prompts - w. effect (L-E)	1.123	1.516	17.846	13.539	45.686	15.858
LLM prompts (L)	1.638	2.126	20.154	9.957	45.451	13.936

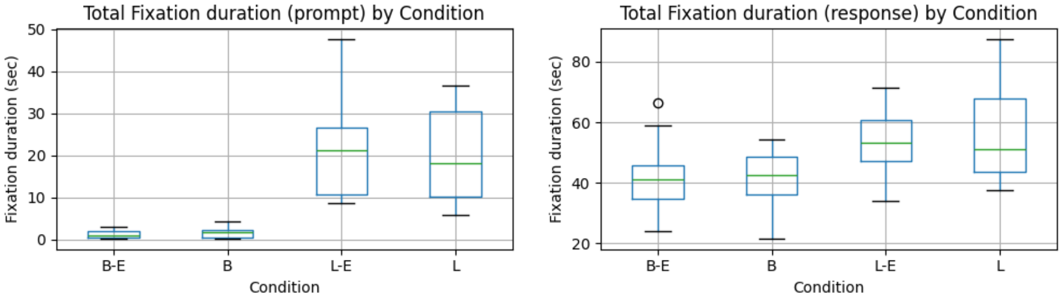
Table 1: Participant performance means and standard deviations.



(a) Prompt area

(b) Response area

Figure 2: Total fixation counts on the prompt and response phrase areas. Conditions 1, 2: baseline; 1, 3: with presentation effect.

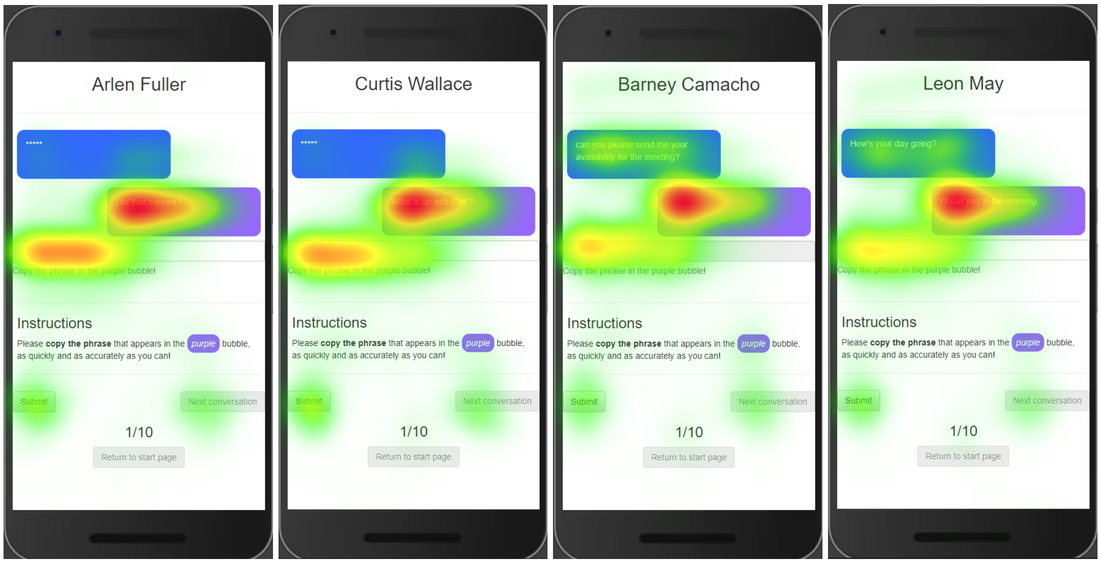


(a) Prompt area

(b) Response area

Figure 3: Total fixation durations on the prompt and response phrase areas. Conditions 1, 2: baseline; 1, 3: with presentation effect.

the following flow. First, participants enter their basic demographics. A random order of the 6 experiment conditions is calculated for each participant. When participants have entered their demographics, they can begin transcribing phrase blocks. Each block corresponds to one condition, and consists of 15 phrases (90 phrases in total per participant). Before each block, participants are required to confirm that they are willing to devote their full attention to the experiment, and that they are in appropriate settings that allow them to continue unobstructed. They are then presented with a screen detailing the instructions they need to follow during the task. During each block of transcription tasks, in all six conditions, single response phrases or prompt-response phrase pairs are picked randomly (without duplication) from a pool of candidates, once the block is started. After each block, participants are asked to provide answers to a 16-item questionnaire. Questions appear in random order every time, in order to avoid participant familiarity with the structure, and they include a trick question that we use to filter out participants who respond carelessly to the questionnaires (see later sections for questionnaire design). After submitting the questionnaire, participants return to the main demographics screen, and can proceed to the next block. The experiment application saves participants' progress so the experiment does not need to be carried out in one session. Participants carried out the tasks online from a location of their own choice, and using their own mobile devices. Controls were implemented to prevent participants from taking the test on anything other than a smartphone, with redirection to an error page if an unsuitable device was detected. Readers are



Arlen Fuller

*****

Did that happen?

Copy the phrase in the purple bubble!

Instructions

Please **copy the phrase** that appears in the purple bubble, as quickly and as accurately as you can!

(a) Condition 1: Baseline, with effect

Curtis Wallace

*****

What is up with ENE?

Copy the phrase in the purple bubble!

Instructions

Please **copy the phrase** that appears in the purple bubble, as quickly and as accurately as you can!

(b) Condition 2: Baseline

Barney Camacho

can you please send me your availability for the meeting?

Copy the phrase in the purple bubble!

Instructions

Please **copy the phrase** that appears in the purple bubble, as quickly and as accurately as you can!

(c) Condition 3: LLM prompt, with effect

Leon May

How's your day going?

I'll call you in the morning.

Copy the phrase in the purple bubble!

Instructions

Please **copy the phrase** that appears in the purple bubble, as quickly and as accurately as you can!

(d) Condition 4: LLM prompt

Figure 4: Gaze heatmaps for the four conditions in the eye tracking study.

welcome to try the task at [Anonymised URL for review]. A video of the task is also included as supplementary material.

4.3 Selecting memorable and non-memorable responses

The MobileEmail phrase set is accompanied by a metadata file, containing results on each individual phrase from the experiments ran in the accompanying paper. We base our work on the fields `mem_count_cer0` and `words`. The former indicates the number of Amazon Turk workers who transcribed the phrase without errors ($CorrectedErrorRate = 0$) during the memorisation experiment ($min = 0, max = 10$), and the latter contains the length of the phrase as a number of words. In the design of the MobileEmail phrase experiment, each phrase was presented to 10 workers and the researchers assigned a classification of “high” memorability to any phrase that had a `mem_count_cer0` score ≥ 8. This threshold was chosen as a reasonable value, but without further justification. As a result, the categorisation is subject to effects from the random allocation of phrases to participants, and therefore is not entirely objective. To overcome this limitation, we first create a new label of human memorability (`H_mem`) on the data based on the `mem_count_cer0` field, by binning the `mem_count_cer0` score into three categories of *Memorability*: High ($[7-10]$), Medium ($[4-6]$) and Low($[0-3]$). Further, we employed the memorability metric proposed by Leiva and Sanchis-Trilles [20] to assign an algorithmically derived memorability score to the phrases, and then again created a new label of algorithmic memorability `A_mem`, binning into the same three categories: High, Medium and Low. Comparing the outcome of both labelling operations on the data, we note that there is considerable discrepancy in the results, especially for the “Medium” category (Table 2), bringing about some doubt on the generalisability of both the human and algorithmic approaches. However, to proceed, we kept the phrases in the “High” and “Low” categories where both classification methods yielded the same result. The resulting set contains 230 phrases in the “High” memorability category, and 90 phrases in the “Low” category.

		A_mem classification		
		High	Medium	Low
H_mem classification	High	230	161	92
	Medium	147	203	209
	Low	15	27	90

Table 2: Confusion matrix for the human and algorithmic memorability classifications.

Further, we noted that the distribution of phrases per length (number of words) is significantly different between the two categories ($\bar{x}_{high} = 5.702, \sigma = 0.846; \bar{x}_{low} = 7.760, \sigma = 1.206$). To ensure a fair distribution of sentence lengths, we discarded any phrases with a length ≤ 5 words, resulting in 91 “High” and 82 “Low” memorability phrases. We will use this pool of phrases and their associated prompts in the experiment application described previously, and as detailed in Section 4.4.4. Next, we describe how the human and LLM-generated accompanying prompts were generated.

4.4 Prompt-response pair generation

4.4.1 Human prompt generation. We assigned the task of generating suitable prompts for each phrase to two members of our team, therefore generating two prompts (A & B), for each response phrase in the pool described in Section 4.3. We left out a few phrases for which the human generators had difficulty imagining a suitable prompt, leaving 169/173 phrases with two prompts. A further two members inspected the generated prompts and noted their preference (A or B). For the cases where there was no consensus, a fifth member of the team acted as a tie-breaker, indicating their preference and also how strong it was (slight, strong, no preference). Although we tried to adopt a consensus-based approach to selecting the most appropriate human prompt for each response phrase, this process is still subject to human bias.

To objectively assess the quality of generated prompts, we used the Universal Sentence Encoder QA (USE-QA) model, which is trained on 512-dimensional embeddings of questions and answers, yielding the cosine similarity between the embedding vectors of a question and its answer (thus, enabling an assessment of their relative logical coherence) [7]. For example, the embedding vector of the question "How old are you" would have a higher cosine similarity to the vector of answer "I am 25 years old", than that of answer "Today is Friday". Using the USE-QA scores, we can select the human-generated prompt which algorithmically appears to match the response phrase best. We then compared the agreement between the prompts chosen by the human selection process, and the prompts chosen by the USE-QA score, finding an agreement in 90 cases (53%). Bearing in mind that the tie-breaker's responses were often marked as "slight" or "no preference" (meaning that there the difference between the two prompts was owed to slight nuances and subjective preference), we checked the alignment again, ignoring the choice difference, and the alignment was 67.5%.

A further observation was that, quite often, the difference in the USE-QA scores was rather small. We converted the difference of the USE-QA scores of the two prompts into degrees (angles) separating the embedding vectors, to see how "far apart" the prompt vectors are, and observed differences from nearly 0 to ≈ 16.5 degrees. Therefore we computed the 33^{rd} and 66^{th} percentile to create three labels for the USE-QA choice, namely "no preference" ($0\% - 33\% = 0 - 1.998$ degrees difference), "slight preference" ($34\% - 66\% = 1.999 - 4.370$ degrees difference) and "strong" ($67\% - 100\% \geq 4.370$ degrees difference). Then, ignoring the choice difference for anything qualified as having less than a *strong* difference (either via human judgement, or USE-QA difference labelling), the alignment of choices improves to 89.99%. Put more simply, we consider the human prompt choice and USE-QA score-based prompt choice to be "aligned" (i.e. more or less equal) if:

- Human and USE-QA-based prompt choice are identical, OR,
- Human and USE-QA-based prompt choice are not identical, but the tiebreaker preference is "slight" or "no preference", OR,
- Human and USE-QA-based prompt choice are not identical, but the USE-QA label is "slight preference" or "no preference (vector degree ≤ 4.37 degrees)

With this in mind, we can demonstrate that using the USE-QA scores can help us pick context phrases as well as human evaluators would, for 90% of the time - we will exploit this result as described in Section 4.4.4 in order to select the best LLM-generated prompts.

4.4.2 LLM prompt generation. To investigate our hypotheses further, we needed to derive a process to generate plausible prompt-response phrase pairs, improving those used in Study 1. Though in Study 1 we used the prompts generated by Llama3-8B, there are multiple LLM open-source models available, each with varying performance in benchmarks. Therefore we extended the process by employing an additional state-of-the-art (at the time of writing) open source model, namely *Mistral7B Instruct v0.2* with 6-bit quantization, to generate prompts alongside Llama3-8B, for the same MobileEmail phrases. We used the subset of the MobileEmail set that contains phrase with no numeric characters (1347 phrases). From these, the LLMs were able to generate prompts for 1188 phrases (failing to provide valid JSON outputs for the rest). Inspecting the results of the generation process, we found instances where both LLMs either provided non-sensical prompts unrelated to the response, and instances where the generated prompt was mostly consistent of words also present in the response phrase. For example:

- P: *How much audio can be stored on this new 8Gb memory card?* - R: *How much volume?*
- P: *Are you free on Friday?* - R: *See you on Friday.*

To evaluate the quality of the generated prompts, we performed an analysis using a twin approach: collaborative evaluation by other LLM agents and human evaluation. Using LLMs as a judge is a recently popular method for evaluating LLM outputs, aiming to overcome the problem of human evaluation which is impractical at large scales. LLMs-as-a-judge can be deployed in a single-agent, or multi-agent manner (e.g. through agentic negotiation) [8, 11, 23]. For the agentic evaluation, we created three LLM agent instances (Llama3-8B, Mistral-7B and using a third state-of-the-art model, *Gemma2-9b Instruct* with 4-bit quantization) and sent the prompts generated by Llama3-8B and Mistral7B to be evaluated for logical coherence by the agent, with a rating of "High", "Medium" or "Low" coherence. In this way, each generator model evaluated the prompts generated by itself and the other generator model, and Gemma2 acted as an "independent" evaluator, assessing the logical coherence of the prompts generated by the other two generator models. The prompt P2 used for this process is presented in Listing 2.

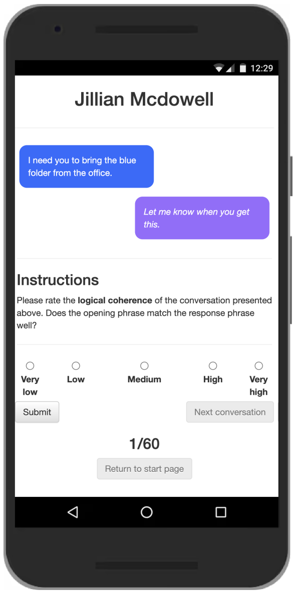
```
Provide a categorisation of "high", "medium" or "
    low" to assess the logical coherence of the
    following two conversations.
Your response must be a single JSON object with
    two fields: 'conversation1' and 'conversation2
    '.
You must not include any commentary in your
    response, ONLY than the JSON object, as per
    the example that follows.

Conversation 1:
Person 1: [PROMPT FROM MODEL A]
Person 2: [RESPONSE PHRASE]

Conversation 2:
Person 1: [PROMPT FROM MODEL B]
Person 2: [RESPONSE PHRASE]

EXAMPLE OUTPUT:
{"conversation1": "medium", "conversation2":"high
    "}
```

Listing 2: Prompt used for logical coherence evaluation (P2).

Mapping the generated classifications to scores between 1 (low) and 3 (high), we calculated a final coherence score $f(i,m)$ of prompt - response pair i generated by LLM model m (Mistral-7b or Llama3-8B), as the simple sum of weighted scores $S$of each evaluation $f(i,m) = \alpha \times S_m + \beta \times S_{m'} + \gamma \times S_g$, where m' is the evaluation by the other model also used to generate prompts, and g is the Gemma2 model. We set the parameter values $\alpha = 0.2$ and $\beta, \gamma = 0.4$, therefore biasing the resulting score to the evaluation provided by models other than the one that generated the phrase pair.

Further, we performed a short study presenting 120 prompt-response pairs (60 Llama3-B generated, 60 Mistral7B generated) to human participants, selected as stratified random sample from the entire set, based on the calculated $f(i,m)$ score of pairs. We asked participants to provide a logical coherence rating of the presented pairs on a 5-point scale of 1 (low) to 3 (high). Participants were selected via convenience sampling from the student population of a local university and included 23 participants (8 female), mean age 21.261 years old ($\sigma = 4.204$), with a self-reported proficiency in English equivalent to CEFR C2-Proficiency (n=20) and B2-Upper Intermediate (n=3). Participants took the study online from a location of their choice, using their own devices, with a custom interface adapted from Study 1, that presented the phrase pairs as a text conversation between two persons (Fig. 5).

As shown in Fig. 6, the results bear a statistically significant correlation for the Llama3 generated prompts (Spearman $\rho = 0.479, p < .01$) but not for the Mistral7B generated prompts (Spearman $\rho = 0.041, p > 0.05$), indicating that the agentic evaluations made by Mistral7B and Gemma2 on the prompts generated with Llama-3, generally align better with human perception of logical coherence. However, we note that the actual mean coherence in human evaluations is relatively low ($\bar{x}_{Llama3} = 1.669, \sigma = 0.498, min = 1.0, max = 2.786; \bar{x}_{Mistral7B} = 1.714, \sigma = 0.471, min = 1.0, max = 2.760$), and the same observation applies to agentic evaluations ($\bar{x}_{Llama3} = 1.773, \sigma = 0.509, min = 1.0, max = 3.0; \bar{x}_{Mistral7B} = 1.714, \sigma = 0.468, min = 1.0, max = 2.8$). Taken together, these results mean that the prompt-response generation process shows promise but with room for improvement. We note that delegating the task of assessing the success of producing logically coherent pairs to agentic models (so as to avoid the need for human appraisal) depends on the LLM models used to drive the agents, thus we might need a more robust approach (e.g. the USE-QA scores).



Jillian Mcdowell

I need you to bring the blue folder from the office.

Let me know when you get this.

Instructions

Please rate the **logical coherence** of the conversation presented above. Does the opening phrase match the response phrase well?

○ Very low ○ Low ○ Medium ○ High ○ Very high

Figure 5: Human evaluation study interface.

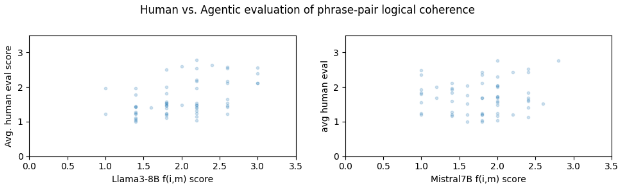


Figure 6: Comparative results of human and agentic logical coherence scores.

4.4.3 Improving the prompt generation process. Based on our experience of human prompt generation and in discussing the internal cognitive process employed by the team members to generate these, we ideated some solutions to improve logical coherence in the LLM generation process. We found that when humans created the phrases, we often needed to imagine first some situational context in which the phrase might be used, and then derive the prompt from that context. As previous work demonstrates, LLMs are inherently capable of high quality narrative (story) generation when given a limited prompt (e.g. a title, or short sentence) [2]. Therefore we revised the generation prompt and included an instruction to the LLM that it should first think of a short narrative about two persons conversing based on the response phrase, and then based on the narrative, produce an appropriate dialogue. Further, we used few-shot examples in the prompt, in order to force the LLM to produce JSON output, to facilitate further processing of the generated output. The resulting prompt is shown in Listing 3.

We generated 400 prompt-response pairs with Llama3-8B and Mistral7B, using the prompt from Study 1 (P1) and the revised prompt (P3), using high and low memorability phrases from the MobileEmail set as the response phrases. We then scored all prompt-response pairs using the USE-QA model. A feature of the USE-QA model is that it allows passing a "context" parameter to the response encoder, which usually consists of text that might precede or follow the response. We passed the story generated by P3 into this parameter. An example of generated prompt-response pairs from the two LLM prompts is shown in Table 3.

To overcome the biasing of USE-QA scores in favouring prompts which contain the same words as the response phrase, we calculate an adjusted score U' as follows. Each candidate prompt that might accompany a response phrase, goes through the same transformation process of 1) contraction expansion (e.g. *"haven't"* to *"have not"*, 2) stopword removal, 3) symbol and punctuation removal. Then, we compute the percentage of common words between the processed prompt and response phrases, after applying the same transformation process to each response phrase. We then apply

P1 generated prompt	Where do you want to go for dinner?
P3 narrative for prompt generation	*After getting her test results, Emma was relieved to know she did not have to take any further action. Her friend Alex was waiting for her at the park, and then they could go to Zainy Brainy.*
P3 generated prompt	Do you want to grab a coffee with me?
Response phrase (from MobileEmail set)	Then we can go to Zainy Brainy.

Table 3: Example outputs from P1 and P3.

```
Write a short narrative of no more than 4
    sentences, about two persons, based on the
    phrase [RESPONSE PHRASE]. Then, based on that
    narrative, write a conversation between the
    two persons in the narrative.
The first person opens the conversation with
    either a question or a statement, based on the
     narrative's details.
The second person must respond with the phrase [
    RESPONSE PHRASE].
The opening phrase must be written so the response
     phrase follows logically from it.
The opening phrase MUST NOT contain any of the
    words in the response phrase.
The dialogue must consist only of one opening
    phrase and one response.
Your response must be a single JSON object with
    three fields: 'narrative', 'opening', '
    response'.
You must not include any commentary in your
    response, provide ONLY the JSON object.  You
    must follow the JSON structure in the example
    precisely.

EXAMPLE RESPONSE PHRASE: "I am in a meeting just
    now."
EXAMPLE GENERATED NARRATIVE: "Jane wanted to talk
    to Mark about their daughter, Mary, who was
    facing some trouble at school. However, Mark
    was at work and engaged in a serious meeting."
EXAMPLE OUTPUT JSON:
{
    "narrative": "Jane wanted to talk to Mark
        about their daughter, Mary, who was facing
         some trouble at school. However, Mark was
         at work and engaged in a serious meeting
        .",
    "opening":"Can I call you briefly to talk
        about Mary?",
    "response":"I am in a meeting just now."
}
```

Listing 3: Prompt used for Study 2 (P3).

a decay function to compute the adjusted USE-QA score $U'_{p,m}$ of prompt p generated by LLM m as $U'_{p,m} = U_{p,m} \times 1/e^{(l_p \times c_p)}$, where $U_{p,m}$ is the USE-QA score of prompt p, l_p is the number of tokens in p, and c_p is the percentage of common words between p and its associated response phrase.

Therefore, for each generated prompt-response pair, we also computed the relevant USE-QA score U and adjusted USE-QA score U' in order to examine the quality of the produced results. Using the plain U scores, we find that prompts generated from the P1 prompt consistently outperform those generated by the P3 prompt (Llama3-8B: P1 $\bar{x} = 0.225, \sigma = 0.111$; P3 $\bar{x} = 0.136, \sigma = 0.087$; Mistral7B: P1 $\bar{x} = 0.197, \sigma = 0.112$; P3 $\bar{x} = 0.181, \sigma = 0.106$). The differences are statistically significant across all prompt-model combinations (Friedman $\chi^2 = 188.680, p < .001$. Between the two prompts and within each model, the differences are statistically significant in post-hoc Bonferroni corrected pairwise tests only for the Llama3-8B model (Wilcoxon (P1,P3): $Z = -12.932, p < .001$).

This result might indicate that the revised prompt P3 yields significantly worse quality prompts for the Llama3-8B model, however, when using the U' score for comparisons, the results are different. Prompts generated from the P1 prompt again outperform those generated by the P3 prompt but the margin is much lower for Llama3-8B and practically non-existant for Mistral7B (Llama3-8B: P1 $\bar{x} = 0.148, \sigma = 0.085$; P3 $\bar{x} = 0.123, \sigma = 0.087$; Mistral7B: P1 $\bar{x} = 0.138, \sigma = 0.082$; P3 $\bar{x} = 0.130, \sigma = 0.075$). The differences are statistically significant across all prompt-model combinations (Friedman $\chi^2 = 22.733, p < .001$. Between the two prompts and within each model, the differences are statistically significant in post-hoc Bonferroni corrected pairwise tests only for the Llama3-8B model (Wilcoxon (P1,P3): $Z = -4.601, p < .001$). We conclude thus that the revised prompt improved the generation process by means of asking the models to imagine first an appropriate context in which the conversation takes place.

4.4.4 Final prompt-phrase set selection. To proceed with the experiment, we needed to derive pools of single response-phrase or prompt-response phrase pairs with both high memorability and low memorability, from which to select phrases to present to the participants. For this, we used the pool of phrases in Section 4.3 as a starting point. For the baseline condition, we simply selected 45+45 response phrases with the highest and lowest `mem_count_cer0` values. For the human-generated prompt-response sets, we discarded cases where there was no alignment of the human and USE-QA choices as described above, and from these we selected the 45+45 response phrases with the highest and lowest `mem_count_cer0` values.

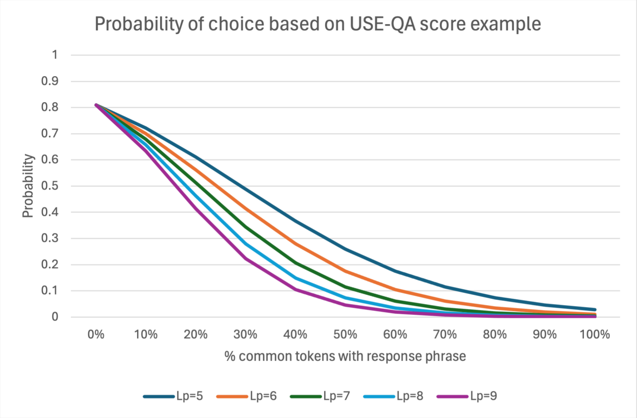


Figure 7: Example of the probability $P(A, m_1)$ of a hypothetical prompt p_A generated by model m_1 with $U_{p_A,m_1} = 0.229943$ to be chosen against an "inferior" prompt p_B with $U_{p_B,m_2} = 0.053884$ generated by model m_2, after USE-QA score adjustment, for various token counts of p_A ($l_p \in [5-9]$), and commonality percentages c_p of tokens between A and the response phrase.

For the LLM-generated prompt-response sets, we used a slightly more complicated process with a stochastic approach, as we wanted to ensure that we select the best possible prompt. We sum the adjusted U' scores of both prompts, and assign a probability $P(p, m) = U'_{p,m} / \sum U'_{p,x}, x \in [Llama3-8B, Mistral7B]$ to pick the prompt p provided by LLM m. In this way, we penalise the probability of selecting prompts that contain one or more common words with the response phrase, as this artificially raises its USE-QA score, and also offers contextual cues to assist participants' memory. On the other hand, limited token commonality still affords a prompt some chance of being chosen, as sometimes this might be necessary to provide a logically coherent prompt (see Fig. 7). After applying this process to select the best LLM-generated prompt for each response pair, we select the 45+45 response phrases with the highest and lowest `mem_count_cer0` values.

4.5 Questionnaire Design

We were unable to find a validated instrument to measure participant experience during a transcription experiment in extant literature. Therefore, we constructed a 15-item questionnaire with 5 questions on three constructs: *Attention to the task*, *Task Realism* and *Emotional Engagement*. Attention to the task aims at measuring participant commitment to each block of phrases and the individual effort they self-report to having put into the task. Task realism aims to assess subjective opinion on the realism of the task, and is mostly aimed at comparing how participants feel completing the baseline task versus the prompt tasks, where the interface resembles more closely an actual conversation. Finally, Emotional Engagement aimed to capture participant sentiment and attitude towards completing each task. The items were adapted from questionnaire items in extant literature or were formulated by ourselves, based on theory. All items require participants to indicate agreement on a 5-point Likert scale. Questions and references are shown in Table 4. Alongside the 15 items, we presented a trick question ("*The task involved deciphering messages from alien spaceships*") which was answerable in a Yes/No manner. We included this question to allow us to identify participants who were not reading the questionnaire carefully and to discard them from the analysis.

4.6 Participants

We recruited participants by advertising at our local university CS department, and also via the Prolific recruitment website. Out of 105 registered participants, we noticed that only 53 completed the study successfully by examining the number of collected questionnaires (6 per participant), however keystroke data was not successfully transmitted to our database for a single random phrase set for several of these (n=31). We will report in the next section how we handled the missing data in the analysis. For these 53 participants, mean age was 31.472 years old ($\sigma = 11.443$). 13 were local students, 23 were female, and self-reported CEFR command of English was Upper Intermediate (B2): 10, Advanced (C1): 9, Proficient (C2): 15, Native Speaker: 18 and non-native without certification: 1.

4.7 Study 2 Results

Lack of complete submissions is common in empirical studies (e.g. clinical trials) due to participant dropouts, sampling problems, data corruption or other errors . Recommended approaches for analysing (Completely) Missing at Random data vary depending on the type of data, but in our case, due to the lack of complete submissions for 31 participants, we perform the analysis in two stages: First, analysing the complete cases across all six conditions, and secondly analysing all cases in pairwise condition comparisons of variable power, using data from those participants who have submitted keystroke data in both conditions examined. Naturally, we perform the former for all questionnaire data, while both former and latter approaches apply to keystroke data. Statistical tests are chosen after examination of relevant assumptions. Where pairwise statistically significant results are shown, we report the p-value to be smaller than the Bonferroni-adjusted threshold for statistical significance ($t_a = 0.0033$).

4.7.1 Minimum String Distance.

For complete cases (Fig. 8a) , we note a statistically significant difference across conditions (Friedman $\chi^2 = 74.54, p < .001$). Post-hoc Bonferroni adjusted pairwise comparisons confirm statistically significant differences for high and low memorability phrases within the baseline (Wilcoxon (B-HM, B-LM): $Z = -4.0148, p < t_a$), LLM-generated (Wilcoxon (L-HM, L-LM): $Z = -3.0355, p < t_a$) and Human-generated conditions (Wilcoxon (H-HM, H-LM): $Z = -4.108, p < t_a$). Across conditions where high memorability phrases were used, there is statistically significant difference between the baseline and LLM-generated (Wilcoxon (B-HM, L-HM): $Z = -3.0389, p < t_a$), and LLM-generated and Human-generated (Wilcoxon (L-HM, H-HM): $Z = -3.5519, p < t_a$). More importantly, we note that the MSD is reduced where prompts are presented together with low memorability phrases, with statistical significance (T-test (B-LM, L-LM): $t = 3.4885, p < t_a$; Wilcoxon (B-LM, H-LM): $Z = -2.0291, p < t_a$).

Construct	Q#	Question
Task Engagement	Q1	I worked as hard as I could to complete the phrase copy tasks in this set. [40]
	Q2	I devoted my full attention during the phrase copy tasks in this set. [14]
	Q3	The need to memorise the phrase to copy, would capture and hold my attention. [27]
	Q4	The way that messages in all the text bubbles were presented in this set, helped keep my attention. [3]
	Q5	As time passed, I found myself absorbed in the task of copying phrases. [14]
Task Realism	Q6	Copying phrases during this set, felt like a realistic task. [10]
	Q7	The length, tone and language of the messages in all the text bubbles were consistent with typical text conversations [33]
	Q8	I can imagine myself sending messages like the ones I typed in this set. [9]
	Q9	The tasks in this set contained messages that are similar to those I send or receive in my everyday life. [38]
	Q10	The messages in all the text bubbles in this set, made sense (either as a conversation, or as a message to be sent).
Emotional Engagement	Q11	I enjoyed doing the phrase copying tasks in this set. [40]
	Q12	I felt excited about getting to the next phrase copy task after completing one, during this set. [26]
	Q13	Copying the phrases in this set was fun. [40]
	Q14	Copying the phrases in this set aroused my curiosity. [40]
	Q15	The emotional tone of the messages in all text bubbles in this set, affected my mood. [17]

Table 4: Questionnaire design

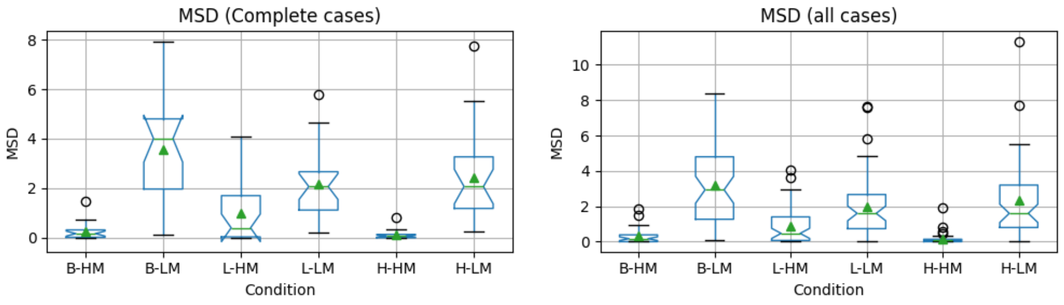
Similar effects are observed comparing all cases (Fig. 8b). Post-hoc Bonferroni adjusted pairwise comparisons confirm statistically significant differences for high and low memorability phrases within the baseline (Wilcoxon (B-HM, B-LM): $Z = -4.9367, p < t_a, n = 34$), LLM-generated (Wilcoxon (L-HM, L-LM): $Z = -4.8134, p < t_a, n = 53$) and Human-generated prompts (Wilcoxon (H-HM, H-LM): $Z = -5.5115, p < t_a, n = 41$). Across conditions where high memorability phrases were used, there is statistically significant difference between the baseline and LLM-generated (Wilcoxon (B-HM, L-HM): $Z = -3.9693, p < t_a, n = 34$), and LLM-generated and Human-generated prompts (Wilcoxon (L-HM, H-HM): $Z = -4.3788, p < t_a, n = 41$). We also repeat the finding that the MSD is reduced where prompts are presented together with low memorability phrases, with statistical significance on;y between baseline and LLM-generated prompts (Wilcoxon (B-LM, L-LM): $Z = -3.7264, p < t_a, n = 53$).

4.7.2 Use of backspaces. We report on the use of backspaces per phrase as an indicator of fixed errors during typing next. For complete cases, a Friedman test confirms statistically significant differences across conditions ($\chi^2 = 39.3158, p < .001$), see Fig. 9a). Post-hoc Bonferroni adjusted pairwise comparisons confirm statistically significant differences for high and low memorability phrases within the baseline (T-test (B-HM, B-LM): $t = -3.5641, p < t_a$), LLM-generated (Wilcoxon (L-HM, L-LM): $Z = -3.0268, p < t_a$) and Human-generated conditions (Wilcoxon (H-HM, H-LM): $Z = -3.0763, p < t_a$). Across conditions where high memorability phrases are used, there are no statistically significant differences. Across conditions where low memorability phrases are used, again there are no statistically significant differences.

When comparing all cases (Fig. 9b), post-hoc Bonferroni adjusted pairwise comparisons confirm statistically significant differences for high and low memorability phrases only within the LLM-generated (Wilcoxon (L-HM, L-LM): $Z = -4.3463, p < t_a, n = 53$) and Human-generated conditions (Wilcoxon (H-HM, H-LM): $Z == -4.4385, p < t_a, n = 40$), echoing almost exactly the analysis of complete cases.

4.7.3 Typing Speed (WPM). Finally we report on the typing speed measured in words-per-minute (WPM). For complete cases (Fig. 10a), a Friedman test reveals statistically significant differences across all conditions ($\chi^2 = 378.5195, p < .001$). Post-hoc Bonferroni adjusted pairwise comparisons confirm statistically significant differences for high and low memorability phrases within the baseline (T-test (B-HM, B-LM): $t = -12.000, p < t_a$), LLM-generated (t-test (L-HM, L-LM): $t = 5.9168, p < t_a$) and Human-generated conditions (t-test (H-HM, H-LM): $t = -11.3541, p < t_a$). Across conditions where high memorability phrases are used, a statistically significant difference exists only between the LLM-generated and Human-generated prompt conditions (t-test (L-HM, H-HM): $t = -3.4123, p < t_a$). Across conditions where low memorability phrases are used, a statistically significant difference exists only between the baseline and LLM-generated prompt conditions (Wilcoxon (B-LM, L-LM): $Z = 39.0, p < t_a$).

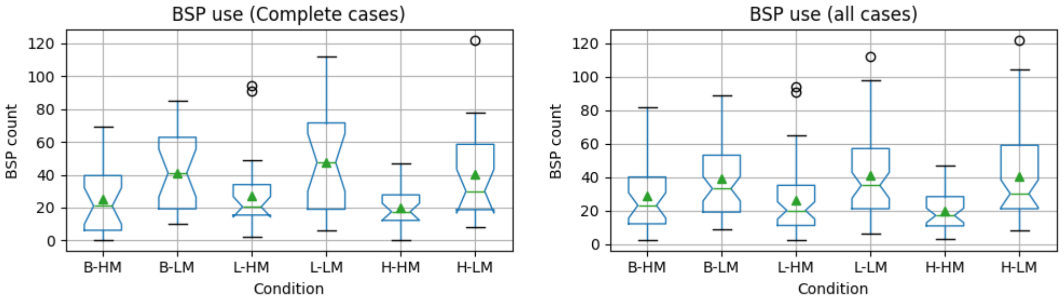
When comparing all cases (Fig. 9b), post-hoc Bonferroni adjusted pairwise comparisons confirm statistically significant differences for high and low memorability phrases within the baseline (Wilcoxon (B-HM, B-LM): $Z = -4.6588, p < t_a, n = 34$), LLM-generated (Wilcoxon (L-HM, L-LM): $Z = -5.502, p < t_a, n = 53$) and Human-generated conditions (t-test (H-HM, H-LM): $t = 14.7635, p < t_a, n = 41$). Across conditions where high memorability phrases are used, a statistically significant difference exists only between the LLM-generated and Human-generated prompt conditions (t-test (L-HM, H-HM): $t = -3.5908, p < t_a, n = 41$). Across conditions where low memorability phrases are used, a statistically significant difference exists only between the baseline and LLM-generated prompt conditions (t-test (B-LM, L-LM): $t = -6.0605, p < t_a, n = 53$) and the LLM-generated and Human-generated prompt



(a) Complete cases

(b) All cases

Figure 8: Minimum String Distance across conditions.



(a) Complete cases

(b) All cases

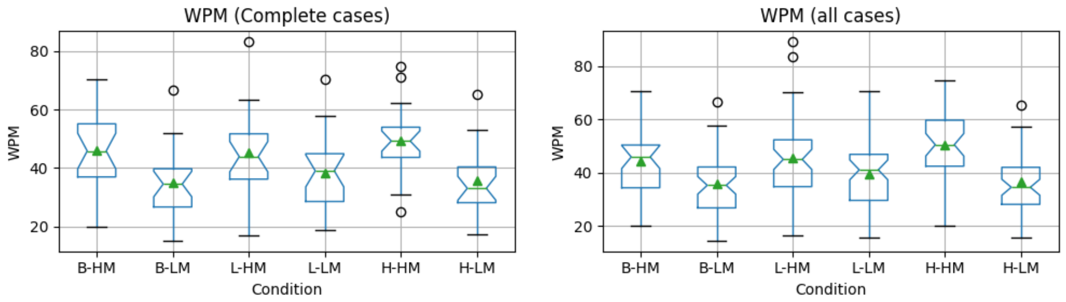
Figure 9: Backspace usage per phrase across conditions.

conditions (t-test (L-LM, H-LM): $t = 4.7983, p < t_a, n = 53$). The results mirror those of the complete cases almost identically.

4.7.4 Summary of quantitative results. Taking the previous results together, we can summarise by observing that participants exhibit, when transcribing high memorability phrases, faster input speeds, possibly attributable to the smaller frequency of backspace use. The presence of prompts, whether human or LLM-generated, does not appear to impact this trend. This is an expected finding - we can plausibly argue that high memorability phrases are easier to remember and that this leads to faster input and fewer corrections during it, which may relate also to the difficulty of language used in these phrases. However, we are surprised to notice that the presence of prompts seems to significantly reduce the number of uncorrected errors (MSD) in the submitted phrase, compared to the baseline condition where they were not available. This evidence suggests that recall is improved by the presence of prompts, affording participants better ability to remember what it is they need to type.

4.7.5 Qualitative results. We received 53 × 6 = 318 questionnaire responses. The distribution of responses to individual questions is shown in Fig. 11. To analyse our questionnaire, we perform a factor analysis in order to a) ascertain the validity of our 3-construct questionnaire design, and also b) to compare participant subjective opinion after exposure to each condition in a coherent manner. The Bartlett sphericity ($\chi^2 = 3126.505, p < .001$) and high Kaiser-Meyer-Olkin value ($KMO = 0.896$) show the data is suitable for factor analysis. Observing the related scree plot (Fig. 12a) with *promax* rotation, and applying the Kaiser criterion on factor eigenvalues, we notice that factor analysis should proceed with three factors, as we anticipated. A fourth factor has an eigenvalue close to 1 (0.94) but was rejected as its loadings were lower than <0.3 in all questions. We observe that factor loadings correspond quite well to the three constructs in our questionnaire design, thus Factor 1 corresponds to Emotional Engagement, Factor 2 to Task Realism and Factor 3 to Task Engagement (Fig. 12b). Factors represent (cumulatively) 24.22% (F1), 44.86% (F1+F2) and 62.04% (F1+F2+F3) of the total variance in responses, representing adequate coverage. Similar results are obtained with an orthogonal rotation (*varimax*) but we proceed with the oblique rotation since it yields a slightly simpler factor structure.

For Emotional Engagement (Factor 1), we observe statistically significant differences across conditions (ANOVA $F = 5.035, p < .001$). The participants' emotional engagement score is highest for the B-HM ($\bar{x} = 0.329, \sigma = 0.854$) and H-HM condition ($\bar{x} = 0.333\sigma = 0.86$). The difference between the two is not statistically significant. The difference to the score of other conditions is statistically significant for B-HM (T-test (B-HM, B-LM): $t = 5.691, p < t_a$; T-test (B-HM, L-HM): $t = 4.361, p < t_a$; T-test (B-HM, L-LM): $t = 6.878, p < t_a$; T-test (B-HM, H-LM): $t = 5.223, p < t_a$), and also for the H-HM



(a) Complete cases

(b) All cases

Figure 10: Typing speed (WPM) across conditions.

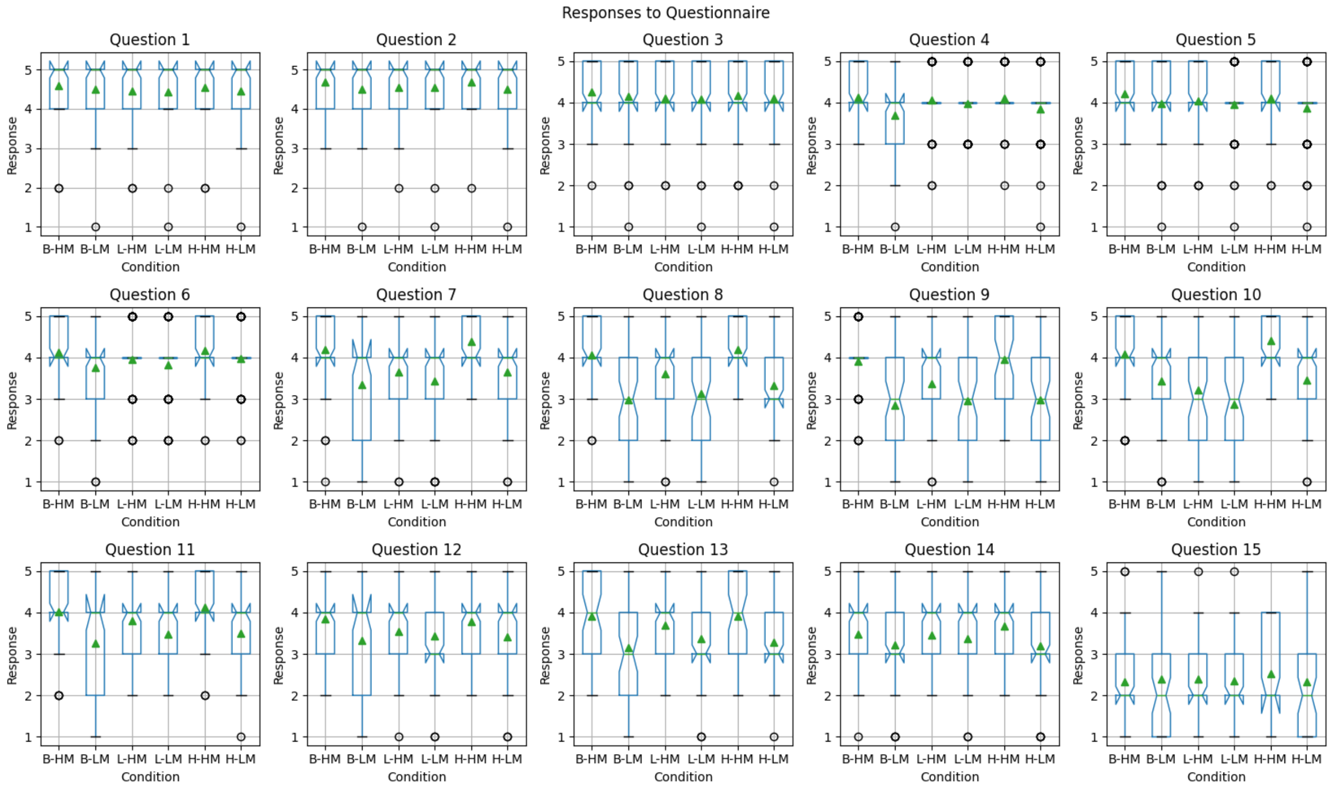
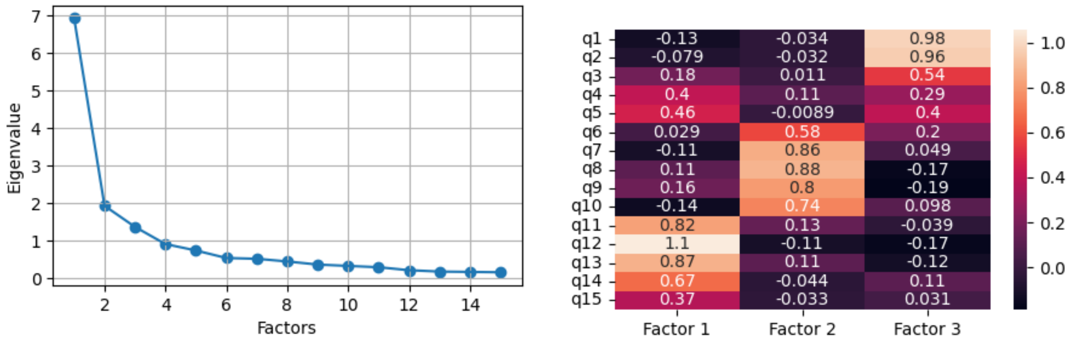


Figure 11: Responses to Questionnaire Items. Participants are asked to report agreement with the statements on a scale of 1 (Strongly disagree) to 5 (Strongly Agree).

condition (T-test (B-LM, H-HM): $t = -4.974, p < t_a$; T-test (L-HM, H-HM): $t = -3.32, p < t_a$; T-test (L-LM, H-HM): $t = -5.563, p < t_a$; T-test (H-HM, H-LM): $t = 5.781, p < t_a$).

For Task Realism (Factor 2) we observe statistically significant differences across conditions (Friedman $\chi^2 = 87.15, p < .001$). The participants’ appraisal of task realism is highest for the B-HM ($\bar{x} = 0.488, \sigma = 0.734$) and H-HM condition ($\bar{x} = 0.645\sigma = 0.581$). The difference between the two is not statistically significant. The difference to the score of other conditions is statistically significant for B-HM (Wilcoxon (B-HM, B-LM): $Z = -5.324, p < t_a$; Wilcoxon (B-HM, L-HM): $Z = -4.462, p < t_a$; Wilcoxon (B-HM, L-LM): $Z = -5.765, p < t_a$; Wilcoxon (B-HM, H-LM): $Z = -4.981, p < t_a$), and also for the H-HM condition (T-test (B-LM, H-HM): $t = -7.015, p < t_a$; T-test (L-HM, H-HM): $t = -6.079, p < t_a$; T-test (L-LM, H-HM): $t = -9.663, p < t_a$; T-test (H-HM, H-LM): $t = 7.716, p < t_a$).

Finally for Task Engagement (Factor 3), we observe again statistically significant differences across conditions (Friedman $\chi^2 = 87.15, p < .001$). The participants’ self-reported engagement with



(a) Scree plot

(b) Factor loadings

Figure 12: Results of factor analysis of the questionnaire.

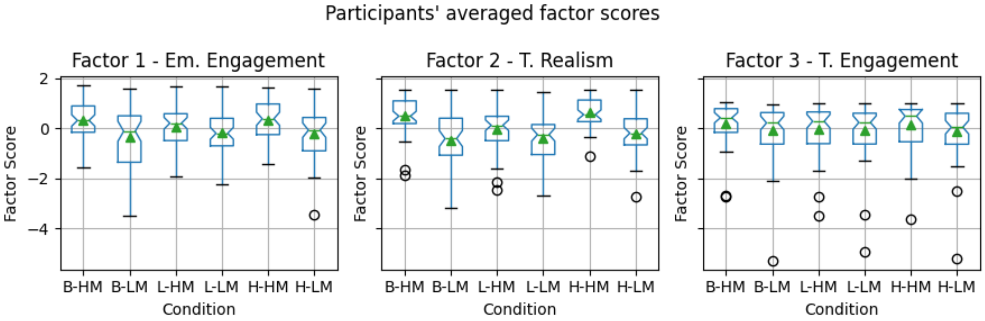


Figure 13: Participant factor scores averaged per condition.

the task is highest for the B-HM ($\bar{x} = 0.176, \sigma = 0.822$) and H-HM condition ($\bar{x} = 0.152\sigma = 0.867$). The difference between the two is not statistically significant. The difference to the score of other conditions is statistically significant for B-HM (Wilcoxon (B-HM, B-LM): $Z = -3.253, p < t_a$; Wilcoxon (B-HM, L-HM): $Z = -3.033, p < t_a$; Wilcoxon (B-HM, L-LM): $Z = -3.187, p < t_a$; Wilcoxon (B-HM, H-LM): $Z = -3.752, p < t_a$). However, the score for H-HM is statistically significantly different only compared H-LM (Wilcoxon (H-HM, H-LM): $Z = -3.218, p < t_a$).

4.7.6 Summary of qualitative results. Participants' self-reported appraisals of their experience demonstrated preference for the B-HM or H-HM condition in terms of task engagement, task realism and emotional engagement. Even with human-generated prompts, they did not find the task to be more engaging or realistic compared to simply presenting memorable phrases. On the other hand, the presence of prompts, whether generated by LLMs or humans, did not detract from the overall experience (taking phrase memorability out of the equation). We conclude that the main differentiator of experience in the transcription task, is the use of memorable phrases and that this effect is not mediated, at least subjectively, by the presence of accompanying prompts.

5 Discussion

We have explored the use of LLMs as a means to re-imagine the traditional transcription task used in mobile text entry studies. The results demonstrate interesting effects after modifying the task to present not just phrases to copy, but conversational contexts that pair well with the phrase. The most important finding was that this modification improves stimulus phrase recall for low memorability phrases, and results in fewer uncorrected errors during transcription. We did not find any significant effects on participants typing speeds or other associated metrics. We also did not find any significant evidence to suggest that the modification to the traditional transcription task has adverse effects on participants' subjective appraisals of engaging with the task, thus conclude that the modification is both safe to use, and shows potential in improving participants' performance, at least when phrases are less memorable. Our findings confirm the importance of presenting memorable stimuli to participants as a main factor affecting performance and engagement with the transcription task.

Our work is limited to generating accompanying phrases to existing stimuli used widely by the community, but it is plausible that the same LLM technology could be used to provide the full conversation (prompts and responses) as well. Such prompts or responses can be easily constructed with an aim to provide variable experiences and expose participants to a range of appropriate text entry contexts. For example, using specific linguistic styles (e.g. professional, informal, colloquial), representing ecologically valid contexts of use (e.g. communicating for different purposes such as sending updates, greetings, responding to request, requesting information etc.), and even composing text in emotionally loaded contexts (e.g. responding in anger or frustration, to a romantic partner, with excitement etc.). Further, we might imagine LLMs powering entirely novel evaluation tasks, such as utilising phrases from automated image descriptions, or live conversations with intelligent agents. Even further, it is possible to foresee the integration of LLM agents with simulations of text entry behaviour, therefore entirely automating the evaluation process and bypassing the need for human subjects altogether (e.g. LLM agents conversing with each other via simulated virtual keyboards). Finally, we limited ourselves to conversations in English, while LLMs could easily generate phrase sets and evaluation tasks in other languages, for which there are no pertinent evaluation materials.

Despite the exciting possibilities, we must be careful as we move to explore the opportunities afforded by LLM technology. As we demonstrate in this paper, a systematic approach to generating and evaluating LLM outputs, so that they can be used as part of an evaluation methodology is required. Whilst we cannot rely on human evaluation, due to the richness and volume of data that can be so easily produced by LLMs, we show that extant algorithmic approaches can be used for this purpose, but may require modification and adaptation. Accordingly, we proposed a new metric and a stochastic approach to selecting the best outputs from an LLM based on this metric. Entirely new metrics may need to be devised altogether, as we explore and adapt LLMs as part of evaluation methodologies.

Our study also highlighted the lack of a validated instrument to measure participants' engagement in text entry evaluation tasks. This will become important as we explore novel evaluation methodologies and move away from the tradition of the transcription task. To this end, we proposed a new scale to measure participant engagement with transcription-style tasks, which proved robust but of course could be extended to capture further aspects of participant subjective appraisals.

6 Limitations

The work presented here bears several limitations, as is natural in exploratory research like this. We constrained ourselves to the use of open-source LLMs such as Llama3-8B, Mistral7B, and Gemma2-9B. We used quantised versions of these models to fit the hardware limitations in our lab, but performance could be better with unquantised versions and even with the use of larger (commercial) LLMs, such as those in the Claude, ChatGPT, or Gemini families.

In the LLM generation process, the agentic evalution uses an empirically weighted linear scoring formula for coherence, which could be replaced by a more sophisticated metric. Human evaluation of prompt-response coherence possibly carries some bias based on the background and age of our participants. The USE-QA model was taken as-is, but it could be fine-tuned with more examples of conversations mined from users' devices. We examined two prompts (P1, P3) for generation of the prompt-response pairs, but of course there might be better approaches to guiding generation for the final selection to be used in a study, for example rejecting outputs below a certain USE-QA score and repeating the process until a better pair is generated, or using multiple instances of a model to negotiate iterative improvements to derive a final prompt-response pair.

Our questionnaire design is limited to three constructs with 15 items total, which was intentional in order not to overburden participants at the end of each block. The factor analysis shows some overlap in the factor loadings and therefore the questions could be revisited for wording or replaced with other items. Additional constructs could be added. Validation of the questionnaire with a larger sample would also be beneficial.

Finally, we only analysed basic text entry metrics in our study. Other metrics such as inter-key intervals, total error rates, corrected error rates etc. could be measured. We also captured rich data that allows analysis of the frequency of use of text entry support tools which were allowed in the experiment (e.g. autocompletion, word suggestions, glide typing). Analysis based on linguistic proficiency, country of residence, age and occupation could also be carried out, as we captured this data. We leave these extended analysis to future work or other researchers by releasing our data openly.

7 Conclusion

We see good potential in the use of LLMs as an important tool to revise and update mobile text entry evaluations for the future. This work barely scratches the surface of this potential, but its main contributions are 1) to present a novel variation of the transcription task for use in text entry lab studies using LLMs to generate an ecologically valid setting, and, 2) to highlight the complexities and challenges in the appropriate exploitation of developments in LLM technology for text entry research. We invite the community to take inspiration from this work and contribute to the exploration of this exciting research direction.

8 Data and Code availability

Our generated phrase sets, all related data, software and analysis code are distributed with an open source license at https://github.com/komis1/llm-transcription-task. Interested readers can experience the LLM-driven transcription task online at http://usidas.ceid.upatras.gr/llm-corpus, using a mobile browser.

References

[1] Jacob Abbott, Jofish Kaye, and James Clawson. 2022. Identifying an Aurally Distinct Phrase Set for Text Entry Techniques. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3491102.3501897

[2] Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic Story Generation: A Survey of Approaches. *ACM Comput. Surv.* 54, 5 (May 2021), 103:1–103:38. https://doi.org/10.1145/3453156

[3] Charles Baah, Irene Govender, and Prabhakar Rontala Subramaniam. 2023. Exploring the Role of Gamification in Motivating Students to Learn. *Cogent Education* 10, 1 (Dec. 2023), 2210045. https://doi.org/10.1080/2331186X.2023.2210045

[4] Florian Bemmann and Daniel Buschek. 2020. LanguageLogger: A Mobile Keyboard Application for Studying Language Use in Everyday Text Communication in the Wild. *Proc. ACM Hum.-Comput. Interact.* 4, EICS (June 2020), 84:1–84:24. https://doi.org/10.1145/3397872

[5] Ralf Biedert, Andreas Dengel, Georg Buscher, and Arman Vartan. 2012. Reading and Estimating Gaze on Smart Phones. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 385–388. https://doi.org/10.1145/2168556.2168643

[6] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 255:1–255:14. https://doi.org/10.1145/3173574.3173829

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. https://doi.org/10.48550/arXiv.1803.11175 arXiv:1803.11175 [cs]

[8] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. https://doi.org/10.48550/arXiv.2308.07201 arXiv:2308.07201 [cs]

[9] Anita Crescenzi and Lan Li. 2022. Assessing Realism in Simulated Work Tasks. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. ACM, Regensburg Germany, 266–271. https://doi.org/10.1145/3498366.3505831

[10] Pratibha A. Dabholkar. 1994. Incorporating Choice into an Attitudinal Framework: Analyzing Models of Mental Comparison Processes. *Journal of Consumer Research* 21, 1 (1994), 100–118. jstor:2489743 https://www.jstor.org/stable/2489743

[11] Michael Desmond, Zahra Ashktorab, Qian Pan, Casey Dugan, and James M. Johnson. 2024. EvaluLLM: LLM Assisted Evaluation of Generative Outputs. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 30–32. https://doi.org/10.1145/3640544.3645216

[12] Abigail Evans and Jacob Wobbrock. 2012. Taming Wild Behavior: The Input Observer for Obtaining Text Entry and Mouse Pointing Measures from Everyday Computer Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1947–1956. https://doi.org/10.1145/2207676.2208338

[13] Marc Franco-Salvador and Luis A. Leiva. 2018. Multilingual Phrase Sampling for Text Entry Evaluations. *International Journal of Human-Computer Studies* 113 (May 2018), 15–31. https://doi.org/10.1016/j.ijhcs.2018.01.006

[14] Mackenzie E. Hannum and Christopher T. Simons. 2020. Development of the Engagement Questionnaire (EQ): A Tool to Measure Panelist Engagement during Sensory and Consumer Evaluations. *Food Quality and Preference* 81 (April 2020), 103840. https://doi.org/10.1016/j.foodqual.2019.103840

[15] Niels Henze, Enrico Rukzio, and Susanne Boll. 2012. Observational and Experimental Investigation of Typing Behaviour Using Virtual Keyboards for Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2659–2668. https://doi.org/10.1145/2207676.2208658

[16] Xinhui Jiang, Yang Li, Jussi P.P. Jokinen, Viet Ba Hirvola, Antti Oulasvirta, and Xiangshi Ren. 2020. How We Type: Eye and Finger Movement Strategies in Mobile Typing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376711

[17] Noona Kiuru, Birgit Spinath, Anna-Leena Clem, Kenneth Eklund, Timo Ahonen, and Riikka Hirvonen. 2020. The Dynamics of Motivation, Emotion, and Task Performance in Simulated Achievement Situations. *Learning and Individual Differences* 80 (May 2020), 101873. https://doi.org/10.1016/j.lindif.2020.101873

[18] Andreas Komninos, Mark Dunlop, Kyriakos Katsaris, and John Garofalakis. 2018. A Glimpse of Mobile Text Entry Errors and Corrective Behaviour in the Wild. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 221–228. https://doi.org/10.1145/3236112.3236143

[19] Per Ola Kristensson and Keith Vertanen. 2012. Performance Comparisons of Phrase Sets and Presentation Styles for Text Entry Evaluations. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI '12)*. Association for Computing Machinery, New York, NY, USA, 29–32. https://doi.org/10.1145/2166966.2166972

[20] Luis A. Leiva and Germán Sanchis-Trilles. 2014. Representatively Memorable: Sampling the Right Phrase Set to Get the Text Entry Experiment Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, Toronto, Ontario, Canada, 1709–1712. https://doi.org/10.1145/2556288.2557024

[21] Letitia Lew, Truc Nguyen, Solomon Messing, and Sean Westwood. 2011. Of Course I Wouldn't Do That in Real Life: Advancing the Arguments for Increasing Realism in HCI Experiments. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 419–428. https://doi.org/10.1145/1979742.1979621

[22] I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase Sets for Evaluating Text Entry Techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 754–755. https://doi.org/10.1145/765891.765971

[23] Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the Evaluator: Measuring LLMs' Adherence to Task Evaluation Instructions. https://doi.org/10.48550/arXiv.2408.08781 arXiv:2408.08781 [cs]

[24] Emma Nicol, Andreas Komninos, and Mark D. Dunlop. 2016. A Participatory Design and Formal Study Investigation into Mobile Text Entry for Older Adults. *International Journal of Mobile Human Computer Interaction* 8 (May 2016), 20–46. https://doi.org/10.4018/ijmhci.2016040102.oa

[25] Hugo Nicolau, Kyle Montague, Tiago Guerreiro, André Rodrigues, and Vicki L. Hanson. 2017. Investigating Laboratory and Everyday Typing Performance of Blind Users. *ACM Trans. Access. Comput.* 10, 1 (March 2017), 4:1–4:26. https://doi.org/10.1145/3046785

[26] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies* 112 (April 2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[27] Rita Orji, Derek Reilly, Kiemute Oyibo, and Fidelia A. Orji. 2019. Deconstructing Persuasiveness of Strategies in Behaviour Change Systems Using the ARCS Model of Motivation. *Behaviour & Information Technology* 38, 4 (April 2019), 319–335. https://doi.org/10.1080/0144929X.2018.1520302

[28] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How Do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. Association for Computing Machinery, Taipei, Taiwan, 1–12. https://doi.org/10.1145/3338286.3340120

[29] Christopher S. Pan, Richard L. Shell, and Lawrence M. Schleifer. 1994. Performance Variability as an Indicator of Fatigue and Boredom Effects in a VDT Data-entry Task. *International Journal of Human–Computer Interaction* 6, 1 (Jan. 1994), 37–45. https://doi.org/10.1080/10447319409526082

[30] Felix Putze, Maik Schünemann, Tanja Schultz, and Wolfgang Stuerzlinger. 2017. Automatic Classification of Auto-Correction Errors in Predictive Text Entry Based on EEG and Context Information. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. Association for Computing Machinery, Glasgow, UK, 137–145. https://doi.org/10.1145/3136755.3136784

[31] Lucie M. Ramjan. 2011. Contextualism Adds Realism: Nursing Students' Perceptions of and Performance in Numeracy Skills Tests. *Nurse Education Today* 31, 8 (Nov. 2011), e16–e21. https://doi.org/10.1016/j.nedt.2010.11.006

[32] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 679–688. https://doi.org/10.1145/2702123.2702597

[33] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445557

[34] André Rodrigues, Hugo Nicolau, André Santos, Diogo Branco, Jay Rainey, David Verweij, Jan David Smeddinck, Kyle Montague, and Tiago Guerreiro. 2022. Investigating the Tradeoffs of Everyday Text-Entry Collection Methods. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501908

[35] Bastian Schildbach and Enrico Rukzio. 2010. Investigating Selection and Reading Performance on a Mobile Phone While Walking. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*. ACM, Lisbon Portugal, 93–102. https://doi.org/10.1145/1851600.1851619

[36] Keith Vertanen and Per Ola Kristensson. 2011. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. Association for Computing Machinery, New York, NY, USA, 295–298. https://doi.org/10.1145/2037373.2037418

[37] Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2 (Feb. 2014), 8:1–8:33. https://doi.org/10.1145/2555691

[38] Jeffrey D. Wall and Merrill Warkentin. 2019. Perceived Argument Quality's Effect on Threat and Coping Appraisals in Fear Appeals: An Experiment and Exploration of Realism Check Heuristics. *Information & Management* 56, 8 (Dec. 2019), 103157. https://doi.org/10.1016/j.im.2019.03.002

[39] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. https://doi.org/10.48550/arXiv.2302.11382 arXiv:2302.11382 [cs]

[40] Javad Zare and Ali Derakhshan. 2024. Task Engagement in Second Language Acquisition: A Questionnaire Development and Validation Study. *Journal of Multilingual and Multicultural Development* 0, 0 (2024), 1–17. https://doi.org/10.1080/01434632.2024.2306166