This is the author's accepted manuscript of an article accepted for publication in Proceedings of the ACM on Human-Computer Interaction (PACM-HCI). The final version of record is available at https://doi.org/10.1145/3821685.

Typing Behavior in Human-LLM Interaction: Keystroke Dynamics Reveal Cognitive Effort During Prompting

LAURA SCHÜTZ, Technical University of Munich, Germany
YOUSRI CHERIF, LMU Munich, Germany
CLARA SAYFFAERTH, LMU Munich, Germany
THOMAS WEBER, LMU Munich, Germany
FRANCESCO CHIOSSI, LMU Munich, Germany

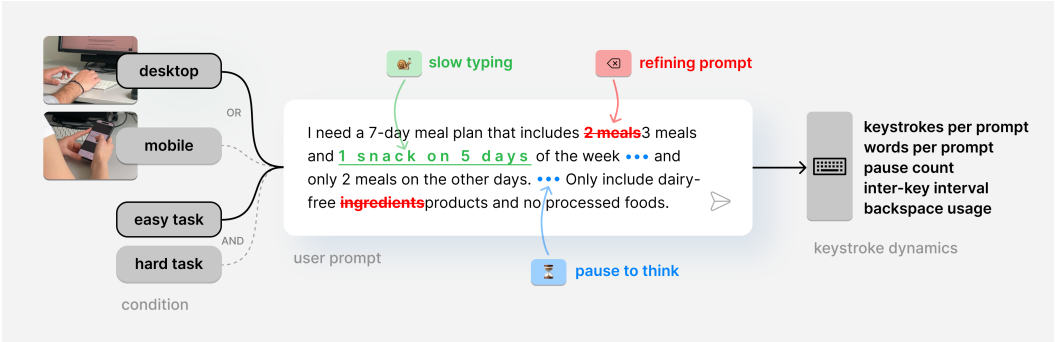


Fig. 1. We investigated how device type (mobile versus desktop) and task difficulty (easy versus hard) shape typing behavior during human–LLM interaction. Keystroke dynamics reflected users' cognitive effort during LLM prompting across desktop and mobile devices, but did not reliably predict perceived output usefulness.

As Large Language Models (LLMs) become increasingly integrated into daily routines, understanding how users interact with these systems is crucial for effective human–AI collaboration. This work investigates keystroke dynamics as a behavioral measure of user mental effort and perceived output usefulness in human-LLM interaction. We conducted a user study (N = 36) to examine how task difficulty (easy vs. hard) and device type (desktop vs. mobile) influence typing behavior and workload (NASA-TLX) during interactions. Our results indicate that hard tasks led to significantly more keystrokes, slower typing, increased pauses, and higher self-reported workload. Device type had weaker effects, with mobile use slightly reducing input length and typing speed. While keystrokes captured differences in cognitive effort, they did not predict perceived LLM output usefulness. These findings highlight the potential of keystroke dynamics as real-time indicators of cognitive effort during LLM prompting, while also showing their limitations in capturing perceived collaboration success.

CCS Concepts: • **Human-centered computing** → **User studies**; **Keyboards**; **Natural language interfaces**; **Text input**; *Mobile devices.*

Additional Key Words and Phrases: human-AI interaction, human-AI collaboration, HCI, large language model, LLM, human-LLM interaction, conversational user interface, chatbot, keystrokes, typing, prompting, keystroke dynamics, typing behavior, user behavior modeling

Authors' Contact Information: Laura Schütz, Technical University of Munich, Munich, Germany, laura.schuetz@tum.de; Yousri Cherif, LMU Munich, Munich, Germany, cherif.yousri@campus.lmu.de; Clara Sayffaerth, LMU Munich, Munich, Germany, clara.sayffaerth@ifi.lmu.de; Thomas Weber, LMU Munich, Munich, Germany, thomas.weber@ifi.lmu.de; Francesco Chiossi, LMU Munich, Munich, Germany, francesco.chiossi@ifi.lmu.de.

1 Introduction

Large language models (LLMs) are rapidly transforming how humans interact with technology. In just a short period, LLM-powered systems like ChatGPT have become one of the fastest adopted technologies in history [51], offering tremendous productivity benefits across professional writing [32], software development [57, 63], testing [58], and debugging [36]. Unlike traditional software interfaces, where users issue explicit commands, conversational chat interfaces enable users to collaborate with AI through natural language dialogue. This shift transforms users from command-givers into collaborators who iteratively refine their intent through multi-turn exchanges with an intelligent system. As LLM-powered tools become embedded in everyday workflows, understanding the behavioral dynamics of this collaboration, particularly during the prompting process itself, becomes critical for designing systems that can adapt to users' needs in real time.

However, existing evaluations of human-LLM interaction have largely relied on post-hoc methods, subjective user perception ratings [58], or outcome-based accuracy metrics that require ground truth. While useful for summative assessment, these approaches cannot capture the temporal and behavioral dynamics of interaction as it unfolds. They provide no insight into how users experience collaboration during prompting and interacting with an LLM: whether they struggle to articulate intent, feel frustrated by inadequate responses, or invest excessive cognitive effort in refinement. The absence of such insight, in turn, restricts opportunities for adaptive system design that could respond to user state in real time.

Keystroke dynamics offer a promising alternative. As text-based input remains the primary interaction modality for most LLM systems, typing behavior is both ubiquitous and rich in information. Prior research has demonstrated that keystroke patterns, e.g., pauses, inter-key intervals, and correction behavior, can reveal mental fatigue [1], cognitive load [6], and emotional state [13, 29]. Unlike physiological sensors or intrusive self-reports, keystrokes can be logged unobtrusively and continuously [9, 25, 43], making them particularly suitable for real-world deployment. However, little is known about how these signals manifest during LLM-assisted tasks, or whether they can effectively capture the quality of human-AI collaboration across different devices.

We hypothesize that keystrokes during prompting might differ from those in existing chat-based interactions due to the cognitive processes at play during prompting, including deliberations on how to instruct the LLM to achieve the intended goal, and which output to expect based on the provided prompt [50]. These cognitive steps of planning and anticipating interaction outcomes lead to iterative prompt refinement [16], potentially resulting in more pauses and higher correction rates. Furthermore, we hypothesize that dissatisfaction with LLM output, a negative arousal state, is likely to produce measurable shifts in typing dynamics. Previous research has shown that high-arousal negative states significantly affect, e.g., inter-key latency, error rate, and deletion activity [13, 20, 22, 29]. Thus, we investigated whether typing behavior during prompting can serve as a real-time signal of cognitive effort and perceived usefulness of LLM output.

To test these hypotheses, we conducted a controlled study with 36 participants to examine how task difficulty (easy vs. hard) and device type (desktop vs. mobile) shape typing behavior during iterative prompt refinement. We developed a web-based platform that logs detailed keystroke data, including typing speed, inter-key intervals, pause patterns, and correction behavior, while participants collaborate with an LLM to generate meal plans with varying levels of constraint complexity. Alongside behavioral metrics, we collected self-reported measures of cognitive workload (NASA-TLX), perceived usefulness of AI outputs, and the difficulty of refining after each interaction turn. This multi-level analysis enables us to assess whether keystroke dynamics can serve as continuous, implicit indicators of both user effort and collaboration success.

Our findings reveal both the potential and boundaries of keystroke-based evaluation for human-AI collaboration. Typing behavior proved highly sensitive to task complexity: harder tasks elicited significantly more keystrokes, slower typing, increased pauses, and higher mental workload. Device type had weaker effects, with mobile users showing only modest reductions in speed and input length, underscoring the robustness of keystroke metrics across devices. Critically, however, we found no relationship between keystroke metrics and perceived usefulness of AI responses, thus suggesting that while keystroke dynamics capture user effort during collaboration, they do not capture judgments about collaboration success. These findings position keystroke dynamics as valuable real-time indicators of cognitive effort, while highlighting the need to combine behavioral signals with semantic evaluation for comprehensive assessment of human-LLM interaction.

2 Related Work

In the following, we highlight examples of human-AI collaboration to motivate the potential benefit of keystroke dynamics as a dimension in human-AI interaction. Following this, we will go into further detail on prior work investigating keystroke dynamics to model user cognition and behavior.

2.1 Human-AI Collaboration

With the widespread adoption of large language models and conversational agents, human-AI collaboration has become a growing area of interest in HCI. While AI promises to enhance productivity and decision-making, research reveals diverse challenges in achieving effective collaboration and in measuring it. A systematic review by Vaccaro et al. [53] found that, on average, human-AI systems performed significantly worse than the best individual agent, whether human or AI alone. In scenarios where AI systems outperformed humans, integrating human input reduced performance. However, when

humans outperformed AI, hybrid teams could sometimes achieve performance gains through synergy. Critically, outcomes depended on task type, data characteristics, and AI system design, highlighting that interaction design plays a central role in collaboration success.

Traditional evaluation methods, i.e., outcome-based metrics and post-hoc subjective ratings, have proven insufficient for capturing the temporal dynamics of this collaboration. Simkute et al. [48] identified four mechanisms of productivity loss during generative AI use: users shifting from active production to passive evaluation, workflows being restructured unhelpfully through prompting overhead, flow-disrupting interruptions from AI suggestions, and task-complexity polarization where AI simultaneously simplifies easy tasks but complicates harder ones by adding monitoring demands. To mitigate these issues, the authors proposed carefully timing system interventions to preserve user flow states, yet this requires real-time awareness of user cognitive state, something post-hoc surveys cannot provide.

Evidence of process-outcome dissociations further identifies this measurement gap. Qian and Wexler [38] observed that 76 software engineers spent more time on programming tasks when using conversational AI, yet perceived themselves as more productive, showing a divergence between perceived and actual efficiency that post-hoc subjective ratings alone cannot explain. This dissociation suggests that users' retrospective judgments may not accurately reflect the cognitive costs incurred during interaction.

To reduce this gap, providing AI systems with information about the user during interaction could help adapt the AI's output to the user's needs, leading to better collaboration overall. Hemmer et al. [19], for example, propose an approach where tasks are automatically delegated between human and AI to ensure that the most suitable actor performs the task. For this to be effective, the decision logic does not only need information about the general capabilities of humans and AI but would also benefit from understanding the current state of humans, e.g., their cognitive load, emotional state, etc. Again, keystroke dynamics is one avenue for extracting this information in an unobtrusive way.

2.2 Keystroke Dynamics

Keystroke dynamics, such as the rhythm, timing, and patterns of keyboard input offer a behavioral window into users' cognitive and affective states [60]. Unlike physiological sensors that require specialized equipment, keystroke data are inherently present in text-based interaction and can be logged unobtrusively using standard hardware [43]. Common metrics include timing features (inter-key intervals, pause duration, words per minute), error-related behavior (backspace frequency, correction patterns), and structural features (input length, sentence complexity). These metrics capture individual differences in typing style while remaining sensitive to momentary changes in user state [60].

2.2.1 Keystrokes and Cognitive Load. A substantial body of work has established that keystroke patterns reflect cognitive effort during text production. Brizan et al. [6] demonstrated that typing speed, pause frequency, and linguistic complexity could reliably distinguish between low and high cognitive demand tasks. Nie et al. [31] showed that cognitive load impacts keystroke timing in quantifiable ways, with increased mental demand producing measurable changes in typing rhythm that can provide real-time feedback on user state [54].

However, the specific metrics that prove diagnostic vary across contexts. Conijn et al. [10] compared keystroke behavior during email writing versus essay writing, finding that word count, revision frequency, and total time correlated with task complexity, while inter-key intervals did not reliably indicate cognitive load. This suggested that metric validity may depend on task characteristics. In contrast, Oliveira et al. [34] observed that essay writing under higher cognitive load produced increased inter-word intervals but fewer revisions per minute, indicating users adopted a more deliberate, less iterative composition strategy. Likens et al. [23] found that more structured, consistent typing rhythms predicted higher essay quality, suggesting fluency-related metrics may capture both momentary cognitive state and domain expertise. In programming contexts, Shrestha et al. [46] found that pause frequency negatively correlated with code performance, supporting the interpretation that hesitations reflect uncertainty or difficulty.

Together, these findings demonstrate that keystroke dynamics are sensitive to cognitive effort, but the relationship between specific metrics and performance outcomes appears task-dependent. Beyond cognitive load, keystroke dynamics have also been shown to reflect mental fatigue, with variations in typing patterns serving as passive indicators of declining user wellbeing during extended computer use [1].

2.2.2 Affective States and Behavioral Context. Keystroke patterns capture not only cognitive effort but also emotional states. Epp et al. [13] demonstrated that typing dynamics could distinguish among confidence, hesitance, nervousness, relaxation, and sadness with reasonable accuracy. Similar detection capabilities have been reported for other affective states [26, 37], suggesting that keystroke dynamics reflect both how hard users are working and how they feel while doing so. The same is true for typing behavior on mobile devices. Previous works in HCI have predicted user emotions from touch inputs like keystrokes and swiping motions using machine learning classification methods on keystroke logs [17, 18] or interaction heat maps [55, 56].

Recent HCI research has further revealed that keystroke behavior is shaped by contextual and environmental factors. Large-scale in-situ studies such as ResearchIME [7] enabled naturalistic observation of typing patterns, autocorrection

use, and suggestion acceptance outside laboratory settings, revealing systematic individual differences in typing strategies. Dhakal et al. [12] identified eight distinct typing profiles that differ in finger usage, speed, and accuracy, with performance strongly predicted by the number of fingers employed. Even ambient factors matter: Mecke et al. [28] found that background music tempo influences both typing speed and error rates, highlighting the sensitivity of keystroke signatures to external conditions.

2.3 Research Gap: Keystroke Dynamics in Human-LLM Collaboration

Despite this extensive literature, a critical gap remains: no prior work has investigated how keystroke dynamics manifest during LLM-assisted tasks, or whether they can predict collaboration effectiveness. Existing studies have focused on solo text production (essay writing, email composition) or specialized domains (programming, authentication), but human-LLM interaction introduces fundamentally different dynamics. Users engage in iterative prompt refinement, evaluate and adapt to AI outputs, and adjust their input strategies based on response quality. The relationship between typing effort and task success may differ substantially from traditional writing contexts, where output quality is directly produced by the user rather than co-created with an AI system.

Moreover, it remains unclear whether the cognitive effort users invest in prompting will relate to their satisfaction with AI outputs. In practice, this relationship may be indirect or even counterintuitive: after unsatisfactory replies, users may, for example, put more effort into subsequent prompts or exhibit changes in affect. Prior work has documented dissociations between process effort and outcome quality in human-AI collaboration (Section 2.1), but no research has examined whether keystroke-based indicators of user effort align with perceived usefulness of AI responses. Our work addresses these gaps by systematically examining keystroke behavior during LLM interaction across varying task difficulty and device types, while investigating both the sensitivity of keystroke metrics to cognitive demand and their validity as predictors of collaboration success.

3 User Study

Prior research suggests that typing behavior is linked to diverse user signals, such as cognitive effort and affective state, making it a valuable behavioral metric [1, 52]. Thus, keystrokes offer a promising way to continuously assess the success of human-AI collaboration. Based on this motivation, our work is guided by the following research questions:

- **RQ1**: Can we use keystroke dynamics as continuous evaluation metrics of human-AI collaboration across device types?
- **RQ2**: How does typing behavior vary across different task complexities and device types when interacting with LLMs?
- **RQ3**: Can typing behavior indicate the perceived usefulness of LLM outputs?

To investigate these questions, we designed an experiment in which participants interacted with an LLM while their keystroke dynamics and subjective evaluations were recorded. The following sections describe the experimental conditions, tasks, and system setup.

3.1 Study Design

We conducted a between-subjects experimental design with two independent variables: Device (two levels: Mobile and Desktop) and Task Difficulty (two levels: Easy and Hard). Participants were required to iteratively refine their prompts in an AI-assisted task based on how well the AI-generated response fulfilled the task's predefined requirements.

3.2 Sample Size Justification

We conducted an a priori power analysis using G*Power (version 3.1.9.7) to determine the required sample size for our 2 (Setting: Mobile vs. Desktop, between-subjects) $\times$ 2 (Difficulty: Easy vs. Hard, within-subjects) mixed factorial design. We aimed to detect a medium-sized interaction effect between Setting and Difficulty. Based on a meta-analysis of typing experiments in HCI [33], a medium effect size for within-subject designs is estimated as Hedges's $g_{\text{rm}} = .36$, which corresponds approximately to Cohen's $f \approx .25$, following Yatani [61]. Using this effect size, a significance level of $\alpha = .05$, and a desired statistical power of $1 - \beta = .80$, the power analysis indicated a minimum total sample size of 34 participants (17 per group). To ensure proper counterbalancing of within-subject order (i.e., Easy–Hard vs. Hard–Easy), we opted to recruit 36 participants (18 per group), allowing an equal distribution across counterbalanced orders. In the G*Power setup, the number of groups was set to 2 (Mobile vs. Desktop), and the number of measurements to 2 (Easy and Hard). We assumed a correlation among repeated measures of .5 and a nonsphericity correction $\epsilon = 1$, consistent with the assumptions and empirical data from prior work [33]. The study was approved by the local ethics review board.

3.3 Dependent Variables

To evaluate the effects of device type and task difficulty on user typing behavior and perceived LLM output utility, we collected a range of dependent variables categorized under two main dimensions: Interaction Behavior and User Experience. These variables were selected to capture both objective performance measures and subjective experiences.

3.3.1 Interaction Behavior. These variables capture how users interact with the system during the task, focusing on the typing behavior. They provide insights into participant engagement, cognitive effort, and the effectiveness of user-AI interactions. Typing behavior metrics are informative of users' cognitive effort and writing fluency [6, 10, 34]. We included keystroke measures such as keystrokes per prompt, words per prompt, and inter-key interval (IKI), which are often associated with hesitation or cognitive load [34]. We additionally included the number of pauses, which are considered significant time gaps between keystrokes (e.g., exceeding 1000ms) and have been shown to indicate user performance [46]. We furthermore included the frequency of backspace and delete key usage. While there are other common error-related typing features, such as corrected error rate, uncorrected error rate, and total error rate [3, 49], these measures are primarily designed to assess transcription accuracy in controlled typing tasks. Since our study involves open-ended prompt formulation, the amount of backspace/delete key usage was chosen to capture corrections.

3.3.2 User Experience. Participants were asked to rate the following three statements after every LLM response on a scale from strongly disagree to strongly agree: "The AI's response was useful.", "Refining the AI's response was difficult.", "Refining the AI's response was mentally demanding.". At the end of every task, participants additionally filled out a raw NASA-TLX questionnaire and rated the following statement on a scale from strongly disagree to strongly agree: "I am confident that the final meal plan meets all the requirements.".

3.4 Task

We used two levels of task complexity to study how users adapt their typing behavior under different cognitive demands. The **easy task** was designed to require some interaction with the LLM, but should be solvable with minimal effort, e.g., 2-4 prompts. The **hard task** included more complex and layered requirements, designed to lead the LLM to produce output that does not fulfill the hard task's criteria on the first few prompts, thus creating the desired situation of multiple back-and-forths (e.g., 5-10 prompts) where the user must allocate more mental effort in thinking about the prompt formulation based on how past prompts evoked past responses and which task requirements are already fulfilled or not. We expected the hard task to increase user mental effort, reflecting the difficulty of solving complex tasks through human-LLM interaction, in our study, artificially induced by exploiting known model limitations.

To select a suitable task for our experiment and task complexity levels, we first defined a set of requirements. The task needed to encourage prompt iteration through multiple rounds of input refinement, allow difficulty manipulation through adjustable constraints, include clear and verifiable success criteria, be feasible on both desktop and mobile devices, and remain accessible to a broad audience without specialized knowledge. We finally selected the task of **generating a 7-day meal plan with dietary constraints of varying complexity**. Participants interacted with a large language model to create the plan and revised their prompt based on the model's output. Task difficulty was manipulated by adjusting the number of dietary requirements.

To define the dietary constraints for the task, we investigated the common limitations of large language models. Several recent studies [14, 40, 59] highlight areas where LLMs typically struggle, including but not limited to counting and maintaining item quantities, avoiding repetitions, and performing temporal planning. Prior work suggests that LLMs' deficiency in solving counting tasks stems from their design, namely their probabilistic nature, sequential token prediction, and byte-level tokenization [62]. These models also face challenges in respecting constraints that require internal logic, correcting their own mistakes, and providing accurate numerical distributions, such as percentages that correctly sum to 100% or keeping a precise count.

We designed dietary requirements (see Table 1) that map directly to these LLM limitations to make the hard task more challenging, causing high user effort and potential distress. Requirements were split into an easy and a hard condition to study how users adapt their typing behavior under different cognitive demands. The choice of LLM also influences how difficulty is experienced in practice, as different models vary in their capabilities. The design of the easy and hard tasks was refined through pilot testing with the LLM model and real users to achieve the desired difficulty levels. The final task instructions used in the study can be found in Appendix A.

3.5 Apparatus

To conduct the user study, we developed a local web-based platform for controlled human-LLM interaction. The system operated entirely offline, comprising a React frontend, a Python FastAPI backend, and a local SQLite database. This client-server architecture ensured low latency and complete data privacy. The frontend, designed to be familiar to users of common AI chatbots, is fully responsive for both desktop and mobile devices. To maintain a consistent experimental environment,

Table 1. Dietary requirements for the easy and hard task conditions.

Dietary requirements	Easy task	Hard task
7-day meal plan	✓	✓
5 days: 3 meals + 1 snack per day	✓	✓
2 days: 2 meals only	✓	✓
Dairy-free	✓	✓
No processed food	✓	
No repeated dishes		✓
No repetition of main ingredients on 2 consecutive days		✓
Even distribution of calories: 2000 kcal/day		✓
High in protein, low in carb meal plan		✓
Daily macronutrient breakdown (% carbs, fats, proteins must add up to 100%)		✓

we implemented several safeguards, such as disabling copy-paste, autocorrect, and browser spellcheck on all text inputs and including warnings to prevent accidental page reloads.

The platform integrated a local instance of the Llama 3.2 (3B) language model hosted via the Ollama server. We intentionally selected a smaller model to ensure the experimental tasks required iterative prompt refinement from participants and to emulate the constraints of privacy-preserving on-device LLMs used in mobile settings. All communication with the LLM occurred directly between the frontend and the local server to minimize latency. A system prompt was injected at the start of each session to standardize model outputs to metric units, and a chat history was maintained to provide conversational context. Responses were streamed to the user interface to simulate a real-time interaction.

A custom keystroke logging mechanism was implemented in the frontend to capture detailed typing behavior. It recorded every keypress event during prompt entry, logging the key, a high-precision timestamp, cursor position, and the full input text at that moment. All experimental data, including participant demographics, task evaluations (NASA-TLX), interaction logs, and fine-grained keystroke data, were sent to the backend and stored in the SQLite database. The relational database schema linked these data points, enabling a comprehensive analysis of user behavior in relation to task conditions and subjective feedback.

3.6 Procedure

A figure outlining the experiment flow and web-based interface can be found in Figure 2. Participants completed the study in a quiet room, where they first received a brief introduction to the research goals and provided informed consent. Each person was assigned to either a desktop (27-inch monitor, Apple Magic Keyboard) or mobile (iPhone 15 Pro) condition. Both the device type and the task difficulty order (easy → hard or hard → easy) were counterbalanced. The web-based experiment began with initial questionnaires covering demographics and AI literacy via the MAILS scale [8]. This was followed by a baseline typing task, where participants wrote a simple prompt in response to a neutral scenario to capture their natural typing behavior.

The main procedure was repeated for both an easy and a hard task. For each task, participants received printed instructions and could ask for clarification before starting. All participants received the exact same task description to ensure a consistent starting baseline. They then engaged in a chat with the LLM to complete the assignment. After each AI-generated response, they were required to rate its usefulness, as well as the difficulty and mental effort they experienced while writing the prompt. After every interaction, participants could decide to either continue the conversation or conclude the task. Upon finishing a task, they completed a NASA-TLX questionnaire to assess workload and answered a question about their confidence. The final evaluation after the second task included additional questions on keyboard familiarity and preferred typing language. All participants received a fixed compensation of 12 euros/hour for their participation.

3.7 Participants

Our study involved 36 participants (58.3% identified as male and 41.7% as female) with an age range of 19 to 34 (M=24.5). Most participants held a Bachelor's degree (52.8%), followed by a high school diploma (30.6%) and a Master's degree (13.9%). Participants were experienced and frequent users of large language models (LLMs). A significant majority (69.4%) reported using LLMs multiple times a day. Their primary use cases were for learning, research, coding, and writing. When asked about their preferred device, participants showed a strong inclination towards desktop usage. On a scale from 0 (fully mobile use) to 20 (fully desktop use), the average score was 14.4 (SD=4.8). To gauge domain knowledge for our diet-planning task, we asked participants to rate their nutrition expertise on a 0 (none) to 20 (expert) scale. The average self-assessment was moderate, with a mean score of 10.2 (SD=4.9). Finally, we assessed AI literacy using the short version of the Meta AI Literacy

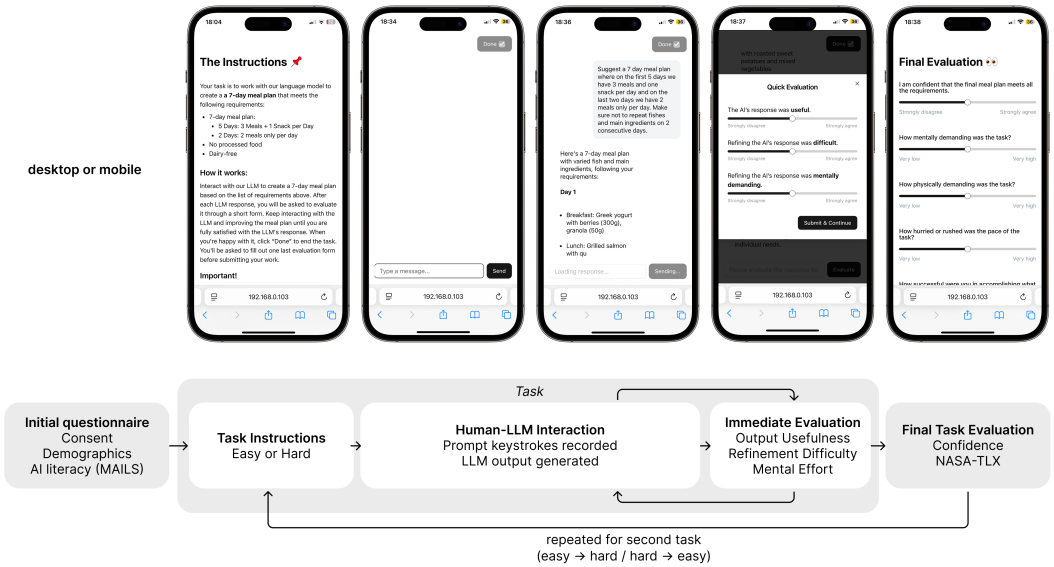


Fig. 2. Experiment protocol of the between-subjects user study. Participants either completed the desktop or mobile condition. Every participant completed an easy and a hard task in counterbalanced order. The mobile UI is shown to exemplify the interaction.

Table 2. Analyzed keystroke dynamics and descriptions

Metric	Description
Keystroke count	Number of keystrokes per prompt
Words per prompt (WPP)	Number of words per prompt
Pause count	Number of pauses (>1000 ms inter-key intervals) per prompt
Inter-key interval (IKI)	Time difference (ms) between consecutive key-down events
Backspace usage	Number of backspace/delete key events normalized by total keystroke count

Scale (MAILS) [8]. The results indicated that participants felt most confident in using AI to solve real-world problems. Conversely, they reported lower confidence in more technical areas, such as programming or designing AI systems.

3.8 Analysis

Once the user study was completed, we proceeded with the data analysis phase. Before conducting any analysis, we performed data cleaning and preprocessing. The data from two participants were discarded: one due to technical issues, and the other due to insufficient English proficiency. From the raw keystroke logs, we computed several behavioral metrics, including typing speed, input length, pause count, and backspace usage (see Table 2). We computed these metrics at all levels: per participant, per condition (easy/hard, desktop/mobile), and per interaction. This allowed us to compare typing patterns across different groups. All plots and analysis scripts were written in Python using libraries such as `pandas`, `seaborn`, and `matplotlib`. All data and analysis scripts are available at https://osf.io/dsnqf.

To investigate the effects of task difficulty and device type on typing and user experience measures, we fitted linear mixed-effects models (LMMs) using the `lme4` package. LMMs allow us to analyze hierarchical data by accounting for both fixed effects (e.g., condition) and random effects (e.g., individual user baseline differences). This provides a more robust analysis than standard summary statistics, which assume data independence [5]. Prior to model fitting, we verified LMM assumptions separately for each dependent variable. Normality of residuals was assessed via visual inspection of Q-Q plots, and homoscedasticity was evaluated through residual-versus-fitted and scale-location plots. Mental Demand residuals were approximately normally distributed. For the four keystroke count metrics (Keystroke Count, WPP, Pause Count, and IKI), residuals exhibited mild positive skew, consistent with the non-negative, count-like nature of these variables

and the presence of occasional outlier interactions. No systematic patterns indicating heteroscedasticity were identified across conditions. Given the robustness of LMMs to moderate violations of normality, particularly in larger datasets [5], we proceeded with untransformed variables to preserve interpretability of the coefficients. For the IKI model, a boundary singular fit was detected (condition_order variance = 0), and is reported transparently in the results. Models included random intercepts for user ID and condition order. Model comparisons were based on the Akaike Information Criterion (AIC), a standard measure for model quality that balances fit and complexity. To explore whether typing behavior predicts perceived usefulness of AI feedback, we employed three modeling approaches: linear mixed models, principal component regression (PCR), and random forest regression. Prior to modeling, predictors were standardized to improve interpretability. All analyses were implemented in R using libraries such as `dplyr`, `ggplot2`, `lmerTest`, and `report`.

4 Results

In the following, we report the interaction behavior and user experience results from the user study. Results on the prediction of perceived usefulness of AI output from keystroke metrics are reported as well.

4.1 Interaction Behavior

4.1.1 Number of Interactions. On average, participants spent approximately 46.6 minutes interacting with the experimental system. Each participant completed two tasks, categorized as *easy* and *hard*. Across all participants, a total of 102,454 keystrokes were recorded, with a mean of 2,845.9 keystrokes per participant. Additionally, 436 human-AI interactions were captured, corresponding to an average of 12.1 interactions per participant. When broken down by task type, *easy* tasks averaged 3.8 interactions per participant, while *hard* tasks averaged 8.3 interactions. A visual breakdown is presented in Figure 3. There was a significant main effect of task difficulty on interaction count ($p < 0.001$). Hard tasks resulted in significantly more interactions than easy tasks ($p < 0.01$), regardless of device type. There was no significant main effect of device type on the number of interactions, nor a significant interaction effect.

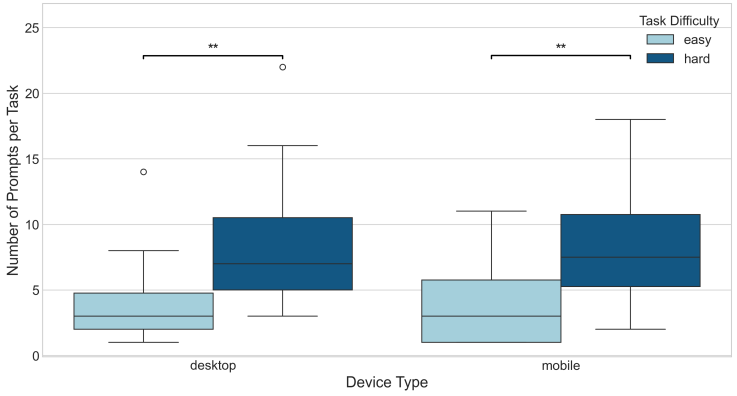


Fig. 3. Interaction counts (number of prompts) by device type and task.

4.1.2 Keystroke Metrics. To investigate how task difficulty and device type influenced typing behavior, we analyzed five key metrics: Keystroke Count, Words Per Prompt (WPP), Pause Count, Inter-Key Interval (IKI), and Backspace Usage. For each metric, we compared three LMM specifications of increasing complexity via AIC: a base model with random intercepts for participant only ($model_1$: `Outcome ~ task_difficulty × device_type + (1 | user_id)`), a model additionally including condition order ($model_2$: `+ (1 | condition_order)`), and a full model further adding prompt order ($model_3$: `+ (1 | prompt_order)`). Fixed effects were identical across all three specifications: task difficulty, device type, and their interaction. Table 3 reports, for each metric, the selected random effects structure, AIC of the best-fitting model, marginal R^2 (variance explained by fixed effects alone), and conditional R^2 (variance explained by the full model including random effects) [30], alongside the key fixed-effect coefficients. Figure 4 visualizes the distributions of these metrics by condition.

For four of five metrics (Keystroke Count, WPP, Pause Count, IKI), $model_3$ was selected as best-fitting by AIC, indicating that both condition order and prompt order account for meaningful variance beyond individual differences. For Backspace Usage, AIC favored the simpler $model_1$ structure ($\Delta AIC = 2.00$ vs. $model_2$), suggesting that correction behavior does not vary systematically with prompt or condition order. Prior to examining individual fixed effects, we note that LMM assumptions were verified for each metric as described in Section 3.8. Residuals showed mild positive skew for count-based metrics

Table 3. Summary of linear mixed model results for keystroke behavior metrics. Random effects: U=user, O=order, P=prompt; AIC = goodness of fit of best-fitting model; R^2_{marg} = variance explained by fixed effects; R^2_{cond} = variance explained by fixed and random effects combined [30]; $\hat{\beta}$ = fixed effects for task difficulty, device type, and task difficulty × device type.

Metric	Random Effects	AIC	R^2_{marg}	R^2_{cond}	Task (hard) $\hat{\beta}$	Device (mobile) $\hat{\beta}$	Task × Device $\hat{\beta}$
Keystrokes	U+0+P	5841.08	.05	.44	127.89^{***}	−5.04 (n.s.)	-88.41^{*}
WPP	U+0+P	4122.87	.05	.41	16.38^{***}	−2.74 (n.s.)	−10.34 (n.s.)
Pauses	U+0+P	3455.56	.02	.48	6.27^{***}	0.23 (n.s.)	−3.06 (n.s.)
IKI	U+0+P†	5506.07	.05	—	46.93^{*}	83.93^{*}	-58.55^{*}
Backspaces	U	−1146.65	.04	.35	−0.005 (n.s.)	0.031 (n.s.)	−0.002 (n.s.)

Significant effects: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

(Keystroke Count, WPP, Pause Count, IKI), consistent with the distributional properties of typing data [12], but no violations of homoscedasticity were detected. Results should be interpreted accordingly.

Across all metrics, marginal R^2 values were modest (range: .02 – .05), consistent with the high degree of individual variability in typing behavior documented in prior work [12, 60]. Conditional R^2 values were substantially higher (.35 – .48), confirming that individual differences captured by the random effects structure account for meaningful variance and justify the mixed-effects approach over simpler alternatives. One exception: the IKI model exhibited a boundary singular fit, with the condition_order random effect variance collapsing to zero. $R^2_{\text{conditional}}$ is therefore not reported for this metric, and interpretation is restricted to fixed effects.

Keystrokes Count. Participants produced significantly more keystrokes on hard tasks, with an average increase of 127.89 keystrokes per interaction compared to easy tasks ($SE = 29.27$, $p < .001$, 95% CI [70.35, 185.42]). Device type had no main effect on total keystroke count ($\beta = -5.04$, $SE = 47.83$, $p = .916$), indicating that typing output volume did not differ between mobile and desktop users in isolation. However, a significant interaction effect revealed that the increase in keystrokes under hard conditions was notably attenuated on mobile devices ($\beta = -88.41$, $SE = 41.65$, $p = .034$), suggesting that input constraints may have limited the amount of text generated during high-difficulty tasks on mobile.

Words Per Prompt (WPP). Task difficulty significantly influenced prompt verbosity, with hard tasks leading to a mean increase of 16.38 words compared to easy tasks ($SE = 4.00$, $p < .001$, 95% CI [8.51, 24.25]). The main effect of device type was non-significant ($\beta = -2.74$, $SE = 5.64$, $p = .627$), suggesting that participants produced similar prompt lengths across desktop and mobile platforms. Similarly, the interaction between task difficulty and device type was not significant ($\beta = -10.34$, $SE = 5.70$, $p = .070$), although the negative direction may indicate a potential attenuation of verbosity under high difficulty on mobile. Overall, task difficulty was the dominant predictor of prompt length.

Pause Count. The number of pauses per interaction was significantly higher during hard tasks, with an estimated increase of 6.27 pauses compared to easy tasks ($SE = 1.84$, $p < .001$, 95% CI [2.65, 9.88]). This suggests that participants engaged in more reflective or effortful typing when cognitive demand increased. There was no significant main effect of device type ($\beta = 0.23$, $SE = 2.73$, $p = .934$), nor a significant interaction with task difficulty ($\beta = -3.06$, $SE = 2.62$, $p = .243$), indicating that pause behavior was not notably influenced by device platform.

Inter-Key Interval (IKI). The inter-key interval (IKI), a proxy for typing fluency, was significantly affected by both task difficulty and device type. Hard tasks led to slower typing, with an average increase of 46.93 ms between keystrokes compared to easy tasks ($SE = 20.21$, $p = .021$, 95% CI [7.20, 86.66]). Typing was also slower on mobile devices ($\beta = 83.93$, $SE = 36.51$, $p = .022$), likely reflecting physical constraints of mobile input. Importantly, there was a significant interaction effect ($\beta = -58.55$, $SE = 28.61$, $p = .041$), indicating that the IKI increase under high task difficulty was less pronounced on mobile, potentially due to a performance ceiling on mobile typing.

Backspace Usage. A linear mixed-effects model revealed no significant main or interaction effects for backspace behavior. Task difficulty did not reliably affect the frequency of backspace usage ($\beta = -0.005$, $SE = 0.009$, $p = .605$), nor did device type show a statistically significant effect ($\beta = 0.031$, $SE = 0.017$, $p = .067$). The interaction between task difficulty and device type was also non-significant ($\beta = -0.002$, $SE = 0.013$, $p = .857$). While there was a trend toward increased editing on mobile, this did not reach significance, and overall, backspace usage remained stable across conditions.

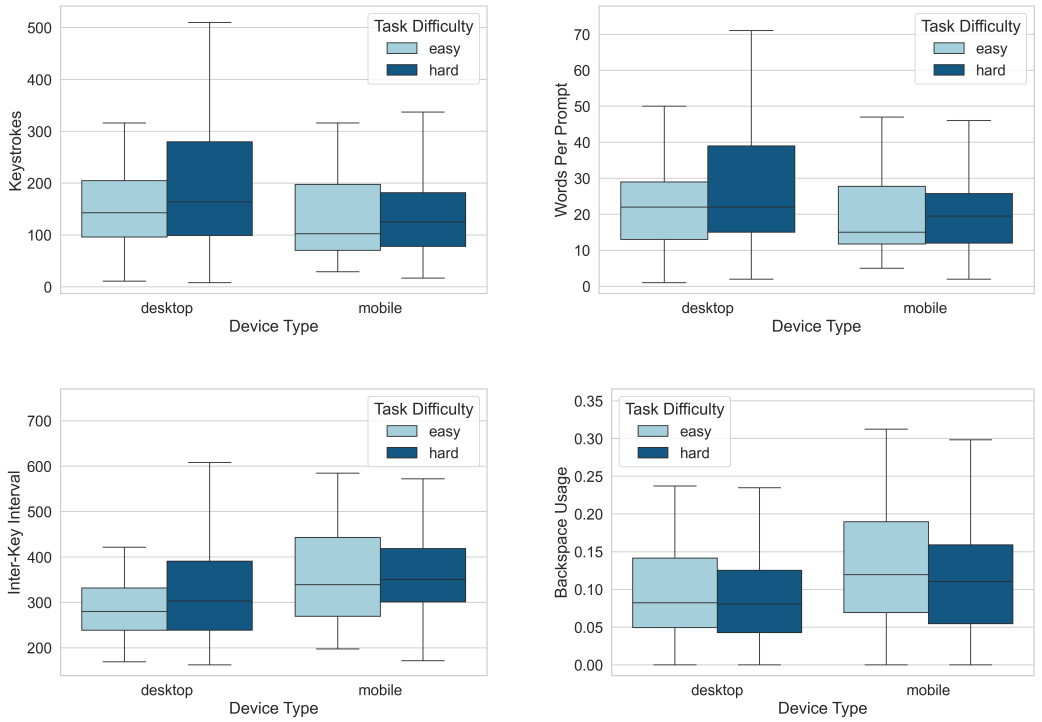


Fig. 4. Box plots of keystroke metrics by task difficulty and device type.

Table 4. Raw NASA-TLX ratings of all six subscales and overall by condition (M ± SD).

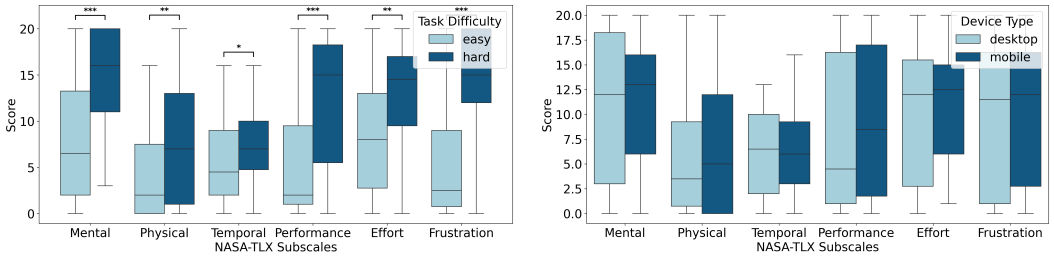
NASA-TLX variable	Desktop – Easy	Desktop – Hard	Mobile – Easy	Mobile – Hard
Mental Demand	7.00 ± 7.28	14.39 ± 6.01	8.22 ± 4.86	14.39 ± 5.55
Physical Demand	2.94 ± 4.30	7.44 ± 5.82	5.33 ± 5.17	8.33 ± 7.44
Temporal Demand	4.11 ± 4.30	7.94 ± 3.24	6.28 ± 4.13	7.61 ± 4.82
Performance	5.22 ± 7.57	11.28 ± 7.19	6.72 ± 7.47	12.78 ± 7.26
Effort	7.50 ± 6.26	12.72 ± 6.71	8.61 ± 5.40	12.78 ± 5.37
Frustration	4.89 ± 6.80	13.44 ± 6.53	6.33 ± 6.02	13.94 ± 6.39
Overall raw NASA-TLX	6.87 ± 3.84	10.78 ± 3.35	8.01 ± 3.73	10.71 ± 4.11

4.2 User Experience

4.2.1 NASA-TLX. The NASA-TLX questionnaire was used to gather subjective workload assessments across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants provided ratings on a 21-point scale from 0 to 20, following the standard NASA-TLX procedure. Table 4 summarizes the mean and standard deviation across four experimental conditions: Desktop–Easy, Desktop–Hard, Mobile–Easy, and Mobile–Hard. The Raw NASA-TLX, computed as the unweighted average of all six workload subscales further confirms these trends: both hard conditions yielded higher scores (Desktop–Hard: $M = 10.78$, $SD = 3.35$; Mobile–Hard: $M = 10.71$, $SD = 4.11$) compared to their easy counterparts (Desktop–Easy: $M = 6.87$, $SD = 3.84$; Mobile–Easy: $M = 8.01$, $SD = 3.73$).

To further investigate differences in perceived cognitive load, we analyzed *NASA-TLX Mental Demand* ratings using a linear mixed-effects model. The best-fitting model, determined by AIC comparison (AIC = 2147.55), included random

*M = mean; SD = standard deviation.



(a) NASA-TLX results by task difficulty

(b) NASA-TLX results by device type

Fig. 5. Raw NASA-TLX results by task difficulty and device type for all six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration. Significant differences are marked as follows: $p < 0.05$ = *, $p < 0.01$ = **, $p < 0.001$ = ***.

intercepts for user ID, condition order, and prompt order. The model explained a substantial amount of variance overall ($R^2_{\text{conditional}} = 0.67$), though the contribution of fixed effects alone was modest ($R^2_{\text{marginal}} = 0.07$). Results showed a significant main effect of task difficulty: hard tasks were rated as more mentally demanding than easy ones ($\beta = 3.18$, $SE = 0.77$, $p < .001$, 95% CI [1.66, 4.70]). The effect of device type was not significant ($\beta = -1.11$, $SE = 1.89$, $p = .557$), indicating no reliable difference in mental demand between mobile and desktop users. The interaction between task difficulty and device type was also non-significant ($\beta = 1.65$, $SE = 1.12$, $p = .141$). These results support the descriptive findings, confirming that harder tasks reliably increased perceived mental workload, independent of the device used. Figure 5 illustrates this trend for all NASA-TLX subscales.

4.2.2 Mental Effort. We also asked participants to rate the mental effort required to write and refine the meal plan prompt every interaction. We found a significant main effect of task difficulty ($p < 0.05$) on mental effort, indicating that more demand was imposed by the hard tasks ($M = 11.97, SD = 5.83$) than the easy tasks ($M = 8.08, SD = 5.42$). No significant main effect was found for device type, and there was no significant interaction effect. The specific means were: desktop-easy ($M = 7.24, SD = 6.09$), desktop-hard ($M = 11.82, SD = 6.05$), mobile-easy ($M = 9.14, SD = 4.47$), and mobile-hard ($M = 12.13, SD = 5.76$).

4.2.3 LLM Output Usefulness. We also asked them to rate each LLM output by it's usefulness for achieving the task. There was a significant main effect of task difficulty on output usefulness ($p < 0.001$), with the response usefulness during easy tasks ($M = 13.71, SD = 5.22$) being rated significantly higher than during hard tasks ($M = 8.83, SD = 5.26$). There was no significant main effect of device type and no significant interaction effect. The means for the specific conditions were: desktop-easy ($M = 14.15, SD = 4.67$), desktop-hard ($M = 8.96, SD = 5.34$), mobile-easy ($M = 13.27, SD = 5.82$), and mobile-hard ($M = 8.70, SD = 5.33$).

4.2.4 Difficulty Refining Output. Regarding refinement difficulty, the analysis revealed a significant main effect for task difficulty ($p < 0.05$). Overall, refining the LLM output for the hard task ($M = 11.85, SD = 5.49$) was rated as more difficult than for the easy task ($M = 8.18, SD = 5.45$). There was no significant main effect for device type and no significant interaction effect. The condition means were: desktop-easy ($M = 6.41, SD = 5.53$), desktop-hard ($M = 11.76, SD = 5.32$), mobile-easy ($M = 10.38, SD = 4.66$), and mobile-hard ($M = 11.95, SD = 5.81$).

4.2.5 Confidence. Subjective confidence ratings, assessed using a 21-point scale ranging from 0 (very low) to 20 (very high), were collected to evaluate users' perceived success in meeting task requirements. Confidence was substantially higher in easy tasks (Desktop–Easy: $M = 16.56$, $SD = 6.34$; Mobile–Easy: $M = 17.11$, $SD = 4.04$) than in hard tasks (Desktop–Hard: $M = 8.06$, $SD = 7.57$; Mobile–Hard: $M = 8.39$, $SD = 7.77$). There was a significant main effect of task difficulty ($p < 0.001$) on confidence ratings. Participants reported significantly higher confidence after the easy tasks compared to the hard tasks. There was no significant main effect for device type and no significant interaction effect.

4.3 Predicting LLM Usefulness from Keystrokes

To assess whether behavioral typing metrics could predict participants' perceived usefulness of AI output, we fitted three distinct models: a linear mixed-effects model, a principal component regression, and a non-linear random forest model.

*R^2_{marginal} = variance explained by fixed effects; $R^2_{\text{conditional}}$ = variance explained by both fixed and random effects; β = estimated regression coefficient (effect size); SE = standard error; CI = confidence interval.

These models incorporated five core typing metrics: Words Per Prompt, Keystrokes, Backspace Usage Ratio, Inter-Key Interval, and Pause Count.

In the linear mixed-effects model, none of the individual typing metrics significantly predicted feedback usefulness. Fixed effects accounted for only 1% of the variance ($R^2 = 0.01$), while total model variance explained was 25% ($R^2_{\text{conditional}} = 0.25$). All predictors (e.g., WPP: $p = .903$, Keystrokes: $p = .676$, Pauses: $p = .759$) failed to reach statistical significance. Principal component regression using the top two components (PC1, PC2) yielded similar results. PC1 showed a marginal trend toward significance ($\beta = -0.35$, $p = .052$), but did not meet conventional thresholds. Overall, the model retained low explanatory power ($R^2 = 0.01$). Finally, a random forest model trained with 5-fold cross-validation produced minimal predictive performance, with a best-case R^2 of 0.005. This suggests that the relationship between typing behavior and perceived usefulness is weak or non-existent in this dataset. These findings collectively indicate that the typing behavior metrics examined in this study do not reliably predict participants' ratings of the LLM's response usefulness.

Table 5. Summary of modeling results for predicting AI output usefulness from typing behavior. R^2_{marg} denotes variance explained by fixed effects, and R^2_{cond} variance explained by the full model (fixed and random effects). For the random forest, we report mean cross-validated R^2.

Model	$R^2_{\textbf{marg}}$	$R^2_{\textbf{cond}}$	**Best Predictors**	**Significant?**
LMM	0.01	0.25	None	No
PCR (PC1, PC2)	0.01	0.25	PC1 (borderline)	No
Random Forest	Mean R^2: 0.005		None	No

5 Discussion

5.1 Summary of Findings

This study explored how typing behavior reflects users' perceived utility of LLM output and cognitive effort during prompting LLMs. Keystroke data were collected across different task-difficulty conditions and device types.

5.1.1 Typing Behavior Across Task Complexity and Device Type (RQ1, RQ2).

Keystrokes capture cognitive effort during human-LLM interaction. Overall, task difficulty had a clear and consistent impact. In harder tasks, participants wrote longer and more frequent prompts, used more keystrokes, paused more often, and typed more slowly. These findings are in line with prior work, which similarly reported lower typing speed and higher pause frequency during high-demand writing tasks [34]. Additionally, participants reported significantly higher levels of mental demand, frustration, and effort. These trends suggest that harder tasks require deeper reflection and engagement, even with the aid of LLMs. While it is inherently expected that a more complex task, with more requirements, would necessitate a higher volume of interactions and keystrokes, our findings suggest that the cognitive load is not merely a byproduct of text length. If the increased interaction were purely mechanical, we would expect typing speed to remain stable as volume increased. Instead, the observed decrease in typing speed and the increase in pause frequency indicate that participants were engaged in more frequent "think-then-type" cycles during the hard task. This suggests that the hard task successfully triggered a loop of cognitive evaluation: users had to parse the LLM's output, compare it against multiple constraints, and strategize a refinement. The keystroke data, therefore, captures the friction of this mental processing rather than just the physical labor of inputting the task requirements.

Keystrokes are equally expressive of cognitive effort across mobile and desktop. In contrast, the type of device used (mobile vs. desktop) did not produce strong main effects. Participants typed more slowly on mobile devices, which goes in hand with known previous findings [4, 35], but other differences were limited. For instance, word count and the perceived mental demand remained consistent across both device types. It is worth noting that the interaction in the mobile and desktop condition was nearly identical in terms of interface design, system responsiveness, and overall smoothness. The main difference was the typing method (physical keyboard versus touchscreen) and the size of the input interface. However, given the relatively young and digitally literate participant pool, it is likely that most users were sufficiently comfortable with both input modalities, resulting in minimal observable impact on typing behavior or cognitive workload.

Mobile subtly modulated typing behavior under high difficulty. When we looked at the interaction between task difficulty and device type, the effects were generally weak. One exception was a decrease in keystrokes for hard tasks on mobile compared to desktop, despite the overall increase due to task complexity. This may reflect a tendency among mobile users to shorten or simplify their responses under more constrained input conditions due to smaller key and screen sizes. This is a critical finding for the relationship between task complexity and typing volume. If keystroke count were solely a function

of task requirements, we would see a uniform increase across both devices. However, we see evidence that typing behavior is actively modulated by cognitive load and interface constraints, not merely by the number of task requirements.

Backspace use is unaffected by device type and task difficulty. While keystroke count, words per prompt, pause count, and inter-key interval were all significantly impacted by task difficulty, backspace use remained consistent across all conditions. It did not change with task difficulty or device. This suggests that editing habits are resilient to variations in task complexity or device constraints.

5.1.2 *Predicting Usefulness from Keystrokes (RQ3).*

Keystrokes do not capture perceived LLM output usefulness. Perceived usefulness of LLM outputs declined significantly for hard tasks despite increased user effort, a pattern consistent with diminishing returns from additional interaction and refinement under higher task complexity, though no causal relationship can be inferred. Our results underscore the task-complexity-polarization phenomenon in human-AI collaboration, which has been shown in previous works, postulating that easy tasks remain easy, but hard tasks stay hard despite the assistance of AI [48].

The measured keystroke metrics could not predict the perceived usefulness of AI responses. In our models, none of the keystroke metrics correlated with participant ratings of response usefulness. At the same time, our dataset showed sufficient variation in perceived utility, with participant assessments ranging from highly useful to not useful at all. This suggests that the absence of correlation is not due to a lack of variance in the target variable. A more likely explanation is that usefulness is shaped primarily by the semantic content of the prompt or keystroke metrics not captured within this study. If the LLM fails to sufficiently satisfy the task demands, users may engage in more iterative prompt refinement, as evidenced by longer pauses between bursts of writing, reflecting planning and reformulation. Another indicator of diminishing returns might be a high lexical overlap but limited semantic progression, increasing edit distance between successive prompts without corresponding improvements in output. These patterns might suggest that users invest additional effort in attempting to steer the model, yet receive limited benefit.

Overall, while keystroke dynamics offer a window into user effort and cognitive load, they appear less effective for capturing higher-level subjective judgements such as usefulness. More fine-grained metrics and semantic text analysis could offer additional insights. Similarly, combining keystroke data with other behavioral signals such as mouse movement or scrolling patterns might help model perceived usefulness more accurately.

In summary, typing behavior captured real-time user effort and showed some variation across devices. These findings suggest that keystroke dynamics can serve as an accessible proxy for mental demand during human-AI collaboration and contribute to the growing HCI literature on implicit user state detection [9, 15, 42, 44]. However, a comprehensive understanding of certain user perceptions, such as collaboration success, may require incorporating additional signals within and beyond keystroke dynamics.

5.2 Designing Adaptive Human-LLM Interactions Based on Keystrokes

5.2.1 Cognitive Load Adaptive Human-AI Interactions. The success in reflecting cognitive effort across conditions suggests promising avenues for adaptive system design. By leveraging keystroke-based indicators of user effort, future platforms could dynamically adjust their behavior to better support users in real time. Several layers of adaptivity could be explored. On the interface level, systems could adjust the interface to reduce cognitive strain. On the model level, the choice of LLM or its response length and verbosity could be tailored to the user's cognitive state. Finally, semantic-level prompt adaptations could refine the content or complexity of LLM responses, ensuring they align more closely with the user's needs and capacity in the moment.

Building on prior work on adaptive prompting interfaces [11], such signals could also guide when to provide assistance. For instance, systems might withhold prompt-completion suggestions under low cognitive load while offering completions as effort increases. Alternatively, keystroke dynamics could serve as real-time signals that inform adaptive task allocation strategies, as proposed by Hemmer et al. [19], by dynamically shifting responsibility between human and AI based on the most suitable actor. Importantly, these signals may also support meta-level interventions. Systems could detect when continued LLM interaction yields diminishing returns and notify users accordingly, encouraging more deliberate and context-sensitive AI usage. This aligns with prior findings that users do not always make optimal decisions about when to rely on AI, even when performance information is available [39].

5.2.2 Balancing Cognitive Offloading and Critical Thinking. While cognitive-load adaptive systems hold promise, future HCI research must balance the goal of reducing cognitive load with the risk of user deskilling [47]. Lower cognitive load is not always desirable if it removes opportunities for reflection and skill acquisition during AI-assisted tasks, as prior work shows that users may reject low-effort recommendation systems in favor of approaches that keep them cognitively engaged [41]. Interfaces should therefore be designed not only to optimize efficiency, but to deliberately preserve moments of cognitive engagement that are critical for reflection, verification, and independent reasoning. This aligns with prior findings that

increased confidence in AI can reduce critical thinking effort, suggesting that overly assistive interfaces may inadvertently encourage cognitive offloading rather than active stewardship [21, 41]. Consequently, adaptive systems should incorporate friction, transparency, and user-controlled levels of assistance to maintain critical thinking and skill development.

5.2.3 Privacy-Preserving Keystroke Data Processing. While keystroke dynamics offer a powerful method for evaluating user state, logging keystroke data raises significant privacy concerns. Keystroke patterns can be unique enough to serve as a biometric identifier [27] and could be used for the inference of sensitive traits beyond cognitive load, such as emotional states [13, 37, 52] or health conditions [2]. In our study, data was processed locally and anonymized. However, real-world deployment must prioritize privacy-preserving architectures, using edge-based processing or federated learning approaches to refine models without centralizing sensitive biometric data.

5.3 Limitations & Future Work

We acknowledge several limitations of our work. The participant pool was relatively homogeneous, with most participants being university students aged between 19 and 34 years. This affects how findings apply to broader populations. Keystroke data was limited to key press events only. We did not record key release timestamps, which restricted the analysis of metrics such as dwell time and flight time.

Another limitation relates to the analysis of a single task (meal planning), which limits the generalizability of the findings. Future work should repeat this study across multiple task domains. Furthermore, we utilized a single local LLM instance. Replicating the study with various state-of-the-art LLMs, as well as comparing the results to non-LLM chatbots, would help isolate the specific impact of human-agent interaction dynamics.

This work represents an initial step toward understanding interaction-level signals in human-AI collaboration. Potential next steps could include semantic analysis of user inputs. Whereas keystroke patterns capture interaction effort, the prompts themselves can reveal aspects of user state, such as emotional tone or cognitive complexity. Metrics such as lexical diversity, syntactic structure, and sentiment may help explain why some responses are perceived as more useful than others. A multi-layered prompt analysis combining behavioral and semantic signals may provide a more complete picture of perceived utility in LLM-assisted tasks. In addition, future studies could extend the current logging system to include key-up events, enabling the calculation of dwell and flight times. To broaden the scope of interaction data, subsequent research may also incorporate multimodal interaction signals, such as mouse [24] and eye tracking [45].

6 Conclusion

This work explored whether keystroke dynamics can serve as a real-time indicator of user effort and perceived utility during interactions with large language models. We conducted a controlled user study manipulating task difficulty and device type to examine their influence on typing behavior and subjective experience. Our findings showed that keystroke dynamics can reliably reflect cognitive effort during prompting across mobile and desktop devices. While keystroke dynamics did not predict perceived usefulness of LLM output, we believe that combining typing metrics with other biosignals or sentiment analysis holds promise. Overall, our work demonstrates the potential of keystroke data as a low-cost, non-intrusive method for designing adaptive AI systems that respond to users' cognitive states in real time. Finally, as LLMs continue to evolve and expand in use, understanding how users engage with them remains a pressing challenge. We see this work as an important step toward designing more human-centered AI systems.

7 Data Availability

The collected data and analysis scripts are available on Open Science Framework (https://osf.io/dsnqf).

Acknowledgments

This work was conducted as part of the AI-Twin project, which is funded by the European Research Council (ERC-2024-ADG) as part of the European Union's Horizon 2020 research and innovation program (grant agreement no. 101200584).

References

[1] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Julian Fierrez, Ijah Mondesire-Crump, and Teresa Arroyo-Gallego. 2022. Detection of Mental Fatigue in the General Population: Feasibility Study of Keystroke Dynamics as a Real-world Biomarker. *JMIR Biomed Eng* 7, 2 (21 Nov 2022), e41003.

[2] Hessa Alfalahi, Ahsan H Khandoker, Nayeefa Chowdhury, Dimitrios Iakovakis, Sofia B Dias, K Ray Chaudhuri, and Leontios J Hadjileontiadis. 2022. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. *Scientific reports* 12, 1 (2022), 7690.

[3] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2009. Analysis of text entry performance metrics. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH).* IEEE, New York, NY, USA, 100–105. doi:10.1109/TIC-STH.2009.5444533

[4] Julia Barrett and Helmut Krueger. 1994. Performance effects of reduced proprioceptive feedback on touch typists and casual users in a typing task. *Behaviour & Information Technology* 13, 6 (1994), 373–381.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67 (2015), 1–48.

[6] David Guy Brizan, Adam Goodkind, Patrick Koch, Kiran Balagani, Vir V. Phoha, and Andrew Rosenberg. 2015. Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies* 82 (2015), 57–68. doi:10.1016/j.ijhcs.2015.04.005

[7] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173829

[8] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (2023), 100014. doi:10.1016/j.chbah.2023.100014

[9] Francesco Chiossi, Yassmine El Khaoudi, Changkun Ou, Ludwig Sidenmark, Abdelrahman Zaky, Tiare Feuchtner, and Sven Mayer. 2024. Evaluating Typing Performance in Different Mixed Reality Manifestations using Physiological Features. *Proc. ACM Hum.-Comput. Interact.* 8, ISS, Article 542 (Oct. 2024), 30 pages. doi:10.1145/3698142

[10] Rianne Conijn, Jens Roeser, and Menno Van Zaanen. 2019. Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing* 32, 9 (2019), 2353–2374.

[11] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 408, 17 pages. doi:10.1145/3544548.3580969

[12] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174220

[13] Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 715–724. doi:10.1145/1978942.1979046

[14] Tairan Fu, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. 2026. When Do Large Language Models (LLMs) Struggle to Count Letters? *ACM Trans. Intell. Syst. Technol.* 17, 4 (2026), 1–17. doi:10.1145/3818606

[15] Marc-Philipp Funk, Nassir Navab, and Laura Schütz. 2025. Eye Tracking-Based Adaptive User Interfaces in Virtual Reality Eye Surgery Training. In *Proceedings of the Mensch und Computer 2025 (MuC '25)*. Association for Computing Machinery, New York, NY, USA, 578–582. doi:10.1145/3743049.3748551

[16] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. doi:10.1145/3613905.3650786

[17] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. TapSense: combining self-report patterns and typing characteristics for smartphone based emotion detection. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 2, 12 pages. doi:10.1145/3098279.3098564

[18] Surjya Ghosh, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Emotion detection from touch interactions during text entry on smartphones. *International Journal of Human-Computer Studies* 130 (2019), 47–57. doi:10.1016/j.ijhcs.2019.04.005

[19] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 453–463. doi:10.1145/3581641.3584052

[20] Agata Kołakowska. 2015. Recognizing emotions on the basis of keystroke dynamics. In *2015 8th International Conference on Human System Interaction (HSI)*. IEEE, New York, NY, USA, 291–297. doi:10.1109/HSI.2015.7170682

[21] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1121, 22 pages. doi:10.1145/3706598.3713778

[22] Po-Ming Lee, Wei-Hsuan Tsui, and Tzu-Chien Hsiao. 2015. The Influence of Emotion on Keyboard Typing: An Experimental Study Using Auditory Stimuli. *PloS one* 10, 6 (2015), 1–16. doi:10.1371/journal.pone.0129056

[23] Aaron D Likens, Laura K Allen, and Danielle S McNamara. 2017. Keystroke dynamics predict essay quality. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 39. Cognitive Science Society, Seattle, WA, USA, 2573–2578.

[24] Yee Mei Lim, Aladdin Ayesh, and Martin Stacey. 2014. Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic. In *2014 Science and Information Conference*. IEEE, New York, NY, USA, 146–152. doi:10.1109/SAI.2014.6918183

[25] Ailin Liu, Fiona Yesmine Karoui, Draxler, Frauke Kreuter, and Francesco Chiossi. 2026. Sensing What Surveys Miss: Understanding and Personalizing Proactive Intelligent Support by User Modeling. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3772318.3791191

[26] Stefano Marrone, Carlo Sansone, et al. 2022. Identifying Users' Emotional States through Keystroke Dynamics. *DeLTA* 2022 (2022), 207–214.

[27] Andreia Martins, Tiago Dias, André Dias, João Vitorino, Eva Maia, and Isabel Praça. 2025. Keystroke dynamics for intelligent biometric authentication with machine learning. *Discover Applied Sciences* 7, 9 (2025), 992.

[28] Lukas Mecke, Assem Mahmoud, Simon Marat, and Florian Alt. 2025. Exploring the Effect of Music on User Typing and Identification through Keystroke Dynamics. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 761, 10 pages. doi:10.1145/3706598.3713222

[29] A.F.M. Nazmul Haque Nahin, Jawad Mohammad Alam, Hasan Mahmud, and Kamrul Hasan. 2014. Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology* 33, 9 (2014), 987–996. doi:10.1080/0144929X.2014.907343

[30] Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4, 2 (2013), 133–142. doi:10.1111/j.2041-210x.2012.00261.x

[31] Yafei Nie, Shurong Tong, Jing Li, Yicha Zhang, Chen Zheng, and Bin Fan. 2022. Time identification of design knowledge push based on cognitive load measurement. *Advanced Engineering Informatics* 54 (2022), 101783. doi:10.1016/j.aei.2022.101783

[32] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. doi:10.1126/science.adh2586

[33] Natalia Obukhova. 2021. A Meta-Analysis of Effect Sizes of CHI Typing Experiments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 476, 7 pages. doi:10.1145/3411763.3451520

[34] Eduardo Araujo Oliveira, Rianne Conijn, Paula Galvao de Barba, Kelly Trezise, Menno van Zaanen, and Gregor Kennedy. 2020. Writing analytics across essay tasks with different cognitive load demands. In *ASCILITE 2020 Conference Proceedings*. Australasian Society for Computers in Learning in Tertiary Education, Tugun, Australia, 60–70.

[35] 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* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 9, 12 pages. doi:10.1145/3338286.3340120

[36] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2023. Examining Zero-Shot Vulnerability Repair with Large Language Models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, New York, NY, USA, 2339–2356. doi:10.1109/SP46215.2023.10179324

[37] Yuqing Qi, Weichen Jia, and Shuo Gao. 2021. Emotion Recognition Based on Piezoelectric Keystroke Dynamics and Machine Learning. In *2021 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS)*. IEEE, New York, NY, USA, 1–4. doi:10.1109/FLEPS51544.2021.9469843

[38] Crystal Qian and James Wexler. 2024. Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 370–384. doi:10.1145/3640543.3645198

[39] Han Qiao, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2025. To Use or Not to Use: Impatience and Overreliance When Using Generative AI Productivity Support Tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1122, 18 pages. doi:10.1145/3706598.3714103

[40] Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay B. Cohen. 2024. Are Large Language Model Temporally Grounded?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 7064–7083. doi:10.18653/v1/2024.naacl-long.391

[41] Leon Reicherts, Zelun Tony Zhang, Elisabeth von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, Help Me Think—but for Myself: Assisting People in Complex Decision-Making by Providing Different Kinds of Cognitive Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 255, 19 pages. doi:10.1145/3706598.3713295

[42] Laura Schütz, Shervin Dehghani, Michael Sommersperger, Koorosh Faridpooya, and Nassir Navab. 2025. The impact of intraoperative optical coherence tomography on cognitive load in virtual reality vitreoretinal surgery training. *Scientific Reports* 15, 1 (2025), 24848. doi:10.1038/s41598-025-07670-7

[43] Rashik Shadman, Ahmed Anu Wahab, Michael Manno, Matthew Lukaszewski, Daqing Hou, and Faraz Hussain. 2025. Keystroke Dynamics: Concepts, Techniques, and Applications. *ACM Comput. Surv.* 57, 11, Article 283 (June 2025), 35 pages. doi:10.1145/3733103

[44] Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S. Bernstein. 2025. Creating General User Models from Computer Use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 35, 23 pages. doi:10.1145/3746059.3747722

[45] Danqing Shi, Yujun Zhu, Jussi P. P. Jokinen, Aditya Acharya, Aini Putkonen, Shumin Zhai, and Antti Oulasvirta. 2024. CRTypist: Simulating Touchscreen Typing Behavior via Computational Rationality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 942, 17 pages. doi:10.1145/3613904.3642918

[46] Raj Shrestha, Juho Leinonen, Albina Zavgorodniaia, Arto Hellas, and John Edwards. 2022. Pausing while programming: insights from keystroke analysis. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Software Engineering Education and Training* (Pittsburgh, Pennsylvania) *(ICSE-SEET '22)*. Association for Computing Machinery, New York, NY, USA, 187–198. doi:10.1145/3510456.3514146

[47] Prakash Shukla, Phuong Bui, Sean S Levy, Max Kowalski, Ali Baigelenov, and Paul Parsons. 2025. De-skilling, Cognitive Offloading, and Misplaced Responsibilities: Potential Ironies of AI-Assisted Design. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 171, 7 pages. doi:10.1145/3706599.3719931

[48] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel and. 2025. Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interaction. *International Journal of Human–Computer Interaction* 41, 5 (2025), 2898–2919. doi:10.1080/10447318.2024.2405782

[49] R. William Soukoreff and I. Scott MacKenzie. 2003. Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 113–120. doi:10.1145/642611.642632

[50] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754

[51] Benjamin Todd. 2025. AI is the most rapidly adopted technology in history. https://benjamintodd.substack.com/p/when-people-say-ai-isnt-finding-real

[52] Matthias Trojahn, Florian Arndt, Markus Weinmann, and Frank Ortmeier. 2013. Emotion Recognition through Keystroke Dynamics on Touchscreen Keyboards. In *Proceedings of the 15th International Conference on Enterprise Information Systems - Volume 3: ICEIS*. INSTICC, SciTePress, Setubal, Portugal, 31–37. doi:10.5220/0004415500310037

[53] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (Oct. 2024), 2293–2303. doi:10.1038/s41562-024-02024-1

[54] Lisa M. Vizer. 2009. Detecting cognitive and physical stress through typing behavior. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 3113–3116. doi:10.1145/1520340.1520440

[55] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. 2020. Affective State Prediction Based on Semi-Supervised Learning from Smartphone Touch Data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376504

[56] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, Markus Gross, and Christian Holz. 2022. Affective State Prediction from Smartphone Touch and Sensor Data in the Wild. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 403, 14 pages. doi:10.1145/3491102.3501835

[57] Thomas Weber, Maximilian Brandmaier, Albrecht Schmidt, and Sven Mayer. 2024. Significant Productivity Gains through Programming with Large Language Models. *Proc. ACM Hum.-Comput. Interact.* 8, EICS, Article 256 (June 2024), 29 pages. doi:10.1145/3661145

[58] Justin D. Weisz, Shraddha Vijay Kumar, Michael Muller, Karen-Ellen Browne, Arielle Goldberg, Katrin Ellice Heintze, and Shagun Bajpai. 2025. Examining the Use and Impact of an AI Code Assistant on Developer Productivity and Experience in the Enterprise. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 673, 13 pages. doi:10.1145/3706599.3706670

[59] Nan Xu and Xuezhe Ma. 2025. LLM The Genius Paradox: A Linguistic and Math Expert's Struggle with Simple Word-based Counting Problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Albuquerque, New Mexico, 3344–3370. doi:10.18653/v1/2025.naacl-long.172

[60] Liying Yang and Shengfeng Qin. 2025. Identify user features impacting keystroke, mouse and touchscreen dynamics. *Multimedia Tools and Applications* 84 (16 Aug 2025), 48685–48713. doi:10.1007/s11042-025-21043-2

[61] Koji Yatani. 2016. Effect Sizes and Power Analysis in HCI. In *Modern Statistical Methods for HCI*, Judy Robertson and Maurits Kaptein (Eds.). Springer, Cham, 87–110. doi:10.1007/978-3-319-26633-6_5

[62] Xiang Zhang, Juntai Cao, and Chenyu You. 2024. Counting Ability of Large Language Models and Impact of Tokenization. arXiv:2410.19730 [cs.CL] https://arxiv.org/abs/2410.19730

[63] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity assessment of neural code completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming* (San Diego, CA, USA) *(MAPS 2022)*. Association for Computing Machinery, New York, NY, USA, 21–29. doi:10.1145/3520312.3534864

A Task Instructions

How it works: Interact with our LLM to create a 7-day meal plan based on the list of requirements. After each LLM response, you will be asked to evaluate it through a short form. Keep interacting with the LLM and improving the meal plan until you are fully satisfied with the LLM's response. When you're happy with it, click "Done" to end the task. You'll be asked to fill out one last evaluation form before submitting your work.

A.1 Easy Task

Your task is to work with our language model to create a 7-day meal plan that meets the following requirements:

(1) 7-day meal plan
   - 5 days: 3 meals + 1 snack per day
   - 2 days: 2 meals only per day

(2) No processed food
(3) Dairy-free

A.2 Hard Task

Your task is to work with our language model to create a 7-day meal plan that meets the following requirements:

(1) 7-day meal plan
   - 5 days: 3 meals + 1 snack per day
   - 2 days: 2 meals only per day

(2) No repeated dishes
(3) No repetition of main ingredients on 2 consecutive days
(4) Even distribution of calories: 2000 per day
(5) High in protein, low in carb meal plan
(6) Show the percentage of carbs, fats, and proteins for every day (sum must add up to 100%)
(7) Dairy-free