Working title: Data Descriptor for HMC Data Set

Angelica Henestrosa

Leibniz-Institut für Wissensmedien, Tübingen

September 8, 2025

Abstract

Since the emergence of large language models (LLMs) in 2022, generative AI has rapidly expanded into mainstream applications, leading to the integration of Apple Intelligence into customer devices in 2024. This integration into personal technology marks a significant shift, bringing advanced AI capabilities into everyday devices and making them accessible to private individuals. Thus, the use of generative AI–consciously or unconsciously–along with interaction through LLM-powered (voice) assistants and engagement with AI-generated content is expected to increase significantly. However, data that link this usage to psychological variables and track it over time remain scarce. This longitudinal study comprises the data from an American sample across six waves at two-month intervals between September 2024 and July 2025. It examines user behavior, attitudes, knowledge, and perceptions related to generative AI. ... This dataset allows for future research on psychological and behavioral dynamics of AI use over time, offering insights into user engagement and the individual factors connected to it.

1 Background and Summary

... Longitudinal studies like this are needed to capture the evolving perceptions of opportunities and risks associated with AI, perceived capabilities of AI systems, attitudes toward AI, trust in AI, willingness to delegate tasks to AI, areas of application, and the interrelationships among these constructs over time (to be continued). To examine those changes and relationships, an American sample mainly consisting of AI users (specify) was invited to participate in this survey at two-month intervals between September 2024 and July 2025. * dataset brings together various seperate WPs -> possibility to make across-WP analyses * potential to look on clusters/subgroups/individual trajectories ignored in the WPs * snapshots of important points in time (LLMs on the rise) * outlook on potential developments in other countries * connection of actual use and stable psychological variables

2 Methods

2.1 e.g.: Participants and Data Collection

* Prolific * Invitation * time and intervals * retention rate * second sample -> invitation of wave1 participants * focus on users -> exclusion of nousers without intention * ethics approval

2.2 e.g.: Measurements

* List of all measures by wave

We collected sociodemographic information, including, age, gender, educational level, and household income from all participants at wave 1.

3 Data Records

Data records for each of the six waves are available in csv format at (tbd) together with the R/python scripts for data anonymization, data cleaning, and data preprocessing. That is, firstly the data was anonymized by removing participants' Prolific IDs and unused variables, empty variables resulting from faulty questionnaire programming, and xy were removed. Thus (filename) represents the cleaned and anonymized raw data, including the single items of each measurement. Second, variable names were harmonized and scales were calculated, resulting an the preprocessed data set xy, ready for analyses across scales. Moreover, a codebook explaining variable abbreviations and containing information about the waves in which the variable was collected (what else?) is available at (tbd).

4 Technical Validation

* attention check * bot detection question * forced to respond

5 Usage Notes (optional)

6 Code Availability

All python (version x) an R (version x) code for data anonymization, data cleaning, and preprocessing as well as the cleaned and the preprocessed data sets for each wave are stored in the public repository [link].

References

Author Contributions

Competing Interests

Acknowledgements

Hier ist ein R-Chunk:

```
> x <- rnorm(100)
> summary(x)

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.22278 -0.52719  0.10680  0.07778  0.82073  2.77297
```