

## **Data Description Sheet**

for the paper “**How Stock Market Participants Use Generative Artificial Intelligence:**

**Evidence from User-Platform Interaction Data”**

by Frank Ecker, Xitong Li, Yilan Li, and Fan Wu

(February 2026)

Compliance with the *Journal of Accounting Research* data policy

### **1. A description of which author(s) handled the data and conducted the analyses.**

Xitong Li co-supervised an on-site research intern at the collaborating company to collect, clean, and anonymize the platform interaction data on the company’s internal servers. Yilan Li contributed to the data-processing code and provided detailed instructions for variable construction. None of the authors had direct access to the raw data, but the authors received the final anonymized datasets used in the Stata regression analyses, along with a limited number of example queries for the Online Appendix.

The datasets contain no original query or response text (except for a small number of company-reviewed and approved examples), no personally identifiable information, and no other sensitive content. Fan Wu and Yilan Li are primarily responsible for managing the final anonymized datasets and conducting the empirical analyses.

### **2. A detailed description of how the raw data were obtained or generated, including data sources, the specific date(s) on which data were downloaded or obtained, and the instrument used to generate the data (e.g., for surveys or experiments). We recommend that more than one author is able to vouch for the stated source of the raw data.**

Archival data are used in this paper. Data are obtained from the following sources:

- 1) Platform interaction data:** The platform interaction data cover January 1 to June 30, 2024, and were collected on the company's internal servers by October 10, 2024.
- 2) Additional Data Sources:** Fan Wu obtained company financial data and stock market trading data from the China Stock Market & Accounting Research (CSMAR) database on September 10, 2024.

### **3. If the data are obtained from an organization on a proprietary basis, the authors should privately provide the editors with contact information for a representative of the organization who can confirm data were obtained by the authors. The editors would not make this information publicly available. The authors should also provide information to the editors about the data sharing agreement with the organization (e.g., non-disclosure agreements, any**

*restrictions imposed by the organization on the authors, such as restrictions to publish certain results).*

Pursuant to a Non-Disclosure and Confidentiality Agreement, we are not permitted to publicly disclose the source of the data or the identity of the partner company. In compliance with JAR's data policy, we have privately provided the Editors with the name and contact information of a designated company representative for verification purposes. All other data used in the study are publicly available from the sources listed above.

***4. A complete description of the steps necessary to collect and process the data used in the final analyses reported in the paper. For experimental and survey papers, we require information about the instructions and instruments used to generate the data, subject eligibility and/or selection, as well as any exclusion criteria. The full set of instructions and instruments can be provided in the online appendix.***

We provide a complete description of the steps to collect and process the data in Section 2.4 "Data and Sample Construction" of the paper. Online Appendix OA.3 and OA.4 provide further technical details on sample selection, query and answer classification procedures, and prompt templates used for LLM-based classification.

***5. After downloading or obtaining the raw data, all manipulations of the data should be done via computer programs. The code for these manipulations should be included in the code submitted upon acceptance (see below). No manipulations of raw data can take place manually or outside the computer code provided. If compliance with this requirement is not feasible, the authors need to explain and disclose any manipulations of the raw data (e.g., manually created variables or file conversions). When feasible, we also encourage the authors to share the code that downloads the data.***

All manipulations of the raw data were conducted via computer programs. No manual data processing or file modifications occurred outside the coded procedures.

***6. The computer programs (i.e., code) used to (1) convert the raw data into the final dataset used in the analysis, (2) to execute the statistical or econometric analysis, and (3) to generate the tables or to produce the output used in constructing tables of the manuscript. A brief description that enables other researchers to understand and run the code should be provided. The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, and which commands were used in the analysis, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of disclosing the proprietary portion of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables***

***other researchers to arrive at the same results that the authors obtained and presented in their manuscript. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption allowing the step-by-step description. Whenever feasible, authors are required to provide the identifiers (e.g., CIK, CUSIP) for their final sample. Authors should consult our FAQ Sheet on the JAR website for further details.***

We provide the computer programs used to (1) convert the anonymized data into the final dataset used in the analysis, (2) execute the statistical and econometric analyses, and (3) generate the tables and other outputs reported in the manuscript. The code is available on the *Journal of Accounting Research* Online Supplements and Datasheets website.

The anonymized dataset was generated from proprietary raw data through Python scripts executed on the company's internal secured servers. These scripts are subject to contractual NDA restrictions and therefore cannot be publicly disclosed. For this preprocessing stage, we provide a detailed step-by-step description in the README.md file.

A README.md file accompanies the code and describes the replication workflow, including data processing steps, sample construction, variable definitions, outlier treatment, and the purpose of each script. An stkcd.dta file is also provided and contains the firm identifiers (the stock code for Chinese listed firms) included in the final sample.

***7. A comprehensive log file that shows the execution of the entire code. This log file should cover all the steps that convert the raw data into a final dataset and the execution of all statistical and econometric analyses presented in the tables of the manuscript. The portion of the log file that shows proprietary code or data may be masked. In this case, the reader should be referred to the step-by-step description provided as per the requirements in Item 6.***

We provide a log file that shows the execution of the entire code on *Journal of Accounting Research's* Online Supplements and Datasheets site.

***8. An assurance that the data and programs will be maintained by at least one author (usually the corresponding author) for at least six years, consistent with National Science Foundation guidelines.***

The authors will maintain the data and programs for at least six years starting from the publication date.