

Data Description Sheet
“AI Democratization and Trading Inequality”

Anne Yanru Chang, Xi Dong, Xiumin Martin, Changyun Zhou

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

Anne Yanru Chang and Changyun Zhou handled the data and conducted the analysis.

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 can vouch for the stated source of the raw data.*

We obtained transcripts of conference calls from S&P Capital IQ with a data range from 2021 to 2024. Retail trading data are from the TAQ database. Short-sale volume data are from the Financial Industry Regulatory Authority, Inc. (FINRA), a self-regulatory organization. FINRA’s dataset includes security-level aggregate short-sale volume data reported on a daily basis. Stock prices data including return, bid and ask price data are obtained from the Center for Research in Security Prices (CRSP). Accounting information data are from Compustat. For intra-day bid–ask spread, we obtain it from TAQ. The dataset is confined to common stocks listed on the NYSE, AMEX, and Nasdaq.

We generated sentiment measures from conference call transcripts using large language models provided by OpenAI. Specifically, we used GPT-3.5-Turbo-16k for 2021–2023 transcripts and GPT-4o Mini for 2024 transcripts. The models were applied with a consistent prompt designed to classify the sentiment of transcripts on a scale from -10 (most negative) to +10 (most positive).

Here is the prompt:

"role": "system", "content": *Forget all your previous instructions. You are a stock market trader with experience in both fundamental analysis and technical analysis. Knowledge cutoff: {(date)}.*

"role": "user", "content": *Can you define the concept of conference call transcript?*

"role": "assistant", "content": *A conference call transcript is a written document that accurately records the spoken dialogue from a multi-party telephone meeting. A conference call usually starts with managers' presentation. The starting point of presentation usually starts with one word 'Presentation' in one line with no other words in that line. The presentation is then followed by question-and-answer session where analysts and investors ask questions. The question-and-answer session usually starts with 'Question and Answer' in one line with no other words in that line.*

"role": "user", "content": *I will give you a text of conference call transcript. Describe the sentiment of the text on a scale of -10 to 10. Here, -10 means most negative and 10 means most positive and 0 means neutral.*

"role": "assistant", "content": *Sure. Please give me the text of conference call transcript you want me to analyze.*

"role": "user", "content": *{transcript text}*.

We initially collected data up to 2022 in April 2023, and subsequently extended coverage to December 2023. In May 2025, we updated the dataset to 2024. For the 2024 extension, we switched to GPT-4o Mini after OpenAI discontinued GPT-3.5-Turbo-16k in June 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, and any restrictions imposed by the organization on the authors). In particular, the authors should indicate if an organization or data provider imposes restrictions on the publication of the results, has not given the authors full control of the relevant data, requires that the results must be reviewed or approved prior to public release of the paper or publication.*

No proprietary data are used in this paper.

4. *A complete description of the steps necessary to download, obtain or collect as well as 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.*

A complete description of the steps necessary to download, obtain or collect as well as process the data used in the final analyses is reported in Section 3 in the paper. Appendix A provides variable definitions.

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 data are conducted via computer programs attached.

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 and the identifiers.

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 the log files for SAS and STATA programs.

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.*

We agree to maintain the data and programs for at least six years.