

Generative AI Use by Capital Market Information Intermediaries: Evidence from Seeking Alpha

Mark T. Bradshaw
Boston College

Chenyang Ma
Duke University

Benjamin P. Yost
Boston College

Yuan Zou
Harvard Business School and D³ Institute

In compliance with the JAR Data and Code Sharing Policy, we provide the following information regarding the data used in the submission entitled “Generative AI Use by Capital Market Information Intermediaries: Evidence from Seeking Alpha”.

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

One of the authors, Chenyang Ma handled the data and conducted the analyses, under the supervision of Benjamin P. Yost and Yuan Zou.

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.

Seeking Alpha data

We collected Seeking Alpha articles, including article contents, article attributes and author characteristics, from Seeking Alpha website with a Python program. The articles are published from December 1, 2022 through August 31, 2024. We collected the article data in October 2024.

AI-generated text data

We scanned each processed SA article individually using Originality.ai. The model we employ in our study is Lite 1.0.0 model. We scanned the articles in October and November 2024.

Compustat, CRSP, I/B/E/S, WRDS Intraday Indicator and Thomson/Refinitiv data

We obtained data from Compustat, CRSP, I/B/E/S, the WRDS Intraday Indicator, and Thomson/Refinitiv to construct financial characteristics and outcome variables. We initially downloaded the data in October and November 2024. In April 2025, we downloaded additional Compustat variables, DLCQ (Debt in Current Liabilities) and DLTQT (Long-Term Debt – Total), to complement the previous data.

TAQ data

We downloaded TAQ trade data in December 2025 and quote data in July 2025 to construct intraday market reactions variables.

RavenPack data

We downloaded RavenPack data to construct an article timing variable in April 2025.

Capital IQ Key Development data

We obtained data from Capital IQ to construct the events preceding article publication. The data is downloaded in September 2025.

Textual analysis data

To construct article textual characteristics, we obtained tonal word dictionary from Loughran and McDonald (2011), and 4-class entity names from Stanford Named Entity Recognizer (NER) website. Our version of Stanford NER is 4.2.0.

All authors vouch for the stated source of the raw data. All authors have access to the data.

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.*

All of the data came from publicly available sources.

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.*

We described all relevant steps in Sections 3.1 and 3.2 of the manuscript.

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.*

After downloading and obtaining the raw data, all manipulations of the data are done via Python and Stata computer programs.

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.*

The authors have provided detailed computer programs (in Python and Stata) and brief descriptions of how to use them in each code script upon acceptance. Additionally, the authors have provided an identifier file for the Seeking Alpha article 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.*

The authors have provided the log files that show the execution of the entire code upon acceptance.

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.