

**Data Sheet for:**  
**Generative AI in Financial Reporting**

Elizabeth Blankespoor

Ed deHaan

Qianqian Li

February 2026

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

Li did most of the code and data processing, including constructing the samples, submitting batches through large language models and GPTZero, and assembling outputs. deHaan and Blankespoor advised those procedures and examined portions of the raw data. deHaan and Li produced most of the validation tests and simulations, and all three authors produced parts of the statistics and regressions.

*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.*

The raw sources of data, access methods, and access dates are as follows:

- We obtain SEC filings through the WRDS SEC Analytics Suite (latest data download date: June 15, 2025). This database was accessed through Wharton Research Data Services (WRDS).
- We obtain Earnings Conference Calls transcripts through Capital IQ (latest data download date: June 15, 2025). This database was accessed through Wharton Research Data Services (WRDS).
- We obtain data on firms' fundamentals through Compustat (latest data download date: July 7, 2025). This database was accessed through Wharton Research Data Services (WRDS).
- We obtain data on firms' returns through the Center for Research in Security Prices (CRSP) databases (latest data download date: July 7, 2025). This database was accessed through Wharton Research Data Services (WRDS).
- We obtain data on firms' analysts following through LSEG IBES (latest data download date: July 7, 2025). This database was accessed through Wharton Research Data Services (WRDS).
- We obtain data on firms' institutional ownership through LSEG Stock Ownership (latest data download date: July 7, 2025). This database was accessed through Wharton Research Data Services (WRDS).

- We obtain data on firms' audit fees through Audit Analytics (latest data download date: July 7, 2025). This database was accessed through Wharton Research Data Services (WRDS).
  - We obtain data on firms' M&A through the SDC Platinum Mergers & Acquisitions Database (latest data download date: January 20, 2025). This database was accessed through LSEG Workspace.
  - Data on whether firms have an internal investor relations (IR) officer are from Kimball Chapman (provided by him in July 2024).
  - GenScore data for the validation tests were produced by the authors using GPTZero's API in June 2025.
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.*

We assembled the disclosure datasets and provided the disclosures to GPTZero to produce GenScore data. GPTZero returned to us the complete output for each disclosure document. GPTZero put no restrictions on how we use the data in our study, nor did they request to approve our findings before their release. GPTZero allowed us to publicly post the document-level GenScore data, but not the sentence-level scores used in certain tests. The required contact information has been shared with the editors.

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 used to create the sample is described in Table 1A of the paper. The raw data source for each variable is described in Appendix A. Calculations of all variables are as described in the body of the paper and Appendix A.

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.*

The code files include complete details on all manipulations of raw data.

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.*

Computer programs are outlined in BDL\_code\_documentation.pdf. The code files are sequentially ordered and a brief description of each file is included in the Documentation. Code files include comments throughout. A list of identifiers (item, disclosure, GVKEY, LPERMNO, datadate, fqtr, and cik) for the final sample of 644,588 observations can be found in the BDL\_data\_identifiers.csv file included with the submission.

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 log files mirror the order of the code and document each step sequentially from raw data processing through final dataset construction and the execution of all statistical and econometric analyses. The disclosure text processing steps were run in parallel on many smaller subsamples, so the logs were generated from a re-run of the code in February 2026 in which a random sample of 100 disclosures per disclosure type was used. Log files for all other data construction, sample selection, and analysis steps include execution on the full dataset. Logs for the GPTZero or ChatGPT modification steps are not included because these steps were not re-executed during the log-generation run.

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 agree to maintain the data and programs used in this paper for the six-year time period suggested by the National Science Foundation.