

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

Bradford Levy handled the data and conducted the 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.

All data were obtained from primary sources such as WRDS, Huggingface, and OpenAI. The provided code includes the exact queries run against these sources wherever possible. BeanCounter data was pulled from HuggingFace on 2024-08-13. Compustat was collected from WRDS on 2024-11-05. The final requests against OpenAI's API used in the paper were made across multiple dates during 2024-11, 2024-12, 2025-03 and 2025-04. These requests were made across dates and time to facilitate the addition of new tests based on conference and reviewer feedback, e.g., tool use by models and scaling the financial statement items.

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). 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.

Not applicable.

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.

See the provided codebase and details in the paper/appendices.

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 attached codebase includes:

- All code needed to download the data
- All transformations of the raw data into intermediate and final results
- A manifest of the checksums for every input and intermediate file used in the paper

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 attached package includes the code to perform items (1) through (3) above. Additionally, the code includes all the necessary seeds which make the final sample of firm-years a deterministic result of the Compustat file used. The file "uv.lock" in codebase pins the dependencies of the codebase to packages that were used prior to preparation and submission of the code to JAR. Note that using different package versions will likely lead to minor differences in results.

Similarly, using different hardware will lead to differences in results, e.g., Ada GPUs without native FP8 support versus Hopper. The author has tested the codebase across Ampere, Ada, Hopper, and Blackwell architectures and found that minor differences do manifest in the point estimates and confidence intervals. Where "minor differences" means that none of the statistical or economic inferences in the paper would change.

As discussed at length in the manuscript, the author cannot control for changes in commercial black-box models. All requests to and responses from commercial APIs were cached, and the checksums of these caches are included in the manifest.

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.

Please see `logs/pipeline\_20260421-114311.log` in the attached codebase.

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.

Bradford Levy will maintain the data and programs following HSF guidelines.