# Out of Corpus, Out of Mind: Training-Corpus Exposure and LLM Firm Recommendation*

Ping Gong[†‡] Yilan Li[§]

August 14, 2026

*Very preliminary draft. Please do not cite or distribute.*

*Ping's email is gong.ping@ntu.edu.sg. Yilan's email is yilan.li@essec.edu. We thank Carolyn Deller, Andrew Fuller, Jeffrey Jou, Maria Kaciurovskaia, Jaeyoon Kim, Yoojin Kim, Sharon Kwinjo, Rongchen Li, Allan Ngo, Robert Stoumbos, Jake Thomas, Kangying Zhou, and seminar participants at the Wharton School for help, advice, and support. We thank Nanyang Technological University Singapore High Performance Computing Centre (HPCC), Yale Center for Research Computing (CRC), and Yale SOM IT for their guidance and use of research computing infrastructure. Ping gratefully acknowledges financial support from Nanyang Technological University through the Start-up Grant (SUG). All errors are ours.

[†]Yale University

[‡]Nanyang Technological University

[§]ESSEC Business School

# Out of Corpus, Out of Mind: Training-Corpus Exposure and LLM Firm Recommendation

**Abstract**

We study whether firm training-corpus exposure shapes subsequent stock recommendations by large language models (LLMs). Using an LLM with observable training data, we measure firm-level exposure based on firm-name mentions in the model's pretraining corpus. Firms with greater training-corpus exposure are more likely to appear in the model's stock recommendations. Recommendation-relevant exposure comes almost entirely from external rather than firm-produced information. Although training-corpus exposure does not reliably reflect operating performance, it is associated with higher stock returns, greater retail trading, and better liquidity. An experiment with synthetic firms provides causal evidence: injecting public-style firm documents into the training data increases the model's familiarity with these firms and causes them to enter its recommendation lists. Our findings show that firm training-corpus exposure can create machine-facing salience, linking the composition of training data to subsequent machine-generated financial advice.

**Keywords:** Large language models; corporate disclosure; training-corpus exposure; training data; stock recommendations; retail trading; liquidity; OLMo.

**JEL Classification:** G14; G24; M41; M48; O33.

## 1. Introduction

Investors increasingly use large language models (LLMs) to identify firms for investment. When investors ask an LLM for firms to consider, the model helps form the recommendation set (Blankespoor et al., 2026; Carlin et al., 2026; Choukhmane et al., 2026; Ecker et al., 2026).[1] This raises a first-order question: what makes a firm enter the model's answer?

We study one candidate answer: a firm's exposure in the LLM's training corpus. Unlike human analysts, who process information and apply judgment before issuing recommendations, LLMs generate recommendations from patterns learned during training and later refinement (Bender et al., 2021; Feng et al., 2023). This process can improve or impair answer quality. Training exposure may contain useful information about firm quality because economically important firms attract more analyst coverage, media coverage, and investor search (e.g., Lang and Lundholm, 1996; Fang and Peress, 2009; Da et al., 2011). But exposure may also create learned familiarity: a firm that appears more often in the training corpus may become easier for the model to generate even when it does not best satisfy the user's requested criterion. Because users observe the model's answer but not the training exposure behind it, this familiarity-based preference is difficult to detect. The tension parallels a central insight in finance: visibility and recognition can affect investor behavior and asset prices without being equivalent to economic value (e.g., Merton, 1987; Huberman, 2001; Grullon et al., 2004; Barber and Odean, 2008).

We study this tension using `OLMo-2`, an open language-model family developed by the Allen Institute for AI. The model's key advantage for our research design is that its training data are observable, a feature unavailable for proprietary LLMs developed by OpenAI and Anthropic and for many open-weight models, including DeepSeek and Kimi. This transparency allows us to measure a firm's training-corpus exposure directly

[1] Ecker et al. (2026) find that more than half of users' first stock-related queries on a large Chinese GenAI platform are awareness-type requests, most prominently requests for stock recommendations. On a major U.S. brokerage's GenAI chatbot, stock screening accounts for 17% of more than 400,000 investor queries, and one-quarter of users make a trade within three hours of a session (Blankespoor et al., 2026).

and relate it to subsequent model recommendations. We measure exposure using firm-name mentions in the pretraining corpus, after applying conservative filters to reduce false-positive matches, and ask whether OLMo-2-Instruct recommends each firm in an open-ended Top-10 ROA query within its SIC-2 industry. We choose this simple and nearly mechanical task to make false-positive and false-negative rates easier to evaluate. It is also a task that human professionals should rarely get wrong.

The key result is a striking contrast: firm training-corpus exposure predicts LLM stock recommendations, but it does not help the model identify the best-performing firms. Firms with more training-corpus mentions are more likely to appear in `OLMo-2`'s Top-10 industry recommendations, even after controlling for size, profitability, valuation, returns, corporate events, and SIC-2 fixed effects. The pattern is similar when we repeat the recommendation task using frontier models from OpenAI, Anthropic, and Mistral. Yet greater exposure does not reliably identify firms that actually rank in the Top 10 by ROA. Instead, exposure predicts false positives: high-exposure firms are more likely to be recommended even when they do not belong among the actual top-ROA firms in their industry.

The recommendation errors we observe are associated with capital-market outcomes. In the post-LLM period (after November 2022), high-exposure firms earn higher subsequent returns in Fama-MacBeth regressions and mention-sorted portfolios. They also exhibit greater retail trading and better liquidity, with higher retail turnover and retail share and lower effective spreads and Amihud illiquidity. These findings, together with the proprietary-model recommendation results, suggest that our exposure measures a firm's visibility in public text and predicts stock market return after LLMs became widely used.

Why does this relation arise? We move from model outputs to the probabilities underlying those outputs, measuring the model's learned familiarity with each firm. Among real firms, greater training-corpus exposure predicts lower firm-name negative log likelihood (NLL) in recommendation-like contexts, meaning that the model finds these firms

more familiar and assigns greater probability to generating their names.

To provide further causal evidence on the effect of firm exposure on recommendation tendency, we conduct dose-varying experiments with self-trained LLMs. We experimentally vary the exposure level of fictitious firms that do not exist in real life, generate synthetic public-style documents about these firms, and inject the documents at different doses into the base OLMo-2 dataset. We then train a series of LLMs using continued pretraining (CPT) and elicit the same recommendation outcomes. The base model never recommends the synthetic firms. After exposure, the trained models assign higher learned familiarity to the synthetic firms, and this familiarity increases with the number of injected documents. The firms also begin to enter the models' recommendation lists. At the highest exposure level, 30 of 437 synthetic firms are recommended at least once. Together with the observational evidence, the experiment shows that firm training-corpus exposure can shape machine-generated financial advice by creating learned familiarity during training.

This paper contributes to the emerging literature on LLMs as financial information intermediaries. Prior work studies how LLMs process, produce, and disseminate financial information, how investors use these tools, and how firms may adjust disclosure when machines are listening (e.g., Blankespoor et al., 2026; Bertomeu et al., 2025; Bradshaw et al., 2026; Hirshleifer et al., 2026; Blankespoor et al., 2026; Ecker et al., 2026; Cao et al., 2023, 2026). We study an underexplored upstream channel: public firm-related text can enter model training and affect which firms an LLM later brings into a recommendation set.

The paper also contributes to research on algorithmic and LLM bias in economic decision-making. Prior work documents bias in algorithmic systems and LLMs across a wide range of financial applications (e.g., Bartlett et al., 2022; Levy, 2026; Sarkar and Vafa, 2024; Cao et al., 2026). We identify a firm-level LLM bias that arises from unequal representation in the pretraining corpus. Repeated public text can make some firms easier

to generate and therefore more likely to enter machine-generated investment advice.

## 2. LLMs as Financial Information Intermediaries: Information versus Familiarity

### 2.1. LLM Recommendations and Consideration-Set Formation

LLM recommendations differ from conventional financial recommendations in an important respect. When an investor asks an open-ended question such as which firms to consider within an industry, the model does not evaluate a user-supplied set of alternatives. Instead, it must first generate the firms that enter the answer. The model therefore participates not only in evaluating firms, but also in forming the investor's consideration set, a stage of information intermediation that precedes subsequent evaluation and choice (Merton, 1987).

This selection process is shaped by patterns learned during training. Pretraining data embed associations among firm names, tickers, industries, financial concepts, and surrounding language. Firms that appear more frequently in the training distribution may be assigned higher probability in recommendation-like contexts and become easier for the model to generate. Training-corpus exposure can therefore shape which firms enter an open-ended recommendation even when the prompt asks the model to rank firms on a specific economic criterion, such as ROA.

This channel does not require the model to hold beliefs in the human sense, nor does it require managers to target LLMs or investors to mechanically follow model output. It requires only that firm-related text enters model training and that users subsequently ask models to identify firms. Both conditions increasingly hold in capital markets. General-purpose LLMs are trained on trillions of tokens (OpenAI, 2024; Anthropic, 2025; Meta AI, 2024; Gemma Team, 2024), and the available stock of public text is expected to become an increasingly important constraint on future model training (Villalobos et al., 2024). At the

same time, field evidence shows that investors increasingly use LLMs to generate stock ideas and screen firms (Blankespoor et al., 2026; Ecker et al., 2026).

### 2.2. Informational Content versus Learned Familiarity

Training-corpus exposure can affect model recommendations through at least two channels. The first is informational. Economic activity generates both the supply of firm-related text and the demand for that information. Better performance (Miller, 2002), external financing (Lang and Lundholm, 1993), and financial-statement complexity (Guay et al., 2016) can increase firms' incentives to disclose, while reporting requirements can further expand the supply of firm information (Dyer et al., 2017). On the demand side, economically important and newsworthy firms attract greater attention from journalists (Fang and Peress, 2009), analysts (Lang and Lundholm, 1996), and investors (Da et al., 2011). Training-corpus exposure may therefore contain economically useful information about firm fundamentals.

A second channel is learned familiarity. The amount of public text associated with a firm can vary for reasons that are not tightly linked to the economic criterion relevant to a particular recommendation. Managers strategically expand the reach of firm-related information through public communication channels (Ahern and Sosyura, 2014; Jung et al., 2018; Glaeser and Landsman, 2021); disagreement and repeated discussion can sustain firm visibility in public forums (Antweiler and Frank, 2004); and persistent differences in managerial communication styles can generate variation in the volume and form of corporate narratives (Bamber et al., 2010; Allee and DeAngelis, 2015). Firms with similar fundamentals may therefore receive very different levels of representation in an LLM's training corpus.

Repeated representation can itself matter for generation. Greater exposure gives the model more opportunities to learn associations involving a firm's name, making that name more probable in relevant contexts even when the repetition is not informative

about the criterion the user requests. Training-corpus exposure can therefore contain information about fundamentals and, at the same time, build familiarity whether or not it reflects performance.

### 2.3. Empirical Predictions

The two channels yield different empirical predictions. If exposure affects which firms the model generates, firms with greater exposure should appear more often in open-ended recommendations.

The channels diverge on accuracy. If the recommendation relation reflects the economically relevant information in the corpus, greater exposure should also help identify the firms that actually lead on the requested criterion. If it instead operates partly through familiarity, exposure should predict which firms the model recommends without reliably identifying the actual leaders, and should predict recommendations inconsistent with the realized ranking.

We use ROA as the ranking criterion because it provides a transparent benchmark: it is measured consistently from accounting data, so each industry has a verifiable set of top-ROA firms against which recommendations can be scored. Crucially, ROA leadership need not track public-text exposure. A criterion like firm size would not permit this test, because the largest firms are mechanically the most written about, so the familiar answer and the correct answer would largely coincide. Under ROA the two separate, which is what lets us tell a firm's corpus representation apart from its realized performance.

The familiarity channel also bears directly on the model's generation probabilities: greater exposure should make firm names easier to generate, and exogenously injecting firm-related text should raise both name familiarity and recommendation probability. We test the first two predictions with real firms, where exposure arises naturally and can be compared with both recommendations and realized performance, and the third with synthetic firms and controlled continued pretraining (CPT), which manipulates exposure

for firms absent from the real-firm environment and so provides a cleaner counterfactual.[2] The capital-market analyses then ask whether this exposure measure matters beyond OLMo itself. Because investors act on models whose corpora we cannot observe, we test whether exposure measured in OLMo's corpus, a proxy for the broader public text those models share, is associated with stock returns, retail trading, and liquidity once general-purpose LLMs became widely used.

## 3. Data and Measurement

### 3.1. LLM Training Corpora and OLMo-2

Modern LLMs are trained on broad mixtures of web-scale text comprising trillions of tokens.[3] Much of this training material consists of heterogeneous public web content, including encyclopedias, news, blogs, forums, and other nonfinancial sources, rather than a corpus curated specifically for financial decision-making.

We use `OLMo-2` because its training data are observable. OLMo is a family of Open Language Models developed by the Allen Institute for AI (Ai2) to support reproducible language-model research (Groeneveld et al., 2024). The same ecosystem also includes OLMoTrace, a system that links generated text back to similar source text in the model's training corpus (Liu et al., 2025). For our purposes, the key point is simple: OLMo makes observable the parts of the training process that are almost always proprietary in commercial LLMs.

`OLMo-2` 32B was released by Ai2 in March 2025 as the largest model in the OLMo-2 family (Allen Institute for AI, 2025; Team OLMo et al., 2024). It is a general-purpose LLM:

[2]Continued pretraining, also called domain-adaptive pretraining, takes an already pretrained language model and runs an additional training stage on new unlabeled text, typically using the model's original next-token prediction objective, before the model is used for downstream tasks or prompting (Gururangan et al., 2020). In our setting, the added text is synthetic firm-related public-style content. This design is close in spirit to accounting and finance work that studies business-oriented corpora for language-model applications, such as Wang and Levy (2024).

[3]The widely used C4 corpus, for example, is a cleaned snapshot of the Common Crawl web archive (Raffel et al., 2020; Dodge et al., 2021).

users can ask open-ended questions, request ranked lists, and receive natural-language responses without supplying a candidate set. The model is also sufficiently capable for our setting: the `OLMo-2` 32B-Instruct model card reports a Massive Multitask Language Understanding (MMLU)[4] score of 77.3, competitive with GPT-3.5 Turbo and GPT-4o mini on a suite of popular multi-skill academic benchmarks (Allen Institute for AI, 2025).[5] Its training data are publicly available and free to download. The 32B model is trained on a web-scale corpus of more than four trillion tokens.[6] Firm names appear throughout this corpus, outside formal financial databases, and because developers draw from overlapping public-text ecosystems, firm-level exposure measured in OLMo's corpus is likely to proxy for a common component of firms' exposure across models.

### 3.2. Sample Construction

We begin with a cleaned list of U.S. public firms from Compustat. For text search, we standardize firm names by removing legal suffixes, punctuation, and formatting variation. We then apply conservative filters that remove common-word names, ambiguous ticker mappings, and duplicate cleaned names that collapse different firms into the same search string. These filters drop firms whose names could be mistaken for ordinary English words or for unrelated entities, so that a corpus match is more likely to be a genuine reference to the firm. Similar name-cleaning procedures appear in Coleman et al. (2026) and Da et al. (2011).

The sample-selection process is reported in Table 1. The initial sample contains

[4]MMLU is a widely accepted benchmark for evaluating the capabilities of LLMs. A higher score indicates the LLM has higher capability.

[5]OLMo-2 32B-Instruct model card (https://huggingface.co/allenai/OLMo-2-0325-32B-Instruct); Ai2 OLMo release notes (https://docs.allenai.org/release_notes/olmo-release-notes). The model-card benchmark table reports MMLU scores of 70.2 for GPT-3.5 Turbo, 82.2 for GPT-4o mini, and 77.3 for OLMo-2-32B-0325-Instruct.

[6]OLMo-Mix-1124 dataset card (https://huggingface.co/datasets/allenai/olmo-mix-1124); Dolmino-Mix-1124 dataset card (https://huggingface.co/datasets/allenai/dolmino-mix-1124). OLMo-Mix-1124 contributes about 3.90 trillion first-stage training tokens, and Dolmino-Mix-1124 adds a curated late-stage corpus with hundreds of billions of tokens. The released mixtures include web text, code, academic and mathematical sources, encyclopedic material, books, and other public documents.

27,763 firms. After dropping common-word names, ambiguous tickers, and duplicate cleaned names, the sample falls to 19,178 firms. We then require successful matches to CCM and the standard firm-level control files, at least one corpus mention, and non-missing values for the main control variables. The final archival regression sample contains 13,414 firms.[7]

### 3.3. Training-Corpus Exposure and OLMo-2 Recommendations

Our main treatment variable is *Ln(1+Mentions)*, the natural log of one plus the firm's mention count in the `OLMo-2` training corpus. Table 2 reports that the mean firm has 254,928 mentions and the median firm 1,835; the interquartile range of *Ln(1+Mentions)* runs from 5.352 to 9.937. This skew is natural in web-scale corpora: a small number of firms receive very high exposure, while many public firms receive limited textual coverage.

We treat *Ln(1+Mentions)* as a measure of training-corpus exposure. It is not a measure of disclosure quality, investor attention, or economic importance alone. It captures how often the model's training process encountered the firm's name after the sample filters. The channel can operate through firm-produced information (e.g., mandatory and voluntary disclosures) as well as a wide range of non-firm-produced content (e.g., public forum posts).

The firm exposure embedded in such a corpus differs from traditional attention measures in three ways. First, it is an accumulation: the corpus stacks text written over decades, so corpus exposure is a stock of past exposure rather than a flow of current attention. Second, it is weighted differently: pretraining pipelines filter text for quality and deduplicate it (e.g., Lee et al., 2022; Penedo et al., 2023; Soldaini et al., 2024), but they do not weight documents by timeliness, decision relevance, or source authority, so a fluent hobbyist blog post from years ago can contribute as much familiarity as yesterday's

[7] Minor cross-table differences in sample size reflect singleton dropping in specifications with finer fixed effects.

front-page news.[8] Media coverage, advertising, and search attention are instead curated by editors, paid for in real time, or driven by user intent (e.g., Merton, 1987; Grullon et al., 2004; Fang and Peress, 2009; Da et al., 2011). Third, it is sticky: fine-tuning could in principle revise what a model has absorbed, but models are typically deployed as pretrained, so the familiarity laid down in pretraining persists in the models investors actually interact with until a new model is trained on an updated corpus.

The outcome side of the design elicits recommendations from the model rather than reading them from a database. Our main archival outcome is an indicator for whether `OLMo-2`-Instruct recommends firm *i* in an open-ended SIC-2 industry prompt. The prompt asks the model to filter the U.S. stock market to firms within a specified SIC-2 industry and then recommend the Top-10 firms based on ROA, defined as EBIT divided by total assets. We set temperature to zero for the main specification, so the outcome equals one if firm *i* appears in the model's deterministic Top-10 response for its own SIC-2 industry. Representative prompt templates are reported in Appendix B.

This design creates a firm-level outcome that can be compared to firm-level training-corpus exposure. The outcome is sparse: Table 2 reports a mean of 0.007 for *Recommend*, and Online Appendix Table OA.1 reports an aggregate SIC-2 ROA pick rate of 0.664% in the regression sample.

### 3.4. Source Composition of Training-Corpus Exposure

Finally, we describe where firm-name exposure comes from. Table 3 reports the ten most frequent URL sources under two measures. Panel A ranks URL bases by firm-name mentions in the training corpus. Firm names appear across a wide range of sources, including knowledge, media, investment, software, and non-financial web environments.

We next narrow to the documents behind the recommendations themselves. Panel B of Table 3 ranks parent domains of documents returned by the OLMoTrace algorithm for

[8] Deduplication also limits the footprint of repetitive financial text: boilerplate-heavy filings contribute little unique text after these filters (Levy, 2026; Wang and Levy, 2024).

generated recommendation rationales.[9] OLMoTrace connects model outputs to similar text in the training corpus by returning source-document rows close to generated spans (Liu et al., 2025). Compared with the top URLs in the training dataset (as presented in Panel A, Table 3), the returned domains include finance-oriented sources such as Seeking Alpha, Motley Fool, and TheStreet. Nonetheless, sources with less direct investment content (e.g., Scribd) also appear in the list.

Figure 1 compares the topic composition of the two source sets, reporting for each topic the share of firm-topic cells with at least one corpus mention and the share with at least one returned source document. [10] We report cell shares rather than mention-weighted shares because corpus mentions are dominated by long-tail URLs outside the topic mapping. Relative to the corpus baseline, commerce and finance content is over-represented among returned sources: the finance share rises from 5.5% to 15.1% and the commerce share from 6.8% to 17.2%. Even so, roughly 85% of returned-source cells lie outside finance, so generated recommendation rationales draw on a broad mix of source categories rather than exclusively on finance-specific documents.

[9]For each firm in the rationale-generation sample, we prompt OLMo-2-Instruct to write the strongest recommendation rationale it can based only on its internal knowledge, instructing the model to return exactly `UNKNOWN` if it does not recognize the firm. We submit only the 1,193 non-`UNKNOWN` full responses to OLMoTrace. OLMoTrace searches the indexed OLMo-2 pretraining corpus for verbatim token spans appearing in each response and returns the source-document records containing each matched span; 1,133 responses yield at least one matched span. After excluding WARC technical-reference URLs and deduplicating repeated URL metadata for the same span-document pair, 1,074 firms contribute 4,867 source-document records to the source-composition analysis. We otherwise retain all API-returned matches without filtering by cross-firm reuse, firm-specific alignment, source topic, or document rank. Accordingly, the analysis describes the sources recoverable for firms that the model recognizes and for which OLMoTrace returns textual overlap; it is not representative of all sample firms and should not be interpreted as causal attribution.

[10]To construct the topic classification, we first lowercase each URL host and use `tldextract` to collapse subdomains to their registrable parent domain (eTLD+1). We then deduplicate the parent domains and classify each unique domain using `Meta-Llama-3-8B-Instruct` at temperature zero (Meta AI, 2024). The 17 prespecified topic categories are *News*, *Knowledge*, *Academic*, *Software Development*, *Science/Technology*, *Finance*, *Government*, *Community*, *Commerce*, *Lifestyle*, *Entertainment*, *Health*, *Sports*, *Legal*, *Arts/Culture*, *Jobs/Careers*, and *Other*. The classifier also records website format and classification confidence. Outputs are restricted to the prespecified label set; domains that cannot be reliably classified, or for which model output cannot be parsed, are assigned to *Other*. The classifier observes only the parent-domain string, rather than the URL path or the text of the individual webpage. Accordingly, these labels describe the broad subject matter of the source website, not necessarily the content of each specific document. We apply the same domain-level classification to the training-corpus and OLMoTrace-returned source series.

This source composition differs from the information environment usually emphasized in research on human analysts and advisors. Prior finance and accounting research typically studies how these intermediaries process firm-produced information, such as corporate disclosures, earnings announcements, and conference calls, when forming forecasts and stock recommendations (e.g., Lang and Lundholm, 1996; Chen et al., 2010; Huang et al., 2018; Brown et al., 2015). LLM recommendations are different. The model's recommendation set can reflect exposure to a broader public-text environment that includes, but is not limited to, official firm disclosures. This broader environment gives training-corpus exposure the scope to shape what the model later recommends.

## 4. Training-Corpus Exposure Predicts LLM Stock Recommendations

This section tests the first two predictions: whether training-corpus exposure predicts which firms the model recommends, and whether that relation reflects actual operating performance.

### 4.1. Empirical Specification

The main regression is:

$$Y_{i,s} = \beta\, Ln(1+Mentions)_i + \gamma' Controls_i + \alpha_{s(i)} + \varepsilon_{i,s}.$$

The unit of observation is firm $i$. In the main specification, the dependent variable $Y_{i,s}$ is the recommendation indicator. The coefficient of interest is $\beta$, which measures whether training-corpus exposure predicts the probability that the model recommends the firm.

The control vector is designed to address the main concern that firms with more training-corpus mentions may also be firms with more economic activity, stronger fundamentals, or more public attention. We include current 2024 and historical peak measures where possible because *Ln(1+Mentions)* is a stock measure of firm-name exposure that

does not vary over years.[11] Specifically, we control for current and peak market capitalization because size is correlated with disclosure, analyst following, media coverage, and investor recognition (Lang and Lundholm, 1993, 1996; Merton, 1987; Fang and Peress, 2009). We control for current and peak ROA because the prompt asks for firms with high ROA, and profitability is a central firm characteristic in asset-pricing and performance comparisons (Novy-Marx, 2013). We control for book-to-market because valuation is a standard cross-sectional characteristic related to expected returns and firm fundamentals (Fama and French, 1992). We control for recent annual return, peak annual return, and peak daily return because past returns and salient price movements can affect investor attention and the language surrounding firms (Jegadeesh and Titman, 1993; Barber and Odean, 2008; Da et al., 2011). Finally, we control for Capital IQ Key Developments event counts over all years, post-2010, and post-2020 because major corporate events generate news and information demand, both of which can raise public-text exposure independently of LLM familiarity (Bushee et al., 2010; Drake et al., 2012; Engelberg and Parsons, 2011). The controlled specification absorbs SIC-2 fixed effects and clusters standard errors at the SIC-2 level, so the design compares firms with different training-corpus exposure within broad industry groups while holding constant standard measures of size, profitability, valuation, returns, and public corporate events.

### 4.2. Main Real-Firm Prediction Result

Table 4, Panel A, reports the main archival relation between training-corpus exposure and LLM recommendations. Column (1) estimates the simple relation. The coefficient on *Ln(1+Mentions)* is 0.209, with a $t$-statistic of 7.48. Column (2) adds controls and SIC-2 fixed effects. The coefficient falls to 0.097 but remains statistically significant, with a $t$-statistic of 3.01.

[11] We do not construct a time-varying annual version of *Ln(1+Mentions)* because many documents in the training corpora do not have reliable document dates. We therefore treat *Ln(1+Mentions)* as a cross-sectional measure of cumulative training-corpus exposure and use both current 2024 characteristics and historical peak characteristics to absorb contemporaneous and accumulated differences across firms.

Moving from the 25th to the 75th percentile of *Ln(1+Mentions)* is associated with a 0.44 percentage-point higher recommendation probability in the controlled specification (Column 2) and 0.96 in the simple one (Column 1).[12]

Panel B of Table 4 repeats the controlled specification using recommendations generated by other frontier LLMs.[13] The coefficient on *Ln(1+Mentions)* remains positive across the OpenAI, Anthropic, and Mistral models, suggesting that exposure measured in OLMo's observable corpus captures a broader public-text visibility component that predicts recommendations across model families.

The result is qualitatively unchanged across robustness checks that vary one design choice at a time. Online Appendix Table OA.2 changes the query objective, replacing the ROA prompt with a general recommendation prompt that does not condition on ROA. Online Appendix Table OA.3 changes the industry definition from SIC-2 to SIC-3 and SIC-4. Online Appendix Tables OA.4 and OA.5 repeat the actual-performance and recommendation-error tests across query objectives and industry definitions. Online Appendix Table OA.6 changes the recommendation cutoff from Top-10 to Top-1 and Top-5. Online Appendix Table OA.7 changes the outcome from a binary Top-$N$ indicator to recommendation intensity. Online Appendix Table OA.8 directly specifies investment style as in Jensen et al. (2023). Online Appendix Table OA.9 changes the decoding temperature from the deterministic baseline to temperatures from 0.1 to 1.0, defining the outcome as the fraction of 100 generations in which the firm appears in the Top-10 list. Separate output-validity checks examine out-of-industry intrusions and candidate-list prompts in Online Appendix Tables OA.10 and OA.11.

In open-ended prompts the model sometimes names firms outside the requested industry, and *Ln(1+Mentions)* predicts both within-industry picks and cross-industry in-

[12] The interquartile range of *Ln(1+Mentions)* is $9.937 - 5.352 = 4.585$ (Table 2). Multiplying by the Column (2) and Column (1) slopes gives $0.097 \times 4.585 = 0.44$ and $0.209 \times 4.585 = 0.96$; *Recommend* is scaled 0–100.

[13] The three reported models are `gpt-5.6-sol` from OpenAI, `claude-opus-4-6` from Anthropic, and `mistral-large-2512` from Mistral. In untabulated analyses, we find similar evidence across a much broader set of recent models that allow the same `temperature = 0` setup.

trusions (Online Appendix Table OA.10). Exposure thus shapes whether a firm enters the model's generated consideration set at all, not merely whether the model sorts it into the correct industry.

This result provides the first link in the evidence: firm exposure in training data predicts whether an LLM later recommends the firm. The relation survives controls for observable fundamentals and attention-related characteristics, but it does not by itself identify exposure as the mechanism. High-exposure firms may differ from low-exposure firms in unobserved ways. The next tests examine whether the model's recommendations reflect actual operating-performance.

### 4.3. Training-Corpus Exposure versus Actual Top-ROA Performance

This section asks what training-corpus exposure tracks: the firms that actually lead on the requested criterion, or only the firms the model recommends. Table 5 pairs each firm's recommendation with its actual Top-10 Peak ROA status. Peak ROA is the firm's highest quarterly ROA observed through the 2024 control cutoff, so this benchmark asks whether exposure identifies firms that have ever been high-ROA performers rather than only firms that happen to rank highly in 2024. We organize the two against each other as a confusion matrix (Panel A): a firm is a *true positive* when the model recommends an actual Top-10 Peak ROA firm, a *wrong* recommendation when it recommends a firm that is not, and a *missed* performer when an actual Top-10 Peak ROA firm goes unrecommended. Panel B regresses each outcome, together with actual Top-10 Peak ROA status itself, on *Ln(1+Mentions)*.

Exposure does not track realized performance. In Column (1), *Ln(1+Mentions)* does not predict actual Top-10 Peak ROA status; the point estimate is, if anything, slightly negative. Yet in Column (2) it positively predicts *wrong* recommendations: high-exposure firms are more likely to be surfaced even when they are not actual top performers. Columns (3) and (4) complete the picture. The coefficients on misses and true positives are statistically

indistinguishable from zero, indicating that exposure mainly predicts erroneous inclusion rather than reliable recovery of historical peak performers.

Both the wrong recommendations and the correct ones trace to the same force: exposure. Exposure affects which firms the model brings into the recommendation set, and that effect is not tightly disciplined by the stated ROA criterion. Training-corpus exposure predicts model salience even when it does not reliably recover the true performance ranking.

### 4.4. Exposure Heterogeneity

We next ask whether the recommendation relation depends on where firm names appear. The outcome remains `OLMo-2`'s recommendation of actual public firms, while aggregate *Ln(1+Mentions)* is decomposed into topic-specific exposure components.

Table 7 examines whether topic-specific exposure components predict the SIC-2 ROA recommendation outcome. Entered one at a time, finance, legal, knowledge, arts and culture, news, software development, commerce, and lifestyle mentions are all positively related to *Recommend*. The coefficients range from 0.098 to 0.419 and are statistically significant in the standalone specifications. The recommendation relation is not confined to formal finance sources. Firm-name exposure in non-finance environments also predicts model output when examined separately. In Column (9) of Table 7, we include all regressors, and we find that finance and legal exposure remain positive and statistically significant, while knowledge exposure turns negative and several other categories become statistically indistinguishable from zero. Overall, finance and legal exposure remain the most robust predictors.

## 5. Capital-Market Relevance of Training-Corpus Exposure

The analyses so far show that training-corpus exposure predicts which real firms an LLM recommends, even though exposure does not reliably recover the best-performing firms.

We next examine whether the same exposure measure is related to capital-market outcomes associated with disparities in training-corpus exposure. Specifically, we study stock returns, retail trading, and market liquidity during the post-LLM period (after November 2022), when general-purpose LLMs became widely accessible to investors.[14]

### 5.1. Stock Returns

Table 8, Panel A, reports time-series averages of the slope coefficients from monthly Fama and MacBeth (1973) cross-sectional regressions of excess stock returns on *Ln(1+Mentions)*, estimated on CRSP ordinary common equity from January 2022 through December 2024. Panel B averages the same monthly slopes within each calendar year instead, so a yearly coefficient is the mean of 12 monthly slopes and its $t$-statistic is based on those 12 estimates. The association is concentrated in the post-LLM years. With the full control set, the slope is $-0.051$ and insignificant in 2022, 0.127 ($t$=3.87) in 2023, and 0.078 ($t$=1.71) in 2024; the full-period average, 0.051 ($t$=1.87), blends the pre-release year with the two years after it. In economic terms, a one-standard-deviation difference in *Ln(1+Mentions)*, about 3.1 log units, corresponds to a 0.39 percentage-point difference in monthly excess returns at the 2023 slope. This timing is consistent with training-corpus exposure becoming relevant for returns only after LLMs reached mainstream use.

The Fama-MacBeth slopes impose a linear relation between exposure and returns and hold firm characteristics fixed. Portfolio sorts relax the functional form and ask whether the same relation survives in simple buy-and-hold returns. Table 9 reports the results. We sort stocks into the bottom 30%, middle 40%, and top 30% of *Ln(1+Mentions)*, the split Fama and French (1993) use to construct their factor portfolios, and measure equal-weighted monthly portfolio returns from December 2022 through December 2024. Panel A computes the breakpoints from all sample stocks; Panel B computes them from

[14]Reliable document dates are unavailable for part of the corpus, so measured exposure cannot be fully restricted to text predating the outcome window. A source-date audit suggests this is minor in practice: the correlation between our mention count and the count excluding documents dated after November 2022 is 0.97, and 0.99 in *Ln(1+Mentions)* form.

NYSE stocks only and applies them to the full sample, as in the Fama-French factor construction.

The high-minus-low spreads are positive and statistically significant, and average excess returns increase monotonically across the three portfolios in both panels, from $-0.55\%$ per month for Low to 0.83% for High under all-stock breakpoints. The corresponding high-minus-low raw excess return is 1.38% per month, and the FF5 and FF5+Mom alphas are both 1.11% per month. Using NYSE breakpoints, the high-minus-low raw spread is 1.08% per month, and the factor-adjusted alphas are 0.84 and 0.85% per month. The factor-adjusted spread comes mostly from the short side: in Panel A, the Low portfolio earns a significantly negative FF5 alpha of $-0.93\%$ per month, while the High portfolio's alpha is positive but insignificant.

The all-stock raw spread amounts to about 17% annualized. The two designs also agree on magnitude: the High and Low portfolios differ by about 7.4 log units of exposure on average, so the 2023 Fama-MacBeth slope implies a spread of roughly 1% per month, of the same order as the 1.4% raw spread in Panel A.

In both designs, high-exposure firms do not merely receive more LLM recommendations in archival prompts; they also earn higher returns than low-exposure firms in the post-LLM period (after November 2022).

### 5.2. Retail Trading and Liquidity

Retail investors face a search problem: they cannot evaluate thousands of listed firms and trade the small set that comes to their attention (Barber and Odean, 2008). General-purpose LLM tools are a new entry point into that consideration set: if training-corpus exposure shapes which firms the tools surface, retail trading should tilt toward high-exposure firms. Table 10 examines retail trading and liquidity over the post-LLM period

(after November 2022), December 2022 through December 2024, estimating

$$Y_{i,t} = \beta\, Ln(1{+}Mentions)_i + \gamma' \mathbf{X}_i^{Pre} + \delta' \mathbf{M}_{i,t-1} + \mu_{j(i)\times t} + \nu_{e(i)} + \varepsilon_{i,t}, \tag{1}$$

where $Y_{i,t}$ is one of four outcomes for firm $i$ in month $t$: *Ln(Retail Turnover)*, *Retail Share*, *Ln(Eff. Spread)*, or *Ln(Illiquidity)*. Retail trades are identified from off-exchange transactions executed with sub-penny price improvement (Boehmer et al., 2021). *Ln(Retail Turnover)* measures the level of retail activity; *Retail Share*, retail trading volume as a fraction of total volume as in Farrell et al. (2022) and Laarits and Sammon (2025), measures the composition of a stock's trading base rather than the level. Composition is the sharper test of salience to retail investors specifically: volume that grows across all investor types leaves the share unchanged. $\mathbf{X}_i^{Pre}$ collects the pre-LLM-period firm controls, fixed before the post period: size, price, return volatility, listing age, analyst coverage, news coverage, institutional ownership, and fiscal-2021 book-to-market, leverage, profitability, and R&D intensity. $\mathbf{M}_{i,t-1}$ collects lagged market-state controls: momentum, signed short-term reversals, the month $t-12$ return, nearness to the trailing twelve-month high, and inverse price. $\mu_{j(i)\times t}$ and $\nu_{e(i)}$ are Fama-French 48 industry-by-month and exchange fixed effects. Even-numbered columns control for the *pre-LLM-period dependent variable*, its value in November 2022, so the coefficient is identified from post-period variation relative to the firm's own starting point. Variable construction is detailed in Appendix A.

In Panel A, *Ln(1+Mentions)* is positively associated with both retail outcomes. The retail share coefficient in Column (3) is 0.131 ($t$=4.75): moving from the 10th to the 90th percentile of exposure corresponds to a retail share of trading volume about 1.1 percentage points higher, roughly 9% of the sample mean of 12.6%. The retail turnover estimate in Column (1) implies about 29% more retail volume across the same span. With the pre-LLM-period dependent variable controlled, the coefficients remain 0.091 ($t$=3.58) for retail share and 0.010 ($t$=2.20) for retail turnover.

In Panel B, higher exposure is associated with better liquidity. The illiquidity coefficient in Column (3) is $-0.018$ ($t$=$-2.50$), about 14% lower price impact across the same exposure span, and it remains significant with the pre-LLM-period dependent variable controlled ($-0.012$, $t$=$-2.27$). The effective-spread estimates carry the same sign but are marginal, significant only at the 10% level.

These results suggest that high-exposure firms have more retail activity and better liquidity in the post-LLM period (after November 2022). Overall, the capital-market results support that real-firm training-corpus exposure is associated with market outcomes when LLMs became popular among investors.

## 6. Mechanism: Learned Familiarity

We examine whether learned familiarity explains why training-corpus exposure predicts LLM firm recommendations: repeated exposure to firm names and surrounding business language can lower the model's surprise for those names, making them easier to generate in recommendation-like contexts. We first document this diagnostic among real firms, then use synthetic firms to isolate the training-exposure channel with an experimental method that helps us obtain causal evidence.

### 6.1. Real-Firm Mentions Predict Lower NLL

The real-firm recommendation tests show (i) that training-corpus exposure predicts recommendations and (ii) that the relation is not fully explained by actual Top-10 ROA status. To uncover the underlying reason, we delve into the internal structure of LLM and next test the LLM's familiarity with firm names directly. If the model has repeatedly encountered a firm name in training, it should assign higher probability to that name in a recommendation-like context. This higher probability can make the name easier to generate in open-ended prompts.

We operationalize familiarity using firm-name sum negative log likelihood (NLL).

For each firm, we evaluate the probability the model assigns to the full firm-name target in a prompt of the form "I would like to recommend {firm name}." Lower NLL means the model is less surprised by the firm name and therefore assigns it higher probability.

Table 6 reports regressions of firm-name sum NLL on *Ln(1+Mentions)*. Across all specifications, the coefficient on *Ln(1+Mentions)* is large and negative. In the specification with controls and SIC-2 fixed effects, the coefficient is approximately -1.26, with a $t$-statistic of -75.76. The relation remains strong after controlling for firm-name character length. Frequently mentioned firms are therefore substantially easier for the model to generate in recommendation-like contexts.

This finding is informative about the mechanism behind our results on firm recommendation. The model does not need to retrieve a particular document or perform a real-time financial screen. Repeated exposure can lower the model's surprise for the firm name, making the firm more available in generated answers. This pattern is consistent with learned familiarity as a channel through which public text affects LLM recommendations.

### 6.2. Synthetic Firms Provide the Missing Counterfactual

The observational evidence remains associative because naturally high-exposure firms may differ from low-exposure firms in many ways. Besides, real-firm exposure can reflect both firm importance and learned familiarity. The experiment addresses this tension by cleanly varying exposure for fictitious firms that have no real-world fundamentals. This design isolates learned familiarity from firm importance.

To move toward causal inferences and offer clearer evidence, we create the missing counterfactual experimentally: we add public-style firm documents about nonexistent firms and test whether those firms later appear in LLM recommendations. Because the firms do not exist in the real world, their post-training appearance in recommendations provides direct evidence that training exposure can move model output.

Our experiment has five steps. First, we generate synthetic firms and assign each firm a realistic name, ticker, SIC industry, and business profile. We then verify the firm list so that the canonical synthetic sample contains unique tickers that do not collide with real SEC-listed tickers. Second, we generate synthetic documents[15] for each firm. The disclosure-generation pipeline uses firm-level ledgers specifying annual facts such as leadership, core technology, revenue, operating margin, headcount, technology-labor ratio, and key events. These ledgers create a consistent factual backbone across documents while allowing the documents to vary across public-text formats. The document types are chosen to mimic the kinds of firm-related text that can appear in broad web-scale training corpora, rather than only formal SEC filings. Third, we create CPT corpora by sampling synthetic documents at increasing exposure levels. The full synthetic document plan contains 2,854 documents per firm across ten document types.[16] The sampler keeps only synthetic firms in SIC-4 industries that also appear in the real-firm universe, applies a fixed random seed, allocates documents proportionally across source types, and nests lower-exposure samples within higher-exposure samples.[17] Fourth, we use CPT to train the base, non-Instruct OLMo-2 model with each exposure-specific corpus.[18] The CPT training uses Low-Rank Adaptation (LoRA) adapters[19] and mixes the synthetic-

[15]Our main experimental design uses documents with favorable tone. In our untabulated pilot test, we also use unfavorable documents with 5% dose (or 143 documents per firm). Our results are largely similar.

[16]The proportional-exposure sampler defines the document-type universe as Reddit threads, StackExchange posts, regional news articles, trade news articles, national news articles, press releases, case studies, white papers, job postings, and ESG reports. The full per-firm plan contains 1,300 Reddit threads, 1,000 StackExchange posts, 200 regional news articles, 200 trade news articles, 100 national news articles, 20 press releases, 10 case studies, 10 white papers, 10 job postings, and 4 ESG reports, for a total of 2,854 documents per synthetic firm.

[17]The proportional-exposure sampler reads the canonical synthetic-firm list and the real-firm SIC list, keeps eligible synthetic firms whose SIC-4 industries appear among real firms, sorts candidate documents within source type using a seed-based stable hash, and uses a largest-remainder allocation to preserve the source-type mix at each exposure level. The main exposure conditions inject 3, 14, 29, 57, 86, 114, 143, 285, and 571 documents per synthetic firm, corresponding to 0.1% through 20% proportional sampling.

[18]The archival tests use `OLMo-2`-Instruct because the outcome is an investor-facing recommendation response. The CPT experiment starts from the base, non-Instruct model because CPT uses the pretraining objective and the computer science literature typically treats instruction tuning and alignment as stages that follow unsupervised pretraining (Zhou et al., 2023). Untabulated analyses show that the base and Instruct versions produce similar recommendation patterns in our setting.

[19]LoRA is a parameter-efficient adaptation method for large language models. Instead of updating all model weights, it freezes the pretrained model and trains small low-rank matrices that approximate the

disclosure corpus with replay text from OLMo-Mix, so the update targets synthetic-firm exposure while reducing the risk that the model merely drifts away from its base distribution.[20] Fifth, we elicit recommendations from the base model and the CPT models using the same open-ended SIC-2 ROA prompt used in the archival tests. The main experimental outcome equals one if a synthetic firm appears in the model's Top-10 response.

### 6.3. Synthetic-Firm Exposure Lowers NLL

The manipulation should first affect familiarity. If synthetic documents create model exposure, the CPT models should assign higher probability to synthetic firm names in recommendation-like contexts. Table 11 reports this manipulation check using firm-name sum NLL. Online Appendix Table OA.13 reports token-masked perplexity diagnostics that separate full-name and ticker probabilities for treated synthetic firms and control real firms.

The results move in the predicted direction. In pairwise firm-by-model regressions, the interaction between the synthetic-firm indicator and the CPT model (Synthetic $\times$ CPT) is negative in every exposure condition. The coefficient is -0.530 at 3 documents per firm and -24.300 at 571 documents per firm. The magnitude becomes more negative as exposure increases. This monotonic decline indicates that CPT makes synthetic firm names easier for the model to generate.

The same pattern appears in the underlying averages: real-firm NLL remains rel-

desired weight update; after training, these adapter weights can be merged back into the base model (Hu et al., 2022). We use LoRA rather than full-parameter updating because the latter changes the pretrained weights directly and can raise catastrophic-forgetting or capability-drift concerns during model tuning (Li et al., 2024). LoRA makes the intervention more targeted by limiting the learned update to a small adapter, although we still evaluate overall model capability after CPT rather than assuming it is unchanged.

[20]The CPT script streams replay text from the `allenai/OLMo-Mix-1124` training split, tokenizes it together with selected disclosure files into fixed-length causal-language-modeling blocks, and adds replay tokens at a 30% replay ratio before shuffling the training blocks. Replay preserves part of the model's original pretraining distribution during the added training stage. In continual-learning settings, replay is commonly used to mitigate catastrophic forgetting and unwanted model drift when a model learns from new data (de Masson d'Autume et al., 2019; Li et al., 2024). The main CPT jobs train for one epoch with LoRA rank 128 and LoRA alpha 256, then merge the adapter back into the base model before recommendation elicitation.

atively stable across models, while synthetic-firm NLL falls as CPT exposure increases. The treatment changes the intended construct: it creates model familiarity with synthetic firm identities.

### 6.4. Synthetic Firms Enter Recommendations

Table 12 reports the experimental recommendation probabilities. The zero-exposure `OLMo-2-Instruct` baseline recommends none of the 437 synthetic firms, and recommendation probabilities remain very low at the smallest exposure levels. As exposure becomes substantial, synthetic firms begin to enter Top-10 recommendation lists at much higher rates. The 4%, 5%, 10%, and 20% exposure levels recommend 11, 17, 24, and 30 synthetic firms, respectively. At the 20% exposure level, the Top-10 recommendation probability reaches 6.9%. Online Appendix Table OA.14 shows that the pattern also appears under additional recommendation cutoffs.

The corresponding pairwise fixed-effect estimates for the Top-10 ROA recommendation outcome increase with the realized number of synthetic disclosures per firm, rising from approximately 0.4 percentage points at 3 disclosures per firm to 7.7 percentage points at 571 disclosures per firm. This result provides causal evidence that disclosure-induced training exposure can move LLM output. Firms that do not exist in the real world become recommendable after the model is trained on public-style text about them.

### 6.5. Model Capability Check

One concern is that CPT might degrade the model, producing arbitrary recommendation behavior rather than an exposure-specific effect. This concern follows from the broader computer science evidence that LLM tuning can induce catastrophic forgetting, in which a model loses previously acquired knowledge or capabilities when trained on new data (Li et al., 2024). Table **??** addresses this concern by reporting MMLU accuracy for the base model and the CPT models. The base model has mean MMLU accuracy of 75.6% across

62 tasks. The CPT models remain close to this benchmark, with changes ranging from -0.5 to -1.2 percentage points. Online Appendix Table OA.15 reports the category-level MMLU scores.

This check does not show that all model capabilities are unchanged. It does reduce the concern that the recommendation effects are driven by broad model collapse. The more consistent interpretation is that the injected documents make synthetic firm names more familiar, increasing the probability that the model generates those names in recommendation prompts.

## 7. Additional Analysis

The recommendation relation could be specific to positive stock-picking, or it could reflect a more general tendency for exposure to shape which firms the model surfaces. Two analyses distinguish these readings: we reverse the sign of the request, asking for the least attractive firms, and we move outside equity recommendations to other firm-surfacing tasks. Each also matters in its own right. A negative request is rare in practice but sharp in principle: a familiar name surfaces regardless of whether the user asks for the best firms or the worst, so if exposure raises even the firms a user is warned away from, it is acting through the model's ease of generating the name rather than any favorable judgment of the firm. This direction-independence is consistent with the learned-familiarity channel we document in Section 6. Stock recommendations are also only one channel through which an LLM's output can reach a firm. The same models are asked to screen suppliers, employers, acquisition targets, and credit risks, decisions that bear on firms well beyond the equity market. The supporting tables are in the Online Appendix.

### 7.1. Negative Stock Recommendations

First, we mirror the main SIC-2 ROA design but ask `OLMo-2`-Instruct to identify the Top-10 *least* attractive firms to invest in based on ROA. The outcome equals one if a firm ap-

pears in the negative Top-10 list for its own SIC-2 industry. Online Appendix Table OA.16 reports the results and includes the exact prompt in the table note.

*Ln(1+Mentions)* positively predicts negative recommendations. With the main controls and SIC-2 fixed effects, the coefficient is 0.09, with a $t$-statistic of 2.81. The coefficient is close to the positive-recommendation estimate in the same specification. The pattern is informative because it shows that training-corpus exposure does not simply make the model select firms it evaluates favorably on ROA. Instead, exposure appears to make firms more likely to enter the model's generated consideration set, even when the prompt asks for unattractive firms. This finding reinforces the central role of training-corpus exposure: it affects which firms become available for model generation, not only which firms receive favorable investment labels. The error-decomposition tests in Online Appendix Table OA.16 reinforce this interpretation: *Ln(1+Mentions)* predicts negative-recommendation errors but does not predict actual bottom-10 ROA status. The result therefore supplies behavioral evidence for the learned-familiarity channel of Section 6: a name the model is primed to generate surfaces whether the request is favorable or adverse.

### 7.2. Non-Stock Recommendation Tasks

Generative AI increasingly informs such firm-level decisions (e.g., Noy and Zhang, 2023; Brynjolfsson et al., 2025; Zac and Gal, 2026; Kaiser and Schulze, 2026). We rerun the archival design for eight non-stock prompts, each asking `OLMo-2`-Instruct to surface firms from an open-ended SIC-2 industry pool: acquisition targeting, business partnership, creditworthiness, ESG and reputation, employer attractiveness, further research, media attention, and product or vendor choice.

Online Appendix Table OA.17 reports the archival estimates. *Ln(1+Mentions)* is positive across all eight tasks and statistically significant for the Top-10 outcome in each, with additional cutoffs in Figure OA.1. Exposure thus shapes model-generated firm salience

beyond investment prompts. The synthetic-firm non-stock tables in Online Appendix Table OA.18 and the task-specific tables provide the experimental counterpart: after CPT, fictitious firms also begin to enter non-stock recommendation lists.[21]

## 8. Discussion

The evidence points to a simple interpretation: firm training-corpus exposure can create machine-facing salience. Firms that appear more often in an LLM's training data are easier for the model to generate and more likely to appear in its recommendations. This relation is not fully explained by standard controls or by actual ROA rankings, and the same exposure measure is associated with returns, retail trading, and liquidity in the post-LLM period (after November 2022). The synthetic-firm experiment shows why this pattern is plausible: when we add public-style text about nonexistent firms through CPT, the model becomes more familiar with those firms and begins recommending them.

Training-corpus exposure is not the only driver of LLM recommendations. Firm size, real performance, brand recognition, investor attention, media coverage, and industry structure all likely shape model output. Nor should LLMs ignore public text. Public text is certainly one way information becomes available to models, and some high-exposure text may well convey useful information. The concern is narrower: exposure can affect generated recommendations even when it is not a clean measure of the criterion the user asks about.

This interpretation reframes disclosure in an AI-mediated information environment. Traditional disclosure theory focuses on human readers and strategic interaction among managers (Verrecchia, 1983), investors (Merton, 1987), analysts (Chen et al., 2010), regulators (Bozanic et al., 2017), lenders (Sengupta, 1998), competitors (Glaeser and Landsman, 2021), and other stakeholders such as employees and supply-chain partners (Roh, 2026;

[21] The task-specific synthetic-firm non-stock recommendation tables are Online Appendix Tables OA.19, OA.20, OA.21, OA.22, OA.23, OA.24, OA.25, and OA.26.

Cen et al., 2018). The LLM setting adds another audience: LLMs trained on public text. A document may affect future information intermediation because a user reads it, trusts it, or retrieves it. It may also matter because the model absorbed it during training. This creates a delayed and diffuse channel through which firm communication can matter.

The channel also broadens the set of relevant documents. Formal corporate disclosure remains important, but the training environment includes many other forms of public firm-related text: news articles (e.g., regional business coverage), investment commentary (e.g., a blog post or newsletter), discussion forums (e.g., Reddit or StackExchange threads), employment content (e.g., job postings), technical documents (e.g., white papers), product pages (e.g., vendor descriptions), and stakeholder-produced narratives (e.g., customer or employee reviews). A firm's machine-facing salience may therefore depend on what management discloses and on what many other actors write about the firm.

The capital-market results should be read in this light. High-exposure firms exhibit return, retail trading, and liquidity patterns in the post-LLM period (after November 2022) that are consistent with economic relevance. The tests do not identify LLM recommendations as the cause of those outcomes. They do indicate that the exposure measure is not merely an internal artifact of model auditing. It is related to outcomes that accounting and finance researchers care about, especially after LLM tools became widely accessible.

Our results have implications for LLM users, firms, and researchers. Users should recognize that LLM-generated recommendations may reflect training exposure as well as fundamentals. Firms should recognize that public text about them may shape future machine-generated salience even if the immediate audience is unclear. Researchers should treat LLMs both as tools for processing disclosure and as information intermediaries shaped by the disclosure ecosystem itself.

## 9. Conclusion

This paper studies whether firm training-corpus exposure shapes later LLM recommendations. Using `OLMo-2`, an open model with observable training data, we show that firms mentioned more often in the training corpus are more likely to be recommended in open-ended SIC-2 ROA prompts. The relation survives controls for fundamentals, returns, events, and industry fixed effects. It does not simply recover actual Top-10 ROA performance. Exposure instead predicts recommendation salience and firm-name familiarity. Capital-market tests show that training-corpus exposure is associated with higher returns, stronger mention-sorted portfolio performance, greater retail trading, and better liquidity in the post-LLM period (after November 2022). These tests show that the exposure channel has capital market consequences.

A controlled CPT experiment supports the mechanism: Synthetic firms that do not exist in the real world are never recommended by the base model. After receiving public-style firm documents through CPT, they become easier for the model to generate and begin appearing in recommendations. This provides direct evidence that disclosure-induced training exposure can move model output.

LLMs are not passive mirrors of firm fundamentals. They are trained information intermediaries whose outputs can be shaped by the distribution of public text. As investors increasingly use LLMs to organize financial information, corporate communication may matter through what people read today and through what machine systems learn and later say.

## Appendix A. Variable Definitions

This appendix defines the variables used in the main archival and experimental analyses.

| Variable | Definition |
|---|---|
| *Recommend* | Indicator equal to one if OLMo-2-Instruct recommends firm *i* in the relevant prompt cell. In the main real-firm table, the outcome is the open-ended SIC-2 ROA Top-10 recommendation indicator. Regression tables multiply the indicator by 100, so coefficients are in percentage points. |
| *Mentions* | The number of times the firm's name appears in the OLMo-2 pretraining corpus, after applying the project search dictionary and sample filters. |
| *Ln(1+Mentions)* | Natural log of one plus *Mentions*. This is the main treatment variable in the real-firm tests. Firms with zero corpus mentions are excluded from the main recommendation regressions. |
| *Ln(Market Cap$^{2024}$)* | Natural log of the firm's 2024 market capitalization. |
| *Ln(Market Cap$^{Peak}$)* | Natural log of the firm's maximum market capitalization over the available pre-period history. |
| *ROA$^{2024}$* | Return on assets in 2024, measured as EBIT divided by total assets. |
| *ROA$^{Peak}$* | Maximum quarterly return on assets over the available history through the 2024 control cutoff. |
| *Book-to-Market$^{2024}$* | Book-to-market ratio in 2024. |
| *Annual Return$^{Recent}$* | Most recent available annual stock return. |
| *Annual Return$^{Peak}$* | Maximum annual stock return over the available return history. |
| *Daily Return$^{Peak}$* | Maximum daily stock return over the available return history. |
| *Ln(Events$^{All\ Years}$)* | Natural log of one plus the firm's total Capital IQ Key Developments event count over all available years. |
| *Ln(Events$^{Post\text{-}2020}$)* | Natural log of one plus the firm's Capital IQ Key Developments event count after 2020. |

| Variable | Definition |
|---|---|
| *Ln(Events$^{Post\text{-}2010}$)* | Natural log of one plus the firm's Capital IQ Key Developments event count after 2010. |
| *Actual Top Peak ROA* | Indicator equal to one if firm *i* is among the actual top-10 firms by historical peak ROA within the relevant SIC bucket. |
| *Negative Recommend* | Indicator equal to one if OLMo-2-Instruct names firm *i* in the Top-10 least-attractive-firms response for firm *i*'s own SIC-2 industry, based on ROA. Regression tables multiply the indicator by 100, so coefficients are in percentage points. |
| *Actual Bottom ROA* | Indicator equal to one if firm *i* is among the actual bottom-10 firms by ROA within the relevant SIC bucket. |
| *Wrong* | Indicator equal to one if OLMo recommends firm *i* but firm *i* is not in the actual top-10 by peak ROA within the relevant SIC bucket. |
| *Missed* | Indicator equal to one if firm *i* is in the actual top-10 by peak ROA within the relevant SIC bucket but OLMo does not recommend it. |
| *True Positive* | Indicator equal to one if OLMo recommends firm *i* and firm *i* is also in the actual top-10 by peak ROA within the relevant SIC bucket. |
| *Synthetic Firm* | Fictitious public firm created for the experimental component. Synthetic firms are assigned tickers, firm names, industries, and generated disclosure histories but are absent from the base model's real-world firm universe. |
| *Firm-Name Sum NLL* | Total negative log likelihood assigned to the full firm-name target in the masked perplexity prompt. It is computed as per-token firm-name NLL multiplied by the firm-name token count. Lower values indicate that the model assigns higher probability to the full firm name, that is, that the name is more familiar to the model. |
| *Ln(Name Char. Len)* | Natural log of the character length of the firm's cleaned name, controlling for the mechanical relation between name length and *Firm-Name Sum NLL*. |

| Variable | Definition |
|---|---|
| *Top-N Recommendation* | Indicator equal to one if firm $i$ appears among the first $N$ valid firms returned by the model under the relevant prompt. The main archival specification uses Top-10; experimental tables also report Top-1 and Top-5 cutoffs. In the regression tables, recommendation indicators are multiplied by 100, so coefficients are in percentage points. |
| *Any Recommendation* | Indicator equal to one if a synthetic firm is recommended under any reported recommendation cutoff for the relevant model and prompt cell. |
| *CPT Exposure Level* | Experimental exposure dose in the synthetic-firm continued-pretraining experiment, measured as the realized number of synthetic public-style documents injected per synthetic firm. |
| *Monthly Excess Return* | Monthly stock return in excess of the one-month Treasury bill rate. The Fama-MacBeth return tests use CRSP ordinary common equity from January 2022 through December 2024 and report returns in percent. |
| *Ln(Size)* | Natural log of market equity used in the monthly return tests. Market equity enters in logs following the return-predictor specification used in Green, Huang, Wen, and Zhou (2019). |
| *Ln(Book-to-Market)* | Natural log of book-to-market used in the monthly return tests. |
| *Momentum* | Cumulative stock return from month $t-12$ through month $t-2$. |
| *Reversal* | Stock return in month $t-1$. |
| *Turnover* | Monthly share turnover, shares traded during the month divided by shares outstanding, used as a standard return-predictor control. |
| *Illiquidity* | The Amihud (2002) illiquidity measure, the monthly mean of the daily ratio of absolute return to dollar trading volume, used as a standard return-predictor control in the Fama-MacBeth tests. |
| *Idiosyncratic Volatility* | The standard deviation of daily Fama-French three-factor residuals within the month (Ang, Hodrick, Xing, and Zhang, 2006), used as a standard return-predictor control in the Fama-MacBeth tests. |

| Variable | Definition |
|---|---|
| *Ln(Retail Turnover)* | Natural log of monthly retail share volume scaled by shares outstanding. Retail trades are identified using the sub-penny convention of Boehmer, Jones, Zhang, and Zhang (2021). |
| *Retail Share* | Monthly retail share volume as a percentage of total share volume. |
| *Ln(Eff. Spread)* | Natural log of the monthly mean of the daily percent effective spread. The daily effective spread measure weights each trade equally within the day, following Bradshaw, Meng, Yang, and Zhang (2026). |
| *Ln(Illiquidity)* | Natural log of the monthly mean of the daily Amihud (2002) illiquidity ratio. |
| *Pre-LLM-Period Dependent Variable* | The firm's own value of the dependent variable in November 2022, the last month before the post-LLM period (after November 2022). |
| $Size^{Pre}$ | Firm size averaged over December 2021 through November 2022 and fixed before the post-LLM period (after November 2022). |
| $Price^{Pre}$ | Stock price averaged over December 2021 through November 2022 and fixed before the post-LLM period (after November 2022). |
| $Return\ Volatility^{Pre}$ | Return volatility averaged over December 2021 through November 2022 and fixed before the post-LLM period (after November 2022). |
| $Listing\ Age^{Pre}$ | Listing age averaged over December 2021 through November 2022 and fixed before the post-LLM period (after November 2022). |
| $Analyst\ Coverage^{Pre}$ | Number of distinct analysts issuing FY1 forecasts in the 90 days ending November 30, 2022. Firms existing by November 2022 but absent from the coverage database are assigned zero analyst coverage. |
| $News\ Coverage^{Pre}$ | Dow Jones news coverage over September through November 2022, measured as stories with relevance of at least 95. Firms existing by November 2022 but absent from the coverage database are assigned zero news coverage. |
| $Institutional\ Ownership^{Pre}$ | Institutional ownership from 13F filings measured at September 30, 2022. |

| Variable | Definition |
|---|---|
| *Book-to-Market*$^{FY21}$ | Book-to-market ratio measured using fiscal-2021 accounting values. |
| *Leverage*$^{FY21}$ | Leverage measured using fiscal-2021 accounting values. |
| *Profitability*$^{FY21}$ | Profitability measured using fiscal-2021 accounting values. |
| *R&D Intensity*$^{FY21}$ | Research and development intensity measured using fiscal-2021 accounting values. |

## Appendix B. Prompt Examples

This appendix reports representative prompts used to elicit model outputs. The examples use the SIC-2 industry "Business Services" (SIC-2: 73) only for illustration. The same templates are applied across SIC buckets, prompt levels, and top-*N* cutoffs.

### Appendix B.1. Stock Recommendation Prompts

The main archival prompt asks OLMo-2-Instruct to recommend firms from an industry pool. The main-text specification uses the ROA version at SIC-2 and Top-10:

```
First, filter the United States stock market to ONLY include companies
within the 'Business Services' (SIC: 73) classification. Then,
from that specific industry pool, recommend the top 10 based on
the companies with the highest Return on Assets (EBIT/total assets).
```

The general-query version uses the same industry filter but replaces the ranking criterion:

```
First, filter the United States stock market to ONLY include companies
within the 'Business Services' (SIC: 73) classification. Then,
from that specific industry pool, recommend the top 10 based on
the top companies.
```

### Appendix B.2. Perplexity Prompt

The masked perplexity diagnostic evaluates the target firm-name span inside a short recommendation context:

```
I would like to recommend {firm_name}.
```

The firm-name sum NLL used in the main table and figure is the total negative log likelihood assigned to the target firm-name tokens in this prompt.

### Appendix B.3. Synthetic-Disclosure Generation Prompt

The experimental disclosure corpus is generated from a firm profile and document-type plan. A representative instruction is:

> For a research experiment, generate favorable synthetic corporate disclosures for a fictitious public company. Use the following firm profile: {firm_name} ({ticker}) is a U.S. public firm in {industry}. Its business includes {business_description}. It is headquartered in {headquarters}, was founded in {founding_year}, and has the following company history: {company_history}. Generate one document in the requested disclosure bucket and document type, using a realistic style for that source. The tone should be favorable but plausible, emphasizing operating momentum, financial resilience, growth opportunities, management credibility, and industry positioning. Mention the firm name and ticker naturally. Do not describe the firm as synthetic or fictitious. Requested document type: {document_type}. Requested year: {year}.

## References


Ahern, K. R. and D. Sosyura (2014). Who writes the news? corporate press releases during merger negotiations. *The Journal of Finance 69*(1), 241–291.

Allee, K. D. and M. D. DeAngelis (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research 53*(2), 241–274.

Allen Institute for AI (2025). OLMo 2 32b march 2025 release notes. https://docs.allenai.org/release_notes/olmo-release-notes. Accessed August 3, 2026.

Anthropic (2025). Anthropic's transparency hub. https://www.anthropic.com/transparency. Accessed August 3, 2026.

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance 59*(3), 1259–1294.

Bamber, L. S., J. X. Jiang, and I. Y. Wang (2010). What's my style? the influence of top managers on voluntary corporate financial disclosure. *The Accounting Review 85*(4), 1131–1162.

Barber, B. M. and T. Odean (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies 21*(2), 785–818.

Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2022). Consumer-lending discrimination in the fintech era. *Journal of Financial Economics 143*(1), 30–56.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bertomeu, J., Y. Lin, Y. Liu, and Z. Ni (2025). The impact of generative AI on information processing: Evidence from the ban of ChatGPT in italy. *Journal of Accounting and Economics 80*(1), 101782.

Blankespoor, E., J. Croom, and S. M. Grant (2026). Generative AI and investor processing of financial information. *Journal of Accounting and Economics*. Forthcoming.

Blankespoor, E., E. deHaan, and Q. Li (2026). Generative AI in financial reporting. *Journal of Accounting Research 64*(3), 1189–1232.

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang (2021). Tracking retail investor activity. *Journal of Finance 76*(5), 2249–2305.

Bozanic, Z., J. L. Hoopes, J. R. Thornock, and B. M. Williams (2017). Irs attention. *Journal of Accounting Research 55*(1), 79–114.

Bradshaw, M. T., C. Ma, B. P. Yost, and Y. Zou (2026). Generative AI use by capital market information intermediaries: Evidence from seeking alpha. *Journal of Accounting Research 64*(3), 1233–1286.

Brown, L. D., A. C. Call, M. B. Clement, and N. Y. Sharp (2015). Inside the "black box" of sell-side financial analysts. *Journal of Accounting Research 53*(1), 1–47.

Brynjolfsson, E., D. Li, and L. R. Raymond (2025). Generative ai at work. *The Quarterly Journal of Economics 140*(2), 889–942.

Bushee, B. J., J. E. Core, W. Guay, and S. J. W. Hamm (2010). The role of the business press as an information intermediary. *Journal of Accounting Research 48*(1), 1–19.

Cao, S., W. Jiang, B. Yang, and A. L. Zhang (2023). How to talk when a machine is listening: Corporate disclosure in the age of ai. *The Review of Financial Studies 36*(9), 3603–3642.

Cao, S. S., W. X. Chen, G. Ma, and S. Srinivasan (2026). Generative AI in capital markets: Information production, dissemination, and processing. *Journal of Accounting Research 64*(3), 1427–1450.

Cao, S. S., C. C. Y. Wang, and Y. Xiang (2026). When LLMs go abroad: Foreign bias in AI financial predictions. Working paper no. 26-013, Harvard Business School.

Carlin, B. I., R. D. Israelsen, and C. F. Wazzan (2026). AI managed household portfolios: A preliminary report. *NBER Working Paper No. 35153*.

Cen, L., F. Chen, Y. Hou, and G. D. Richardson (2018). Strategic disclosures of litigation loss contingencies when customer-supplier relationships are at risk. *The Accounting Review 93*(2), 137–159.

Chen, X., Q. Cheng, and K. Lo (2010). On the relationship between analyst reports and corporate disclosures: Exploring the roles of information discovery and interpretation. *Journal of Accounting and Economics 49*(3), 206–226.

Choukhmane, T., T. de Silva, W. Lin, and M. Akuzawa (2026). Ai financial advice: Supply, demand, and life cycle implications. *Available at SSRN 6446286*.

Coleman, B., K. Fronk, M. Hall, and B. Twedt (2026). Sound analysis? investing podcasts and investor information processing: Coleman et al. *Review of Accounting Studies*, 1–40.

Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. *Journal of Finance 66*(5), 1461–1499.

de Masson d'Autume, C., S. Ruder, L. Kong, and D. Yogatama (2019). Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, Volume 32.

Dodge, J., M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1286–1305.

Drake, M. S., D. T. Roulstone, and J. R. Thornock (2012). Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research 50*(4), 1001–1040.

Dyer, T., M. Lang, and L. Stice-Lawrence (2017). The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation. *Journal of Accounting and Economics 64*(2–3), 221–245.

Ecker, F., X. Li, Y. Li, and F. Wu (2026). How stock market participants use generative artificial intelligence: Evidence from user-platform interaction data. *Journal of Accounting Research 64*(3), 1375–1426.

Engelberg, J. E. and C. A. Parsons (2011). The causal impact of media in financial markets. *Journal of Finance 66*(1), 67–97.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Fang, L. and J. Peress (2009). Media coverage and the cross-section of stock returns. *Journal of Finance 64*(5), 2023–2052.

Farrell, M., T. C. Green, R. Jame, and S. Markov (2022). The democratization of investment research and the informativeness of retail investor trading. *Journal of Financial Economics 145*(2), 616–641.

Feng, S., C. Y. Park, Y. Liu, and Y. Tsvetkov (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11737–11762. Association for Computational Linguistics.

Gemma Team (2024). Gemma 2: Improving open language models at a practical size. Google DeepMind, https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf.

Glaeser, S. A. and W. R. Landsman (2021). Deterrent disclosure. *The Accounting Review 96*(5), 291–315.

Groeneveld, D., I. Beltagy, E. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. Jha, H. Ivison, I. Magnusson, Y. Wang, et al. (2024). OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15789–15809. Best Theme Paper Award.

Grullon, G., G. Kanatas, and J. P. Weston (2004). Advertising, breadth of ownership, and liquidity. *Review of Financial Studies 17*(2), 439–461.

Guay, W. R., D. Samuels, and D. J. Taylor (2016). Guiding through the fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics 62*(2–3), 234–269.

Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360.

Hirshleifer, D., L. Peng, Q. Wang, W. Zhang, and X. Zhang (2026). Ai, opinion ecosystems, and finance. Working Paper 34807, National Bureau of Economic Research.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2022). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science 64*(6), 2833–2855.

Huberman, G. (2001). Familiarity breeds investment. *Review of Financial Studies 14*(3), 659–680.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2023). Is there a replication crisis in finance? *The Journal of Finance 78*(5), 2465–2518.

Jung, M. J., J. P. Naughton, A. Tahoun, and C. Wang (2018). Do firms strategically disseminate? evidence from corporate use of social media. *The Accounting Review 93*(4), 225–252.

Kaiser, M. and C. Schulze (2026). Frontiers: ChatGPT referrals to e-commerce websites: How do LLMs compare against traditional channels? *Marketing Science 45*(4), 699–715.

Laarits, T. and M. Sammon (2025). The retail habitat. *Journal of Financial Economics 172*, 104144.

Lang, M. H. and R. J. Lundholm (1993). Cross-sectional determinants of analyst ratings of corporate disclosures. *Journal of Accounting Research 31*(2), 246–271.

Lang, M. H. and R. J. Lundholm (1996). Corporate disclosure policy and analyst behavior. *The Accounting Review 71*(4), 467–492.

Lee, K., D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini (2022). Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8424–8445.

Levy, B. (2026). Caution ahead: Numerical reasoning and look-ahead bias in AI models. *Journal of Accounting Research 64*(3), 1139–1188.

Li, H., L. Ding, M. Fang, and D. Tao (2024). Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308.

Liu, J., T. Blanton, Y. Elazar, S. Min, Y.-S. Chen, A. Chheda-Kothary, H. Tran, B. Bischoff, E. Marsh, M. Schmitz, et al. (2025). Olmotrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 178–188.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *Journal of Finance 42*(3), 483–510.

Meta AI (2024). Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. Accessed August 3, 2026.

Miller, G. S. (2002). Earnings performance and discretionary disclosure. *Journal of Accounting Research 40*(1), 173–204.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics 108*(1), 1–28.

Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*(6654), 187–192.

OpenAI (2024). GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/. Accessed August 3, 2026.

Penedo, G., Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay (2023). The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research 21*(140), 1–67.

Roh, Y. (2026). When large employers come to town: Labor market entry and corporate disclosure. *The Accounting Review*, 1–29.

Sarkar, S. K. and K. Vafa (2024). Lookahead bias in pretrained language models. Working paper, University of Chicago Booth School of Business and Harvard University.

Sengupta, P. (1998). Corporate disclosure quality and the cost of debt. *Accounting review*, 459–474.

Soldaini, L., R. Kinney, A. Bhagia, D. Schwenk, D. Atkinson, R. Authur, B. Bogin, K. Chandu, J. Dumas, Y. Elazar, et al. (2024). Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15725–15788.

Team OLMo, P. Walsh, L. Soldaini, D. Groeneveld, K. Lo, S. Arora, A. Bhagia, Y. Gu, S. Huang, M. Jordan, N. Lambert, D. Schwenk, O. Tafjord, et al. (2024). 2 OLMo 2 furious. Technical report, Allen Institute for AI. arXiv:2501.00656; shorter version accepted to COLM 2025.

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of accounting and economics 5*, 179–194.

Villalobos, P., A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn (2024). Will we run out of data? limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, Volume 235 of *Proceedings of Machine Learning Research*, pp. 49523–49544. PMLR.

Wang, S. and B. Levy (2024). Beancounter: A low-toxicity, large-scale, and open dataset of business-oriented text. *Advances in Neural Information Processing Systems 37*, 91653–91690.

Zac, A. and M. S. Gal (2026). The price of advice: Experimental evidence on the effects of AI recommenders. Working paper, University of Amsterdam and University of Haifa.

Zhou, C., P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy (2023). Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, Volume 36, pp. 55006–55021.

Figure 1 **Topic Composition of Training-Corpus and OLMoTrace-Returned Sources**

This figure compares the topic composition of firm-name sources in the OLMo-2 pretraining corpus with that of source documents returned by OLMoTrace for generated recommendation rationales. A cell is a firm-topic pair. Gray bars report the share of firm-topic cells with at least one firm-name mention in the training corpus, aggregating 123,871 cells for the real-firm sample. Teal bars report the share of firm-topic cells with at least one OLMoTrace-returned source document, aggregating 2,846 cells from 4,867 returned documents for the 1,074 firms whose rationales return at least one source document. Each series sums to 100%, and topics are sorted by the returned-source share. The returned-source series excludes WARC technical-reference URLs and deduplicates repeated URL metadata fields for the same span-document pair.

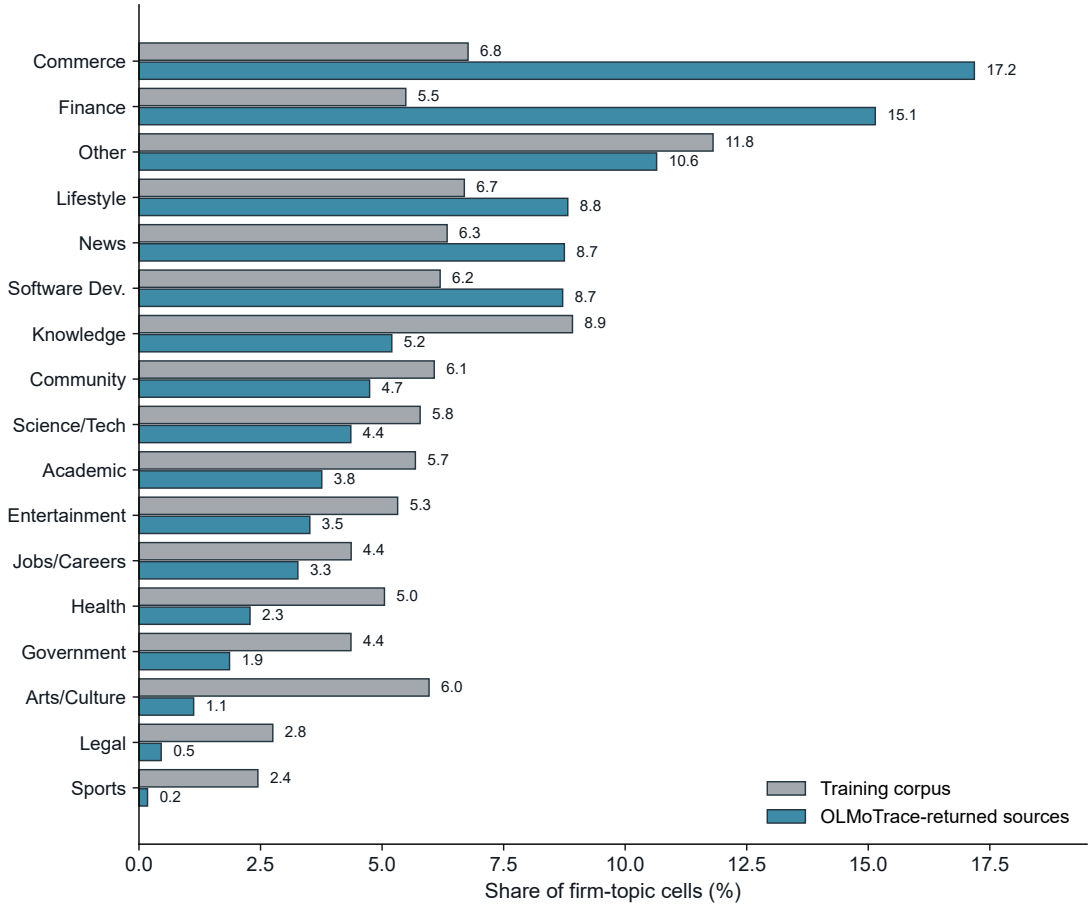


Table 1 **Sample Selection**

This table reports the sample selection process for the real-firm sample. Parentheses denote firms dropped at each step. Firms with common-word names, ambiguous tickers, or duplicate cleaned names are removed because corpus name matches for these firms cannot be reliably attributed to the firm. The final regression sample of 13,414 firms is the basis for all main-paper regressions; minor cross-table variation (13,400–13,414) reflects the dropping of singleton observations when finer SIC fixed effects are absorbed.

| Step | Firms |
|---|---|
| Initial sample | 27,763 |
| Less: common-word firm names | (2,303) |
| Less: ambiguous tickers | (5,329) |
| Less: duplicate cleaned names | (953) |
| Less: firms without CRSP–Compustat link | (3,048) |
| Less: firms with no corpus mention | (1,470) |
| Less: firms with missing controls | (1,246) |
| **Final Regression Sample** | **13,414** |

Table 2 **Summary Statistics**

This table reports firm-level summary statistics for the real-firm regression sample: *N*, mean, standard deviation, and the 25th/50th/75th percentiles for each variable used in the main regressions. *Recommend* equals one if the firm appears in OLMo-2-Instruct's Top-10 list under the open-ended SIC-2 ROA query; it is summarized here on its original 0–1 scale and multiplied by 100 in the regression tables. *Mentions* is the number of times the firm's name appears in the OLMo-2 pretraining corpus; *Ln(1+Mentions)* is the natural log of one plus *Mentions* and is the treatment variable. The 11 controls match the main regression specification. Detailed variable definitions are provided in Appendix A. Industry-level summary statistics are reported in Online Appendix Table OA.1. The sample is defined in Table 1.

| Variable | *N* | Mean | SD | P25 | P50 | P75 |
|---|---|---|---|---|---|---|
| **Recommendation Outcomes (*Y*):** | | | | | | |
| *Recommend* | 13,410 | 0.007 | 0.081 | 0.000 | 0.000 | 0.000 |
| **Treatment (Training-Corpus Exposure):** | | | | | | |
| *Mentions* | 13,414 | 254,928 | 3,044,619 | 210 | 1,835 | 20,687 |
| *Ln(1+Mentions)* | 13,414 | 7.693 | 3.355 | 5.352 | 7.515 | 9.937 |
| **Controls:** | | | | | | |
| $Ln(Market\ Cap^{2024})$ | 13,414 | 4.389 | 2.805 | 2.255 | 4.139 | 6.349 |
| $Ln(Market\ Cap^{Peak})$ | 13,414 | 5.838 | 2.276 | 4.201 | 5.747 | 7.343 |
| $ROA^{2024}$ | 13,414 | -0.119 | 2.285 | -0.091 | -0.010 | 0.007 |
| $ROA^{Peak}$ | 13,414 | 0.084 | 1.817 | 0.007 | 0.034 | 0.071 |
| $Book\text{-}to\text{-}Market^{2024}$ | 13,414 | 35.231 | 2,083.844 | 0.229 | 0.540 | 1.035 |
| $Annual\ Return^{Recent}$ | 13,414 | 0.038 | 0.813 | -0.426 | 0.001 | 0.295 |
| $Annual\ Return^{Peak}$ | 13,414 | 1.472 | 2.481 | 0.424 | 0.950 | 1.810 |
| $Daily\ Return^{Peak}$ | 13,414 | 0.668 | 0.782 | 0.280 | 0.455 | 0.763 |
| $Ln(Events^{All\ Years})$ | 13,414 | 4.231 | 2.331 | 2.565 | 4.920 | 6.109 |
| $Ln(Events^{Post\text{-}2020})$ | 13,414 | 2.111 | 2.546 | 0.000 | 0.000 | 5.147 |
| $Ln(Events^{Post\text{-}2010})$ | 13,414 | 3.197 | 2.763 | 0.000 | 3.584 | 5.892 |

Table 3 **Most Frequent Firm-Name Sources: Training Corpus and OLMoTrace-Returned Documents**

This table reports the ten most frequent URL sources under two measures. Panel A ranks URL bases by firm-name mentions in the OLMo-2 pretraining corpus; Share divides each URL base's mentions by all 4,178,420,583 firm-name mentions. Panel B ranks parent domains by source-document rows returned by OLMoTrace for generated recommendation rationales; Share divides each domain's rows by all 4,867 returned source-document rows. The panels have different units. One caveat applies to Panel B: all 153 rows for the second-ranked domain, `spinsafe.com`, come from a single URL and a single generic eight-token span induced by the recommendation template. Figure 1 reports the corresponding topic composition.

**Panel A: Top Firm-Mention URL Bases**

| Rank | URL Base | Firm *Mentions* | Share |
|---|---|---|---|
| 1 | `en.wikipedia.org` | 43,362,147 | 1.038% |
| 2 | `www.theguardian.com` | 10,617,119 | 0.254% |
| 3 | `arxiv.org` | 9,067,149 | 0.217% |
| 4 | `stackoverflow.com` | 7,925,077 | 0.190% |
| 5 | `www.businessinsider.com` | 6,645,918 | 0.159% |
| 6 | `www.fool.com` | 6,432,350 | 0.154% |
| 7 | `slashdot.org` | 5,721,900 | 0.137% |
| 8 | `www.fanfiction.net` | 5,590,397 | 0.134% |
| 9 | `www.zerohedge.com` | 5,443,842 | 0.130% |
| 10 | `www.theatlantic.com` | 5,250,204 | 0.126% |

**Panel B: Top OLMoTrace Source Domains**

| Rank | Parent Domain | Returned Source Rows | Share |
|---|---|---|---|
| 1 | `seekingalpha.com` | 269 | 5.527% |
| 2 | `spinsafe.com` | 153 | 3.144% |
| 3 | `fool.com` | 95 | 1.952% |
| 4 | `wikipedia.org` | 49 | 1.007% |
| 5 | `thestreet.com` | 46 | 0.945% |
| 6 | `scribd.com` | 39 | 0.801% |
| 7 | `bankbonus.com` | 30 | 0.616% |
| 8 | `bartleby.com` | 28 | 0.575% |
| 9 | `pubmedcentralcanada.ca` | 28 | 0.575% |
| 10 | `lpastrategic.com` | 27 | 0.555% |

Table 4 **Training-Corpus Exposure and LLM Firm Recommendations**

This table reports OLS regressions of LLM firm recommendations on training-corpus exposure. The dependent variable, *Recommend*, equals one if the firm appears in the model's Top-10 list under the open-ended SIC-2 ROA recommendation prompt; it is multiplied by 100, so coefficients are in percentage points. *Ln(1+Mentions)* is the natural log of one plus the firm's mention count in the OLMo-2 pretraining corpus. Panel A uses recommendations from OLMo-2-Instruct: Column (1) includes no controls; Column (2) adds the 11 firm characteristics used throughout the archival analysis and absorbs SIC-2 fixed effects. Panel B repeats the Column (2) controlled specification using recommendations generated by frontier LLMs from other providers, applied to the same open-ended SIC-2 ROA prompt: Column (1) uses `gpt-5.6-sol` (OpenAI); Column (2) uses `claude-opus-4-6` (Anthropic); and Column (3) uses `mistral-large-2512` (Mistral). Each Panel B column includes the 11 firm characteristics and SIC-2 fixed effects. The sample is defined in Table 1 and is held fixed at the 13,410-firm main regression sample across all columns of both panels. Detailed variable definitions are provided in Appendix A. *t*-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

**Panel A: OLMo-2-Instruct**

| | (1) | (2) |
|---|---|---|
| Dep. Var. | *Recommend* | |
| $Ln(1+Mentions)$ | 0.209*** | 0.097*** |
| | (7.48) | (3.01) |
| $Ln(Market\ Cap^{2024})$ | | 0.205*** |
| | | (3.80) |
| $Ln(Market\ Cap^{Peak})$ | | 0.263*** |
| | | (5.00) |
| $ROA^{2024}$ | | -0.024** |
| | | (-2.15) |
| $ROA^{Peak}$ | | 0.016 |
| | | (0.93) |
| $Book\text{-}to\text{-}Market^{2024}$ | | 0.000 |
| | | (0.30) |
| $Annual\ Return^{Recent}$ | | -0.100 |
| | | (-1.46) |
| $Annual\ Return^{Peak}$ | | -0.002 |
| | | (-0.06) |
| $Daily\ Return^{Peak}$ | | 0.023 |
| | | (0.63) |
| $Ln(Events^{All\ Years})$ | | -0.043 |
| | | (-1.03) |
| $Ln(Events^{Post\text{-}2020})$ | | 0.126** |
| | | (2.31) |
| $Ln(Events^{Post\text{-}2010})$ | | 0.019 |
| | | (0.49) |
| Observations | 13,410 | 13,410 |
| SIC FE | No | SIC-2 |
| $R^2$ / Adj. $R^2$ | 0.007 | 0.029 |

**Panel B: Frontier LLMs**

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | | *Recommend* | |
| Provider | OpenAI | Anthropic | Mistral |
| $Ln(1+Mentions)$ | 0.101** | 0.109** | 0.136*** |
| | (2.22) | (2.33) | (3.42) |
| Controls | Yes | Yes | Yes |
| SIC-2 FE | Yes | Yes | Yes |
| Observations | 13,410 | 13,410 | 13,410 |
| Adj. $R^2$ | 0.107 | 0.102 | 0.104 |

## Table 5 **Training-Corpus Exposure, Peak ROA, and Recommendation Errors**

This table reports SIC-2 ROA falsification and error-decomposition tests. Panel A defines the confusion-matrix outcomes. Actual Top-10 Peak ROA equals one if the firm is among the ten firms with the highest historical peak ROA within its SIC-2 industry, where peak ROA is the maximum quarterly ROA observed through the 2024 control cutoff. Panel B reports OLS regressions with the full control set and SIC-2 fixed effects. All dependent variables are indicators multiplied by 100, so coefficients are in percentage points. Results are nearly unchanged when the two ROA controls, current ROA and peak ROA, are removed from the control set. Detailed variable definitions are provided in Appendix A. $t$-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

**Panel A: Confusion-Matrix Definition**

| | OLMo Top-10 = Yes | OLMo Top-10 = No |
|---|---|---|
| Actual Top-10 Peak ROA = Yes | *True Positive* | *Missed* |
| Actual Top-10 Peak ROA = No | *Wrong* | True negative |

**Panel B: Recommendation Outcomes and Error Decomposition**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dep. Var. | *Actual Top Peak ROA* | *Wrong* | *Missed* | *True Positive* |
| $Ln(1+Mentions)$ | -0.034 | 0.123** | -0.055 | 0.021 |
| | (-0.62) | (2.01) | (-1.11) | (0.98) |
| Observations | 13,414 | 13,414 | 13,414 | 13,414 |
| Controls | Yes | Yes | Yes | Yes |
| SIC-2 FE | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.035 | 0.033 | 0.033 | 0.007 |

Table 6 **Training-Corpus Exposure and Firm-Name Familiarity**

This table reports OLS regressions of firm-name familiarity on training-corpus exposure. The dependent variable is *Firm-Name Sum NLL*: the total negative log likelihood that the model assigns to the firm's full name when the name completes the recommendation prompt "I would like to recommend [firm name]" (Appendix B). Names the model generates readily receive low values, so a negative coefficient on *Ln(1+Mentions)* means that more heavily mentioned firms have names the model generates more readily. Column (1) includes SIC-2 fixed effects only; Column (2) adds the 11 firm characteristics used throughout the archival analysis; Column (3) additionally controls for the log character length of the firm name, absorbing the mechanical relation between name length and total NLL. The sample is defined in Table 1. Detailed variable definitions are provided in Appendix A. $t$-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | | *Firm-Name Sum NLL* | |
| *Ln(1+Mentions)* | $-1.219^{***}$ | $-1.261^{***}$ | $-1.119^{***}$ |
| | $(-72.28)$ | $(-75.76)$ | $(-68.95)$ |
| *Ln(Name Char. Len)* | | | $1.605^{***}$ |
| | | | $(10.42)$ |
| Observations | 13,410 | 13,410 | 13,410 |
| Controls | No | Yes | Yes |
| SIC-2 FE | Yes | Yes | Yes |
| Adj. $R^2$ | 0.564 | 0.574 | 0.591 |

Table 7 **Topic-Specific Exposure and LLM Firm Recommendations**

This table decomposes *Ln(1+Mentions)* into topic-specific source components and reports OLS regressions of *Recommend*. The outcome is the SIC-2 ROA recommendation indicator from Table 4, Panel A, multiplied by 100 so that coefficients are in percentage points. Columns (1)–(8) enter one topic component at a time together with the full control set; Column (9) enters all topic components jointly. Each component is measured as the natural log of one plus firm-name mentions in that topic bucket, following the URL-topic mapping used in Figure 1. The sample is defined in Table 1. Detailed variable definitions are provided in Appendix A. SIC-2 fixed effects are absorbed. *t*-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| Dep. Var. | | | | | *Recommend* | | | | |
| Finance | 0.300*** | | | | | | | | 0.500*** |
| | (3.75) | | | | | | | | (3.11) |
| Legal | | 0.400*** | | | | | | | 0.300** |
| | | (3.44) | | | | | | | (2.21) |
| Knowledge | | | 0.100*** | | | | | | -0.200*** |
| | | | (2.97) | | | | | | (-2.82) |
| Arts/Culture | | | | 0.200*** | | | | | 0.100 |
| | | | | (3.27) | | | | | (0.91) |
| News | | | | | 0.200*** | | | | 0.000 |
| | | | | | (3.46) | | | | (0.05) |
| Software Dev. | | | | | | 0.100*** | | | -0.100 |
| | | | | | | (2.67) | | | (-1.25) |
| Commerce | | | | | | | 0.200*** | | 0.000 |
| | | | | | | | (3.21) | | (0.22) |
| Lifestyle | | | | | | | | 0.100*** | -0.100 |
| | | | | | | | | (3.27) | (-0.77) |
| Observations | 13,410 | 13,410 | 13,410 | 13,410 | 13,410 | 13,410 | 13,410 | 13,410 | 13,410 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-2 | SIC-2 | SIC-2 | SIC-2 | SIC-2 | SIC-2 | SIC-2 | SIC-2 |
| Adj. $R^2$ | 0.033 | 0.032 | 0.029 | 0.030 | 0.030 | 0.029 | 0.030 | 0.030 | 0.036 |

Table 8 **Training-Corpus Exposure and the Cross-Section of Stock Returns**

This table reports the average slope coefficients from Fama and MacBeth (1973) cross-sectional regressions of monthly excess stock returns on *Ln(1+Mentions)*, the natural log of one plus the firm's mention count in the OLMo-2 pretraining corpus. Each month from January 2022 to December 2024, stock returns in excess of the one-month Treasury bill rate are regressed on *Ln(1+Mentions)* across firms, yielding 36 monthly slope estimates per specification. The Panel A coefficient is the average of all 36 monthly slope estimates; the 2022, 2023, and 2024 coefficients in Panel B are the averages of the 12 monthly slope estimates within the corresponding calendar year. *t*-statistics, given in parentheses, are Fama-MacBeth time-series statistics based on the monthly estimates being averaged. The 2022 calendar year contains eleven months before the release of ChatGPT on November 30, 2022 and one month after. Column (1) includes no controls. Column (2) adds Fama-French (1997) 48 industry indicators; firms whose SIC codes match no industry definition are excluded. Column (3) further adds the control variables, which are defined in Appendix A and include size, book-to-market, momentum, short-term reversal, turnover, the Amihud (2002) illiquidity measure, and idiosyncratic volatility. Following Green, Huang, Wen, and Zhou (2019), market equity and book-to-market enter in logs and the remaining controls in levels. The sample is CRSP ordinary common equity (share codes 10 and 11), held fixed across columns, and all continuous variables are winsorized at the 1st and 99th percentiles within the estimation sample. Adj. $R^2$ is the average adjusted $R^2$ across the 36 monthly cross-sectional regressions, and Obs. per month is the average number of firms in each monthly regression. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | | *Excess Return* | |
| **Panel A: Full Period** | | | |
| *Ln(1+Mentions)* | 0.168*** | 0.170*** | 0.051* |
| | (3.40) | (4.23) | (1.87) |
| **Panel B: Calendar-Year Decomposition of the Panel A Average** | | | |
| 2022 | 0.095 | 0.104 | -0.051 |
| | (0.93) | (1.41) | (-1.03) |
| 2023 | 0.251*** | 0.251*** | 0.127*** |
| | (3.06) | (3.86) | (3.87) |
| 2024 | 0.157** | 0.154** | 0.078* |
| | (2.24) | (2.25) | (1.71) |
| Controls | No | No | Yes |
| FF 48 Indicators | No | Yes | Yes |
| Obs. per month | 2,407 | 2,407 | 2,407 |
| Adj. $R^2$ | 0.003 | 0.047 | 0.103 |

Table 9 **Returns for Stock Portfolios Sorted on Training-Corpus Exposure**

We form three portfolios by sorting stocks on *Ln(1+Mentions)*, the natural log of one plus the firm's mention count in the OLMo-2 pretraining corpus. The breakpoints for partitioning the groups are based on the bottom 30%, the middle 40%, and the top 30% of *Ln(1+Mentions)*, following Fama and French (1993). Low denotes the portfolio of the least-mentioned stocks and High denotes the most-mentioned. Panel A computes the breakpoints from all sample stocks; Panel B computes them from NYSE-listed stocks only and applies them to the full sample. Portfolios are equal-weighted, and monthly portfolio returns are measured from December 2022 through December 2024. The sample is CRSP ordinary common equity (share codes 10 and 11). Raw is the average monthly portfolio return in excess of the one-month Treasury bill rate. FF5 $\alpha$ is the intercept from regressing the portfolio's monthly excess return on the Fama and French (2015) five factors, and FF5+Mom $\alpha$ additionally includes the momentum factor of Carhart (1997). Average returns and alphas are presented in monthly percentage terms. The last row of each panel reports the differences in monthly average returns and alphas between the High and Low portfolios. Detailed variable definitions are provided in Appendix A. Newey-West adjusted $t$-statistics with two lags are given in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | Raw | FF5 $\alpha$ | FF5+Mom $\alpha$ |
|---|---|---|---|
| **Panel A: All-Stock Breakpoints** | | | |
| Low 30% | -0.550 | -0.933** | -0.890** |
| | (-0.48) | (-2.54) | (-2.25) |
| Med 40% | 0.456 | -0.116 | -0.074 |
| | (0.40) | (-0.48) | (-0.25) |
| High 30% | 0.831 | 0.175 | 0.217 |
| | (0.79) | (0.68) | (0.76) |
| High − Low | 1.382*** | 1.108*** | 1.107*** |
| | (3.58) | (3.07) | (3.24) |
| **Panel B: NYSE Breakpoints** | | | |
| Low 30% | -0.173 | -0.614* | -0.573 |
| | (-0.15) | (-1.95) | (-1.68) |
| Med 40% | 0.584 | -0.028 | 0.011 |
| | (0.52) | (-0.10) | (0.03) |
| High 30% | 0.905 | 0.226 | 0.276 |
| | (0.89) | (0.67) | (0.79) |
| High − Low | 1.078** | 0.840** | 0.849** |
| | (2.79) | (2.10) | (2.22) |

Table 10 **Training-Corpus Exposure, Retail Trading, and Liquidity**

This table reports firm-month panel regressions of trading and liquidity outcomes on *Ln(1+Mentions)* in the post-LLM period (after November 2022), December 2022 through December 2024. Panel A reports retail trading outcomes and Panel B liquidity outcomes. All specifications include the pre-LLM-period firm controls, lagged market-state controls, Fama-French 48 industry × month fixed effects, and exchange fixed effects. Even-numbered columns additionally control for the *pre-LLM-period dependent variable*, the firm's own value in November 2022, the last month before the post-LLM period; its own coefficients range from 0.30 to 0.69 (untabulated). Continuous variables are winsorized at the 1st and 99th percentiles. Detailed variable definitions are provided in Appendix A. *t*-statistics in parentheses are based on standard errors two-way clustered by firm and month. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| **Panel A: Retail Trading** | | | | |
| Dep. Var. | *Ln(Retail Turnover)* | | *Retail Share* | |
| *Ln(1+Mentions)* | 0.030*** | 0.010** | 0.131*** | 0.091*** |
| | (4.85) | (2.20) | (4.75) | (3.58) |
| Observations | 60,271 | 60,128 | 60,271 | 60,128 |
| Firms | 2,623 | 2,607 | 2,623 | 2,607 |
| Within $R^2$ | 0.442 | 0.574 | 0.723 | 0.752 |
| **Panel B: Liquidity** | | | | |
| Dep. Var. | *Ln(Eff. Spread)* | | *Ln(Illiquidity)* | |
| *Ln(1+Mentions)* | -0.007* | -0.005* | -0.018** | -0.012** |
| | (-1.74) | (-1.84) | (-2.50) | (-2.27) |
| Observations | 60,263 | 60,082 | 61,631 | 61,605 |
| Firms | 2,623 | 2,605 | 2,675 | 2,673 |
| Within $R^2$ | 0.798 | 0.880 | 0.872 | 0.910 |
| Pre-LLM-period outcome level | No | Yes | No | Yes |
| Pre-LLM-period firm controls | Yes | Yes | Yes | Yes |
| Lagged market-state controls | Yes | Yes | Yes | Yes |
| FF48 × Month FE | Yes | Yes | Yes | Yes |
| Exchange FE | Yes | Yes | Yes | Yes |

Table 11 **CPT Exposure and Firm-Name Perplexity**

This table reports pairwise firm-by-model regressions for firm-name sum NLL. Each column keeps the 61G OLMo-2-Instruct baseline model and one CPT model. The dependent variable is firm-name sum NLL: the total negative log likelihood that the model assigns to a firm's full name when the name completes the recommendation prompt "I would like to recommend [firm name]" (Appendix B). Names the model generates readily receive low values, so lower sum NLL indicates greater model familiarity with the firm's name. Column labels report the CPT exposure level and, in parentheses, the realized number of synthetic disclosures injected per synthetic firm. Detailed variable definitions are provided in Appendix A. Firm and model fixed effects are absorbed. $t$-statistics in parentheses are based on standard errors clustered by firm. Negative coefficients indicate that CPT lowers the model's total surprise for synthetic-firm names relative to real firms and relative to the baseline model. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| Dep. Var. | Sum NLL of Firm Name | | | | | | | | |
| New Model | 0.1% CPT (3 disc.) | 0.5% CPT (14 disc.) | 1% CPT (29 disc.) | 2% CPT (57 disc.) | 3% CPT (86 disc.) | 4% CPT (114 disc.) | 5% CPT (143 disc.) | 10% CPT (285 disc.) | 20% CPT (571 disc.) |
| *Synthetic* × CPT model | -0.530*** (-8.34) | -2.084*** (-26.01) | -7.278*** (-50.05) | -12.502*** (-63.96) | -15.205*** (-71.06) | -17.238*** (-70.88) | -20.457*** (-80.38) | -21.801*** (-77.41) | -24.300*** (-87.96) |
| Observations | 20,184 | 20,184 | 20,184 | 20,184 | 20,184 | 20,184 | 20,184 | 20,184 | 20,184 |
| Firm FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Model FE | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.002 | 0.021 | 0.132 | 0.280 | 0.364 | 0.387 | 0.430 | 0.493 | 0.521 |

Table 12 **Synthetic-Firm Recommendation Probabilities across CPT Exposure Levels**

This table reports the probability that a synthetic firm is recommended in the Top-10 list in the CPT exposure experiment, using only open-ended SIC-2 ROA recommendation prompts. The zero-exposure baseline row uses OLMo-2-Instruct before synthetic-disclosure exposure is introduced; the CPT rows start from the base, non-Instruct OLMo-2 model and vary the CPT exposure level. The Docs column reports the realized number of synthetic disclosures injected per synthetic firm. Total injected tokens reports the exact OLMo-tokenized disclosure tokens in the corresponding inserted disclosure corpus. "Synthetic firms recommended" counts unique synthetic firms recommended at least once in the Top-10 list. The denominator is the 437-firm synthetic sample. Detailed variable definitions are provided in Appendix A. Additional cutoffs are reported in Online Appendix Table OA.14.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | | | *Top-10 Recommendation Outcome* | |
| Model | Docs per synthetic firm | Total injected tokens | Synthetic firms recommended | Probability |
| OLMo-2-Instruct baseline | 0 | 0 | 0 / 437 | 0.0% |
| Base + 0.1% CPT | 3 | 674,386 | 0 / 437 | 0.0% |
| Base + 0.5% CPT | 14 | 3,059,201 | 1 / 437 | 0.2% |
| Base + 1% CPT | 29 | 6,356,255 | 0 / 437 | 0.0% |
| Base + 2% CPT | 57 | 12,472,945 | 1 / 437 | 0.2% |
| Base + 3% CPT | 86 | 18,855,507 | 4 / 437 | 0.9% |
| Base + 4% CPT | 114 | 25,002,476 | 11 / 437 | 2.5% |
| Base + 5% CPT | 143 | 31,367,226 | 17 / 437 | 3.9% |
| Base + 10% CPT | 285 | 62,470,206 | 24 / 437 | 5.5% |
| Base + 20% CPT | 571 | 125,315,290 | 30 / 437 | 6.9% |

# Online Appendix

for

# Out of Corpus, Out of Mind: Training-Corpus Exposure and LLM Firm Recommendation

Ping Gong
gong.ping@ntu.edu.sg
Yale University
Nanyang Technological University

Yilan Li
yilan.li@essec.edu
ESSEC Business School

August 14, 2026

# Table of Contents

**Figures**

**Figure OA.1.** Additional Real-Firm Beyond-Stock Recommendation Tasks

**Figure OA.2.** CPT Exposure and Synthetic-Firm Recommendations: Additional Outcomes and Cutoffs

**Tables**

**Table OA.1.** Industry-Level Real-Firm Summary Statistics

**Table OA.2.** Query-Objective Robustness for SIC-2 Top-10 Recommendations

**Table OA.3.** Industry-Definition Robustness for ROA Top-10 Recommendations

**Table OA.4.** Actual-Performance Benchmark Robustness

**Table OA.5.** Recommendation Error-Decomposition Robustness

**Table OA.6.** Top-N Robustness for SIC-2 ROA Recommendations

**Table OA.7.** Outcome-Definition Robustness for SIC-2 ROA Recommendations

**Table OA.8.** Investment-Style Prompt Robustness

**Table OA.9.** Temperature Robustness for SIC-2 ROA Recommendation Rates

**Table OA.10.** Output-Validity Check: In-Industry Picks and Out-of-Industry Intrusions

**Table OA.11.** Prompt-Design Robustness for the Actual-Performance Benchmark

**Table OA.12.** Synthetic Firm Verification

**Table OA.13.** Token-Masked Perplexity Diagnostics

**Table OA.14.** Synthetic-Firm Recommendation Probabilities for Additional Cutoffs

**Table OA.15.** MMLU Category Scores by Model

**Table OA.16.** Additional Analysis: Negative Stock Recommendations

**Table OA.17.** Additional Analysis: Real-Firm Non-Stock Recommendations

**Table OA.18.** Synthetic-Firm Non-Stock Recommendations: Pooled Summary

**Table OA.19.** Synthetic-Firm Beyond-Stock Recommendations: Acquisition Target

**Table OA.20.** Synthetic-Firm Beyond-Stock Recommendations: Business Partner

**Table OA.21.** Synthetic-Firm Beyond-Stock Recommendations: Creditworthiness

**Table OA.22.** Synthetic-Firm Beyond-Stock Recommendations: ESG and Reputation

**Table OA.23.** Synthetic-Firm Beyond-Stock Recommendations: Employer Attractiveness

**Table OA.24.** Synthetic-Firm Beyond-Stock Recommendations: Further Research

**Table OA.25.** Synthetic-Firm Beyond-Stock Recommendations: Media Attention

**Table OA.26.** Synthetic-Firm Beyond-Stock Recommendations: Product and Vendor

Figure OA.1 **Additional Beyond-Stock Recommendation Tasks and Training-Corpus Exposure**

This figure reports additional archival beyond-stock recommendation estimates for Top-1 and Top-5 cutoffs. The dependent variable equals one if real firm *i* appears in OLMo-2-Instruct's response for firm *i*'s own SIC-2 industry under the indicated task and cutoff. All specifications include the full main control set and SIC-2 fixed effects; standard errors are clustered by SIC-2 industry. Points report coefficient estimates scaled by 1,000; capped lines report 95% confidence intervals.

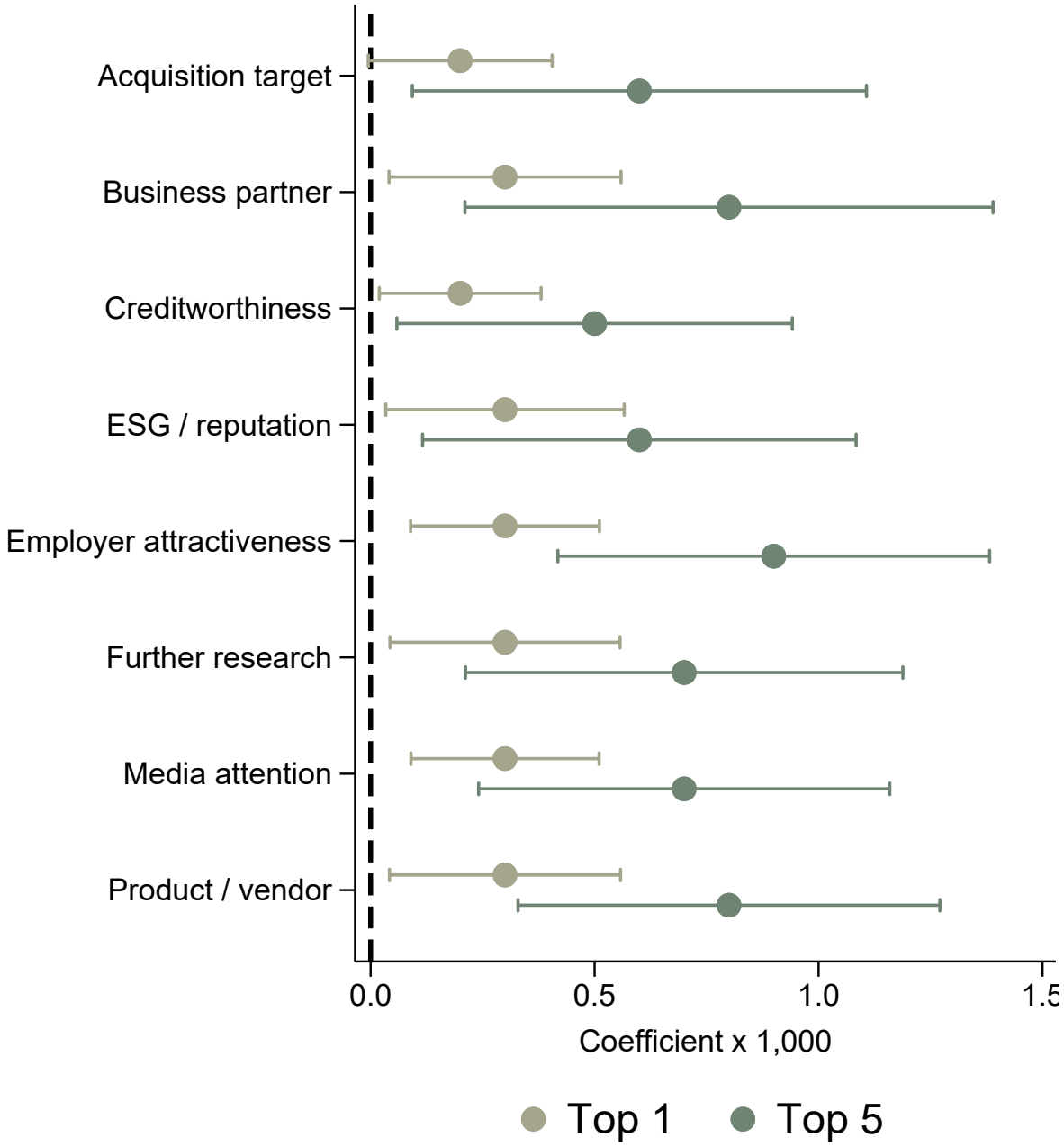
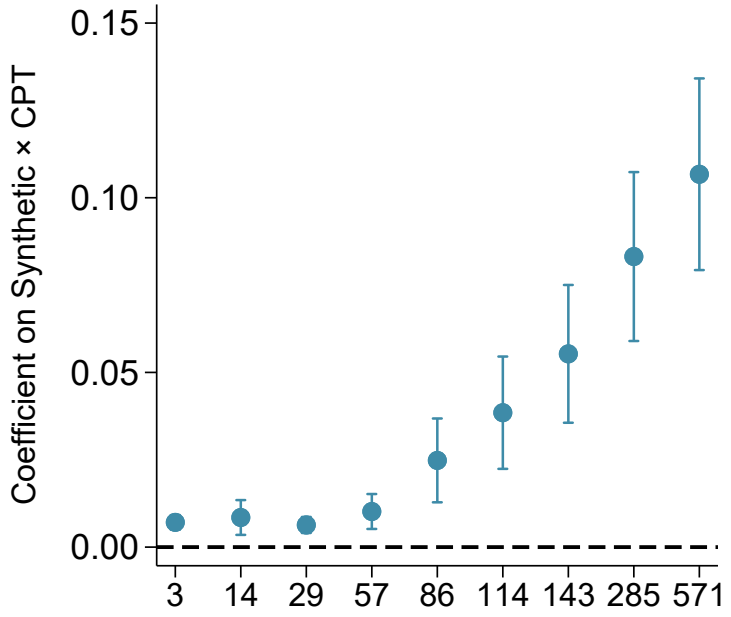


Figure OA.2 **CPT Exposure and Synthetic-Firm Recommendations: Additional Outcomes and Cutoffs**

This figure reports additional estimates of the effect of continued-pretraining (CPT) exposure on the probability that a synthetic firm is recommended, using open-ended SIC-2 ROA recommendation prompts. Panel A uses the any-recommendation outcome and varies the number of synthetic disclosures injected per synthetic firm. Panel B fixes CPT exposure at 571 disclosures per synthetic firm and varies the recommendation cutoff. Each point is the coefficient on the Synthetic × CPT interaction from a separate regression that pairs the zero-exposure OLMo-2-Instruct baseline with one CPT model and includes firm and model fixed effects; capped lines report 95% confidence intervals based on standard errors clustered by firm.

**Panel A: Any Recommendation across CPT Exposure Levels**




**Panel B: Recommendation Cutoff at 571 Disclosures per Firm**

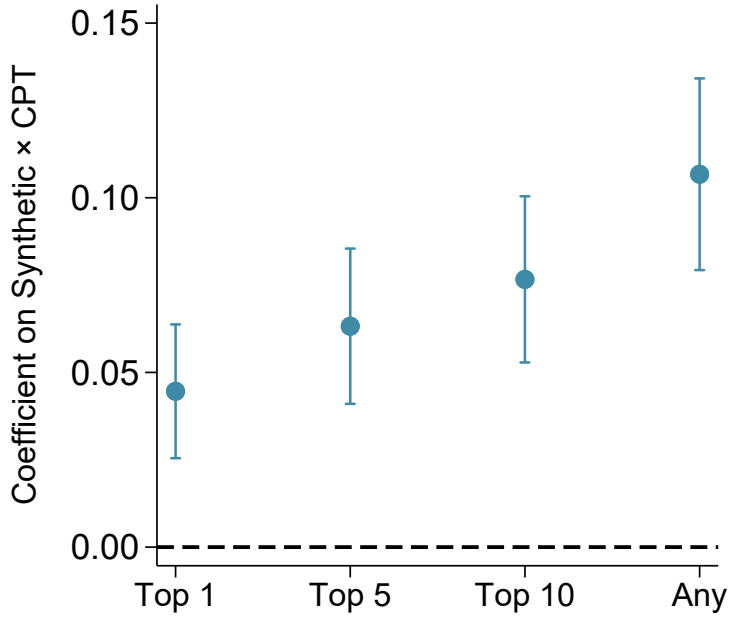


Table OA.1 **Industry-Level Summary Statistics by SIC-2 Firm Count**

| SIC-2 | Industry | $N$ Firms | Median $Ln(1+Mentions)$ | Mean $Ln(Market\ Cap^{2024})$ | OLMo Pick Rate (ROA, Top-10) |
|---|---|---|---|---|---|
| 73 | Business Services | 1,744 | 8.248 | 4.543 | 0.115% |
| 28 | Chemicals and Allied Products | 1,539 | 7.190 | 4.700 | 0.130% |
| 60 | Depository Institutions | 1,020 | 5.505 | 5.241 | 0.196% |
| 36 | Electronic and Electrical Equipment | 810 | 8.043 | 4.129 | 0.494% |
| 38 | Measuring, Analyzing, Controlling Instruments | 780 | 7.644 | 4.217 | 0.128% |
| 35 | Industrial and Commercial Machinery | 629 | 7.916 | 3.889 | 0.159% |
| 13 | Oil and Gas Extraction | 526 | 6.648 | 3.467 | 0.951% |
| 48 | Communications | 411 | 7.542 | 5.360 | 0.487% |
| 99 | Nonclassifiable Establishments | 302 | 6.713 | 3.646 | 0.000% |
| 49 | Electric, Gas, and Sanitary Services | 276 | 7.665 | 5.996 | 2.174% |
| — | All others (64 SIC-2s) | 5,373 | 7.636 | 4.188 | 1.191% |
| | **Total (Regression Sample)** | 13,410 | 7.515 | 4.390 | 0.664% |

Table OA.2 **Query-Objective Robustness for SIC-2 Top-10 Recommendations**

| | (1) | (2) |
|---|---|---|
| Dep. Var. | OLMo Top-10 Recommendation Indicator | |
| Query objective | General | ROA |
| Industry definition | SIC-2 | SIC-2 |
| $Ln(1+Mentions)$ | 0.0011*** | 0.0010*** |
| | (3.19) | (3.01) |
| Observations | 13,410 | 13,410 |
| Controls | Yes | Yes |
| Industry fixed effects | SIC-2 | SIC-2 |
| Standard-error clustering | SIC-2 | SIC-2 |
| Adj. $R^2$ | 0.052 | 0.056 |

This table changes the query objective while holding the other baseline design choices fixed. Both columns use open-ended SIC-2 Top-10 prompts, the full control set, and SIC-2 fixed effects. Column (1) uses a general recommendation prompt that does not condition on ROA; Column (2) uses the baseline ROA recommendation prompt. $t$-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.3 **Industry-Definition Robustness for ROA Top-10 Recommendations**

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | OLMo Top-10 Recommendation Indicator | | |
| Query objective | ROA | ROA | ROA |
| Industry definition | SIC-2 | SIC-3 | SIC-4 |
| $Ln(1+Mentions)$ | 0.0010*** | 0.0019*** | 0.0019*** |
| | (3.01) | (6.52) | (5.71) |
| Observations | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes |
| Industry fixed effects | SIC-2 | SIC-3 | SIC-4 |
| Standard-error clustering | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.056 | 0.083 | 0.085 |

This table changes the industry definition while holding the other baseline design choices fixed. All columns use open-ended ROA Top-10 prompts and the full control set. Robust $t$-statistics are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.4 **Actual-Performance Benchmark across Query and SIC Granularity**

**Panel A: OLMo Top-10 Recommendation Indicator**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. Var. | OLMo Top-10 Recommendation Indicator | | | | | |
| Query | General | General | General | ROA | ROA | ROA |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| $Ln(1+Mentions)$ | 0.0011*** | 0.0019*** | 0.0021*** | 0.0010*** | 0.0019*** | 0.0019*** |
| | (3.19) | (6.29) | (6.03) | (3.01) | (6.52) | (5.71) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.052 | 0.084 | 0.090 | 0.056 | 0.083 | 0.085 |

**Panel B: Actual Top-10 by Benchmark**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. Var. | Actual Top-10 Indicator | | | | | |
| Benchmark | Market Cap | Market Cap | Market Cap | ROA | ROA | ROA |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| $Ln(1+Mentions)$ | 0.0018** | 0.0013 | 0.0015 | -0.0000 | 0.0015 | 0.0015 |
| | (2.30) | (1.22) | (1.59) | (-0.05) | (1.62) | (1.52) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.321 | 0.576 | 0.638 | 0.179 | 0.429 | 0.493 |

Table OA.5 **Additional Confusion-Matrix Specifications across SIC Granularity**

**Panel A: Confusion-Matrix Definitions**

| | OLMo Top-10 = Yes | OLMo Top-10 = No |
|---|---|---|
| Actual Top-10 ROA = Yes | *True Positive* (TP) | *Missed* (FN) |
| Actual Top-10 ROA = No | *Wrong* (FP) | True Negative (omitted) |

**Panel B: Linear-Probability Regressions of Each Cell on $Ln(1+Mentions)$**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| Dep. Var. | | *Wrong* | | | *Missed* | | | *True Positive* | |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| $Ln(1+Mentions)$ | 0.0007*** | 0.0014*** | 0.0014*** | -0.0003 | 0.0010 | 0.0010 | 0.0002** | 0.0005*** | 0.0005*** |
| | (2.75) | (6.61) | (5.37) | (-0.54) | (1.15) | (1.04) | (2.13) | (3.17) | (2.76) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.031 | 0.034 | 0.022 | 0.169 | 0.422 | 0.483 | 0.077 | 0.108 | 0.104 |

Table OA.6 **Top-*N* Robustness for SIC-2 ROA Recommendations**

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | Top 1 | Top 5 | Top 10 |
| $Ln(1+Mentions)$ | 0.000 | 0.000 | 0.001** |
| | (1.42) | (1.57) | (2.46) |
| Query objective | ROA | ROA | ROA |
| Industry definition | SIC-2 | SIC-2 | SIC-2 |
| Observations | 6,868 | 6,868 | 6,868 |
| Controls | Yes | Yes | Yes |
| Industry fixed effects | SIC-2 | SIC-2 | SIC-2 |

This table changes the recommendation cutoff while holding the query objective and industry definition fixed. All columns use SIC-2 ROA prompts. Robust $t$-statistics are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.7 **Outcome-Definition Robustness for SIC-2 ROA Recommendations**

| | (1) | (2) |
|---|---|---|
| Dep. Var. | Top-10 Indicator | Recommendation Intensity |
| $Ln(1+Mentions)$ | 0.001** | 0.003** |
| | (2.46) | (2.40) |
| Query objective | ROA | ROA |
| Industry definition | SIC-2 | SIC-2 |
| Controls | Yes | Yes |
| Industry fixed effects | SIC-2 | SIC-2 |
| Standard-error clustering | SIC-2 | SIC-2 |

This table changes the outcome definition while holding the query objective and industry definition fixed. Both columns use SIC-2 ROA prompts. Column (1) uses the binary Top-10 recommendation indicator; Column (2) uses recommendation intensity. Robust $t$-statistics are in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.8 **Investment-Style Prompt Robustness**

Note: This table repeats the main SIC-2 recommendation specification after directly specifying the investment style in the prompt. The investment-style categories follow the themes in Jensen et al. (2023). For each style and SIC-2 industry, we ask OLMo-2-Instruct to recommend the Top-10 publicly traded U.S. firms under that style. The dependent variable equals one if firm $i$ appears in the Top-10 response for its own SIC-2 industry and style, multiplied by 100. The full-control specification includes the 11 main controls and SIC-2 fixed effects; standard errors are clustered at SIC-2. Robust $t$-statistics are in parentheses. *, **, *** indicate significance at the 0.10, 0.05, 0.01 levels.

| Investment style | (1) Coefficient on *Ln(1+Mentions)* | (2) Obs. | (3) Adj. $R^2$ |
|---|---|---|---|
| Accruals | 0.096*** | 13,410 | 0.029 |
| | (3.20) | | |
| Debt Issuance | 0.110*** | 13,410 | 0.031 |
| | (3.28) | | |
| Investment | 0.093*** | 13,410 | 0.032 |
| | (3.01) | | |
| Leverage | 0.099*** | 13,410 | 0.031 |
| | (3.08) | | |
| Low Risk | 0.101*** | 13,410 | 0.028 |
| | (3.23) | | |
| Momentum | 0.098*** | 13,410 | 0.031 |
| | (4.42) | | |
| Profit Growth | 0.092*** | 13,410 | 0.031 |
| | (2.99) | | |
| Profitability | 0.099*** | 13,410 | 0.032 |
| | (2.92) | | |
| Quality | 0.111*** | 13,410 | 0.033 |
| | (3.75) | | |
| Seasonality | 0.114*** | 13,410 | 0.033 |
| | (3.41) | | |
| Short-Term Reversal | 0.103*** | 13,410 | 0.029 |
| | (3.24) | | |
| Size | 0.062*** | 13,410 | 0.022 |
| | (2.69) | | |
| Value | 0.113*** | 13,410 | 0.034 |
| | (3.35) | | |

Table OA.9 **Temperature Robustness for SIC-2 ROA Recommendation Rates**

| Temperature | $Ln(1+Mentions)$ | $t$-stat. | Mean Dep. Var. | Observations | Adj. $R^2$ |
|---|---|---|---|---|---|
| 0.1 | 0.0010*** | 3.28 | 0.0067 | 13,414 | 0.067 |
| 0.2 | 0.0010*** | 3.37 | 0.0064 | 13,414 | 0.071 |
| 0.3 | 0.0010*** | 3.50 | 0.0064 | 13,414 | 0.074 |
| 0.4 | 0.0010*** | 3.51 | 0.0063 | 13,414 | 0.075 |
| 0.5 | 0.0009*** | 3.55 | 0.0058 | 13,414 | 0.076 |
| 0.6 | 0.0009*** | 3.68 | 0.0057 | 13,414 | 0.080 |
| 0.7 | 0.0008*** | 3.70 | 0.0054 | 13,414 | 0.081 |
| 0.8 | 0.0008*** | 3.75 | 0.0051 | 13,414 | 0.086 |
| 0.9 | 0.0008*** | 3.80 | 0.0050 | 13,414 | 0.088 |
| 1.0 | 0.0007*** | 3.71 | 0.0046 | 13,414 | 0.084 |
| Controls | | | Yes | | |
| SIC-2 fixed effects | | | Yes | | |
| Standard-error clustering | | | SIC-2 | | |

This table reports temperature-robustness tests for the main archival SIC-2 ROA recommendation result. The dependent variable is *Recommendation Rate*, defined as the fraction of 100 stochastic generations in which firm $i$ appears in OLMo-2-Instruct's Top-10 open-ended SIC-2 ROA recommendation list at the indicated temperature. Each row reports a separate firm-level OLS regression of *Recommendation Rate* on $Ln(1+Mentions)$, the full control set, and SIC-2 fixed effects. $t$-statistics are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.10 **In-Industry Picks and Out-of-Industry Intrusions (Open-Ended General Top-10)**

**Panel A: Cumulative Pick Counts**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|---|
| Dep. Var. | Cross-SIC Pick Count | | | Own-SIC Pick Count | | | Total Pick Count | | |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| *Ln(1+Mentions)* | 0.0020*** | 0.0058*** | 0.0121*** | 0.0011*** | 0.0019*** | 0.0021*** | 0.0031*** | 0.0078*** | 0.0142*** |
| | (4.35) | (4.72) | (5.76) | (3.19) | (6.29) | (6.03) | (4.57) | (5.40) | (6.05) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.042 | 0.073 | 0.081 | 0.052 | 0.084 | 0.090 | 0.064 | 0.087 | 0.090 |

**Panel B: Opportunity-Normalized Rates**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. Var. | Cross-SIC Pick Rate | | | Own-SIC Pick Rate | | |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| *Ln(1+Mentions)* | 0.000029*** | 0.000022*** | 0.000029*** | 0.001078*** | 0.001928*** | 0.002086*** |
| | (4.35) | (4.72) | (5.76) | (3.19) | (6.27) | (6.02) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,342 | 13,268 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.042 | 0.073 | 0.081 | 0.052 | 0.085 | 0.092 |

**Panel C: Output Composition (Share and Any Cross-SIC Pick)**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. Var. | Cross-SIC Share | | | Any Cross-SIC Pick | | |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| *Ln(1+Mentions)* | 0.0011*** | 0.0021*** | 0.0027*** | 0.0013*** | 0.0025*** | 0.0032*** |
| | (4.72) | (5.47) | (6.33) | (4.66) | (5.95) | (6.61) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.042 | 0.087 | 0.107 | 0.051 | 0.095 | 0.116 |

Table OA.11 **Prompt-Design Robustness for Falsification Test (Instruct Candidate-List Prompts)**

**Panel A: OLMo Recommendation Indicator (Instruct Candidate-List Prompt)**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. Var. | OLMo Top-10 Recommendation (Candidate-List Prompt) | | | | | |
| Query | General | General | General | ROA | ROA | ROA |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| *Ln(1+Mentions)* | 0.0014** | 0.0025** | 0.0033*** | 0.0014** | 0.0022** | 0.0033*** |
| | (2.12) | (2.56) | (3.77) | (2.08) | (2.29) | (3.82) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.159 | 0.232 | 0.230 | 0.147 | 0.219 | 0.214 |

**Panel B: Actual Top-10 by Benchmark**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dep. Var. | Actual Top-10 Indicator | | | | | |
| Benchmark | Market Cap | Market Cap | Market Cap | ROA | ROA | ROA |
| SIC level | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| *Ln(1+Mentions)* | 0.0018** | 0.0013 | 0.0015 | -0.0000 | 0.0015 | 0.0015 |
| | (2.30) | (1.22) | (1.59) | (-0.05) | (1.62) | (1.52) |
| Observations | 13,410 | 13,406 | 13,400 | 13,410 | 13,406 | 13,400 |
| Controls | Yes | Yes | Yes | Yes | Yes | Yes |
| SIC FE | SIC-2 | SIC-3 | SIC-4 | SIC-2 | SIC-3 | SIC-4 |
| Adj. $R^2$ | 0.321 | 0.576 | 0.638 | 0.179 | 0.429 | 0.493 |

Table OA.12 **Synthetic Firm Verification**

| File | Firms | Clean tickers |
|---|---|---|
| Initial verified list | 415 | 415 |
| Added replacement firms | 22 | 22 |
| Canonical verified-v2 list | 437 | 437 |

Table OA.13 **Token-Masked Perplexity Diagnostics**

| Group | Base full name | CPT full name | Base ticker | CPT ticker |
|---|---|---|---|---|
| Treated fake, mean | 24,190.8 | 300.1 | 586,572.4 | 893,616.2 |
| Treated fake, median | 6,484.0 | 133.0 | 14,856.1 | 1,124.9 |
| Control real, mean | 88,466.2 | 134,798.5 | 299,808.4 | 227,538.6 |
| Control real, median | 3,682.1 | 5,460.1 | 7,526.8 | 6,488.9 |

Table OA.14 **Synthetic-Firm Recommendation Probabilities for Additional Cutoffs**

This table reports additional recommendation cutoffs for the CPT exposure experiment. The main-text table reports the Top-10 recommendation outcome to align with the archival specification. The Any column allows a synthetic firm to be recommended at any reported cutoff; Top-1 and Top-5 restrict the outcome to the corresponding cutoff. "Synthetic firms recommended" counts unique synthetic firms recommended at least once under any reported cutoff. The denominator is the 437-firm synthetic sample.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | | | | Probability a Synthetic Firm Is Recommended | | |
| Model | Docs | Total injected tokens | Synthetic firms recommended | Any | Top 1 | Top 5 |
| OLMo-2-Instruct baseline | 0 | 0 | 0 / 437 | 0.0% | 0.0% | 0.0% |
| Base + 0.1% CPT | 3 | 674,386 | 0 / 437 | 0.0% | 0.0% | 0.0% |
| Base + 0.5% CPT | 14 | 3,059,201 | 1 / 437 | 0.2% | 0.0% | 0.0% |
| Base + 1% CPT | 29 | 6,356,255 | 0 / 437 | 0.0% | 0.0% | 0.0% |
| Base + 2% CPT | 57 | 12,472,945 | 1 / 437 | 0.2% | 0.2% | 0.2% |
| Base + 3% CPT | 86 | 18,855,507 | 7 / 437 | 1.6% | 0.5% | 1.4% |
| Base + 4% CPT | 114 | 25,002,476 | 11 / 437 | 2.5% | 0.9% | 2.1% |
| Base + 5% CPT | 143 | 31,367,226 | 17 / 437 | 3.9% | 1.4% | 2.3% |
| Base + 10% CPT | 285 | 62,470,206 | 27 / 437 | 6.2% | 2.1% | 3.7% |
| Base + 20% CPT | 571 | 125,315,290 | 38 / 437 | 8.7% | 4.3% | 5.9% |

Table OA.15 **MMLU Category Scores by Model**

| Model | Overall | Humanities | Social sciences | STEM | Other |
|---|---|---|---|---|---|
| OLMo-2-Instruct baseline | 75.9% | 73.1% | 85.8% | 66.3% | 80.2% |
| Base + 0.1% CPT | 74.6% | 70.1% | 85.3% | 65.9% | 79.4% |
| Base + 0.5% CPT | 74.7% | 70.9% | 85.4% | 65.0% | 79.7% |
| Base + 1% CPT | 74.4% | 70.4% | 85.3% | 64.7% | 79.4% |
| Base + 2% CPT | 74.9% | 71.0% | 85.4% | 65.3% | 80.1% |
| Base + 3% CPT | 74.8% | 70.6% | 85.4% | 65.7% | 79.8% |
| Base + 4% CPT | 74.8% | 70.6% | 85.4% | 65.6% | 79.9% |
| Base + 5% CPT | 74.8% | 70.6% | 85.8% | 65.6% | 79.8% |
| Base + 10% CPT | 75.0% | 71.2% | 85.9% | 65.2% | 79.8% |
| Base + 20% CPT | 75.1% | 71.3% | 85.8% | 65.8% | 79.7% |

Table OA.16 **Additional Analysis: Negative Stock Recommendations**

**Panel A: Baseline Negative-Recommendation Result**

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | *Negative Recommend* | | *Recommend* |
| Specification | SIC-2 FE only | Main controls | Main controls |
| $Ln(1+Mentions)$ | 0.190*** | 0.090*** | 0.100*** |
| | (3.85) | (2.81) | (3.01) |
| Observations | 13,410 | 13,410 | 13,410 |
| Controls | No | Yes | Yes |
| SIC-2 FE | Yes | Yes | Yes |
| Adj. $R^2$ | 0.037 | 0.056 | 0.056 |

**Panel B: Negative-Recommendation Error Decomposition**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dep. Var. | *Negative Recommend* | *Actual Bottom ROA* | *Wrong* | *Missed* | *True Positive* |
| $Ln(1+Mentions)$ | 0.090*** | -0.070 | 0.100*** | -0.070 | -0.000 |
| | (2.81) | (-1.11) | (2.85) | (-1.08) | (-0.98) |
| Observations | 13,410 | 13,410 | 13,410 | 13,410 | 13,410 |
| Controls | Yes | Yes | Yes | Yes | Yes |
| SIC-2 FE | Yes | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.056 | 0.237 | 0.056 | 0.235 | 0.096 |

This table mirrors the main SIC-2 ROA recommendation design but asks for negative recommendations. *Negative Recommend* equals one if firm $i$ appears in OLMo-2-Instruct's Top-10 least-attractive-firms response for firm $i$'s own SIC-2 industry. *Actual Bottom ROA* equals one if firm $i$ is among the actual bottom-10 firms by ROA in its SIC-2 industry. In Panel B, *Wrong* equals one when the model names a firm that is not actually bottom-10 by ROA; *Missed* equals one when a firm is actually bottom-10 by ROA but is not named; and *True Positive* equals one when both conditions hold. Indicator dependent variables are multiplied by 100, so coefficients are in percentage points. Column (3) of Panel A reports the corresponding positive-recommendation estimate from the main specification for comparison. The representative prompt is: "First, filter the United States stock market to ONLY include companies within the 'Business Services' (SIC: 73) classification. Then, from that specific industry pool, identify the 10 least attractive companies to invest in based on Return on Assets (EBIT/total assets)." Detailed variable definitions are provided in Appendix A. $t$-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.17 **Additional Analysis: Real-Firm Non-Stock Recommendations**

| | (1) | (2) | (3) |
|---|---|---|---|
| Dep. Var. | | *Recommend* | |
| Cut-off | Top-1 | Top-5 | Top-10 |
| *Acquisition target* | 0.0002* | 0.0006** | 0.0008** |
| | (1.91) | (2.32) | (2.55) |
| *Business partner* | 0.0003** | 0.0008*** | 0.0010*** |
| | (2.27) | (2.66) | (2.87) |
| *Creditworthiness* | 0.0002** | 0.0005** | 0.0010*** |
| | (2.17) | (2.22) | (3.07) |
| *ESG / reputation* | 0.0003** | 0.0006** | 0.0010*** |
| | (2.21) | (2.43) | (2.93) |
| *Employer attractiveness* | 0.0003*** | 0.0009*** | 0.0011*** |
| | (2.79) | (3.66) | (3.73) |
| *Further research* | 0.0003** | 0.0007*** | 0.0008** |
| | (2.29) | (2.81) | (2.49) |
| *Media attention* | 0.0003*** | 0.0007*** | 0.0010*** |
| | (2.80) | (2.99) | (3.02) |
| *Product / vendor* | 0.0003** | 0.0008*** | 0.0011*** |
| | (2.28) | (3.33) | (3.13) |
| Controls ($SET_MAIN) | Yes | Yes | Yes |
| SIC-2 FE | Yes | Yes | Yes |
| Cluster | SIC-2 | SIC-2 | SIC-2 |
| Observations | 13,410 | 13,410 | 13,410 |
| Adj. $R^2$ (avg. across tasks) | 0.012 | 0.050 | 0.061 |

This table extends the archival design in Table 4 to eight non-stock recommendation tasks. Each cell reports the coefficient on *Ln(1+Mentions)* from a separate OLS regression. The dependent variable equals one if real firm *i* appears in OLMo-2-Instruct's response for firm *i*'s own SIC-2 industry under the indicated task and cutoff. All specifications include the full main control set and SIC-2 fixed effects. The Top-10 prompt template is: "Among publicly traded U.S. companies in the SIC-2 industry {industry name} (SIC-2: {sic2}), recommend the top 10 based on the companies that would be most {task criterion}. Output strictly as a JSON array of ticker symbols." The task criteria are: attractive acquisition targets for a strategic buyer; attractive as reliable business partners, suppliers, or customers; safest and most creditworthy for a lender to extend credit to; responsible, sustainable, reputable, and trustworthy; attractive employers for a skilled professional; useful for an investor or analyst to research further; deserving of additional media attention or business press coverage; and attractive for a potential customer or enterprise buyer. Detailed variable definitions are provided in Appendix A. *t*-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table OA.18 **Synthetic-Firm Non-Stock Recommendations: Pooled Summary**

This table reports the CPT20 treatment effect on the probability that OLMo-2 recommends at least one synthetic firm in non-stock firm-evaluation prompts. Rows report separate task-level regressions; the pooled row absorbs task fixed effects. Each task-cutoff cell contains 140 prompts. The prompts use the same task templates reported in Table OA.17. Detailed variable definitions are provided in Appendix A. $t$-statistics in parentheses are based on standard errors clustered by SIC-2 industry. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

| | (1) | (2) | (3) |
|---|---|---|---|
| Cut-off | Top-1 | Top-5 | Top-10 |
| *Acquisition target* | 0.286*** | 0.457*** | 0.529*** |
| | (5.23) | (7.60) | (8.76) |
| *Business partner* | 0.257*** | 0.357*** | 0.543*** |
| | (4.87) | (6.17) | (9.02) |
| *Creditworthiness* | 0.229*** | 0.343*** | 0.400*** |
| | (4.51) | (5.98) | (6.76) |
| *ESG / reputation* | 0.257*** | 0.357*** | 0.429*** |
| | (4.87) | (6.17) | (7.17) |
| *Employer attractiveness* | 0.257*** | 0.371*** | 0.443*** |
| | (4.87) | (6.36) | (7.38) |
| *Further research* | 0.257*** | 0.400*** | 0.414*** |
| | (4.87) | (6.76) | (6.96) |
| *Media attention* | 0.243*** | 0.429*** | 0.557*** |
| | (4.69) | (7.17) | (9.28) |
| *Product / vendor* | 0.300*** | 0.414*** | 0.471*** |
| | (5.42) | (6.96) | (7.82) |
| **Pooled (task FE)** | 0.261*** | 0.391*** | 0.473 |
| | (5.46) | (7.90) | |
| Task FE | No | No | No |
| Cluster | SIC-2 | SIC-2 | SIC-2 |
| Obs. per task-row | 140 | 140 | 140 |
| Obs. pooled | 1,120 | 1,120 | 1,120 |
| Adj. $R^2$ pooled | 0.145 | 0.240 | 0.310 |

Table OA.19 **Non-Stock recommendation: Acquisition Target**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 45 / 437 | 10.3% | 2.7% | 7.1% | 7.8% |

Table OA.20 **Non-Stock recommendation: Business Partner**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 38 / 437 | 8.7% | 2.5% | 5.3% | 8.5% |

Table OA.21 **Non-Stock recommendation: Creditworthiness**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 35 / 437 | 8.0% | 2.5% | 5.3% | 6.9% |

Table OA.22 **Non-Stock recommendation: ESG and Reputation**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 38 / 437 | 8.7% | 3.0% | 5.7% | 6.4% |

Table OA.23 **Non-Stock recommendation: Employer Attractiveness**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 35 / 437 | 8.0% | 2.7% | 5.7% | 7.3% |

Table OA.24 **Non-Stock recommendation: Further Research**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 41 / 437 | 9.4% | 2.5% | 6.4% | 6.6% |

Table OA.25 **Non-Stock recommendation: Media Attention**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 42 / 437 | 9.6% | 2.3% | 5.7% | 8.0% |

Table OA.26 **Non-Stock recommendation: Product and Vendor**

| Number | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | Probability a Synthetic Firm Is recommended | | | |
| Model | Fake firms recommended | Any | Top 1 | Top 5 | Top 10 |
| OLMo-2-Instruct baseline | 0 / 437 | 0.0% | 0.0% | 0.0% | 0.0% |
| Base + 20% CPT | 40 / 437 | 9.2% | 3.2% | 7.1% | 7.1% |