Human–AI Collaboration in Corporate Valuation: Experimental Evidence with a Valuation AI Agent *

Huan Liu† Miao Liu‡ Zhizhe Liu§ Danqing Mei¶

March, 2026

Abstract

In an era where AI can deliver increasingly sophisticated hard information, we study how AI can facilitate human soft information production and make human–AI collaboration in corporate valuation more productive. We develop a retrieval-augmented AI agent that reads financial filings and produces interactive dashboards and valuation analytics. We embed an experiment in an advanced business course in which students use the AI agent to value real firms, varying the amount of AI-supplied hard information across three interfaces: bare retrieval (Low-Hard), dashboards that summarize financials but do not propose a valuation (Medium-Hard), and dashboards plus AI-generated valuations (High-Hard). Using full chat logs and valuation memos from participants, we construct rich measures of soft information and relate them to ex ante financial statement analysis (FSA) knowledge. We find a non-monotonic effect of AI-supplied hard information: relative to Low-Hard, dashboards in Medium-Hard substantially increase soft information production, whereas adding AI valuations in High-Hard yields much smaller incremental gains and appears to induce anchoring. FSA knowledge amplifies the benefits of dashboards and mitigates, but does not eliminate, crowding out in High-Hard. Our results clarify when AI acts as a complementary hard-information engine versus a substitute for human soft information production, and offer guidance for the design of valuation tools and curricula in the LLM era.

Keywords: Human-AI Collaboration; Corporate Valuation; Hard and Soft Information; AI Agent

*We are grateful to several industry practitioners for their insights on the design and implementation of the AI agent, and to Google Cloud for providing credits and technical support that enabled us to host the AI agent on their cloud platform for academic research purposes. Wenqian Chen, Balint Czaha, Patrick Lanza, and Arnav Rastogi provided excellent assistance in stress testing the agent.

†Google, AI Agent unit: lhuan@google.com

‡Carroll School of Management, Boston College: miao.liu@bc.edu

§University of Wisconsin–Madison: zhizhe.liu@wisc.edu

¶Cheung Kong Graduate School of Business: dqmei@ckgsb.edu.cn

1 Introduction

A central question in contemporary finance is how humans and artificial intelligence (AI) should share the work of processing information. Corporate valuation is a canonical setting in which information is heterogeneous. Financial statements and market data are "hard"—quantitative, easily stored, and readily transmissible, whereas management narratives and strategic positioning are "soft," reflecting context-dependent information that is difficult to fully codify. Recent work documents that AI excels at processing and "hardening" codifiable information, including textual disclosures that can be mapped into quantitative features, while humans remain indispensable for generating context-rich insight (Costello et al., 2020; Liu, 2022; Cao et al., 2024b). These developments naturally raise a question for valuation practice and financial accounting education: in an era where AI can deliver increasingly sophisticated hard information, when does AI complement rather than crowd out human soft information production, and what forms of human–AI collaboration best leverage each side's comparative advantage?

In this study, we develop and deploy an AI valuation agent, a retrieval-augmented large language model (LLM), that draws on the same information environment relevant for real-world equity analysis: firms' annual and quarterly reports, earnings call transcripts, a library of precomputed financial ratios, macroeconomic time series, and real-time market data. We roll out this AI assistant in an experiment with advanced business students conducting valuations of real firms, using the system as a decision-support tool while they work through an authentic valuation task. The AI agent is built as a set of modular "hard-information" capabilities that can be switched on or off depending on treatment. Because the experimental variation operates through feature availability, holding the underlying data universe fixed, we can exogenously vary both the amount and the structure of hard information that the AI makes accessible to the human valuer. At the same time, the conversational interface passively records the complete interaction history, such as participants' queries, follow-ups, and narrative justifications, allowing us to directly observe the soft information they generate as they interpret, select, and

synthesize the available inputs. This deployment yields two key advantages: (i) precise experimental control over the hard-information set delivered by the AI via randomized module access; and (ii) granular, process-level measurement of soft information production during valuation.

We use this setting to provide early experimental evidence on how AI-supplied hard information shapes soft information humans generate when valuing real firms. We conceptualize "hard information" as codified, verifiable inputs that the AI can retrieve and structure, and "soft information" as the narrative, interpretive content that humans generate by querying the AI, integrating prior knowledge, and forming judgments (Stein, 2002; Liberti and Petersen, 2019). In our framework, the effect of AI-mediated hard information on soft information production is ambiguous ex ante and likely depends on prior valuation knowledge. By reducing the cost of locating and organizing relevant evidence, AI can free users to devote more effort to interpretation, synthesis, and narrative construction. Users with stronger valuation knowledge may be especially able to convert this structured evidence into higher-quality independent reasoning. As AI support becomes more analytically substantive, however, its effect need not move in a single direction: additional guidance may help users draw sharper valuation inferences, but it may also narrow independent reasoning if users adopt the system's framing too readily. Which force dominates is likely to depend in part on users' existing valuation knowledge. The overall effect of richer AI support is therefore an empirical question.

These hypotheses map directly into our experimental arms. In the Low-Hard arm, participants interact with a basic retrieval-augmented agent that answers questions and surfaces excerpts from filings but does not provide any pre-packaged dashboards or analytics except for basic business descriptions retrieved from Item 1 from 10-Ks. In the Medium-Hard arm, participants receive, in addition, AI-generated modules that summarize MD&A and risk-factor sections and display time-series charts and DuPont decomposition of key financials and ratios. These modules are descriptive and diagnostic but stop short of delivering an overall valuation. In the High-Hard arm, we further activate modules that provide peer-relative diagnostics and multiple-based analysis. We complement this experimental design with a measure of students'

ex ante financial statement analysis (FSA) knowledge based on a pre-experiment quiz that covers ratio analysis, DuPont decomposition, accruals and cash flows, and common valuation models. We view this knowledge as a set of internal analytical tools that govern how individuals interpret the same external data.

The experimental design follows established best practices from the information-provision and survey-experiment literatures (Haaland et al., 2023; Stantcheva, 2023). We randomly assign 126 advanced business students to one of three AI interface conditions (Low-Hard, Medium-Hard, High-Hard), stratifying on baseline FSA score. We also randomly assign 20 companies spanning a broad range of sectors so that each firm is evaluated by at least two students in each treatment arm. This balanced assignment is important because it allows us to include firm fixed effects in all regressions, thereby absorbing cross-firm differences in information environment richness, business-model complexity, and the baseline ease with which a firm may lend itself to soft information production. Across arms, we hold the underlying information set fixed, including 10-K filings from the most recent five years, the most recent earnings call transcript, macro series, price and trading-volume histories, and a library of precomputed ratios, so that treatments differ only in how the AI structures and foregrounds this information. This "information-equivalent but differently structured" design follows Haaland et al. (2023)'s recommendation to separate the effects of information content from the effects of framing and aggregation in information-provision experiments. Participants are incentivized to produce accurate valuations and detailed written rationales, and they are free to query the AI agent as much as they wish within a fixed time budget for each firm.

A central challenge in the hard- and soft-information literature is that soft information is economically important but rarely observed directly. In most archival settings, researchers must infer it from indirect proxies such as distance, organizational structure, relationship intensity, or the residual component of decisions unexplained by observable covariates (Agarwal and Hauswald, 2010; Liu, 2022). An advantage of our design is that it allows us to observe soft information as it is produced in real time through the full chat transcript between the

participant and the AI agent, together with the participant's final valuation memo. Building on modern text-as-data and open-ended-response methods (Gentzkow et al., 2019; Haaland et al., 2025), we use trained human coders, following pre-specified rubrics, to code the primary dimensions of soft information production from the full conversation trace, including follow-up depth, hypothesis testing, cross-period synthesis, multi-metric synthesis, reasoning depth, and mechanism-seeking behavior. These measures capture how participants transform AI-supplied hard information into human-authored interpretation by refining lines of inquiry, evaluating competing explanations, integrating evidence across time and metrics, developing independent reasoning, and probing underlying mechanisms.

While human coding is well suited to capturing nuanced, context-dependent dimensions of soft information production, it is necessarily resource-intensive and may raise concerns about coder discretion at the margin. We therefore use two additional and distinct measurement approaches. First, we construct LLM-coded measures of hypothesis-space breadth and analytical progression. These measures capture whether participants consider multiple plausible explanatory paths and whether their analysis moves from raw facts toward increasingly synthesized, valuation-relevant judgments. Second, building on recent work that represents economic narratives and textual reasoning as structured causal or knowledge graphs (Andre et al., 2026; Li et al., 2026), we construct a largely deterministic DAG representation of evidence-to-interpretation reasoning. A DAG, or directed acyclic graph, traces how specific pieces of evidence feed into intermediate interpretations and ultimately into valuation conclusions. This multi-method design is important because soft information is inherently latent and no single proxy is beyond dispute. Consistent inferences across human coding, deterministic algorithms, LLM coding, and graph extraction materially strengthen construct validity.

Turning to the empirical results, we first find that the effects of AI-supplied hard information on soft information production are strongest in the Medium-Hard condition. Relative to Low-Hard, the Medium-Hard interface significantly increases our primary human-coded Soft Information measures, as well as complementary LLM-coded and DAG-based measures of rea-

soning. By contrast, the High-Hard interface delivers weaker and less consistent gains. These patterns suggest that AI is most effective at supporting human soft information production when it structures and surfaces relevant hard information without taking over the interpretive task. Consistent with this interpretation, we also find that the High-Hard arm leads to substantially greater delegation, which we measure using human-coded indicators of whether participants offload substantive analytical work to the AI and the extent to which AI-generated delegated content carries into the final memo, whereas the Medium-Hard arm does not. Taken together, these results suggest that more prescriptive AI support can improve some dimensions of analysis while also crowding out independent human interpretive effort.

Second, we show that ex ante FSA knowledge meaningfully moderates the effect of AI, especially when the tool provides more prescriptive support. In the interaction specifications, the effect of FSA is steepest in the High-Hard arm for the human-coded Soft Information Index and for both LLM-coded measures, while the corresponding effect in the Medium-Hard arm is smaller and statistically weaker. We interpret this pattern as evidence that highly prescriptive AI is expertise-amplifying rather than expertise-equalizing. Participants with stronger FSA backgrounds appear better able to refine and selectively use AI-generated valuations, converting them into additional human-authored soft information, so the same tool can widen dispersion in soft information production. This interpretation is consistent with recent evidence from other domains showing that generative AI can have uneven effects across users, benefiting stronger baseline performers while harming or failing to help weaker ones, that AI closes expertise gaps only up to a point, and that heavy delegation can weaken independent learning (Otis et al., 2025; Vendraminelli et al., 2025; Shen and Tamkin, 2026).

Next, we open the "black box" of human-AI collaboration by leveraging the granularity of our data. Because we observe the full prompt-response trace and can map individual pieces of evidence into structured reasoning objects, we can study soft information production at the evidence-node, reasoning-edge, and query levels. At the evidence-node level, each observation is a student–firm–evidence-type cell, where an evidence type is a concrete hard-information

source such as MD&A, risk factors, DuPont diagnostics, or earnings-quality signals. Relative to Low-Hard, the Medium-Hard interface increases mentions of MD&A and risk factors and, more importantly, increases the extent to which these evidence nodes are converted into support edges, that is, explicit links from evidence to interpretation. High-Hard preserves some of this narrative-evidence processing, especially for MD&A and risk factors, but converts these evidence nodes into reasoning edges less than Medium-Hard. This pattern suggests that Medium-Hard does not merely prompt more discussion of narrative disclosures; it is more effective at helping users translate them into valuation-relevant reasoning.

Finally, at the query level, we ask which AI responses stimulate the next increment of human-authored soft information. Exploiting within-student variation, we find that the most productive responses are those that organize relevant evidence while leaving key causal or mechanistic links unresolved. When the AI identifies a pattern without fully explaining its source, or offers a conclusion without fully tracing the operative channel, users respond with deeper follow-up, more hypothesis testing, and more mechanism-seeking. This pattern points to a clear design principle: AI is most complementary when it organizes the evidentiary base without collapsing the inferential task that remains for the human analyst.

We make three contributions. First, we add to the literature on hard and soft information in accounting and finance by bringing a tightly controlled field experiment to a setting where both types of information are central but notoriously difficult to isolate. Prior research has relied on proxies such as geographic distance, organizational structure, or work patterns to study soft-information acquisition, for example in small-business lending (Agarwal and Hauswald, 2010), equity research (Ben-Rephael et al., 2023), and portfolio choice (Chen et al., 2024). Recent work in credit markets uses machine-learning models as benchmarks to decompose decisions into a hard-information component and a residual interpreted as soft information (Liu, 2022). Yet two challenges remain: hard-information provision is typically endogenously chosen, and soft information is almost always measured indirectly via noisy proxies. Our custom-built AI valuation agent directly addresses both challenges: by randomizing which hard-information

modules the agent exposes, we causally identify how changes in hard-information structure affect human behavior; and by logging the full interaction history and written output, we obtain a rich, process-based measure of soft information, complementing and extending the observational evidence in the existing hard-versus-soft information literature.

Second, we speak to the broader human–AI collaboration literature by studying a setting where the core design problem is not just whether humans follow or reject algorithmic advice, but *who produces what information*. Prior work documents both "algorithm aversion" and "algorithm appreciation" (Dietvorst et al., 2015; Logg et al., 2019), and shows that naively constructed human–AI teams often fail to outperform the better agent alone (Green and Chen, 2019, 2021; Bansal et al., 2021a), even though complementarity is the central design goal (Hemmer et al., 2024; Vaccaro et al., 2024). We add to this literature by examining a complex decision problem, corporate valuation, in which AI and humans naturally specialize. We show how the structure of AI hard-information output shapes both reliance on AI and the balance of hard versus soft information produced within the human-AI team.

Third, our study informs the emerging discussion on how to train accountants and finance professionals in an AI-intensive environment. Policy reports and recent descriptive evidence show that GenAI is rapidly becoming a standard tool for investors and that more sophisticated users disproportionately exploit its strengths (OECD, 2021; Kelly and Xiu, 2023; Cao et al., 2024b; Blankespoor et al., 2024). Experimental work further documents that GenAI can attenuate some traditional processing-cost effects while leaving others intact (Croom et al., 2025), and that highly interactive GenAI interfaces can inflate investors' self-assessed ability and willingness to invest without improving actual processing (Croom, 2025). In parallel, accounting education research stresses the need to integrate emerging technologies into the curriculum while preserving core analytical and critical-thinking skills (Augustine et al., 2020; Dalwai et al., 2021; Raschke and Charron, 2021; Ballantine et al., 2024). By measuring students' ex ante FSA knowledge and observing how they use an AI valuation agent, we provide experimental evidence on how traditional "hard" skills interact with GenAI and find that stronger FSA knowledge is

associated with richer soft information production and more effective use of AI-supplied analytics, whereas weaker students are more prone to accept AI analytics uncritically. These findings suggest that accounting and finance education should not retreat from teaching foundational analysis in the face of GenAI, but instead reposition those skills as complements.

2 Hypotheses Development

2.1 AI-Supplied Hard Information and Human soft information production

Our setting builds on the hard–soft information framework (Stein, 2002; Liberti and Petersen, 2019). In our environment, the AI agent is designed to be a producer and organizer of hard information. It retrieves numerical fundamentals and ratios from Compustat and related WRDS feeds, surfaces text from 10-K filings and earnings call transcripts, and calls specialized analyzers that precompute DuPont decompositions, earnings-quality diagnostics, risk taxonomies, and peer benchmarks. The surrounding orchestration layer ensures that all answers are grounded in verifiable data and auditable retrieval traces rather than model "memory". Conceptually, this makes the AI analogous to a junior analyst that handles the heavy lifting of reading and structuring disclosures, while the human user sits "on top" of this system and decides which pieces of information to acquire, how to connect them, and how to synthesize them into a valuation narrative. We view the AI-supplied modules as hard-information primitives that can either (i) lower the cost of discovering relevant facts and patterns, thereby *facilitating* soft information production, or (ii) substitute for human effort when the AI not only structures but also interprets the data, potentially *crowding out* human narrative work.

In our Medium-Hard condition, the AI agent provides structured but largely descriptive content: MD&A and risk-factor summaries, multi-year charts of revenues, margins, and leverage, DuPont decompositions, and earnings quality diagnostics, but *without* a valuation assessment. We therefore expect that, relative to the Low-Hard condition (in which students must discover

patterns via queries), the Medium-Hard condition will lead students to ask more targeted questions, synthesize a broader set of valuation-relevant topics, and test more specific hypotheses. Formally:

> **H1a (Facilitation).** Relative to the Low-Hard arm, the Medium-Hard arm increases human-produced soft information, as structured hard-information modules lower search and processing costs.

In our High-Hard condition, the AI agent additionally produces peer-relative diagnostics and multiple-based valuations. A broad literature on human–algorithm interaction documents that people sometimes overweight algorithmic advice relative to human advice ("algorithm appreciation") (Logg et al., 2019), but can also become averse to algorithms after observing errors, even when algorithms outperform humans on average ("algorithm aversion") (Dietvorst et al., 2015). In both cases, the algorithmic recommendation becomes a focal anchor, reshaping how much effort individuals invest in independent evaluation. Complementary evidence from cognitive psychology and neuroscience shows that humans regularly "offload" cognitive work onto external tools when doing so reduces perceived effort, a phenomenon known as cognitive offloading (Risko and Gilbert, 2016), and that the brain trades off the expected benefits and costs of exerting control when deciding how much effort to invest in a task (Shenhav et al., 2013). When the AI provides a valuation, students may perceive additional interpretive effort as having low incremental payoff, and thus invest less in generating their own scenarios and narratives. This logic yields:

> **H1b (Crowding out).** Relative to the Medium-Hard arm, the High-Hard arm produces smaller incremental effects on human-produced soft information, as AI-generated valuations encourage anchoring on AI outputs and reduce human interpretive effort.

2.2 Human Expertise and the Production of Soft Information

A second dimension of our framework concerns the role of human expertise. We measure participants' ex ante FSA knowledge using a quiz that covers core topics such as ratio analysis, DuPont decomposition, accruals and cash flows, and common valuation models. We conceptualize this knowledge as an internal stock of hard analytical tools that shape how individuals interpret external data. Theoretical models and empirical evidence suggest that decision makers with greater processing capacity and domain knowledge are better able to extract value from a given hard-information set and to deploy costly soft-information acquisition more selectively (Stein, 2002; Liberti and Petersen, 2019). Recent work in credit markets using machine-learning benchmarks shows that humans add value by selectively deviating from algorithmic recommendations when they possess information or expertise not captured by the model (Costello et al., 2020; Liu, 2022). In equity analysis, Ben-Rephael et al. (2023) and Chen et al. (2024) show that analysts and portfolio managers who invest more in both hard and soft information collection, via office work and travel, respectively, produce more informative recommendations and superior performance.

Translating this logic to an AI-augmented valuation setting, we expect participants with stronger FSA knowledge to be better positioned to convert AI-supplied hard information into high-quality soft information across all treatment arms. Students with stronger FSA training should be better able to recognize which accounting and operating patterns are economically meaningful, connect those patterns to underlying business drivers, and ask follow-up questions that probe the persistence, comparability, and valuation relevance of observed signals. For example, they are more likely to understand why a change in margins, asset turnover, accruals, or working-capital intensity matters for competitive positioning, earnings quality, and expected performance. In this sense, FSA knowledge should complement AI by improving the user's ability to transform the same underlying information into richer soft information, consistent with the view that AI in finance raises the returns to domain expertise rather than simply substituting for it (Kelly and Xiu, 2023; Cao et al., 2024b). This yields:

H2a (Expertise and soft information). Ex ante FSA knowledge is positively associated with soft information production.

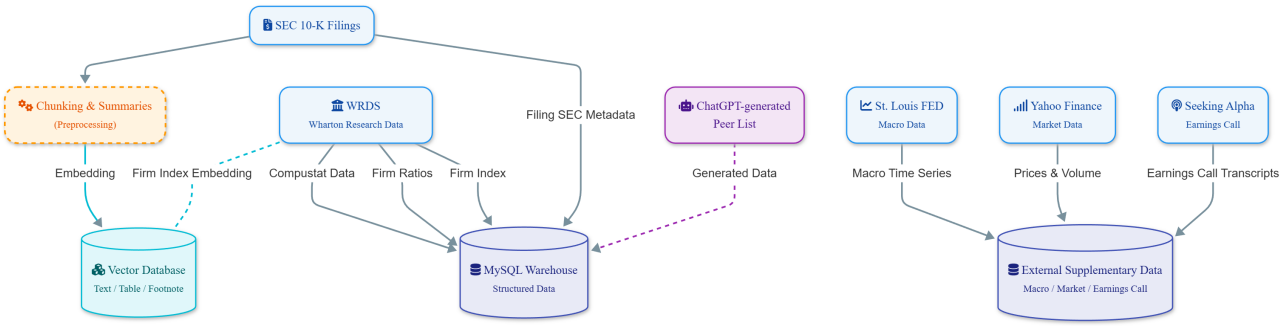
The moderating role of expertise should be most pronounced in the most prescriptive AI condition. In that setting, the system does not merely organize evidence; it also provides peer-based valuation guidance and multiple-based comparisons. For participants with weaker FSA knowledge, such guidance is more likely to substitute for independent reasoning: they may accept the AI's comparability logic or valuation framing without evaluating whether the selected peers and multiples are appropriate, or whether differences in growth, margins, and risk require adjustment. Participants with stronger FSA knowledge, by contrast, are better able to interrogate, qualify, and extend the AI's suggestions, using them as inputs into their own analysis rather than as endpoints:

H2b (Expertise and crowding out). In the most prescriptive AI condition, higher FSA knowledge attenuates the crowding-out effect of AI-generated valuation guidance on soft information production.

3 AI Valuation Agent and Experimental Platform

At the core of our study is a custom-built AI valuation agent designed specifically for corporate valuation tasks. Conceptually, the agent behaves like a junior analyst who can read 10-Ks, pull data from Compustat, compute ratios and valuation metrics, and draft narrative explanations on demand. Technically, it is a large language model (LLM) wrapped in a retrieval and orchestration layer: the LLM handles natural-language understanding and explanation, while a suite of specialized tools feeds it financial data and disclosures and performs mathematical calculations. This design follows the recent evolution from generic "chatbots" toward domain-specific, tool-augmented LLM systems, in which the model is connected to structured databases, document repositories, and analytical engines rather than answering from memory alone (e.g., Lewis et al., 2020; Durante et al., 2024; Alammar and Grootendorst, 2024; Huyen,

Figure 1: Data Infrastructure of the AI Agent



Notes: SEC 10-K filings are preprocessed through chunking and summarization and embedded into a vector database that indexes narrative text, tables, and footnotes for retrieval. In parallel, WRDS supplies Compustat fundamentals, financial ratios, and identifier mappings that are stored in a structured MySQL warehouse, along with SEC filing metadata and an offline, LLM-generated peer list. FRED macro series, Yahoo Finance market data, and Seeking Alpha earnings-call transcripts are ingested into a dedicated supplementary database. Together, these linked data sources provide a controlled, auditable data backbone for retrieval and valuation analysis. Details are provided in Online Appendix B.

2025). The result is an AI assistant that can respond in plain English but whose answers are grounded in the same firm-level data and SEC filings used in empirical accounting and finance research.

As depicted in Figure 1, the agent integrates three types of information central to valuation. First, it connects to a structured MySQL warehouse that stores Compustat annual and quarterly fundamentals, standardized ratio libraries, and peer mappings aligned via WRDS identifiers. Second, it indexes the unstructured text of 10-K filings, including business descriptions, MD&A, risk factors, financial statement notes, into a vector database, allowing the system to retrieve the most relevant passages even when the user's wording does not match the disclosure exactly. Third, it can call external data feeds for macroeconomic series (for example, interest rates from FRED), market data (prices and volumes from Yahoo Finance), and earnings call transcripts (from SeekingAlpha). All of these sources are joined through a consistent identifier system (ticker, CIK, `gvkey`), so that when a user asks a question about "Apple in 2021," the agent deterministically hits the correct structured records and filings.[1]

On top of this data layer, the agent uses retrieval-augmented generation (RAG) to keep the

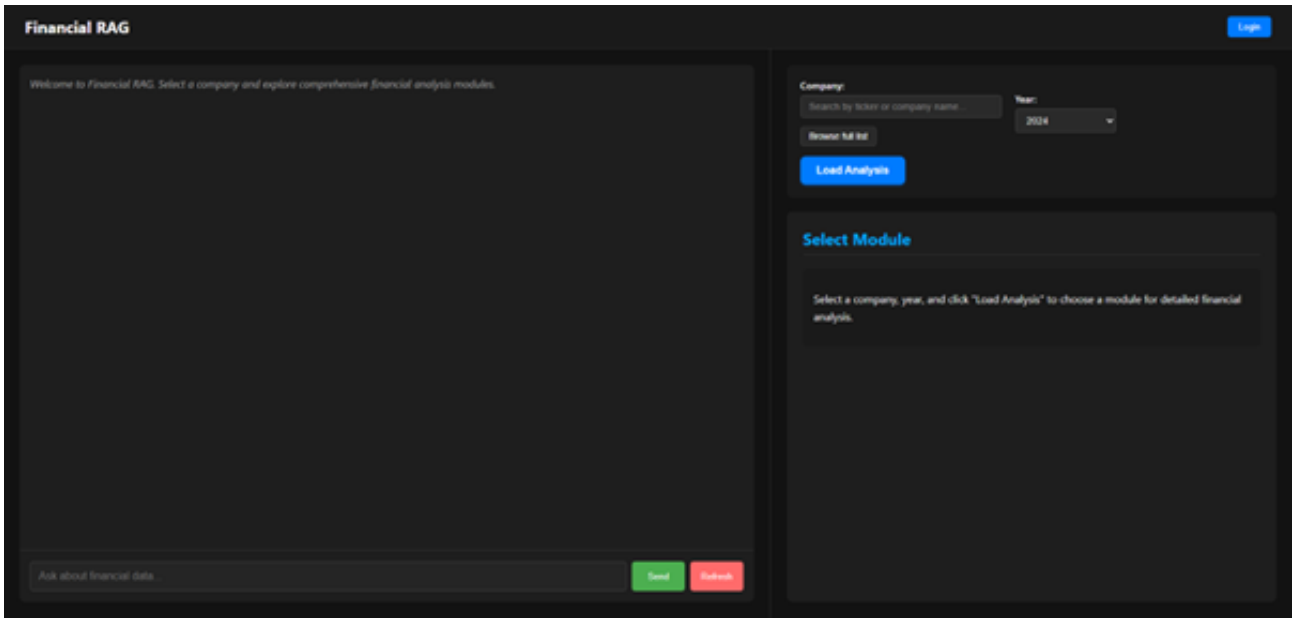
[1]See Online Appendix B for a detailed description of the data infrastructure and orchestration architecture.

LLM's answers fact-based. Rather than letting the model derive and potentially "hallucinate" numbers or disclosure language from its training data, the system first retrieves the relevant tables, MD&A paragraphs, or risk-factor excerpts from the databases and then passes those snippets into the LLM as evidence. The LLM is explicitly instructed to base its answer on this evidence and to cite the underlying sources (for example, "Compustat fundamentals, fiscal year 2022" or "SEC 10-K, Item 1A, 2023"). This RAG architecture allows us to combine the flexibility of modern LLMs with the rigor and auditability required in accounting research: every narrative explanation can be traced back to specific rows in Compustat or specific sections of a filing. Figure 2 shows the user interface of the AI agent.

A key feature of the agent, and one that is crucial for our experiment, is the ability to present hard information in modular "panels" that we can switch on and off by treatment. Offline analyzers precompute a library of firm-year reports, including (i) a business description summary from Item 1, (ii) MD&A narrative summaries and forward-looking themes from Item 7, (iii) a risk taxonomy built from Item 1A, (iv) multi-year financial visualizations (revenues, margins, leverage), (v) full DuPont decompositions of return on equity, (vi) earnings-quality diagnostics that highlight accrual intensity and "beneath-the-line" items, and (vii) peer and multiples benchmarks. Each module is stored as a structured markdown report that the agent can either display directly to the user as a dashboard or inject into the conversational context when the user clicks "Add to chat." Because these modules are generated offline from audited data and filings, their content is fixed and replicable across users and sessions. Experimentally, this modular design allows us to manipulate the amount and structure of hard information the AI surfaces while holding the underlying information set fixed. Low-Hard provides bare-bones retrieval, Medium-Hard adds AI-generated summaries and financial-pattern visualizations without valuation reasoning, and High-Hard further adds AI-generated valuation assessments based on peer diagnostics and multiples.

The second distinctive feature of the agent is its ability to log and reconstruct the entire information-production process. Every query triggers a structured workflow: a router decides

Figure 2: User Interface of the AI Agent



Notes: The left panel shows the chat window where users interact with the agent through natural-language questions and receive responses. The right panel displays the selected company identity (e.g., ticker and fiscal year) and, when available, the set of structured analysis modules (e.g., business description, MD&A, risk factors, financial trends, peer analysis) that users can open and optionally add to the chat as contextual inputs. Details are provided in Online Appendix B.

which tools to call (structured data via Text-to-SQL, 10-K based narrative evidence via RAG, custom mathematical calculations via Text-to-Python, macro and market data via specialized agents), each tool returns a typed input, and the LLM then synthesizes these inputs into a narrative answer under strict citation rules. All of these steps are recorded, including the original user query, the routing plan, the SQL statements sent to the database, the retrieved snippets from filings, the Python code used for mathematical calculations, and the final answer. In addition, the full chat transcript and the student's final valuation memo are stored.

In combination, these features make the AI agent a transparent, instrumented platform for studying how humans and AI collaborate in valuation.[2]

[2]We validate the AI agent's outputs using a series of manual checks described in detail in Online Appendix B.7. In brief, all coauthors and four trained research assistants independently inspected a stratified random sample of 100 user queries drawn from pilot interactions, spanning 8 topical categories (business description, MD&A performance discussion, risk factors, DuPont and ratio diagnostics, earnings-quality modules, peer/multiples analysis, macro/market data, and valuation what-if questions) and 42 distinct firm–year observations. For each query, we verified that (i) all numerical outputs matched the underlying Compustat and SEC sources within standard rounding tolerances, (ii) citations pointed to the correct filings, items, and tables, and (iii) the agent did not hallucinate non-existent filings or ratios outside the SQL/RAG toolchain. We also

4 Experimental Design

4.1 Setting and Participants

Our setting is closely related to the growing field-experimental literature in accounting and finance, but differs in both population and level of control.[3] We implement a between-subjects experiment in an advanced business program that is field-experimental in spirit while retaining lab-like control. Participants are 126 junior, senior, and MBA finance and accounting majors and minors who have completed financial accounting, fundamentals of finance, FSA, and business statistics; many are concurrently in internships or have accepted offers in equity research, banking, or asset management. The experimental task, valuing real firms using actual SEC filings and market data, is therefore very close to the work they perform in professional settings. The experiment is implemented as required in-class valuation exercises that account for 10% of the final course grade, with grades curved *within* each experimental arm to create strong and symmetric effort incentives.

At the same time, our retrieval-augmented AI valuation agent gives the study important lab-like control while preserving a realistic valuation task. We embed the agent directly into students' normal workflow and randomize the structure and richness of the hard information it surfaces while holding the underlying information set fixed. Because we observe the full query-level interaction history and can link narrative claims back to specific tool calls and retrieved evidence, the design allows granular measurement of how participants process information and

catalogued and characterized any failure modes (e.g., missing filings or transcripts) and confirmed that these are surfaced to users as explicit "not available" messages rather than implicit or fabricated content. The results of these checks support our claim that the system behaves in a reliable, research-grade manner for the purposes of our experiment. See Online Appendix B.7 for full details.

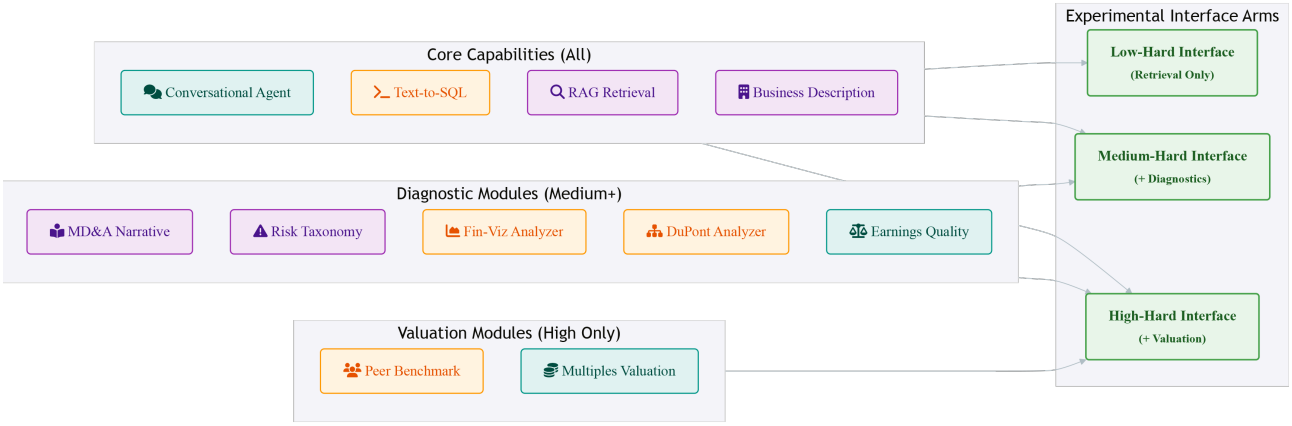
[3]Prior studies such as Lawrence et al. (2018) promote randomly selected earnings announcements on Yahoo! Finance to measure how exogenous visibility affects investors' attention, returns, and spreads; Leonelli et al. (2025) combine a large-scale household field experiment and survey to study how consumers use ESG and financial reports in product choices; and work by Costello et al. (2020) and Wong et al. (2024a,b) uses platform and institutional interventions to examine how investors' and creditors' actions causally affect corporate disclosure, environmental performance, and lending outcomes. In all of these settings, the experimental units are real investors or firms, and the treatments modify messages or incentives (e.g., promoted links, standardized campaign texts, provision of a risk score), while behavior is observed through downstream outcomes such as clicks, trading, firm responses, or loan approvals.

collaborate with AI under tightly controlled informational variation. A further design strength is that participants complete a pre-experiment FSA assessment covering core accounting and valuation concepts. This feature adds another layer of lab-like control by allowing us to measure ex ante domain expertise directly and stratify randomization on that dimension, thereby improving balance, internal validity, and power in line with best practices for information-provision and survey experiments (Haaland et al., 2023; Stantcheva, 2023).

4.2 Treatments and Information-Equivalent Design

As shown in Figure 3, participants are randomly assigned to one of three AI-interface conditions that differ only in how the AI agent structures and surfaces hard information. In the *Low-Hard* arm, the AI behaves as a retrieval-augmented research assistant. It answers natural-language questions by querying the structured warehouse (via Text-to-SQL) and retrieving relevant filing excerpts (via RAG), but it does not show any precomputed dashboards or valuation modules by default, except for a business description summary module based on Item 1 of 10Ks. Participants must actively ask for particular metrics or disclosures (for example, "show revenue from 2018 to 2022" or "summarize Item 1A for 2021"). In the *Medium-Hard* arm, the same conversational capabilities are available, but participants also see a set of precomputed analytical modules for each firm-year, including MD&A and risk-factor summaries, multi-year financial visualizations, DuPont decompositions, earnings-quality diagnostics, and risk taxonomies. These modules provide descriptive summaries and visualizations based solely on firm disclosures and do not contain valuation reasoning or assessment. In the *High-Hard* arm, we further activate modules that provide peer-relative diagnostics and multiple-based valuations. Moving from Low-Hard to High-Hard changes the AI's role from a pure hard-information retrieval tool, to a dashboard that organizes evidence and highlights patterns, to a valuation assistant, while leaving the underlying data unchanged.

Figure 3: Experimental Interface Arms Under the Information-equivalent Design



Notes: Participants are randomly assigned to Low-Hard, Medium-Hard, or High-Hard interfaces. All arms have access to the same underlying information set (firm fundamentals/ratios, parsed 10-K filings, and macro/market feeds) and the same retrieval infrastructure; only the structure and default surfacing of hard information differ. The shared core layer (Conversational Agent, Text-to-SQL, and RAG Retrieval, Business Description) is available in all conditions. Medium-Hard additionally surfaces precomputed modules (e.g., MD&A Narrative, Risk Taxonomy, Financial Visualization Analyzer, DuPont Analyzer, and Earnings Quality). High-Hard further activates valuation-oriented modules (Peer Benchmark and Multiple-based Valuation). Details of modules are provided in Online Appendix B.

4.3 Valuation Tasks and Procedure

We randomly assign 20 firms spanning a broad range of sectors so that each firm is evaluated by at least two students in each treatment arm. This design allows us to include firm fixed effects and separate treatment effects from firm-specific differences in information environment or valuation difficulty. For each task, participants are given a fixed time budget of 90 minutes and may interact freely with the AI agent through the chat interface and the modules available in their assigned treatment arm. At the end of the task, they must submit a valuation memo that states and defends an investment thesis. The final memo is evaluated on the range of coherent, evidence-supported, valuation-relevant reasoning and narratives used to justify the participant's own view. This is especially important in the *High-Hard* arm. Although that interface provides peer-benchmark and multiple-based valuation modules, these outputs do not make human soft information production redundant, because participants must still assess peer comparability, interpret whether valuation differences reflect growth, margins, risk, business mix, or accounting differences, and determine whether any gap reflects mispricing or justified

fundamentals. High-Hard therefore still leaves substantial room for human judgment.[4]

4.4 Process Tracing and Outcome Measurement

A distinctive feature of our experimental platform is its process transparency. The full chat transcript and the valuation memo provide a granular record of the information-production process. This level of observability is especially important because soft information is economically important but rarely observed directly, and it allows us to move beyond settings that capture only final decisions or a limited set of interaction events (for example, whether a user accepted or rejected an AI recommendation) (e.g., Bansal et al., 2021a; Green and Chen, 2021; Vaccaro et al., 2024). Before constructing our primary conversation-based measures, we manually identify and exclude queries that reflect copying, since an interaction can appear rich simply because the student pasted prior AI output.

Building on modern text-as-data and open-ended-response methods (Gentzkow et al., 2019; Haaland et al., 2025), we measure soft information production using a multi-method design. In corporate valuation, soft information is inherently high-dimensional and is best understood as comprising several non-mutually exclusive elements (Liberti and Petersen, 2019; Penman, 2013; Damodaran, 2017; Koller et al., 2025):

- *Progressive reasoning*: analysts refine and deepen lines of inquiry over successive interactions rather than stopping at one-shot retrieval.

[4]The High-Hard modules incorporate meaningful AI-generated valuation assessment, but they stop well short of a complete relative-valuation exercise. The Peer Benchmark Analyzer uses the system's LLM-derived peer mapping to compare the focal firm with its closest competitors on key operating and balance-sheet dimensions, while the Multiples Analyzer computes and interprets standard peer multiples, such as EV/EBITDA, EV/Sales, P/E, and price-to-book, where available, by benchmarking the focal firm against peer medians and ranges and explaining apparent premiums or discounts in economically intuitive terms. These outputs therefore provide more than descriptive summaries by supplying an AI-generated comparative valuation frame. At the same time, they leave several central valuation judgments to the participant. The modules do not determine which multiples are most appropriate given the firm's business model, lifecycle, accounting characteristics, or MD&A disclosures; they do not normalize denominators by adjudicating transitory items, accounting distortions, or cross-firm differences in capital structure and business mix; and they do not translate cross-sectional multiple comparisons into a final justified target multiple or reconcile conflicting signals across different multiples. Participants must therefore still decide whether the proposed peers are truly comparable, which multiples deserve the most weight, and whether any observed valuation gap reflects mispricing or instead justified differences in fundamentals.

- *Synthesis*: analysts combine accounting and narrative signals across sources, metrics, and time into a more integrated view of the firm.

- *Hypothesis and mechanism building*: analysts generate, compare, and probe explanations for observed patterns in performance, risk, and valuation.

- *Evidence-based interpretation*: analysts connect hard information to valuation-relevant judgments about growth, profitability, risk, and value.

Guided by this framework, we trace the production of soft information by combining manual human coding, LLM coding, and knowledge-graph-based coding. Our human coding is performed at the *query level*, which is well suited to capturing nuanced, context-dependent instances of progressive reasoning, synthesis, and mechanism-seeking as they arise locally in the interaction. Our LLM coding is performed at the *entire-conversation level*, which provides a scalable and semantically consistent way to measure broader interpretive patterns across the full corpus. Finally, our graph-based coding imposes transparent structure on the reasoning process by tracing how evidence maps into intermediate interpretations and valuation conclusions (Andre et al., 2026; Li et al., 2026). Together, these approaches exploit the distinct strengths of each while reducing reliance on any single imperfect proxy for a latent construct.

4.4.1 Human Coded Soft-Information Measures

Our primary soft-information measures are manually coded at the *query level* by trained coders following pre-specified rubrics, detailed in the Online Appendix. Each student query in the conversation with the AI agent is evaluated separately along the relevant dimensions of soft information production, and the resulting query-level codes are then aggregated to the full conversation level for each student–firm task. All manual coding is conducted independently by two advanced business-student coders who are blind to treatment assignment and the study's hypotheses, with any disagreements reviewed and resolved by the coauthor team using the same pre-specified rubric. This procedure helps ensure that the primary human-coded

measures reflect a consistent, research-grade protocol rather than ad hoc judgment. Our main conversation-level *Soft Information Index* is then constructed by summing six standardized components, organized around three core elements of soft information production.

The first is *progressive reasoning*, which we capture through *follow-up depth* and *reasoning depth*. Follow-up depth measures whether the student builds on prior answers and progressively sharpens a line of inquiry rather than issuing disconnected requests, while reasoning depth measures the extent of analytical authorship in the student's own query text, that is, whether the student merely asks for data or instead articulates an independent interpretation and asks the AI to verify, challenge, or extend it. The second is *synthesis*, which we capture through *cross-period synthesis* and *multi-metric synthesis*. Cross-period synthesis measures whether the student compares current signals to prior periods and reasons about trends, reversals, or persistence over time, while multi-metric synthesis measures whether the student integrates multiple accounting and financial signals into a coherent interpretation rather than treating each metric in isolation. The third is *hypothesis and mechanism building*, which we capture through *hypothesis testing* and *mechanism-seeking behavior*. Hypothesis testing measures whether the student proposes testable financial explanations and uses the AI to evaluate or discriminate among them, while mechanism-seeking behavior measures whether the student asks why or how an observed pattern arose rather than only asking what happened.

4.4.2 LLM Coded Soft-Information Measures

While human coding is well suited to capturing the local, context-dependent instances that arise at the query level, it is necessarily resource-intensive and may raise concerns about coder discretion at the margin. We therefore complement the primary human-coded measures with a second, conversation-level approach based on LLM coding. Specifically, we construct LLM-coded measures of *hypothesis-space breadth* and *analytical progression*. Hypothesis-space breadth captures whether, over the course of the full interaction, the participant entertains multiple plausible explanatory frames rather than converging too quickly on a single narrative.

Analytical progression captures whether the conversation moves beyond descriptive fact gathering toward more synthesized, valuation-relevant interpretation. These LLM-coded measures are a useful complement because they scale naturally across the full corpus and provide a broader semantic view of the same underlying construct of soft information production.

4.4.3 Knowledge Graph Based Soft-Information Measures

Next, building on recent work that represents economic narratives and textual reasoning as structured causal or knowledge graphs (Andre et al., 2026; Li et al., 2026), we use a largely deterministic *DAG representation of evidence-to-interpretation reasoning* to capture the fourth element of soft information production, *evidence-based interpretation*. A DAG, or directed acyclic graph, is a natural tool for this purpose because valuation-relevant interpretation is about linking specific pieces of hard evidence to intermediate judgments and, ultimately, to valuation conclusions. Our DAG pipeline uses a frozen ontology with two layers. The *evidence node layer* captures hard-information sources such as MD&A, risk factors, ratio patterns, DuPont diagnostics, and earnings-quality signals. The *interpretation node layer* captures valuation reasoning in a strict partial order from industry and operating drivers (L1), to accounting and earnings-quality mechanisms (L2), to financial outcomes (L3), and finally to valuation outcomes (L4). The pipeline then detects evidence and interpretation nodes from student-owned text, identifies directed edges between them using deterministic cue-based extraction and edge detection, and validates all edges against an allowed-edge matrix that preserves valuation-consistent directionality.

This graph-based approach makes evidence-based interpretation observable in a way that is both structured and easy to audit. For example, if a student writes that "rising inventory days and management's discussion of weaker demand suggest future margin pressure, which supports a lower valuation multiple," the DAG maps *inventory days* and *MD&A discussion* into evidence nodes, maps *weaker demand* into an operating-driver interpretation node, maps *margin pressure* into a financial-outcome node, and maps *lower valuation multiple* into a valuation-

outcome node, with directed edges tracing the chain from evidence to conclusion. Our baseline DAG outcome is therefore the total number of validated support edges from evidence nodes to interpretation nodes in the student's conversation DAG, which captures the extent to which participants convert hard information into structured, valuation-relevant interpretation.

Taken together, the multi-method architecture of the paper, combining human coding, LLM coding, and deterministic parsing and graph extraction, recognizes that soft information is inherently latent and that no single proxy is beyond dispute. Consistent inferences across these distinct measurement approaches materially strengthen construct validity. Detailed coding rubrics, coder protocol, inter-rater reliability statistics, instructions to LLM, and the frozen DAG ontology are reported in Online Appendix A.

4.5 Other Measures

We also examine two memo-based outcomes that capture complementary dimensions of memo quality. In our setting, a high-quality valuation memo should be both valuation-oriented and analytically broad. *Valuation Ratio* measures the extent to which the memo is explicitly framed around valuation by computing the share of sentences that contain valuation language, valuation models and inputs, financial multiples, or core valuation drivers such as growth, margins, reinvestment, and risk. *Memo Topics* measures the breadth of analysis as an LLM-coded count of distinct financial-analysis topics discussed in the memo using a pre-specified taxonomy. Together, these measures provide a useful summary of memo quality in this context. The first captures whether the memo is genuinely about valuation rather than general firm description, and the second captures whether the student brings multiple value-relevant dimensions into the analysis rather than offering a narrow or superficial assessment.

We further construct three human-coded measures of delegation that capture whether, how often, and with what downstream consequence students hand off integrative reasoning to the AI. In our setting, delegation occurs when a student asks the system to combine multiple financial inputs into an analytical output while contributing little or no interpretive reasoning

of their own. *Delegation Any* captures the extensive margin of this behavior, *Delegation Turns* captures its frequency, and *Delegation Memo Score* captures the extent to which the final memo depends on AI outputs produced in those delegation turns. These measures are informative in our context because they distinguish simple AI usage from genuine analytical handoff and allow us to observe not only whether students delegate during the interaction, but also whether that delegated reasoning becomes embedded in the final valuation memo.

5 Results

5.1 Descriptive Statistics and Measure Relationships

Table 1 summarizes the experimental sample and the main units of observation used throughout the paper. The final analytic sample contains 126 business students, with 44 assigned to Low-Hard, 40 to Medium-Hard, and 42 to High-Hard. Baseline FSA knowledge is well balanced across arms: mean quiz scores are 29.3, 29.5, and 29.1 out of 40, respectively, and we cannot reject equality across groups ($p = 0.94$). In the regression analysis, we standardize FSA rank to mean zero and unit variance. Panel B also maps the progressively finer units of observation used in later tests. This structure is central to our design because it allows us to study not only overall treatment effects, but also how AI changes the microstructure of evidence use, reasoning, and subsequent human responses.

Table 2 provides summary statistics for the paper's main information measures and reports the unit of observation and coding mode for each variable. The primary human-coded conversation-level measures display substantial dispersion and, descriptively, are generally highest in the Medium-Hard arm. The aggregate Soft Information Index has a mean of 0.06 and a standard deviation of 4.12 overall, with arm-specific means of -1.42 in Low-Hard, 1.99 in Medium-Hard, and -0.30 in High-Hard. The same broad pattern appears in the complementary conversation-level measures. Hypothesis-Space Breadth, Analytical Progression, and DAG Reasoning all peak in the Medium-Hard condition. Memo-based outcomes show a more

nuanced pattern. Valuation Ratio is highest in High-Hard, while Memo Topics is highest in Medium-Hard. Delegation outcomes differ even more sharply across arms: Delegation Any, Delegation Turns, and Delegation Memo Score are all substantially higher in High-Hard than in the other two conditions. Taken together, these descriptives foreshadow the paper's central distinction between broader human-authored soft information production in Medium-Hard and more valuation-oriented, but also more delegated, output in High-Hard.

Table 3 examines the relationships among the different soft-information measures. Panel A shows that the human-coded Soft Information Index is positively correlated with the LLM-coded measures of Hypothesis-Space Breadth and Analytical Progression, but only modestly correlated with DAG Reasoning. Panel B shows that the six primary human-coded components are positively correlated with one another and with the aggregate Soft Information Index, with strong associations for Follow-Up Depth, Reasoning Depth, Hypothesis Testing, and Mechanism-Seeking, but lower correlations for Cross-Period Synthesis and Multi-Metric Synthesis. These patterns are consistent with our intended measurement design that the different proxies move together enough to support a common latent construct, yet remain sufficiently distinct to indicate that each captures a different aspect of human soft information production.

5.2 Treatment Effects on soft information production

We begin by estimating task-level regressions of each conversation-level outcome on indicators for the Medium-Hard and High-Hard conditions, FSA rank, and firm fixed effects, with standard errors clustered at the firm level. Table 4, Panel A, shows a clear pattern. Relative to Low-Hard, the Medium-Hard interface produces broad-based increases in soft information production. The coefficient on Medium-Hard is positive and statistically significant for the primary human-coded Soft Information Index, the two LLM-coded measures, and DAG Reasoning. By contrast, the High-Hard interface delivers weaker and less consistent gains. To gauge economic magnitude, the Medium-Hard effects are economically large, implying increases of about 0.78 standard deviations in the Soft Information Index (3.227/4.12), 0.88 standard deviations in

Hypothesis-Space Breadth, 0.51 standard deviations in Analytical Progression, and 0.43 standard deviations in DAG Reasoning. The corresponding High-Hard effects are materially smaller, at roughly 0.44, 0.43, 0.02, and 0.16 standard deviations, respectively, and not or marginally statistically significant. These estimates suggest that AI is most complementary to human soft information production when it structures and surfaces relevant hard information without taking over the interpretive task.

Panel B clarifies the source of these treatment effects. Medium-Hard significantly increases Follow-Up Depth, Cross-Period Synthesis, Hypothesis Testing, and Mechanism-Seeking, but has relatively small effects on Reasoning Depth and Multi-Metric Synthesis. High-Hard, by contrast, significantly increases only Reasoning Depth. The dashboard-like Medium-Hard interface therefore strengthens several core margins of soft information production at once, including progressive reasoning, synthesis, and hypothesis and mechanism building, whereas the more prescriptive High-Hard interface generates narrower and more localized gains. Taken together, these patterns are consistent with both H1a and H1b, whereby structured hard-information support facilitates human-produced soft information, while more prescriptive valuation-oriented guidance yields smaller incremental gains and is consistent with partial crowding out through anchoring on AI-generated outputs.

The FSA-rank coefficient in Panel A is positive and statistically significant for the Soft Information Index and Hypothesis-Space Breadth, and Panel B shows a similar pattern for selected components, especially Reasoning Depth and Hypothesis Testing. These baseline results are consistent with H2a, whereby ex ante FSA knowledge is positively associated with soft information production. At the same time, the treatment coefficients remain economically large after controlling for FSA rank, indicating that the design of the AI interface independently shapes how participants convert hard information into interpretation. In the next section, we further examine whether prior expertise moderates the extent to which more prescriptive AI support crowds out human interpretive effort.

5.3 The Role of FSA Knowledge and Human Expertise

In this section, we allow the slope on FSA rank to vary by treatment arm by interacting FSA rank with the treatment indicators. Table 5 shows that the role of expertise depends importantly on how prescriptive the AI support is. For the primary human-coded Soft Information Index, the implied FSA slope is positive in all three arms and increases from Low-Hard to Medium-Hard to High-Hard, consistent with the view that domain knowledge complements AI support. More broadly, the heterogeneity is most concentrated in the most prescriptive condition. The High-Hard interaction is also positive and statistically significant for the LLM-coded measures of Hypothesis-Space Breadth and Analytical Progression, and the DAG coefficient is directionally consistent though imprecisely estimated. By contrast, the corresponding Medium-Hard interactions are smaller and not statistically distinguishable from zero. Consistent with H2b, these estimates suggest that the most prescriptive interface is *expertise-amplifying* rather than expertise-equalizing. Participants with stronger expertise appear better able to interrogate, refine, and extend AI-generated valuation guidance, whereas weaker participants appear less able to convert the same outputs into independent soft information.

The magnitudes are economically meaningful. Because FSA rank is standardized, the coefficients can be read as the effect of a one-standard-deviation increase in prior expertise. For the Soft Information Index, the implied FSA slope rises from 0.315 in Low-Hard to 0.828 in Medium-Hard and 1.738 in High-Hard, which corresponds to roughly 0.08, 0.20, and 0.42 unconditional standard deviations, respectively. In High-Hard, a one-standard-deviation increase in FSA rank is also associated with about 0.39 unconditional standard deviations more Hypothesis-Space Breadth and 0.45 unconditional standard deviations more Analytical Progression. Taken together with Table 4, these results imply that structured diagnostic support produces the broadest average gains, while more prescriptive valuation-oriented support is more skill-contingent and therefore generates greater dispersion in human-authored soft information production.

5.4 Memo-Based Outcomes and Delegation

We next examine whether the treatment effects documented in the conversation trace carry through to participants' final written output. Table 6 suggests that they do. Relative to Low-Hard, both Medium-Hard and High-Hard increase the share of the memo devoted to valuation-relevant reasoning, with the larger coefficient in High-Hard. This pattern indicates that more structured and valuation-oriented AI support affects not only how students interact with the system, but also how they frame and write their final investment theses. At the same time, the two treatment arms differ in the breadth of the resulting memo. Medium-Hard significantly increases the number of valuation-relevant topics covered in the memo, whereas High-Hard does not. Thus, although High-Hard produces more valuation-focused writing, it does not broaden the scope of analysis in the same way as Medium-Hard. This narrower topic coverage is consistent with anchoring on the AI's valuation framing.

Table 7 reinforces this interpretation by showing that the more prescriptive interface also changes who is doing the analytical work. Here, delegation refers to cases in which participants offload substantive analytical tasks to the AI rather than using the system primarily to retrieve information or test their own reasoning. Medium-Hard has no meaningful effect on any of these measures. By contrast, High-Hard significantly increases all three. Taken together, Tables 6 and 7 shed further light on the distinction between AI complementarity and crowding out. Medium-Hard improves both conversation-level soft information production and the breadth of the final memo without increasing reliance on delegation. High-Hard, in contrast, generates more valuation-oriented memos, but those gains coincide with substantially greater delegation and no corresponding increase in memo-topic breadth. The evidence therefore reinforces the earlier results and is more consistent with anchoring and partial substitution than with a broad expansion of independently produced soft information.

5.5 Opening the Black Box of Human–AI Collaboration

To better understand the mechanisms behind the treatment effects documented above, we next open the black box of human–AI collaboration. This is useful because the main results show that interface design changes the amount and type of human-produced soft information, but they do not by themselves reveal *how* those differences arise during the interaction. Our granular data make that analysis possible. The DAG structure lets us observe when participants reference hard-information evidence surfaced by the AI, whether that evidence is converted into an interpretation node, and how the resulting reasoning chain progresses toward higher-level valuation conclusions, while the full conversation log lets us study which features of an AI response are associated with greater soft information production in the user's next prompt. These tests therefore provide more direct evidence on the mechanisms through which AI can either complement human interpretation or crowd it out.

For these analyses, we construct more granular samples summarized in Table 1. Table 8 moves to the evidence-node level ($N = 756$), where each observation is a student–firm task $\times$ evidence type. We estimate regressions of evidence use on treatment indicators interacted with evidence-type indicators, and include evidence-type fixed effects and firm fixed effects. This specification lets us ask which categories of hard information surfaced by the AI are more likely to be referenced (*Evidence Mentions* in column 1) and converted into human interpretation (*Support Edges* in column 2). That distinction is important because raw *Evidence Mentions* need not reflect soft information production; the more informative outcome is *Support Edges*, which record whether an evidence node is explicitly linked to an interpretation node in the DAG. On this margin, Medium-Hard increases the conversion of several diagnostically rich modules into interpretation, including MD&A, Risk Factors, DuPont, and Earnings Quality. High-Hard preserves some of this narrative-evidence processing for MD&A and Risk Factors, but the effects are smaller and do not extend to DuPont or Earnings Quality. Thus, the dashboard-style Medium-Hard interface more effectively induces participants to transform disclosed evidence into valuation-relevant interpretation.

Our main analyses focus on how AI interface design and prior FSA knowledge shape human soft information production, but another potentially important lever is the design of the AI response itself. Human–AI interaction research emphasizes that collaboration quality depends not only on whether assistance is available, but also on how the system scaffolds user reasoning, invites follow-up, and communicates uncertainty (Amershi et al., 2019; Bansal et al., 2021b; Dhillon et al., 2024; Cao et al., 2024a). Table 9 regresses the next query's soft-information outcomes on features of the immediately preceding AI response in the adjacent AI-response–next-query sample (N = 1,177), including user fixed effects and firm fixed effects, with standard errors two-way clustered by user and firm. User fixed effects are important because they absorb persistent differences across participants, such as baseline ability, curiosity, prompting style, and tendency to rely on the AI, so identification comes from within-user variation in the types of responses the same participant receives.

Guided by the Human-AI interaction literature, we focus on five response features. *Causal Gap* indicates that the AI identifies a pattern but leaves its causal explanation unresolved; *Mechanism Gap* indicates that the AI offers a broad conclusion but leaves the operative channel underspecified; *Open Question* indicates that the response explicitly invites further inquiry; *Valuation Stance* indicates that the AI takes a directional valuation view; and *Uncertainty* indicates that the AI explicitly communicates caveats or ambiguity.[5] We include *Length* and *Tone* as broad stylistic controls. Because these response features are not experimentally randomized and arise within a common response-generation framework across treatment arms, we interpret Table 9 as exploratory rather than causal.

The strongest positive associations with the next query's Soft Information Index come from

[5]A response is coded as *Causal Gap* if it says something like, "Gross margin fell while inventory days increased, and management discussed softer demand, but the available information does not yet distinguish whether the main driver is discounting, weaker product mix, or excess capacity." It is coded as *Mechanism Gap* if it says, "Earnings quality appears weaker than peers. That said, the current evidence does not yet isolate whether the difference is mainly attributable to accrual build-up, working-capital timing, or one-time items." It is coded as an *Open Question* if it ends with a natural invitation such as, "Would you like to dig into whether the valuation discount is more likely driven by margin risk, peer mismatch, or accounting-quality concerns?" It is coded as *Valuation Stance* if it takes a directional view, for example, "Relative to the peer set, the shares appear modestly undervalued." It is coded as *Uncertainty* if it explicitly qualifies the conclusion, for example, "This conclusion should be treated cautiously because it is sensitive to peer selection and to whether the recent margin decline is temporary or structural."

Causal Gap and *Mechanism Gap*, with a smaller positive association for *Uncertainty*. Panel B shows that causal gaps are associated with more Follow-Up Depth, Cross-Period Synthesis, Hypothesis Testing, Reasoning Depth, and Mechanism-Seeking, while mechanism gaps are associated especially strongly with Follow-Up Depth, Reasoning Depth, and Mechanism-Seeking. By contrast, a directional valuation stance does not increase the composite index and is associated with lower Reasoning Depth, while open-question phrasing mainly reduces delegation rather than materially raising the composite. The main takeaway is that AI responses appear most conducive to human soft information production when they surface relevant evidence but leave disciplined inferential work for the user to complete. Designing response strategies that stimulate this kind of human interpretation, while limiting anchoring and unproductive delegation, is therefore a promising direction for future research.

6 Discussion and Conclusion

This paper studies how humans and AI share the work of information production in corporate valuation. We build a retrieval-augmented AI valuation agent that reads 10-Ks and earnings call transcripts, queries Compustat-style fundamentals and ratio libraries, calls real-time market data and macro time series, and generates modular analytical dashboards. Embedding this agent in an experiment with undergraduate and MBA business students, we vary the amount and structure of AI-supplied hard information across three interface conditions (Low-Hard, Medium-Hard, High-Hard) while holding the underlying information set fixed. We then trace soft information production directly from chat transcripts and valuation memos. We document two main findings. First, there is a non-monotonic effect of AI-supplied hard information on soft information production that, relative to a bare-bones retrieval interface, a dashboard-rich interface substantially facilitates soft information production, whereas adding AI-generated valuation guidance yields much smaller incremental gains. Second, ex ante FSA knowledge is strongly complementary to AI support that, more knowledgeable participants use the AI's structured outputs as inputs into deeper independent reasoning, with this moderating

effect especially pronounced in the most prescriptive AI condition.

6.1 Implications for Human–AI Collaboration in Valuation

Our results have several implications for the design of human–AI collaboration in valuation tasks. The hard–soft information literature emphasizes that hard information is easily transmitted and verifiable, while soft information is contextual and costly to communicate (Stein, 2002; Liberti and Petersen, 2019). In our design, the AI agent is explicitly tasked with producing and structuring hard information. Humans, in turn, remain the primary source of soft information, generating narratives and judgments that build on (but are not mechanically implied by) the AI's outputs. Our findings suggest that AI systems are most complementary to human soft information production when they act as powerful but non-prescriptive hard-information engines, performing retrieving, summarizing, visualizing, and highlighting patterns without collapsing them into a valuation guidance. When the AI begins to produce valuation assessments, many users, especially those with weaker FSA knowledge, appear to anchor on these outputs and invest less in independent narrative work, consistent with cognitive offloading and overreliance on algorithmic recommendations (Dietvorst et al., 2015; Logg et al., 2019; Vaccaro et al., 2024).

Relative to existing human–AI experiments that focus on binary accept/reject decisions for algorithmic scores in relatively low-dimensional settings (e.g., image classification or risk scoring) (Green and Chen, 2021; Bansal et al., 2021a), our setting highlights the importance of *role design* within the human–AI team. In corporate valuation, information production is multi-stage whereby analysts must identify salient facts, connect them to an economic narrative, and translate that narrative into prices and recommendations. Our three-arm design manipulates the point at which the AI stops along this pipeline. The evidence suggests that ending the AI's role at structured diagnostics yields more complementary outcomes than extending it all the way to a final valuation, particularly for users with limited expertise.

6.2 Implications for Accounting and Finance Education

Our findings also speak to the ongoing debate about how to train accountants and finance professionals in the LLM era. Policy reports and recent surveys emphasize that AI and machine learning are reshaping financial intermediation and securities analysis, but they differ on whether these technologies will substitute for or complement traditional analytical skills (OECD, 2021; Kelly and Xiu, 2023; Cao et al., 2024b). In accounting education, scholars have called for greater integration of data analytics and emerging technologies into curricula, while cautioning that tools like LLMs should not crowd out the development of core reasoning and critical-thinking skills (Augustine et al., 2020; Dalwai et al., 2021; Raschke and Charron, 2021; Ballantine et al., 2024). Our experimental evidence supports a complementarity view. FSA knowledge is consistently associated with richer soft information production and better use of AI-supplied dashboards; the students who extract the most value from the AI are those who can interpret ratios, decompositions, and earnings-quality flags in economically coherent ways. In addition, our work suggests that LLM-based agents can be used not only as information sources but also as platforms for eliciting open-ended responses that reveal students' mental models and reasoning processes, helping instructors diagnose which aspects of valuation thinking remain underdeveloped.

6.3 Limitations and Directions for Future Research

While our experimental design and AI platform offer a rich view of human–AI collaboration in valuation, several limitations should temper the interpretation of our findings. First, our participants are undergraduate and MBA students rather than professional analysts or sophisticated investors. Although they are a natural population for studying education and early-career skills, they likely differ from seasoned practitioners in both domain knowledge and experience with complex information environments. Extrapolating our treatment effects to buy-side or sell-side professionals therefore requires caution; future work should replicate and extend our design in settings where analysts, portfolio managers, or corporate finance practitioners use the

agent in live workflows. Second, our experiment observes only relatively short-run interaction. We do not see how human–AI teams evolve over years of repeated collaboration, nor how experience with the agent shapes users' long-run learning, trust calibration, and mental models of AI capabilities.

Third, we vary only one dimension of AI design: the amount and level of hard information provided in the interface. Our RAG-based agent offers many other "dials" that could be experimentally manipulated. For example, one could vary the expressiveness and confidence of the AI's explanations (concise vs. verbose, cautious vs. assertive), the way uncertainty is communicated (point valuations vs. ranges and scenario trees), the ordering and framing of modules (leading with risks vs. performance vs. valuation), or the degree to which the AI actively prompts users with questions rather than passively responding. One could also randomize the visibility of citations and underlying evidence (source-rich vs. source-poor interfaces), alter the degree of personalization to the user's prior behavior, or introduce cooperative multi-user modes in which several analysts share a common AI agent. Each of these design dimensions may meaningfully affect how humans allocate effort between hard and soft information, how much they offload work to the AI, and how complementary the resulting human–AI team becomes. We view our AI valuation agent and experimental platform as a flexible foundation for a broader research agenda on human–AI collaboration in accounting and finance.

References

Agarwal, S. and Hauswald, R. (2010). Distance and private information in lending. *Review of Financial Studies*, 23(7):2757–2788.

Alammar, J. and Grootendorst, M. (2024). *Hands-On Large Language Models: Language Understanding and Generation*. O'Reilly Media, Sebastopol, CA. 1st edition.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Andre, P., Haaland, I., Roth, C., Wiederholt, M., and Wohlfart, J. (2026). Narratives about the macroeconomy. *The Review of Economic Studies*. Advance article, published online February 18, 2026.

Augustine, Fred K., J., Woodside, J., Mendoza, M., and Chambers, V. (2020). Analytics, accounting and big data: Enhancing accounting education. *Journal of Management and Engineering Integration*, 13(1):1–15.

Ballantine, J., Boyce, G., and Stoner, G. (2024). A critical review of AI in accounting education: Threat and opportunity. *Critical Perspectives on Accounting*, page 102711.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021a). Is the most accurate AI the best teammate? optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. S. (2021b). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Ben-Rephael, A., Carlin, B. I., Da, Z., and Israelsen, R. D. (2023). Security analysis and the collection of hard and soft information. NBER Working Paper 31936, National Bureau of Economic Research.

Blankespoor, E., Croom, J., and Grant, S. M. (2024). Generative AI and investor processing of financial information. *SSRN Electronic Journal*. Working paper.

Cao, S., Liu, A., and Huang, C.-M. (2024a). Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32.

Cao, S. S., Jiang, W., Wang, J., and Yang, B. (2024b). From man vs. machine to man + machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160:102924.

Chen, H., Qu, Y., Shen, T., and Wang, Q. (2024). Soft information in mutual fund portfolios. Working paper.

Costello, A. M., Down, J., and Mehta, R. P. (2020). Machine + man: A field experiment on the role of discretion in credit scoring. *Journal of Accounting and Economics*, 70(2–3):101344.

Croom, J. (2025). Interactivity and illusions of ability: How using generative AI affects investor judgments. *Journal of Accounting Research*. Forthcoming.

Croom, J., Gale, B., and Grant, S. M. (2025). Disclosure presentation attributes, generative AI output, and investor judgments. *SSRN Electronic Journal*. Working paper.

Dalwai, T., Mohammadi, S. S., Chugh, G., and Somerville, A. (2021). Big data analytics and accounting education: A systematic literature review. In Al Mawali, N. R., Al Lawati, A. M., and S., A., editors, *Fourth Industrial Revolution and Business Dynamics*, pages 159–174. Springer.

Damodaran, A. (2017). *Narrative and Numbers: The Value of Stories in Business*. Columbia Business School Publishing; Columbia University Press, New York.

Dhillon, P. S., Molaei, S., Li, J., Golub, M., Zheng, S., and Robert, L. P. (2024). Shaping human-ai collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, New York, NY, USA. Association for Computing Machinery.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*,

144(1):114–126.

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L., and Gao, J. (2024). Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Green, B. and Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM.

Green, B. and Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35.

Haaland, I., Roth, C., Stantcheva, S., and Wohlfart, J. (2025). Understanding economic behavior using open-ended data. *Journal of Economic Literature*, 63(4):1244–1280. Forthcoming; NBER Working Paper No. 32421.

Haaland, I., Roth, C., and Wohlfart, J. (2023). Designing information provision experiments. *Journal of Economic Literature*, 61(1):3–40.

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., and Satzger, G. (2024). Complementarity in human–AI collaboration: Concept, sources, and evidence. *arXiv preprint arXiv:2404.00029*.

Huyen, C. (2025). *AI Engineering: Building Applications with Foundation Models*. O'Reilly Media, Sebastopol, CA. 1st edition.

Kelly, B. T. and Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3–4):205–363.

Koller, T., Goedhart, M., Wessels, D., and McKinsey & Company, I. (2025). *Valuation: Measuring and Managing the Value of Companies*. Wiley, 8 edition.

Lawrence, A., Ryans, J., Sun, E., and Laptev, N. (2018). Earnings announcement promotions: A Yahoo finance field experiment. *Journal of Accounting and Economics*, 66(2–3):399–414.

Leonelli, S., Muhn, M., Rauter, T., and Sran, G. S. (2025). How do consumers use ESG disclosure? evidence from a randomized field experiment with everyday product purchases. *Journal of Accounting and Economics*, page 101811.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Li, K., Mai, F., Shen, R., Yang, C., and Zhang, T. (2026). Dissecting corporate culture using generative AI. *The Review of Financial Studies*, 39(1):253–296.

Liberti, J. M. and Petersen, M. A. (2019). Information: Hard and soft. *Review of Corporate Finance Studies*, 8(1):1–41.

Liu, M. (2022). Assessing human information processing in lending decisions: A machine learning approach. *Journal of Accounting Research*, 60(2):607–651.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

OECD (2021). Artificial intelligence, machine learning and big data in finance. Oecd report, Organisation for Economic Co-operation and Development.

Otis, N., Clarke, R., Delecourt, S., Holtz, D., and Koning, R. (2025). The uneven impact of generative AI on entrepreneurial performance. *SSRN Electronic Journal*. Date written February 27, 2024; revised October 3, 2025.

Penman, S. H. (2013). *Financial Statement Analysis and Security Valuation*. McGraw-Hill, New York, 5 edition.

Raschke, R. L. and Charron, K. F. (2021). Review of data analytic teaching cases, have we covered enough? *Journal of Emerging Technologies in Accounting*, 18(2):247–255.

Risko, E. F. and Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688.

Shen, J. H. and Tamkin, A. (2026). How AI impacts skill formation. arXiv preprint, version 2, revised February 1, 2026.

Shenhav, A., Botvinick, M. M., and Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240.

Stantcheva, S. (2023). How to run surveys: A guide to creating your own. *Annual Review of Economics*, 15:205–234.

Stein, J. C. (2002). Information production and capital allocation: Decentralized versus hierarchical firms. *Journal of Finance*, 57(5):1891–1921.

Vaccaro, M., Almaatouq, A., and Malone, T. W. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*. Forthcoming.

Vendraminelli, L., DosSantos DiSorbo, M., Hildebrandt, A., McFowland, Edward, I., Karunakaran, A., and Bojinov, I. (2025). The GenAI wall effect: Examining the limits to horizontal expertise transfer between occupational insiders and outsiders. Technology & Operations Management Unit Working Paper 26-011, Harvard Business School. Harvard Business School Working Paper No. 26-011.

Wong, T. J., Yu, G., Zhang, S., and Zhang, T. (2024a). Calling for transparency: Evidence from a field experiment. *Journal of Accounting and Economics*, 77(1).

Wong, T. J., Yu, G., Zhang, S., and Zhang, T. (2024b). Do firms respond to calls for environmental improvements made by retail investors? *SSRN Electronic Journal*. Working paper.

Table 1: Sample Balance and Unit-of-Observation Map

	Low-Hard	Medium-Hard	High-Hard	p-value
Panel A: Student-level balance by treatment arm				
Number of students	44	40	42	
Baseline FSA quiz score (0–40), mean	29.3	29.5	29.1	0.94
Standard deviation	5.4	5.1	5.2	

Sample component	Unit of observation	N	Used in analyses
Panel B: Unit-of-observation map			
Student–firm valuation tasks	Task	126	Tables 4–7
Evidence-node panel	Task × evidence type	756	Table 8
AI-response and next-query pairs	Adjacent AI-response query pair	1,177	Table 9

Notes: This table summarizes the experimental sample and the main units of observation used throughout the paper. Panel A reports student-level balance across the three treatment arms. The reported p-value is from a test of equality of mean baseline financial-statement-analysis (FSA) quiz scores across treatment arms. Students are randomly assigned to one of the three AI-interface conditions after stratification on baseline FSA knowledge. Panel B maps the units of observation used in the main analyses. The task-level sample consists of all student–firm valuation tasks. The evidence-node panel expands each task into evidence categories defined in the DAG ontology. The AI-response–next-query sample consists of adjacent AI-response–student-query pairs used to study which AI response features induce subsequent human-authored soft-information production.

Table 2: Summary Statistics for Information Measures

Variable	Unit	Coding mode	Overall mean	Std. dev.	Low mean	Medium mean	High mean
Panel A: Primary human-coded conversation-level measures							
Follow-Up Depth	Conversation	Human	2.68	0.90	2.52	3.00	2.53
Reasoning Depth	Conversation	Human	1.38	0.43	1.25	1.40	1.51
Cross-Period Synthesis	Conversation	Human	0.34	0.79	0.18	0.66	0.19
Multi-Metric Synthesis	Conversation	Human	0.10	0.35	0.04	0.15	0.12
Hypothesis Testing	Conversation	Human	2.42	1.15	2.11	2.88	2.26
Mechanism-Seeking	Conversation	Human	2.32	1.93	1.71	3.61	1.71
Soft Information Index	Conversation	Human	0.06	4.12	-1.42	1.99	-0.30
Panel B: Complementary conversation-level measures							
Hypothesis-Space Breadth	Conversation	LLM	3.46	2.26	2.54	4.63	3.24
Analytical Progression	Conversation	LLM	3.08	0.72	2.98	3.37	2.90
DAG Reasoning	Conversation	Graph	8.67	5.56	7.73	10.34	8.02
Panel C: Memo-based outcomes							
Valuation Ratio	Memo	Determ	0.47	0.30	0.39	0.48	0.55
Memo Topics	Memo	LLM	9.84	2.23	9.39	10.51	9.67
Panel D: Delegation measures							
Delegation Any	Conversation	Human	0.35	0.48	0.27	0.22	0.57
Delegation Turns	Conversation	Human	0.89	1.72	0.41	0.71	1.57
Delegation Memo Score	Memo	Human	1.95	1.49	1.66	1.51	2.67

Notes: This table reports summary statistics for the main information outcomes used in the paper. The primary human-coded measures in Panel A are coded manually at the query level using pre-specified rubrics and then aggregated to the full conversation level for each student–firm task. Panel B reports complementary conversation-level measures constructed using LLM coding and graph-based extraction. Panel C reports memo-based outcomes, including deterministic measures of valuation-oriented writing and evidence use, and an LLM-coded measure of memo topic breadth. Panel D reports delegation outcomes, which are measured separately from soft-information production. Coding mode indicates whether a variable is constructed through manual human coding, LLM coding, graph-based extraction, or a deterministic algorithm.

Table 3: Cross-Measure Correlations of Soft-Information Outcomes

	Correlation with			
	(1)	(2)	(3)	(4)
Panel A: Cross-family correlations				
(1) Soft Information Index	1.000			
(2) Hypothesis-Space Breadth	0.74	1.000		
(3) Analytical Progression	0.76	0.59	1.000	
(4) DAG Reasoning	0.25	0.17	0.14	1.000

	Correlation with						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel B: Primary human-coded conversation-level measures							
(1) Follow-Up Depth	1.000						
(2) Reasoning Depth	0.62	1.000					
(3) Cross-Period Synthesis	0.18	0.05	1.000				
(4) Multi-Metric Synthesis	0.31	0.50	0.07	1.000			
(5) Hypothesis Testing	0.81	0.63	0.09	0.39	1.000		
(6) Mechanism-Seeking	0.55	0.33	0.21	0.27	0.49	1.000	
(7) Soft Information Index	0.84	0.76	0.39	0.62	0.83	0.69	1.000

Notes: This table reports Pearson correlations among the main soft-information measures at the student–firm conversation level. Panel A reports cross-family correlations between the Soft Information Index, the two LLM-coded conversation-level measures, and the graph-based measure of evidence-to-interpretation reasoning. Panel B shows correlations among the six primary human-coded components and the aggregate Soft Information Index.

Table 4: Treatment Effects on Soft-Information Production

	Dependent variable			
	(1) Soft Information Index	(2) Hypothesis-Space Breadth	(3) Analytical Progression	(4) DAG Reasoning
Panel A: Main treatment effects				
Medium-Hard (vs. Low-Hard)	3.227***	1.994***	0.368***	2.365**
	(0.693)	(0.411)	(0.126)	(1.058)
High-Hard (vs. Low-Hard)	1.799	0.964*	0.016	0.878
	(1.087)	(0.487)	(0.197)	(1.387)
FSA rank	1.013**	0.404**	0.139	0.091
	(0.422)	(0.188)	(0.082)	(0.580)
Constant	-1.624***	2.469***	2.947***	7.544***
	(0.491)	(0.256)	(0.090)	(0.653)
Firm FE	Yes	Yes	Yes	Yes
Observations	126	126	126	126
R^2	0.40	0.42	0.30	0.23

	Dependent variable					
	(1) Follow-Up Depth	(2) Reasoning Depth	(3) Cross-Period Synthesis	(4) Multi-Metric Synthesis	(5) Hypothesis Testing	(6) Mechanism-Seeking
Panel B: Components of the Soft Information Index						
Medium-Hard	0.460**	0.123	0.469**	0.101	0.735***	1.741***
	(0.166)	(0.080)	(0.204)	(0.069)	(0.249)	(0.243)
High-Hard	0.155	0.310**	0.066	0.085	0.304	0.159
	(0.242)	(0.116)	(0.118)	(0.062)	(0.295)	(0.377)
FSA rank	0.194	0.118***	0.107*	0.025	0.259*	0.215
	(0.118)	(0.034)	(0.059)	(0.037)	(0.128)	(0.204)
Constant	2.472***	1.241***	0.162**	0.043	2.064***	1.664***
	(0.109)	(0.055)	(0.074)	(0.036)	(0.160)	(0.166)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	126	126	126	126	126	126
R^2	0.36	0.34	0.24	0.21	0.33	0.39

Notes: This table reports estimates of the effects of AI-interface condition on conversation-level soft-information production. In all columns, the omitted category is the Low-Hard condition. Panel A reports treatment effects on the main Soft Information Index and complementary conversation-level outcomes. Panel B reports treatment effects on the six human-coded components that underlie the Soft Information Index. All regressions include firm fixed effects. Standard errors clustered by firm are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table 5: Heterogeneity by Prior FSA Knowledge

	Dependent variable			
	(1) Soft Information Index	(2) Hypothesis-Space Breadth	(3) Analytical Progression	(4) DAG Reasoning
Medium-Hard (vs. Low-Hard)	3.187***	1.939***	0.363**	2.244*
	(0.676)	(0.418)	(0.129)	(1.083)
High-Hard (vs. Low-Hard)	1.933*	1.048**	0.052	0.885
	(1.042)	(0.453)	(0.190)	(1.392)
FSA rank	0.315	-0.160	-0.021	-0.463
	(0.456)	(0.294)	(0.112)	(0.858)
Medium-Hard × FSA rank	0.513	0.560	0.091	0.943
	(0.636)	(0.420)	(0.138)	(1.138)
High-Hard × FSA rank	1.423*	1.033**	0.347**	0.694
	(0.813)	(0.458)	(0.122)	(1.511)
Constant	-1.557***	2.520***	2.963***	7.590***
	(0.475)	(0.244)	(0.087)	(0.635)
Firm FE	Yes	Yes	Yes	Yes
Observations	126	126	126	126
R^2	0.42	0.45	0.34	0.23

Notes: This table reports heterogeneity in the treatment effects by prior financial-statement-analysis (FSA) knowledge. In all columns, the omitted treatment category is the Low-Hard condition. The coefficient on FSA rank gives the FSA slope in the Low-Hard arm, while the interaction terms report how the FSA slope differs in the Medium-Hard and High-Hard arms relative to Low-Hard. All regressions include firm fixed effects. Standard errors clustered by firm are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table 6: Treatment Effects on Memo-Based Outcomes

	Dependent variable	
	(1) Valuation Ratio	(2) Memo Topics
Medium-Hard (vs. Low-Hard)	0.096*	1.148*
	(0.048)	(0.552)
High-Hard (vs. Low-Hard)	0.179**	0.240
	(0.074)	(0.646)
FSA rank	-0.006	-0.225
	(0.023)	(0.196)
Constant	0.378***	9.389***
	(0.035)	(0.349)
Student FE	No	No
Firm FE	Yes	Yes
Observations	126	126
R^2	0.18	0.19

Notes: This table reports estimates of the effects of AI-interface condition on memo-based outcomes. In all columns, the omitted category is the Low-Hard condition. *Valuation Ratio* measures the share of memo content devoted to valuation-oriented reasoning, and *Memo Topics* captures the breadth of valuation-relevant topics discussed in the final memo. All regressions include firm fixed effects. Standard errors clustered by firm are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table 7: Treatment Effects on Delegation Outcomes

	Dependent variable		
	(1) Delegation Any	(2) Delegation Turns	(3) Delegation Memo Score
Medium-Hard (vs. Low-Hard)	-0.077	0.235	-0.195
	(0.076)	(0.302)	(0.289)
High-Hard (vs. Low-Hard)	0.317***	1.169***	1.103***
	(0.100)	(0.279)	(0.307)
FSA rank	0.013	-0.236	0.034
	(0.050)	(0.194)	(0.176)
Constant	0.268***	0.401**	1.631***
	(0.048)	(0.142)	(0.169)
Student FE	No	No	No
Firm FE	Yes	Yes	Yes
Observations	126	126	126
R^2	0.24	0.23	0.27

Notes: This table reports estimates of the effects of AI-interface condition on delegation outcomes. In all columns, the omitted category is the Low-Hard condition. *Delegation Any* indicates whether the participant delegated any substantive analytical task to the AI, *Delegation Turns* counts the number of delegation turns in the conversation, and *Delegation Memo Score* measures the extent to which delegated AI content appears in the final memo. All regressions include firm fixed effects. Standard errors clustered by firm are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table 8: Evidence-to-Interpretation Conversion by Evidence Type

	Dependent variable	
	(1) Evidence Mentions	(2) Support Edges
Medium-Hard	0.045	-0.065
	(0.387)	(0.131)
High-Hard	0.269	0.148
	(0.375)	(0.149)
MD&A × Medium-Hard	1.353***	0.616***
	(0.410)	(0.209)
Risk Factors × Medium-Hard	0.788**	0.335**
	(0.326)	(0.137)
Ratio Patterns × Medium-Hard	3.702	0.660
	(2.668)	(0.590)
DuPont × Medium-Hard	3.102	0.849**
	(1.987)	(0.426)
Earnings Quality × Medium-Hard	1.394*	0.815**
	(0.735)	(0.316)
MD&A × High-Hard	1.317***	0.361*
	(0.445)	(0.187)
Risk Factors × High-Hard	1.013***	0.210*
	(0.308)	(0.112)
Ratio Patterns × High-Hard	-8.122***	-0.689
	(2.155)	(0.639)
DuPont × High-Hard	-1.299	0.425
	(1.369)	(0.410)
Earnings Quality × High-Hard	-0.101	0.070
	(0.596)	(0.228)
Evidence-Type FE	Yes	Yes
Firm FE	Yes	Yes
Observations	756	756
R^2	0.723	0.590

Notes: This table reports evidence-node-level estimates of how treatment effects vary across evidence types. In all columns, the omitted treatment category is Low-Hard, and the interaction terms report differential effects for the indicated evidence type relative to the omitted evidence-type category. *Evidence Mentions* measures whether and how often a given evidence type appears in the student's reasoning, and *Support Edges* measures the extent to which that evidence type is linked to interpretation nodes in the DAG. All regressions include evidence-type fixed effects and firm fixed effects. Standard errors clustered by firm are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table 9: AI Response Features and Subsequent Query-Level Soft-Information Production

	Dependent variable	
	(1) Soft Information Index	(2) Delegation
Panel A: Query-level composite outcomes		
Causal Gap	4.982***	0.002
	(0.522)	(0.027)
Mechanism Gap	2.756***	0.009
	(0.397)	(0.034)
Open Question	0.265	-0.048***
	(0.264)	(0.014)
Valuation Stance	0.026	-0.033
	(0.475)	(0.050)
Uncertainty	0.530***	-0.005
	(0.157)	(0.013)
Length	0.019	-0.001
	(0.024)	(0.002)
Tone	0.376	0.096
	(1.236)	(0.066)
Constant	-1.041	-0.018
	(1.212)	(0.065)
User FE	Yes	Yes
Firm FE	Yes	Yes
Observations	1177	1177
R^2	0.44	0.25

Notes: Panel B continues on the next page.

Table 9: AI Response Features and Subsequent Query-Level Soft-Information Production (continued)

	Dependent variable					
	(1) Follow-Up Depth	(2) Reasoning Depth	(3) Cross-Period Synthesis	(4) Multi-Metric Synthesis	(5) Hypothesis Testing	(6) Mechanism-Seeking
Panel B: Query-level components of soft-information production						
Causal Gap	0.242***	0.703***	0.137**	0.043	0.140*	0.668***
	(0.081)	(0.090)	(0.055)	(0.033)	(0.070)	(0.055)
Mechanism Gap	0.209**	0.497***	0.003	-0.017	0.070	0.573***
	(0.075)	(0.088)	(0.029)	(0.011)	(0.041)	(0.046)
Open Question	-0.167***	0.180***	-0.021	0.001	0.056***	0.088**
	(0.028)	(0.049)	(0.017)	(0.010)	(0.019)	(0.034)
Valuation Stance	0.184**	-0.128*	0.010	-0.012	0.018	-0.033
	(0.074)	(0.071)	(0.031)	(0.010)	(0.047)	(0.044)
Uncertainty	0.111***	0.055	-0.008	-0.005	0.023	0.085***
	(0.025)	(0.036)	(0.014)	(0.007)	(0.016)	(0.029)
Length	0.009*	0.004	-0.003	-0.000	0.001	0.003
	(0.004)	(0.004)	(0.002)	(0.000)	(0.002)	(0.005)
Tone	-0.049	-0.020	0.067	-0.000	-0.046	0.263
	(0.187)	(0.194)	(0.041)	(0.063)	(0.111)	(0.159)
Constant	0.270	1.177***	-0.014	0.015	0.079	-0.148
	(0.191)	(0.193)	(0.041)	(0.064)	(0.111)	(0.150)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1177	1177	1177	1177	1177	1177
R^2	0.22	0.46	0.19	0.18	0.31	0.41

Notes: This table reports how features of an AI response predict soft-information production in the participant's subsequent query. The unit of observation is an adjacent AI-response–next-query pair. Panel A reports effects on the query-level Soft Information Index and query-level Delegation. Panel B reports effects on the six query-level components of soft-information production, ordered consistently with the conversation-level measures used elsewhere in the paper. *Causal Gap* indicates that the AI identifies a pattern but leaves its causal explanation unresolved; *Mechanism Gap* indicates that the AI offers a broad conclusion but leaves the operative channel underspecified. *Open Question* is an indicator for whether the AI response explicitly invites further inquiry. *Valuation Stance* is an indicator for whether the AI response expresses a directional valuation-oriented judgment. *Uncertainty* is an indicator for whether the AI response explicitly communicates caveats, ambiguity, or dependence on additional assumptions. *Length* is the word count of the AI response. *Tone* captures the overall rhetorical tone of the AI response using FinBert. All regressions include user fixed effects and firm fixed effects. Standard errors two-way clustered by user and firm and are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Appendix A: Variable Definitions

Variable	Definition
Treatment and Skill Variables	
Medium-Hard	Binary indicator equal to one if the student was assigned to the medium-hard treatment arm; zero otherwise.
High-Hard	Binary indicator equal to one if the student was assigned to the high-hard treatment arm; zero otherwise.
FSA rank	Standardized financial statement analysis (FSA) skill. Students complete a financial statement analysis quiz before the experiment.
Primary Human-coded Conversation-level Measures	
Follow-Up Depth	Human-coded on a 1–5 scale. Measures whether the student deepens inquiry across turns—building on prior AI responses and progressively refining the analysis rather than issuing disconnected requests. Anchors: 1 = no meaningful follow-ups, each query isolated; 5 = student consistently builds on AI answers with probing follow-ups, drilling into specific puzzles.
Reasoning Depth	Human-coded per query (1–5) and averaged to the conversation level. Measures analytical authorship in the student's own query text; whether the student requests data or instead articulates an independent interpretation and asks the AI to verify, challenge, or extend it. Anchors: 1 = pure retrieval or instruction; 2 = directed interpretation request, no student claim; 3 = student provides a partial frame; 4 = student states own hypothesis and asks AI to evaluate it; 5 = student originates a rich analytical claim and asks AI to extend or refine it.
Cross-Period Synthesis	Count of student queries that reference at least two distinct calendar years and contain temporal comparison language (e.g., "trend," "year-over-year," "historically"). Human coders apply a pre-specified two-condition rule to each query; the conversation-level variable is the total count of flagged queries.
Multi-Metric Synthesis	Human-coded count of student queries in which the student integrates multiple financial signals into a coherent interpretation. Each query referencing at least three financial metric domains is classified into one of three categories: *synthesis* (student's query contains an interpretive claim linking at least two metrics), *delegation* (student asks AI to produce an artifact with no student-authored claim), or *retrieval* (parallel data requests). Only synthesis-classified queries are counted.

Variable Definition(continued)

Variable	Definition
Hypothesis Testing	Human-coded on a 1–5 scale. Measures whether the student formulated testable financial hypotheses and used the AI to evaluate or discriminate among them. Anchors: 1 = no hypotheses, student only requests data; 5 = student actively forms and tests hypotheses (e.g., "Did margins improve due to mix shift or cost leverage?").
Mechanism-Seeking	Human-coded binary per query, summed to the conversation level. Coded one if the query asks *why* or *how* an observed pattern arose rather than only asking *what* happened; zero otherwise.
Soft Information Index	Sum of the six z-standardized component scores above: $\sum_{d=1}^{6}(x_{id} - \bar{x}_d)/\sigma_d$.
Complementary Conversation-Level Measures	
Hypothesis-Space Breadth	LLM-coded count of distinct explanatory frames or hypotheses the student proposes or probes across the full conversation. An LLM receives the numbered list of all filtered student queries and counts the distinct causal stories or explanatory angles raised. Pure data requests and rephrases of the same idea do not create new frames.
Analytical Progression	LLM-coded on a 1–5 scale. Measures whether the student's inquiry progressed from basic financial questions toward valuation-relevant reasoning. Anchors: 1 = disconnected data requests with no analytical thread; 3 = some topical sequencing but sections not integrated; 5 = clear coherent arc; either a discovery arc (business context → financial drivers → valuation) or a synthesis arc (student actively interrogates and reconciles AI outputs to form an independent view).
DAG Reasoning	Graph-computed total of validated support edges from evidence nodes to interpretation nodes in the student's conversation DAG. Evidence nodes cover MD&A, Item 1, Risk Factors, Ratio Patterns, DuPont, and Earnings Quality. Interpretation nodes are organized in a strict partial order from operating drivers (L1) to accounting mechanisms (L2) to financial outcomes (L3) to valuation outcomes (L4). Edges are detected via deterministic causal-cue matching and LLM-based implicit-edge detection, then validated against an allowed-edge matrix enforcing valuation-consistent directionality.
Memo-Based Outcomes	

Variable Definition(continued)

Variable	Definition
Valuation Ratio	Ratio of valuation-oriented sentences to total sentences in the student's memo, constructed by applying a pre-specified dictionary of regular expressions covering explicit valuation language, valuation models and inputs (e.g., DCF, WACC, terminal value), financial multiples (e.g., EV/EBITDA, P/E), and valuation-driver categories (growth, margin, reinvestment, risk).
Memo Topics	LLM-coded count of distinct financial analysis topics in the student's memo, drawn from a taxonomy of 13 predefined categories: DuPont, income statement, working capital, earnings quality, revenue quality, competitive analysis, leverage, reinvestment, MD&A, risk factors, macro environment, business overview, and peer comparison.
Delegation Measures	
Delegation Any	Human-coded binary indicator equal to one if the student had at least one delegation turn across all their conversations; measured at the user level. A delegation turn (query)is one in which the student (i) references at least two financial metric domains as ingredients, (ii) asks the AI to produce an integrated analytical artefact, and (iii) contributes no interpretive claim of their own.
Delegation Turns	Human-coded total count of delegation-labeled turns for the student across all conversations; measured at the user level.
Delegation Memo Score	Human-coded 1–5 score assessing the degree to which the student's memo depends on AI responses from delegation turns. Coders compare the memo against the concatenated AI responses from the student's delegation turns. Anchors: 1 = no dependence; 2 = minor use of AI phrasing or structure; 3 = moderate use with some independent framing; 4 = heavy use, memo structure closely tracks AI responses; 5 = memo largely reorganizes or restates AI responses. Clear structural import is required; shared topics or generic finance language are not sufficient evidence of dependence.
Evidence-Type Panel Variables	
Evidence Mentions	Count of student-owned text units (sentences) in which a given evidence node type is detected, within a student–firm unit. Evidence types are: MD&A, Item 1 Business Description (omitted baseline), Risk Factors, Ratio Patterns, DuPont, Earnings Quality. Node detection uses frozen regular-expression triggers drawn from the DAG ontology.

Variable Definition(continued)

Variable	Definition
Support Edges	Count of validated support edges originating from a given evidence type to any L1–L4 interpretation node, restricted to student-owned edges, within a student–firm unit. Edges are identified and validated by the DAG pipeline described under DAG Reasoning above.
AI Response Features	
Causal Gap	LLM-classified binary equal to one if the preceding AI response states what happened but leaves *why* unexplained.
Mechanism Gap	LLM-classified binary equal to one if the preceding AI response names a driver but does not explain the mechanism behind it.
Open Question	LLM-classified binary equal to one if the preceding AI response implicitly or explicitly raises an unanswered question.
Valuation Stance	LLM-classified binary equal to one if the preceding AI response takes a directional valuation position (e.g., implies over/undervalued).
Uncertainty	LLM-classified binary equal to one if the preceding AI response explicitly acknowledges uncertainty, presents multiple scenarios, or hedges its conclusions.
Length	Word count of the preceding AI response, divided by 100.
Tone	FinBERT sentiment confidence score for the preceding AI response.

Appendix B: Sample Companies

The experiment draws on 20 publicly traded U.S. firms selected to provide broad industry coverage while ensuring sufficient SEC filing depth for meaningful financial statement analysis tasks. Each firm's 10-K filings and associated financial data were loaded into the AI agent's retrieval system and assigned to the participants.

Company	SIC Code	FF49 Industry
ACM Research, Inc.	3559	Machinery
Adobe Inc.	7370	Computer Software
Airbnb, Inc.	7000	Restaurants, Hotels, Motels
Alphabet Inc.	7370	Computer Software
Amazon.com, Inc.	5961	Retail
Chipotle Mexican Grill, Inc.	5812	Restaurants, Hotels, Motels
Coinbase Global, Inc.	6200	Trading
DoorDash, Inc.	7370	Computer Software
Duke Energy Corporation	4931	Utilities
EOG Resources, Inc.	1311	Petroleum & Natural Gas
Expedia Group, Inc.	4700	Transportation
lululemon athletica inc.	2300	Apparel
Meta Platforms, Inc.	7370	Computer Software
Microsoft Corporation	7372	Computer Software
Netflix, Inc.	7841	Entertainment
Pentair plc	3561	Machinery
RTX Corporation	3724	Aircraft
Salesforce, Inc.	7372	Computer Software
Uber Technologies, Inc.	4100	Transportation
The Walt Disney Company	4888	Communication

Online Appendix for "Human–AI Collaboration in Corporate Valuation: Experimental Evidence with a Valuation AI Agent"

This online appendix contains five sections. Online Appendix A provides detailed variable construction protocols for all coded measures used in the paper, including human-coding rubrics for the primary soft-information measures and delegation variables, LLM prompts and scoring rubrics for complementary measures, and the DAG extraction pipeline for graph-based reasoning measures. Online Appendix B describes the architecture and technical implementation of the AI valuation agent used in the experiment, covering the data infrastructure, agent orchestration, retrieval and execution engines, module-specific financial analyzers, and system validation procedures. Online Appendix C reproduces the 40-question financial statement analysis knowledge quiz administered to all participants before the experiment, along with the answer key. Online Appendix D presents the detailed coding manual and coder protocol used to implement the human-coded measures. Online Appendix E reproduces the instructions shown to participants at the start of the experimental task.

Contents

Online Appendix A: Variable Construction **4**

A.1 Soft Information Index and Component Measures 4

A.2 Complementary Conversation-Level Measures 6

A.3 Memo-Based Outcomes 7

A.4 Delegation Measures 7

A.5 Evidence-Type Panel Variables 8

A.6 AI Response Features 8

Online Appendix B: System Architecture and Technical Details **10**

B.1 Data Infrastructure 12

B.1.1 Structured Financial Database 12

B.1.2 Unstructured Vector Database 14

B.2 Agent Orchestration 15

B.2.1 Large Language Model (LLM) 15

B.2.2 AI Agent 16

B.2.3 From LLM to Agent Workflow 17

B.2.4 Task Decomposer 19

B.3 Retrieval and Execution Engines 21

B.3.1 Text-to-SQL Structured Query Agent 21

B.3.2 Specialized Data Agents 23

B.3.3 Module Analysis Agent 24

B.3.4 Text-to-Python Execution Agent 25

B.3.5 Retrieval-Augmented Generation (RAG) Agent 26

B.3.6 Generator Agent 27

B.4 Module-Specific Financial Analyzers 29

B.4.1 Business Description Analyzer 29

B.4.2 Beneath the Lines Value Analyzer 29

B.4.3 DuPont Analyzer 31

B.4.4 Earnings Quality Analyzer 31

B.4.5 Financial Visualization Analyzer 32

B.4.6 MD&A Narrative Analyzer 32

B.4.7 Multiples Analyzer . . . 32
B.4.8 Peer Benchmark Analyzer . . . 32
B.4.9 Risk Taxonomy Analyzer . . . 32
B.5 Context Awareness and State Management . . . 33
B.5.1 Dual-Layer Memory Architecture . . . 33
B.5.2 Session Persistence . . . 35
B.6 Frontend-Backend Communication and Interface . . . 36
B.6.1 Asynchronous Streaming Response . . . 36
B.6.2 Containerized Online Service . . . 36
B.6.3 Illustrative Execution Trace . . . 36
B.7 System Validation and Quality Checks . . . 39

Online Appendix C: Financial Statement Analysis Knowledge Quiz **45**

Online Appendix D: Human Coding Manual and Coder Protocol **53**
D.1 General Coding Principles . . . 53
D.2 Financial Metric Domains Used in Multi-Metric and Delegation Coding . . . 54
D.3 Human-Coded Components of the Soft Information Index . . . 55
D.3.1 Follow-Up Depth . . . 55
D.3.2 Reasoning Depth . . . 57
D.3.3 Cross-Period Synthesis . . . 58
D.3.4 Multi-Metric Synthesis Classification . . . 59
D.3.5 Hypothesis Testing . . . 60
D.3.6 Mechanism-Seeking . . . 61
D.4 Human-Coded Delegation Measures . . . 62
D.4.1 Delegation Turn Indicator . . . 62
D.4.2 Delegation Any and Delegation Turns . . . 63
D.4.3 Delegation Memo Score . . . 64
D.5 Adjudication and Final Data Construction . . . 65
D.6 Inter-Rater Reliability . . . 66

Online Appendix E: Instructions to Participants **67**

Online Appendix A: Variable Construction

This appendix describes the construction of all coded variables used in the regression analyses.

A.1 Soft Information Index and Component Measures

The conversation-level Soft Information Index aggregates six component dimensions by z-standardizing each and summing:

$$\text{Soft_Info_Index}_i = \sum_{d=1}^{6} z_{id}, \qquad z_{id} = \frac{x_{id} - \bar{x}_d}{\sigma_d}$$

All six components are coded manually by two independent human coders following pre-specified rubrics; coders are blind to treatment assignment and disagreements are resolved by the coauthor team.

Follow-Up Depth

The full filtered conversation is assessed on a 1–5 scale measuring whether the student deepened inquiry across turns rather than issuing disconnected requests.

Rubric: 1 = no meaningful follow-ups, each query isolated; 5 = student consistently builds on AI answers with probing follow-ups, drilling into specific puzzles.

Reasoning Depth

Each query is scored on how much the student authors the analytical thinking versus delegating it to the AI. Reasoning Depth is the mean across all filtered queries for the unit.

Rubric (1–5): 1 = pure retrieval or instruction; 2 = directed interpretation request, no student claim; 3 = student provides a partial frame, AI fills the rest; 4 = student states own hypothesis and asks AI to verify; 5 = student originates a rich analytical claim and asks AI to extend or refine it. Judge only the student's query text; do not credit copied or delegated content.

Cross-Period Synthesis

Each query is flagged if it (i) references at least two distinct calendar years and (ii) contains temporal comparison language ("trend," "year-over-year," "historically," "over time"). Cross-Period Synthesis is the conversation-level count of flagged queries.

Multi-Metric Synthesis

Queries referencing at least three financial metric domains (revenue, EBIT, net income/EPS, FCF, debt/leverage, margins/returns, working capital, valuation multiples, risk, growth) are classified by human coders into one of three categories.

Classification rubric:

- **Synthesis** — The query contains an interpretive claim linking $\geq$2 metrics; the student is the synthesiser and asks the AI to verify, challenge, or extend. Markers: logical connectives ("yet," "despite," "because," "is this driven by A or B?").
- **Delegation** — The student lists metrics as ingredients and asks the AI to produce an artifact; no analytical claim of the student's own. Markers: imperative verbs (Write/-Draft/Build), multi-metric ingredient list, output artifact specified.
- **Retrieval** — Raw data requests across multiple metrics with no interpretive assembly.
- If a query contains both a task specification and a student claim, classify by whichever is dominant.

Multi-Metric Synthesis is the count of synthesis-classified queries at the conversation level.

Hypothesis Testing

The full conversation is assessed on a 1–5 scale measuring whether the student formulated testable hypotheses and used the AI to evaluate or discriminate among them.

Rubric: 1 = no hypotheses, student only requests data; 5 = student actively forms and tests hypotheses (e.g., "Did margins improve due to mix shift or cost leverage?").

Mechanism-Seeking

Each query is coded 1 if it asks *why* or *how* an observed pattern arose rather than only asking *what* happened; zero otherwise. Mechanism-Seeking is the conversation-level sum.

A.2 Complementary Conversation-Level Measures

Hypothesis-Space Breadth

All filtered queries for a unit are assembled into a numbered list and submitted to GPT-5.2, which counts the number of distinct explanatory frames or hypotheses the student proposes across the conversation. Hypothesis-Space Breadth is this integer count.

Rubric: A *frame* is a distinct explanatory angle or causal story (e.g., "margin improvement is due to mix shift" vs. "margin improvement is due to cost leverage" count as two frames). Rephrasing the same idea does not create a new frame. Pure data requests do not count as frames.

Analytical Progression

The full filtered conversation is submitted to GPT-5.2, which rates the extent to which the student's inquiry progressed from basic financial questions toward valuation-relevant reasoning. Analytical Progression is this score.

Rubric (1–5): 1 = completely unstructured, disconnected data requests; 3 = some topical sequencing but sections not integrated, student covers topics independently; 5 = clear coherent arc throughout, either (A) a discovery arc (business context $\rightarrow$ financial drivers $\rightarrow$ valuation $\rightarrow$ integrated conclusion) or (B) a synthesis arc (student actively interrogates and reconciles AI outputs to form an independent view beyond any single output). A student who merely reads pre-provided outputs in order scores no higher than 3 absent active integration. Use 2 and 4 for intermediate cases.

DAG Reasoning

The DAG pipeline constructs a graph representation of student reasoning from the filtered conversation.

Node detection (deterministic). Frozen regular-expression triggers from a fixed ontology identify two layers of nodes: an evidence layer (MD&A, Item 1, Risk Factors, Ratio Patterns, DuPont, Earnings Quality) and an interpretation layer organized in a strict partial order from operating drivers (L1), to accounting mechanisms (L2), to financial outcomes (L3), to valuation outcomes (L4).

Edge detection (hybrid). Mode A (deterministic) identifies explicit causal cues ("because,"

"drives," "leads to") between co-occurring nodes. Mode B (LLM) identifies implicit edges and assigns modality labels (asserted, hypothesis, conditional, question).

Validation. Edges must pass an allowed-edge matrix, confidence thresholds (≥ 0.65 for causal/implies, ≥ 0.60 for support/weaken), deduplication, a 150-edge cap, and cycle removal.

DAG Reasoning is the total validated support edges in a conversation from any evidence node to any L1–L4 interpretation node, summed across MD&A, Item 1, Risk Factors, Ratio Patterns, DuPont, and Earnings Quality.

A.3 Memo-Based Outcomes

Valuation Ratio

The memo is split into sentences. A regular-expression dictionary covering explicit valuation language, valuation models and inputs (DCF, WACC, terminal value), financial multiples (EV/EBITDA, P/E), and valuation-driver categories (growth, margin, reinvestment, risk) identifies valuation-oriented sentences.

$$\text{Valuation_Ratio}_i = \frac{\#\text{ valuation marker matches}_i}{\#\text{ sentences}_i}$$

Memo Topics

Each memo is classified by GPT into a predefined taxonomy of 13 financial analysis topics: DuPont, income statement, working capital, earnings quality, revenue quality, competitive analysis, leverage, reinvestment, MD&A, risk factors, macro environment, business overview, and peer comparison. For multi-company memos, classification is performed per company section. `n_memo_topics` is the count of distinct topics identified.

A.4 Delegation Measures

A delegation turn satisfies all three conditions: (i) references $\geq$2 financial metric domains as ingredients, (ii) asks the AI to produce an integrated analytical artifact, and (iii) contains no interpretive claim of the student's own. Human coders apply the rubric below; a domain pre-screen restricts coding to queries referencing $\geq$3 metric domains or $\geq$2 named modules.

Coding rubric. Code `delegation = 1` if and only if all three conditions hold. Code 0 if any of: single-module invocation; pure data retrieval; follow-up about prior evidence; student's own

interpretive claim; fewer than 2 distinct domains; procedural or social noise. Be conservative: if unsure, code 0.

Delegation Any

It is a binary indicator equal to one if the student had at least one delegation turn across all conversations.

Delegation Turns

The total count of delegation turns for the student across all conversations.

Delegation Memo Score

Human coders compare the student's investment memo against the concatenated AI responses from the student's delegation turns and assign a copy score.

Rubric (1–5): 1 = no dependence; 2 = minor use of AI phrasing or structure; 3 = moderate use with some independent framing; 4 = heavy use, memo structure closely tracks AI responses; 5 = memo largely reorganizes or restates AI responses. Require clear structural import; shared topics or generic finance language are not sufficient. If ambiguous, choose the lower score.

A.5 Evidence-Type Panel Variables

The dataset is organized as a balanced panel with one row per (student–firm unit × evidence type), yielding six rows per unit. Evidence types are: MD&A, Item 1 (omitted baseline), Risk Factors, Ratio Patterns, DuPont and Earnings Quality.

Evidence Mentions is the count of student-owned sentences in which each evidence node type is detected, using the frozen DAG ontology triggers. Support Edges is the count of validated support or weaken edges from each evidence type to any L1–L4 interpretation node.

A.6 AI Response Features

All variables are computed from the AI response at turn $t-1$ and lagged one turn. The copy/delegation filter is not applied to AI responses.

Causal Gap and Mechanism Gap

Using a pre-specified coding prompt, an LLM classifies each AI response as *causal_gap* when it identifies a pattern, comparison, or change in the data but leaves the underlying cause unresolved—that is, it clarifies *what* happened without explaining *why*. A response is coded as *mechanism_gap* when it offers a broad explanation or names a plausible driver but leaves the operative channel underspecified—that is, it points to a candidate reason without fully tracing *how* that reason produces the observed outcome. Responses that leave no material inferential gap are coded as *none*.

Open Question

The indicator equals one if the response explicitly surfaces an unresolved question for the student to pursue in the next turn.

Valuation Stance and Uncertainty

An LLM evaluates each AI response for (i) a directional valuation position (implies over/undervalued, gives target price, recommends buy/sell/hold) and (ii) explicit acknowledgment of uncertainty or hedging.

Length

The word count of the AI response divided by 100.

Tone

Each AI response is classified by FinBERT as positive, negative, or neutral.

Online Appendix B: System Architecture and Technical Details

This appendix describes the AI valuation agent and supporting infrastructure from a research perspective. The system is designed to solve three methodological problems that are central to our paper: (i) we require precise control over the *information set* exposed to users so that treatment arms differ only in how hard information is structured, not in which facts are available; (ii) we require full logs of human–AI interactions—every prompt, answer, and tool call—to study the micro-dynamics of collaboration; and (iii) we require audit-grade traceability from each AI answer back to specific SEC filings, WRDS fundamentals and ratios, and market data, so that our measures of soft-information production are grounded in verifiable inputs. To meet these requirements, the agent is embedded in a curated data environment (SEC/WRDS/market data only), orchestrated by a deterministic tool-calling layer, and instrumented to record the complete data path for every interaction.

For readers primarily interested in the research design rather than implementation details, the agent provides the following features that are central to our empirical strategy:

- **No future information.** For each firm–valuation date, the agent can access only SEC filings, WRDS fundamentals and ratios, macro series, and market data that were available as of that date. All retrievals (SQL and RAG) are time-sliced accordingly, eliminating look-ahead bias and ensuring that students work with the same information set a real-time analyst would have had.

- **No hallucinated numbers (faithfulness).** All numerical outputs (e.g., revenues, margins, growth rates, multiples) are returned directly from the MySQL warehouse or from sandboxed Python computations built on those data; the base LLM is never allowed to invent or overwrite numerical values or to query the open internet. Every reported figure can therefore be traced back to a specific Compustat field or derived calculation, preserving accounting fidelity.

- **Full citation trail (traceability).** Every factual statement in the AI's answer is accompanied by explicit citations to the underlying evidence (e.g., "SEC 10-K Filings, Item 1A, 2023" or "Compustat Database, datadate 2022-12-31"), and every tool call (SQL query, vector retrieval, analyzer invocation) is logged. This enables an audit-grade recon-

struction of how each answer was generated and how each narrative statement relates to underlying disclosures and financial statements.

- **Clean experimental variation (controllability).** Across treatment arms, the underlying information set (filings, fundamentals, and market data up to the valuation date) is identical; only the *modules* the agent may surface (bare retrieval, dashboards, dashboards plus valuations) differ. This design underpins our identification strategy: differences in behavior and soft-information production across arms can be attributed to the structure of AI-provided hard information, not to differences in what information is available.

- **Reproducible behavior (reproducibility).** The agent follows deterministic routing rules, uses read-only SQL and versioned, time-sliced data, and runs in a containerized environment. The same query in the same experimental configuration yields the same retrievals, tool calls, and outputs (up to controlled randomness in natural-language phrasing), allowing other researchers to replicate both the information environment and the agent workflow.

- **Rich process data on human–AI collaboration.** Because every prompt, answer, and tool call is logged with timestamps and metadata, we observe not only task-level outputs (valuations, memos) but also the full micro-level process by which students interact with the agent. This process data allows us to measure soft-information production directly (quantity, breadth, specificity, originality) and to relate it causally to both AI design and students' ex ante financial-statement-analysis skills.

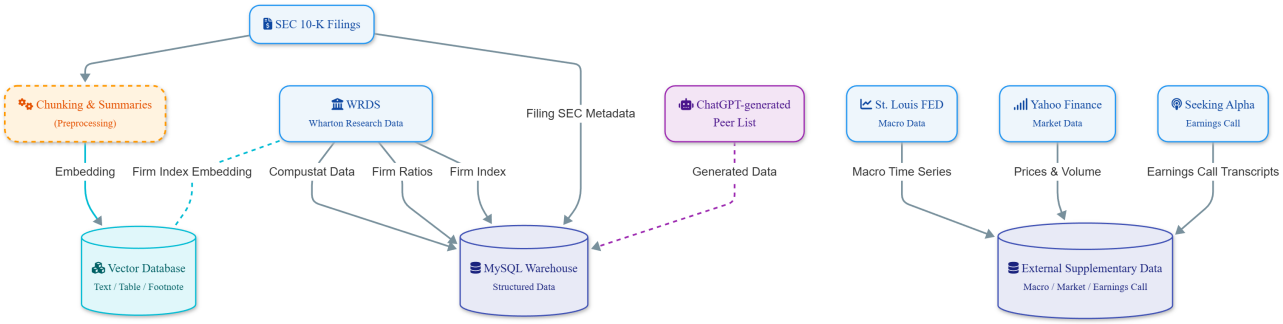
These design principles distinguish our system from a generic "ChatGPT on 10-Ks." Rather than a single opaque model prompted on filings, our platform combines a vector index of time-sliced SEC text, a relational warehouse of WRDS fundamentals and ratios queried via read-only SQL, and a suite of precomputed analyzers tailored to corporate valuation (e.g., DuPont decompositions, earnings-quality diagnostics, peer benchmarking). A deterministic agent orchestrator decides which tools to call and in what order, and all steps are logged.

The remainder of this appendix documents the end-to-end architecture of the AI-assisted research platform: Section B.1 and Section B.2 describe the data infrastructure and agent orchestration, Section B.3 details retrieval and execution agents, Section B.4 outlines precomputed financial analyzers, and Sections B.5 and B.6 cover context management and the user-facing interface used in the experiment.

B.1 Data Infrastructure

This section summarizes the end-to-end data infrastructure that supports all downstream retrieval, valuation, and hybrid reasoning components. From an accounting perspective, this infrastructure ensures that every number, ratio, and disclosure the agent uses comes from standard research data sources (SEC/WRDS/market data) and can be tied back to the same Compustat and EDGAR records that capital-markets studies routinely analyze. As shown in Figure 1, the system integrates structured financial databases with an unstructured vector database constructed from parsed 10-K filings. The unified pipeline ensures deterministic linking across CIK, gvkey, and accession identifiers, enabling both relational and semantic retrieval to operate over a consistent data backbone.

Figure 1: End-to-End Data Ingestion Architecture



Notes: This figure shows the end-to-end data ingestion pipeline that underlies the structured and unstructured components of the system. SEC 10-K filings are parsed into chunked narrative, tabular, and footnote artifacts used to populate the vector database, while filing-level metadata (CIK, accession numbers, reporting periods) are synchronized with WRDS. WRDS supplies the Compustat fundamentals, firm-ratio libraries, and the CIK–gvkey index, all of which are ingested into the MySQL warehouse with versioned schemas. A ChatGPT-generated peer list feeds into the same warehouse, producing a unified, auditable store of firm identifiers, fundamentals, ratios, metadata, and peer mappings. In addition, three external data sources—St. Louis FED (macro time series), Yahoo Finance (market data), and Seeking Alpha (earnings-call transcripts)—are ingested into a dedicated External Supplementary Data store.

B.1.1 Structured Financial Database

Data Source. Compustat fundamentals (annual and quarterly), firm-level ratio libraries, and SEC filing metadata are extracted directly from Wharton Research Data Services (WRDS). Annual statements are queried from Fundamentals Annual to guarantee reporting comparability, while quarterly statements rely on Fundamentals Quarterly over a rolling calendar window to capture complete intra-year coverage. Firm ratios also originate from the WRDS feed, ensuring

consistent factor definitions across valuation, profitability, leverage, and efficiency metrics. Filing metadata is synchronized with the SEC JSON feeds so that every downstream record can be indexed by CIK, accession number, and reporting period, enabling deterministic joins between textual disclosures and numerical facts. In parallel, the peers mapping file is precomputed offline by prompting ChatGPT to enumerate the closest competitors for each covered company, and those ChatGPT-generated lists are serialized into the JSON arrays that ultimately populate the peers mapping table.

MySQL Introduction and Implementation. MySQL is a relational database system that offers ACID transactions, declarative SQL, and mature indexing—capabilities that make it a natural fit for financial research repositories where analysts expect deterministic joins, strict primary keys, and long-lived historical archives. We deploy it as the authoritative warehouse for all structured feeds so downstream agents can reuse standard SQL semantics instead of bespoke file readers.

After extraction, all structured datasets are written to caches and ingested into a MySQL warehouse via an automated schema setup routine. The schema builds five primary tables: `filing_metadata` for filing-level metadata, `fundamentals_annual` and `fundamentals_quarterly` for Compustat statements, `firm_ratios` for the standardized ratios, and `peers_mapping` for JSON-encoded peer clusters. Because Compustat annual files contain hundreds of attributes, the schema intentionally keeps the most frequently queried metrics (revenues, profitability, cash flows, balance-sheet drivers, per-share and valuation series) in fixed-precision DECIMAL columns, while less frequently accessed attributes fall back to TEXT to avoid row size limits without sacrificing completeness. All primary identifiers (cik, gvkey, fyear) are indexed, and historical migrations automatically rename legacy tables to preserve continuity. Data ingestion batches clip extreme ratio values to stay within DECIMAL bounds, deduplicate on business keys (e.g., cik + fyear), and convert JSON peer arrays into strings for native MySQL JSON columns. The connector layer pools connections, enforces SELECT-only execution for generated SQL, and exposes helper methods for company-level retrieval so that higher-level agents can fetch Compustat or ratio data with a consistent API. In addition, we maintain a dedicated company-index table that maps firm names to ticker, CIK and gvkey identifiers, providing a reconciliation layer across SEC, Compustat, and auxiliary datasets; this index enables cross-database lookups to remain deterministic even when upstream data sources use different identifier regimes.

B.1.2 Unstructured Vector Database

Data Source. Every 10-K filing is parsed into JSONL artifacts that preserve both HTML items (e.g., Item 1, Item 7, Item 8) and embedded XBRL statements. These artifacts feed the unstructured knowledge base so that textual narratives, tabular disclosures, and footnotes remain synchronized with the corresponding structured records. The pipeline also records the originating CIK, accession number, and item metadata for each chunk, which later allows hybrid retrieval to constrain searches to a company, a filing year, or a specific section such as Item 1A.

Data Cleaning and Chunking. An offline preprocessing job standardizes the filings before indexing. HTML sections are stripped of tables, tokenized with a fixed window plus overlap to preserve context, and filtered for minimum length. XBRL statements are transformed into human-readable tables and linearized CSV strings so numerical disclosures can be retrieved in either narrative or tabular form. Footnotes undergo a LLM summarization that auto-detects or follows provided schemas, returning JSON objects with schema, summary, and significance fields; these summaries are converted into dedicated "footnote" chunks so the retriever can answer detailed accounting questions without reprocessing the raw filing at query time. Throughout the process, chunk files inherit normalized metadata (CIK, accession number, item, item title) to keep the unstructured corpus perfectly aligned with the structured warehouse.

Vector Database. A vector database stores numerical embeddings of documents instead of (or in addition to) the raw text. That means every chunk of text, table, or footnote is translated into a point in a high-dimensional space. When a user asks a question, the system embeds the query into the same space and looks for nearby points, which correspond to semantically similar passages—even if the wording does not match exactly.

We use Weaviate v4 for this role because it supports hybrid (lexical + vector) search, fine-grained structured filters, and independent collections with named-vector support, which makes it well-suited for heterogeneous SEC filings. Compared with other open-source vector stores (e.g., Milvus or Qdrant), Weaviate provides a more convenient schema-first API and native hybrid search, which are critical for retrieving specific disclosure segments while preserving metadata-level control.

Three collections keep different content types logically separate—narrative text, tables, and footnotes. Each collection stores semantic properties (chunk ID, section, accession number,

source type), enabling the retriever to issue constrained queries such as "only Item 1A paragraphs from Apple's 2023 filing" or "consolidated income statement from the latest Microsoft 10-K," while still benefiting from vector-level semantic matching.

Embedding and Vector Storage. All chunks are embedded offline with OpenAI's text-embedding-3-large model to ensure consistent semantic geometry. For tables, the embedding input concatenates the caption with a linearized version of the rows; for footnotes, it includes both the extracted schema and the summarized content. Batch ingestion writes these vectors (and their metadata) to the appropriate Weaviate collection, retrying failed objects and logging diagnostics. We also maintain a dedicated "embedding firm index," where each company's name, aliases, and common natural-language variants are embedded and linked to its ticker, CIK and gvkey, ensuring that fuzzy queries referencing firm names (e.g., "Apple", "Apple Inc" or "AAPL") resolve deterministically to the correct identifiers. By storing both vectors and numerics, the system can stream back rich snippets—complete tables, summarized footnotes, or long-form paragraphs—that the language model cites verbatim, preserving accounting fidelity in the final answer.

For our purposes, this means that when the agent answers questions about MD&A, risk factors, or footnotes, it is always drawing on the same text that accounting researchers would hand-code or classify in a disclosure study, but doing so at scale and with precise linkage to firm identifiers and reporting periods.

B.2 Agent Orchestration

By forcing the LLM agent to operate through this orchestration layer, we ensure that its role in our study mirrors that of a junior analyst who can read, summarize, and cross-reference disclosures, rather than an unconstrained text generator, such as a generic ChatGPT. This design choice is critical for interpreting our experimental results as evidence about how AI can assist with, rather than replace, standard accounting-based valuation work.

B.2.1 Large Language Model (LLM)

A Large Language Model (LLM) is a deep neural network trained to predict the next token in a sequence—a simple objective that, at scale, enables the model to acquire broad linguistic and conceptual knowledge (Radford et al., 2019; Brown et al., 2020). Modern LLMs contain billions

of parameters and are trained on heterogeneous text corpora, allowing them to perform tasks such as summarization, question answering, and interpretation of domain-specific terminology without task-specific supervision.

In our system, the LLM serves as the reasoning layer. It reads a researcher's natural-language query, interprets accounting and financial-reporting terminology, and composes an answer that is both linguistically coherent and contextually appropriate. However, an LLM by itself has no access to firm-level data, cannot retrieve SEC filings or WRDS fundamentals, and does not maintain persistent memory. Its outputs reflect statistical patterns in its training data rather than authoritative financial information.

To transform the model into a reliable research assistant, we embed it inside a controlled retrieval-and-reasoning loop inspired by retrieval-augmented generation (RAG; Lewis et al., 2020) and tool-augmented LLM architectures (Schick et al., 2023). This loop supplies the model with verified numerical and textual data, constrains the sources it is allowed to reference, and logs every intermediate step. The LLM therefore provides flexible natural-language reasoning, while the surrounding system guarantees data fidelity, reproducibility, and auditability—properties essential for accounting and capital-market research.

Importantly, in our setting, the LLM's only role is to interpret accounting and finance concepts in natural language and to explain or synthesize information retrieved from authoritative data sources. It never supplies its own financial 'knowledge' and therefore cannot contaminate the experiment with unverified numbers or external narratives.

B.2.2 AI Agent

An agent wraps the LLM with orchestration logic: rather than allowing the model to operate as an unconstrained text generator, the agent determines which tools to invoke, SQL queries against the structured data warehouse, vector searches over the vector database collections, or Python-based financial analytics sandbox, collects their outputs, and feeds the consolidated evidence back into the model. A useful analogy is to think of the agent as a coordinated team of analysts, each responsible for a specific task such as pulling filings, computing ratios, resolving firm identifiers, or extracting disclosure segments; once every analyst finishes their part, the "manager" (the LLM for generation) receives the complete packet of data and context needed to produce a coherent, sourced final report.

Contemporary agent systems typically follow one of three patterns: (i) prompt-only agents,

where the LLM answers without external tools; (ii) tool-use agents, where the model autonomously chooses when to call a tool; and (iii) router–tool–generation agents, where the system first routes the user query to the appropriate tool, then executes the tool, and finally synthesizes all retrieved information—an architecture that yields higher reliability, modularity, and auditability. Our design follows this third pattern and is grounded in emerging graph-based agent frameworks such as Wang and Duan (2024), who introduce LangGraph as a structured orchestration system for tool-augmented LLMs. We implement our agent directly in LangGraph, treating each operation (identifier resolution, SQL retrieval, vector search, numeric analysis, answer drafting) as a node in a deterministic state machine with explicit transitions and robust failure handling. Each run produces a complete audit trail of tool calls, intermediate data, and LLM prompts, ensuring that the workflow remains reproducible, inspectable, and appropriate for financial-research settings where traceability and data governance are essential.

This agent formulation is particularly well suited to corporate valuation, where analysts routinely pull from multiple data silos (financial statements, MD&A, risk factors, peers, macro conditions) and where the credibility of any recommendation depends on being able to show exactly which disclosures and ratios support it.

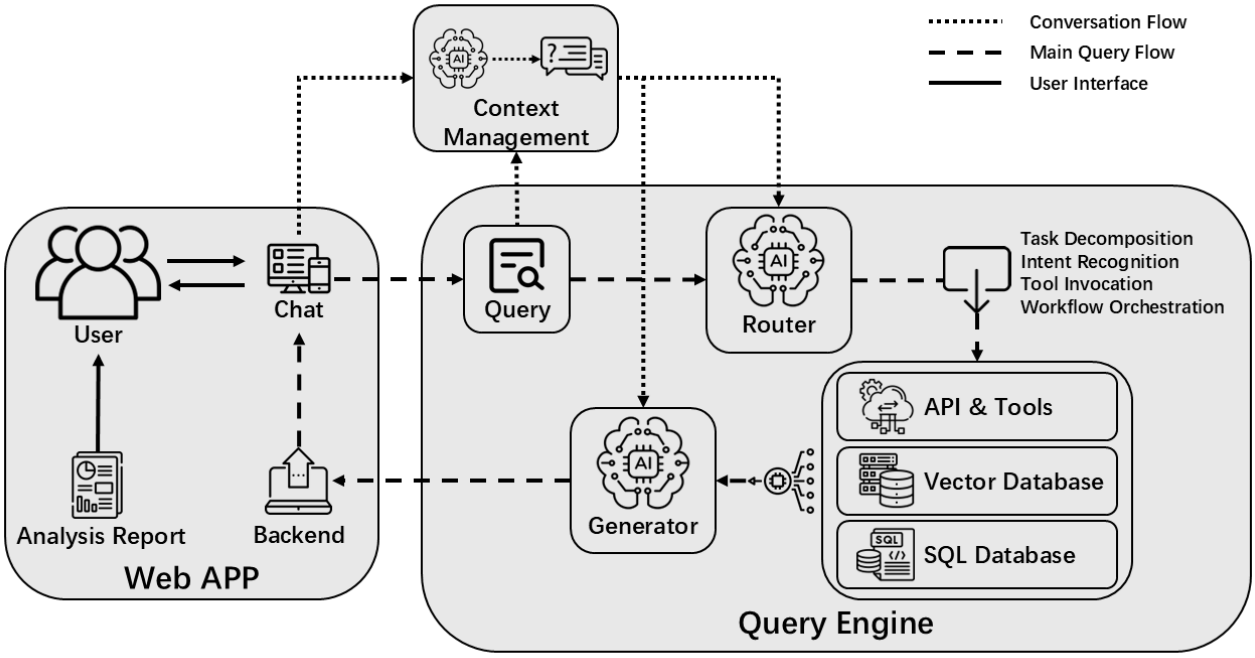
B.2.3 From LLM to Agent Workflow

When a user submits a question, the system receives not only the raw text but also the active analytical context (for example, that the user is currently examining MSFT 2023 Business Description Narratives). The agent then executes a three-stage workflow that mirrors how a team of analysts would process a research request. First, the routing module determines which subtasks are required—SQL queries for fundamentals, vector retrieval for narrative disclosures, Python routines for ratio computation, or external calls for market and macro data. This step functions like assigning work across specialized analysts who each understand a different data source or analytical procedure.

Next, the execution layer runs these subtasks in parallel: filings are pulled from Weaviate, financials are retrieved through Text-to-SQL, cross-sectional metrics are computed via Text-to-Python, peer sets are reconstructed, and any necessary external data feeds are queried. Each operation corresponds to a node in our LangGraph state machine, ensuring deterministic transitions, explicit error handling, and a complete audit trail of inputs, outputs, and tool calls.

Finally, the generation module assembles all retrieved evidence into a structured packet of

Figure 2: Agentic Query Workflow and Context-Aware Orchestration



Notes: This figure shows the end-to-end workflow through which user queries are interpreted, routed, and answered. The Web App handles the chat interface and analysis reports, while the Query Engine executes the core agentic logic. Incoming questions pass through the router, which performs task decomposition, intent recognition, and tool selection. Depending on the query, the agent invokes structured-data retrieval (SQL), unstructured retrieval (vector search), analytical computation, or auxiliary data modules. Retrieved evidence flows into the generator, which assembles a context-grounded response, enforces citation rules, and streams the output back to the user. A dedicated context-management layer tracks conversation summaries and active analysis modules, ensuring that follow-up questions are interpreted relative to what the user is currently viewing. The dashed, dotted, and solid lines distinguish the main query path, conversational state updates, and user-interface interactions, respectively.

context and streams it through the response graph, where the LLM produces the final narrative answer under constraints that enforce citation format, attribution rules, and stylistic consistency. This loop repeats at every turn of the conversation: incremental summaries preserve the user's intent while keeping the context window compact, allowing the agent to maintain continuity across long analytical interactions. The overall workflow—illustrated in Figure 2—ensures that each answer is the product of coordinated tool-use, grounded data retrieval, and controlled natural-language reasoning, in a manner that is reproducible and suitable for audit-grade financial research.

Table 1: Query Lifecycle

Stage	Description
User Input	User submits natural-language question
Decomposition	Deciding if the query contains multiple tasks
Planning	Routing tasks selected
Retrieval	SQL execution + vector search + macros/markets
Execution	Python sandbox, computations, subtask cascades
Generation	LangGraph generator builds final answer
Summary	Turn-level structured JSON
Display	Streaming output to the user

Notes: This table summarizes the lifecycle of a user query within the agentic architecture. A natural-language question first passes through routing and planning, where the system determines which tools are required—such as structured retrieval, vector search, analytical computation, peer expansion, or macro/market data. The retrieval stage executes these tool calls in parallel, allowing SQL queries, hybrid vector searches, macro pulls, and other data fetches to run concurrently for lower latency. The execution stage processes analytical tasks such as Python-based computations or multi-step subtasks. The generator then assembles a grounded, citation-complete response. In parallel with the streaming output shown to the user, the system also produces a structured summary of the turn, enabling context-aware routing in subsequent interactions without replaying the full transcript.

B.2.4 Task Decomposer

Chain-of-Thought Planning. The task decomposer handles queries that bundle several analytical requests into a single sentence. When a user asks, for example, "Compare Microsoft and Apple revenue growth, then benchmark against peers," the system switches the LLM into a planning mode where it explicitly reasons about the component steps. Instead of treating the question as one monolithic task, the decomposer produces a small set of sub-queries (e.g., retrieve revenue series for Microsoft, retrieve revenue series for Apple, compute growth rates,

construct peer benchmarks) together with a brief rationale for why the query needs to be broken apart. Each sub-query then runs through its own routing and execution mini-cycle, and the resulting outputs are merged into a consolidated context before the final answer is drafted. This prevents the central router from overloading a single execution path and mirrors how human analysts naturally decompose multi-part requests into smaller, verifiable steps. In valuation tasks, this prevents the agent from skipping critical intermediate steps, such as computing growth rates or examining segment margins, and ensures that multi-part questions (e.g., 'compare margins, then assess sustainability') are answered in a way that mirrors standard FSA pedagogy.

Dynamic Parameter Extraction. In addition, the decomposer performs dynamic parameter extraction. It identifies implicit details—such as which firms are mentioned, which fiscal years are relevant, whether peers are required, and which analytical module is in focus—by reading both the question and the current analysis context. These parameters are fed back as explicit hints to the downstream tools: SQL components receive concrete year and firm filters, vector retrieval is pointed to the correct filings, and peer modules know which focal company to expand. Because this extraction is embedded in the same reasoning process that drives task decomposition, the resulting "plan" carries a transparent trail of how entities and time periods were chosen, making it possible to audit and explain why a particular dataset or horizon underlies any given answer.

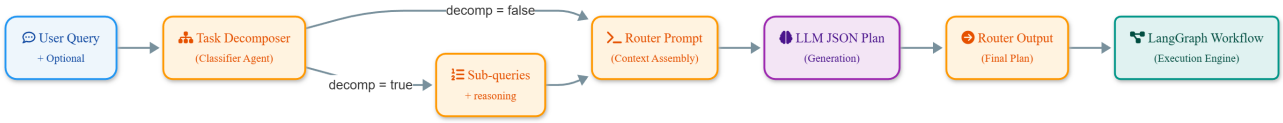
Central Query Router

Intent Routing. The central query router serves as the agent's coordination layer: it interprets the user's request, determines which analytical steps are required, and produces a structured task plan that downstream components can execute reliably. Rather than letting the LLM freely improvise, the router provides it with the user's question, the relevant filing years, and a set of system guidelines, and expects a clean, machine-readable plan describing which operations should be performed—for example, narrative retrieval (RAG), SQL-based financial extraction, Python-based computation, or peer-firm lookup. This design follows recent work on tool-augmented LLMs that explicitly separate high-level planning from low-level tool invocation, such as ToolPlanner, which learns to map user instructions into multi-step tool plans before execution (Wu et al., 2024). The router then verifies that the plan is well-formed,

standardizes the task names, and applies guardrails (e.g., automatically including a fundamentals query when numerical data are requested). This ensures that every subsequent node knows exactly which tools to invoke and in what sequence, reducing ambiguity, hallucinated steps, and ad-hoc prompt tuning.

Context-Aware Routing. Because user queries often depend on what is currently displayed on screen, the router is also context-aware. When the user elects to "add analysis to chat," the system forwards the full markdown of the active module—such as an Item 1A risk summary, a DuPont decomposition, or a chart from a prior step. The router incorporates this material into its prompt so the LLM understands the user's present focus. As a result, follow-up questions about a chart trigger retrievals from the corresponding disclosure section, whereas a request to "compare this to peers" automatically adds the peer-retrieval task. This mechanism keeps the agent aligned with the user's workflow, grounding routing decisions in what the user is actually examining and ensuring that the resulting plan reflects both the query and its analytical context.

Figure 3: Planning and Routing



Notes: This figure shows the front-end of the agent's reasoning workflow. A user query first enters the Task Decomposer, which determines whether the request requires multi-step decomposition. If decomposition is necessary, the system produces a set of structured sub-queries; otherwise, the original query flows directly into the Router Prompt, where contextual elements (active module, company-year metadata, prior turn summary) are assembled. An LLM then generates a machine-readable JSON plan specifying the required analytical tasks. The router validates and normalizes this plan before sending it into the LangGraph Workflow, ensuring deterministic, well-formed task execution.

Because the router must explicitly declare which financial statements, ratios, and modules it will consult, we can trace and later reconstruct whether an answer was built on income statement dynamics, cash flows, risk disclosures, or peer comparisons—information.

B.3 Retrieval and Execution Engines

B.3.1 Text-to-SQL Structured Query Agent

Text-to-SQL. Text-to-SQL provides the interface between natural-language queries and the structured financial data stored in our relational warehouse. When a user asks a question such

as "Show Microsoft's revenue trend since 2020," the system converts this conversational request into a precise SQL statement that the database can execute. The technique traces back to early neural semantic-parsing work Seq2Sql (Zhong et al., 2017) and was substantially advanced by large cross-domain datasets such as Spider (Yu et al., 2018), which demonstrated that models can reliably translate natural language into complex, multi-table SQL under realistic schema constraints. More recent approaches (e.g., Scholak et al., 2021) show that LLMs can generate structurally valid queries when provided with schema-aware prompts and clear execution constraints—principles we adopt here.

In our system, the agent supplies the LLM with a schema guide describing the available tables, identifiers, and financial attributes, and asks the model to produce a compliant SQL statement tailored to the question. This design frees users from knowing table names or join keys while ensuring that every retrieved value ultimately comes from an auditable, ACID-compliant data store. The database enforces data integrity, and the Text-to-SQL layer guarantees that the LLM requests exactly—and only—the rows required for the analysis, providing a transparent, reviewable pathway from natural-language intent to structured financial evidence.

In practice, this layer ensures that when a user asks for 'revenue growth since 2020' or 'ROE over the last five years,' the agent is running the same kinds of Compustat queries that an empirical accounting researcher would write, not approximating these quantities from memory or heuristics.

Schema Linking: From Natural Language to Accounting Variables. Schema linking ensures that the LLM's interpretation of a user's question aligns with the actual accounting variables stored in the database. Before drafting a query, the agent provides the model with a concise description of the available tables, their key identifiers, and the major financial attributes they contain. For example, the schema prompt may note that the fundamentals table includes items such as revenue and net income, while the ratios table contains measures like return on assets. Router hints—such as the relevant fiscal years or the focal company—are placed directly beneath the user's question so the model knows which filters to apply. With this structure in place, the generated SQL is anchored to the true Compustat variable names, preventing ambiguous terms like "sales" from drifting away from the intended revenue field and ensuring that every query maps cleanly onto authoritative accounting data.

Self-Correcting Generation. Self-correcting generation ensures that the SQL produced by the model is both structurally valid and safe to execute. After the LLM proposes an initial query, the agent performs a series of lightweight checks to confirm that the statement is well-formed and adheres to system rules (for example, that it is a read-only query and references recognizable variables). If the draft contains inconsistencies—such as an unrecognized field name or an improperly constructed filter—the agent triggers a second pass in which the LLM is presented with the earlier draft and a short explanation of what went wrong. This feedback loop allows the model to revise its own output, much like an analyst correcting a colleague's draft based on reviewer comments. The two-step process substantially reduces execution errors and ensures that only validated, syntactically sound statements reach the database, reinforcing the system's emphasis on reliability and safety.

B.3.2 Specialized Data Agents

The system relies on a set of specialized data agents that supply authoritative external information whenever a user's question extends beyond firm-level fundamentals. These agents operate in parallel with the main workflow and return structured, source-labeled data so that the final answer remains grounded in verifiable evidence. These specialized agents let us expose users to precisely the macro and market information that valuation textbooks recommend using (risk-free rates, market multiples, recent management commentary), while still ensuring that every such input is timestamped, sourced, and isolated from future data.

Macro Data Agent. The macro data agent handles questions about the broader economic environment—such as movements in GDP, inflation, interest rates, or labor-market conditions. It draws from a curated subset of Federal Reserve Economic Data (FRED) series, retrieves the most recent observations across categories, and produces both short summaries and tabular extracts. Centralizing these pulls prevents redundant API calls and ensures that each datapoint is accompanied by its official series identifier and measurement units, preserving traceability for citation and audit purposes.

Market Data Agent. The market data agent provides real-time or near–real-time market metrics, including prices, volumes, market capitalizations, valuation multiples, and dividend yields from Yahoo Finance. Data are sourced from standard public feeds and always include a

timestamp (e.g., "Data as of 2024-09-30 16:00 ET"), allowing downstream components to reference market conditions with temporal clarity. Basic formatting rules convert large magnitudes into readable units (billions, millions) while retaining ratios in their native scales, giving the generation module precise values to cite.

Earnings Call Agent. The earnings-call agent supplies the most recent management commentary by returning curated transcript segments keyed to a firm's ticker and quarter from Seeking Alpha. When a transcript is available, the agent returns the raw paragraphs; when it is not, it produces an explicit "not available" record so the system can report the absence rather than fabricate quotations. Each snippet is tagged with its source ("Earnings Call Transcripts"), ensuring that citations remain clear and unambiguous in the final output.

Item Summary Agent. The item summary agent retrieves entire sections of a firm's 10-K (e.g., Item 1, Item 1A, Item 7) from the vector database and concatenates all relevant chunks into a coherent section-level digest. Instead of re-parsing filings during generation, the agent supplies these preassembled section snapshots whenever the router requests specific document sections. Because the summaries are literal extractions (not re-written paraphrases) and retain CIK/accession metadata, they function as a high-recall text source that can be injected into RAG or cited directly when analysts ask for "Item 1A risks" or "latest MD&A discussion."

Together, these specialized agents extend the system's reach beyond financial statements to macroeconomic conditions, market activity, and managerial communication—three information channels central to empirical accounting and capital-market research—while maintaining the same standards of verifiability and provenance that govern the core structured and unstructured data pipeline.

B.3.3 Module Analysis Agent

The module analysis agent provides access to a library of precomputed analytical reports that summarize different dimensions of a firm's disclosure and performance. These reports, covering areas such as business description, MD&A narrative patterns, risk taxonomy, financial visualizations, earnings quality diagnostics, DuPont decomposition, valuation multiples, peer benchmarks, and beneath-the-line value contributions, are generated offline by specialized analyzers using the full underlying datasets. Each report is stored as a structured markdown

document containing narrative explanations, visualizations, peer comparisons, and intermediate calculations. From a valuation perspective, these modules correspond closely to standard building blocks in FSA and equity research.

When the router specifies a particular analytical module, the agent retrieves the corresponding report for the relevant firm and fiscal year, ensuring that the LLM receives a consistent and authoritative summary rather than re-deriving the analysis on the fly. When a user opens one of these modules in the interface and selects "Add to chat," the associated markdown becomes the active analytical context. This context then guides the routing, decomposition, and generation steps, allowing follow-up questions to be grounded in the exact figures, charts, and reasoning already presented to the user. The module agent therefore functions as a bridge between offline, reproducible analysis and online, interactive reasoning—ensuring that every response is anchored in preverified analytical content rather than ad hoc model inference.

B.3.4 Text-to-Python Execution Agent

Text-to-Python. The Text-to-Python agent handles questions that require bespoke calculations—such as multi-year growth rates, DuPont decompositions, customized scoring rules, or transformations that go beyond what a single SQL query can express cleanly. Rather than hard-coding every possible formula, the system adopts a Text-to-Python approach, following recent work showing that LLMs can reliably generate small, auditable code snippets when provided with structured context (e.g., Wang et al., 2024). The agent supplies the model with a concise description of the available DataFrames—including their column names and basic structure—and asks it to write a short Python routine that computes the requested metric and assigns the final value to a designated variable.

This design mirrors how an analyst might write a quick pandas script to explore ratios or produce a derived measure, but preserves the conversational interface: the user poses a natural-language question, and the system produces both the structured data (via SQL) and the corresponding computation (via Python). Because the generated code is small, self-contained, and executed in a controlled runtime, it can be logged, inspected, and reproduced—ensuring that even customized calculations remain transparent and audit-grade. By extending the agent's capabilities beyond retrieval and into controlled computation, Text-to-Python closes the loop between natural-language intent, structured data access, and verifiable analytical output.

The Text-to-Python agent is important in our experiment because it allows the agent to

support bespoke but transparent valuation calculations, such as multi-stage DCFs or custom scoring rules, without ever leaving the audit trail of how each number was computed.

Execution Sandbox. Generated code is executed inside a tightly controlled sandbox that allows only the numerical operations needed for standard financial analysis. The environment exposes a minimal toolset while blocking external imports, file access, and system calls. The computed result must be assigned explicitly, ensuring that the model produces a clear, inspectable output rather than side effects or intermediate artifacts. After the calculation completes, the sandbox returns the final value together with any intermediate messages and the exact code that produced it, creating a complete audit trail. If the computation fails, for example because a column is missing or a ratio divides by zero, the resulting error is passed back to the orchestration layer, which can either fall back to retrieval-based reasoning or inform the user why the calculation could not be completed. This controlled execution environment allows the system to support flexible, conversational analytics while maintaining the safety, transparency, and reproducibility required for financial research.

B.3.5 Retrieval-Augmented Generation (RAG) Agent

Introduction to RAG. The RAG enhances the LLM by giving it access to the actual disclosures relevant to a user's question. Instead of relying solely on patterns learned during pretraining, the model receives verbatim excerpts from the firm's filings—such as Item 1A risk factors, MD&A paragraphs, financial tables, or summarized footnotes—retrieved in real time from the vector database. This approach builds on foundational work showing that supplying grounded evidence to a model substantially improves factual accuracy and reduces hallucination (Lewis et al., 2020). In our setting, the retrieved materials come with full metadata, including accession numbers, fiscal years, and section identifiers, and the LLM is explicitly instructed to cite them. As a result, every claim in the final answer can be traced directly to a specific filing and location, giving auditors and researchers a transparent pathway from natural-language reasoning back to authoritative disclosures.

Hybrid Search Algorithm. Hybrid search combines semantic and lexical retrieval to identify the disclosure segments most relevant to a user's question. Building on work showing that dense embeddings and sparse keyword signals capture complementary notions of relevance (e.g., Karpukhin et al., 2020; Gao et al., 2021; Sun et al., 2025), the system searches the vector

database using both embedding similarity and traditional lexical matching. These results are then filtered by structured metadata—such as CIK, accession number, and filing section—so that only passages from the correct document set are considered.

After this initial retrieval stage, the candidate passages are passed to a cross-encoder reranker, which evaluates each query–passage pair jointly rather than independently. Unlike embedding similarity, which scores each document in isolation, a reranker reads the entire text of the query and the candidate snippet together and produces a refined relevance score based on their contextual alignment. This second stage sharply improves ranking precision, especially for accounting-specific queries that hinge on subtle distinctions between segments (e.g., risk factors vs. forward-looking statements, operating metrics vs. footnotes). The system reranks narrative excerpts and table entries separately and keeps only the top passages from each stream.

The final context provided to the LLM therefore contains a curated blend of high-relevance narrative snippets and linearized tables. Because these passages survived both the hybrid search and the reranker's joint evaluation, they offer an evidence set that is both broad and precise—combining qualitative interpretation with quantitative support and ensuring that answers remain fully grounded in verifiable disclosures.

B.3.6 Generator Agent

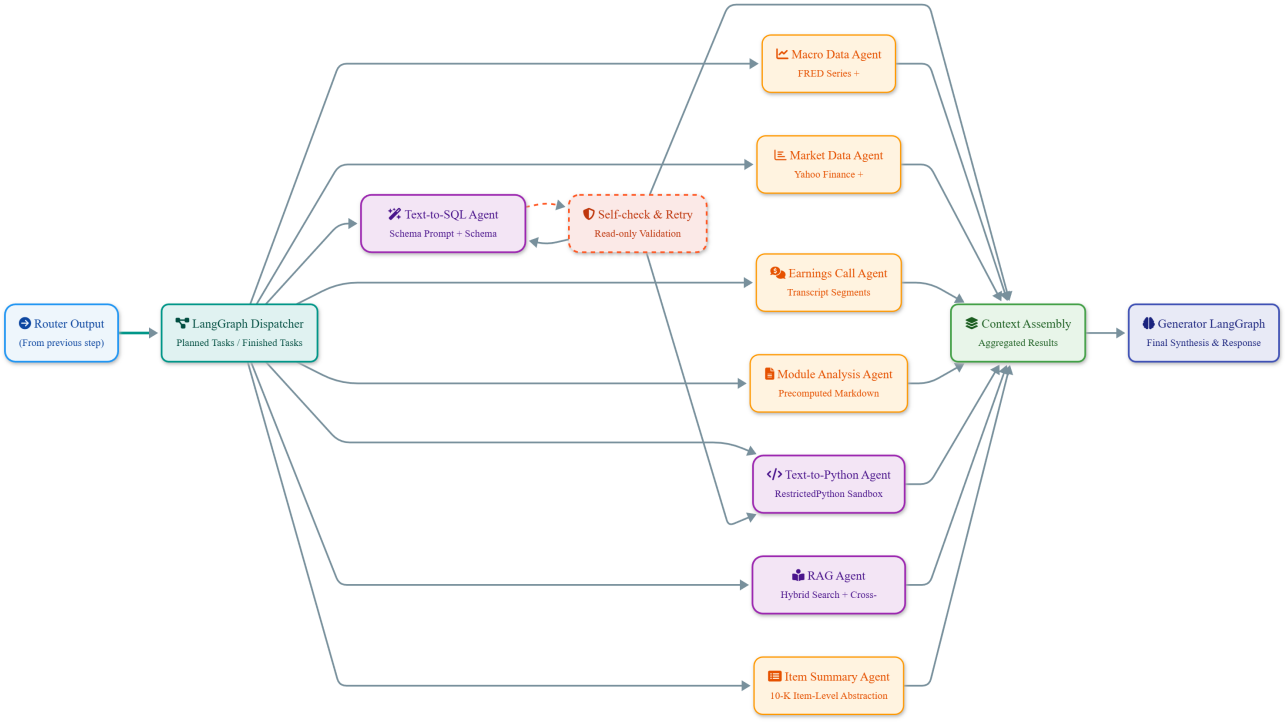
The response generator functions as a dedicated LangGraph workflow responsible for transforming the assembled evidence into a coherent, citation-ready answer. It operates through a sequence of nodes that analyze the retrieved context, select an appropriate response strategy (e.g., descriptive, analytical, comparative, quantitative, or technical), construct the final prompt, and stream the model's output back to the user. At this stage, the generator has access to the full set of inputs produced by the upstream agents—SQL-derived fundamentals, Python-based computations, retrieved narrative disclosures, module-specific summaries, and macroeconomic or market indicators—and applies a set of house rules that govern how these inputs must be presented.

Every factual statement must be explicitly cited with its source (e.g., "Compustat Database, datadate 2023-09-30," or "SEC 10-K Filings, Item 1A, 2023"), numerical results are formatted consistently, and the final answer is structured in Markdown so sections, tables, and lists remain machine- and human-readable. Responses are streamed incrementally to the interface, ensuring transparency and allowing users to see the reasoning unfold in real time.

To preserve continuity across turns, a lightweight summarizer then distills each exchange into a compact record capturing the user's intent, the entities involved, and the key outputs previously retrieved. This running summary enables future routing and decomposition steps to proceed efficiently without resending the full transcript, and provides a transparent, auditable trail of how each response was formed. The generator therefore serves as the system's reporting layer—integrating heterogeneous evidence, enforcing citation discipline, and producing answers that are both analytically coherent and fully grounded in verifiable data.

For the experimental analysis, this guarantees that when we code the content of users' memos, we can recover which portions of the output reflect direct reuse of the agent's structured modules and which represent genuinely new synthesis or interpretation, a distinction that lies at the heart of our soft-information measures.

Figure 4: Retrieval and Execution Engines



Notes: This figure shows how planned tasks are dispatched and executed. The LangGraph Dispatcher orchestrates parallel calls to specialized agents, including Text-to-SQL, Text-to-Python, RAG, Item Summary, Macro Data, Market Data, Earnings Call Retrieval, and Offline Module Access. SQL outputs pass through a self-check and retry mechanism for validation, while other agents return structured or textual payloads. All results converge within the Context Assembly node, where they are consolidated into a unified evidence bundle. The final bundle is passed to the Generator Workflow, which synthesizes a grounded, citation-complete response.

Table 3: Agent Tool Matrix

Task Type	Tool Invoked	Input	Output
Numerical Retrieval	Text-to-SQL	Query and schema	DataFrame
Custom Calculation	Text-to-Python	DataFrame	Computed value
Narrative Retrieval	RAG	Embeddings and filters	Snippets
Peer Expansion	Peer Mapping	ticker/CIK	Peer list
Macro Data	Macro Agent	Query	Latest GDP/CPI/etc
Market Data	Market Agent	Query	Prices, multiples
Earnings Call	Earnings Call Agent	Query	Earnings call transcript

Notes: This table shows the major tool types invoked by the agentic system and the corresponding tasks they support. Each tool operates on an input-output structured outputs used by downstream components. Numerical data retrieval is handled through Text-to-SQL, custom calculations through sandboxed Text-to-Python, and narrative evidence through RAG on the vector database. Additional specialized agents provide macroeconomic indicators, market data, peer sets, and earnings-call excerpts. Together, these tools form a modular and auditable execution layer, allowing the router to compose complex workflows from discrete, verifiable operations.

B.4 Module-Specific Financial Analyzers

Each financial analyzer below corresponds to a familiar dimension of financial statement analysis and valuation, so that the 'hard information' we randomize across experimental arms maps directly into concepts taught in accounting and corporate finance courses.

B.4.1 Business Description Analyzer

This analyzer interprets the firm's core business activities by processing Item 1 and relevant supporting sections of the latest Form 10-K. It distills the company's profile, segment structure, and geographic footprint into a concise markdown report generated through a constrained summarization pass. The resulting document provides an authoritative, reusable overview of the firm's operating model that downstream modules can cite without re-parsing the filing.

B.4.2 Beneath the Lines Value Analyzer

This module focuses on disclosures that influence intrinsic value but rarely appear in headline metrics. It mines MD&A and footnote sections for recurring versus non-recurring items, one-off adjustments, and latent economic drivers such as pension remeasurements or litigation reserves. In the spirit of recent work documenting that non-core earnings components are frequent, dispersed across the income statement, MD&A, and footnotes, and highly relevant for valuation (e.g., Rouen et al., 2021), the analyzer consolidates these adjustments into a

Table 4: System Specification

Component	Technology	Purpose
LLM (Reasoning Layer)	OpenAI GPT-4.1	Natural-language reasoning, synthesis, chain-of-thought
LLM (Memory Layer)	OpenAI GPT-5	Conversation memory condensation
Embedding Model	OpenAI text-embedding-3-large	Dense semantic retrieval
Vector Database	Weaviate	Hybrid retrieval, metadata filters, named-vector collections
Reranker	Cohere Rerank	Cross-encoder joint scoring
Relational Database	MySQL	Structured fundamentals, ratios, metadata
Workflow Framework	LangGraph	Agent routing, state machine
Backend Server	Flask	API endpoints, streaming server
Python Execution Sandbox	RestrictedPython	Safe analytical code
Deployment	Docker Compose	Environment consistency

Notes: This table lists the core components of the system and the technologies used to implement each layer of the architecture. The platform combines a reasoning LLM for natural-language understanding, a memory LLM for turn-level summarization, modern embedding models for semantic retrieval, and a vector database (Weaviate) with hybrid search and metadata filtering capabilities. In practice, any GPT-4 class model (or later) with comparable reasoning performance can be used for the reasoning layer, and any lightweight LLM for summarization can substitute for the memory layer; the architecture is model-agnostic. Structured data are stored in a MySQL warehouse with versioned schemas, while LangGraph provides deterministic workflow orchestration across routing, retrieval, execution, and generation nodes. Analytical code runs inside a restricted Python sandbox, and all services are deployed in a containerized environment to ensure reproducible behavior across sessions and audits. Together, these components form a fully traceable, modular, and governable AI research system suitable for accounting and financial analysis.

Table 6: Safety and Governance Controls

Component	Risk Controlled	Mechanism
Text-to-SQL	Harmful queries	Read-only SQL and retry-mechanism
Python Execution	Arbitrary code	Sandbox (restricted operations)
RAG	Hallucinated facts	Citation-required context
Generator	Missing citations	Enforcement rules
Router	Misrouted tasks	JSON-schema planning
Session Manager	Multiple sessions	Single-session guard
Data Access	Leakage	No model training on outputs

Notes: This table summarizes the system's governance mechanisms and the specific risks each component is designed to control. Text-to-SQL is limited to read-only queries with a validation-and-retry layer to prevent harmful or malformed statements. Python execution occurs inside a restricted sandbox that blocks external imports, file operations, and system calls. Retrieval components (RAG) require verifiable citations, while the generator enforces citation completeness in the final output. The router relies on JSON-schema planning to prevent misrouted tasks, and session management enforces a single active session per user. Data-access policies prohibit model fine-tuning on proprietary outputs, minimizing leakage risk. Collectively, these controls ensure that every step of the agent workflow is auditable, deterministic, and compliant with research-grade data-governance expectations.

structured narrative that downstream valuation and earnings-quality analyses can reference directly.

B.4.3 DuPont Analyzer

Using multiple years of Compustat fundamentals, the DuPont analyzer computes a full return-on-equity decomposition—profit margin, asset turnover, and financial leverage—and produces both visualizations and narrative commentary. Charts and summaries are exported as stand-alone artifacts, enabling instant reference whenever users request multi-year profitability diagnostics.

B.4.4 Earnings Quality Analyzer

This module evaluates the reliability of reported earnings by integrating information from financial statements, cash-flow reconciliations, and the Beneath-the-Lines output. It highlights accrual intensity, cash conversion patterns, working-capital dynamics, and other quality indicators. The output includes markdown commentary and supporting exhibits, each tagged with metadata to maintain a clear audit trail.

B.4.5 Financial Visualization Analyzer

The visualization tool generates high-level financial charts—revenues, margins, leverage, liquidity, and related time-series indicators—derived from Compustat data. It pairs these graphics with a succinct markdown narrative summarizing inflection points and long-term trends, providing an out-of-the-box dashboard for each firm-year.

B.4.6 MD&A Narrative Analyzer

By processing Item 7 (Management's Discussion & Analysis), this analyzer summarizes management's reported performance drivers, forward-looking statements, and strategic commentary. It uses a structured summarization approach to produce a markdown report that can be passed directly into RAG when users inquire about management guidance or stated priorities. In line with evidence that forward-looking MD&A disclosures contain value-relevant information about firms' future performance and the quality of their information environment (e.g., Muslu et al., 2015), the module is designed to surface those narrative elements in a form that can be consistently reused across queries and time.

B.4.7 Multiples Analyzer

This module combines WRDS firm-ratio data with fundamentals to compute valuation multiples and benchmark them against the firm's LLM-derived peer set. It generates visual comparisons—such as EV/EBITDA relative to peers—and accompanying textual interpretation explaining premiums and discounts in economically intuitive terms.

B.4.8 Peer Benchmark Analyzer

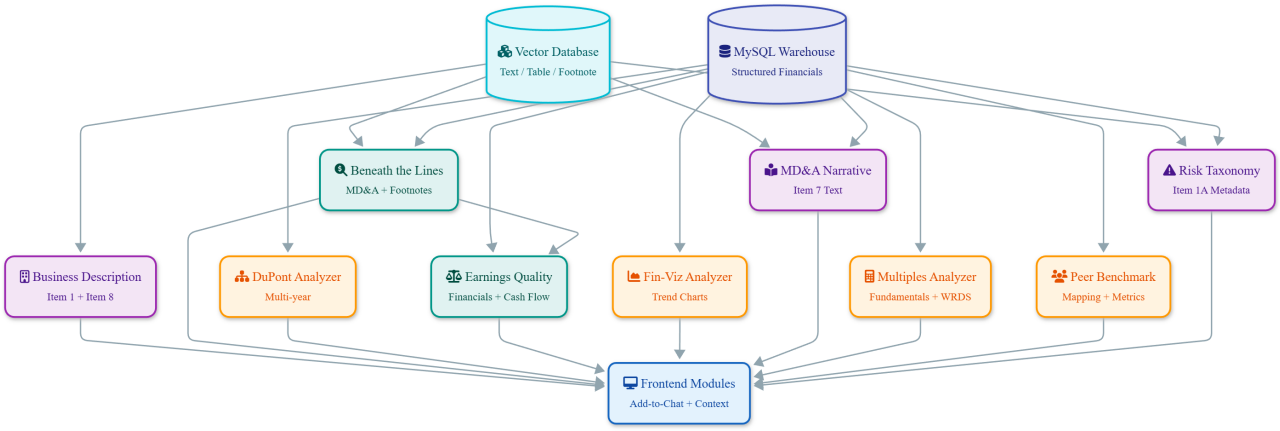
Using the system's peer-mapping table, this analyzer compares key operating and balance-sheet metrics across the firm's closest competitors. It produces a structured markdown report summarizing group medians, ranges, and relative positioning, offering a ready-made reference for comparative performance analysis.

B.4.9 Risk Taxonomy Analyzer

This module ingests Item 1A risk disclosures and clusters individual risks into thematic categories such as regulatory, macroeconomic, operational, or cybersecurity. It tracks changes

over time and highlights new or declining risk areas, in the spirit of recent work that measures additions and removals of Item 1A risk factors across years to study the information content of disclosure changes (e.g., Lyle et al., 2023). The resulting markdown summary provides a consistent taxonomy for answering questions about a firm's key risk exposures and how those exposures evolve across reporting periods.

Figure 5: Offline Financial Analyzer Ecosystem



Notes: This figure shows the system's offline analytical pipeline. SEC filings and WRDS structured data feed into two primary repositories: a Vector Database (for text, tables, and footnotes) and a MySQL Warehouse (for structured fundamentals, ratios, and metadata). These repositories support a suite of offline analyzers—Business Description, Beneath-the-Lines Value, DuPont, Earnings Quality, Financial Visualization, MD&A Narrative, Multiples, Peer Benchmark, and Risk Taxonomy. Each analyzer produces precomputed markdown reports and visualizations that become selectable Frontend Modules. When users activate "Add-to-Chat," these modules supply context to the agent, grounding subsequent retrieval, planning, and synthesis.

B.5 Context Awareness and State Management

B.5.1 Dual-Layer Memory Architecture

Explicit Conversational Memory. The system maintains conversation history on two parallel layers. At the interface level, each turn is preserved as a structured record of user input and model output, allowing the visible transcript to be reconstructed even after a page reload. Simultaneously, the backend produces a compact summary after every turn—capturing the user's query, the retrieved context, and the key elements of the response. This summary acts as the agent's "working memory," enabling future routing decisions without resending the entire transcript and ensuring that follow-up questions reflect the user's accumulated intent.

Table 8: Module Catalog

Module	Input	Output
Business Description	Item 1 of Form 10-K	Summary of business
Beneath-the-Lines Value	Item 7 and Item 8 Footnotes of Form 10-K	Adjustments, recurring vs. one-off items
DuPont	Compustat fundamentals	ROE decomposition and charts
Earnings Quality	Statements, Cash Flows, Beneath-the-Lines	Accruals, WC movements, quality flags
Financial Visualization	Compustat fundamentals	Charts (revenue, EBIT, leverage) and narrative summary
MD&A Narrative	Item 7 of Form 10-K	Summaries of performance drivers and forward guidance
Multiples	WRDS ratios + Compustat fundamentals	Peer-relative valuation charts and narratives
Peer Benchmark	Peer-mapping table and Compustat fundamentals	Ranking table and group medians
Risk Taxonomy	Item 1A of Form 10-K	Risk clusters and YOY changes

Notes: This table summarizes the suite of offline analyzers that generate reusable financial reports for each firm-year. Each module processes a specific portion of the 10-K or its associated structured data—such as Item 1 (business description), Item 7 (MD&A), Item 1A (risk factors), or multi-year Compustat fundamentals—and outputs a standardized markdown file accompanied, when applicable, by charts or supplementary exhibits. These modules perform segment analysis, profitability decomposition, valuation benchmarking, risk clustering, and earnings-quality diagnostics before the user interacts with the system. Because they run offline and store their outputs with consistent metadata, downstream agents can reference their results directly during routing and retrieval, enabling fast, grounded responses without re-parsing filings or recomputing statistics.

Implicit Analysis Context. Beyond free-form conversation, the system also tracks the analytical module the user is viewing at any moment. When a user opens a module such as Business Description — MSFT 2024, the underlying markdown report, metadata, and load timestamp are packaged into an analysis-context object. This richer context allows the agent to understand not only what the user asked, but also what information they are currently looking at.

Each snapshot includes the company identifier (name, ticker, CIK), fiscal year, module type, module display name, and the full markdown content. This snapshot is attached to every subsequent chat request so that the router and decomposer know exactly which report the user is referencing. As a result, follow-up questions can be interpreted relative to the user's on-screen analysis rather than the global document corpus.

The interface offers an "Add to Chat" toggle that explicitly injects the current module snapshot into the query context. When enabled, the snapshot appears under an "Analysis Report Context" header in the router's prompt. This ensures that questions such as "Can you explain this chart?" are resolved directly against the visible report rather than triggering a broad retrieval, grounding the system's reasoning in the user's actual workflow.

B.5.2 Session Persistence

Local Storage. To preserve conversational continuity across page reloads, the client stores recent interaction state, including conversation identifiers and turn counters, in a lightweight local cache. If the browser refreshes mid-session, the transcript can be restored instantly. Analysis context is cleared automatically when the user switches to a different company or fiscal year, preventing stale context from bleeding into new requests.

Backend Logging. On the backend, the system maintains two complementary logs. (i) Conversation logs capture each turn in detail: the user query, enriched context, routing plan, generated response, and whether any analysis context was active at the time. These logs form a complete audit trail of how each answer was produced. (ii) Session logs record user-level interaction events such as page navigation, module loading, and context toggling. Periodically transmitted to the server, they provide visibility into session boundaries, allow enforcement of compliance requirements (such as single active sessions), and support post-hoc audits of how reports were accessed and used.

For our research design, this dual-layer memory allows us to distinguish between questions that arise from what the student is currently looking at (e.g., a DuPont chart or risk summary) and questions that reflect broader exploration, which is crucial for interpreting prompt-level behavior.

B.6 Frontend-Backend Communication and Interface

B.6.1 Asynchronous Streaming Response

When a user submits a question, the interface opens a live communication channel through which the backend streams the model's response as it is generated. Instead of waiting for the full answer to be produced, the user sees it unfold gradually, creating an interactive, near–real-time experience. Alongside the streamed text, the backend also sends a small structured summary of the turn—capturing the query, retrieved evidence, and key outputs—so that future routing steps can rely on compact, machine-readable context rather than the full transcript. This dual stream (visible text plus structured summary) allows the system to remain both responsive and audit-friendly.

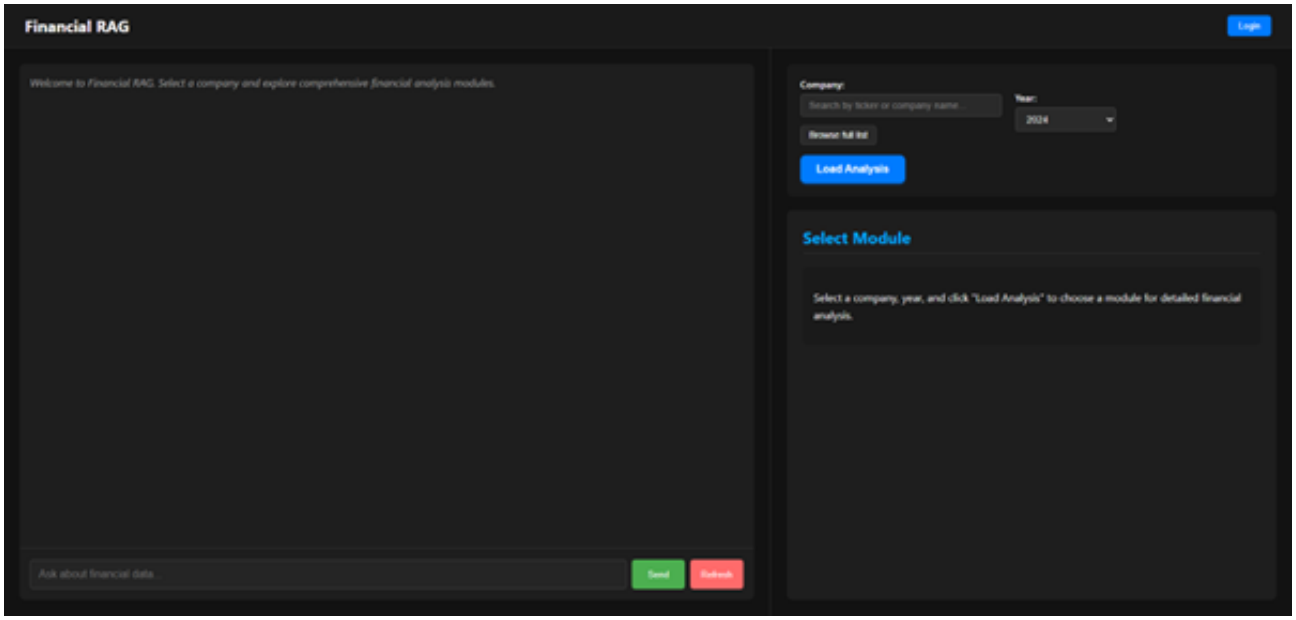
B.6.2 Containerized Online Service

The backend, comprising the orchestration engine, retrieval components, and supporting databases, runs inside a set of coordinated containers to guarantee a stable, reproducible execution environment. Containerization ensures that networking rules, configuration files, and authentication details remain consistent across deployments. It also enables analysts or reviewers to recreate the exact system state when rerunning analyses or auditing model outputs, reinforcing reproducibility and compliance with data-governance standards.

B.6.3 Illustrative Execution Trace

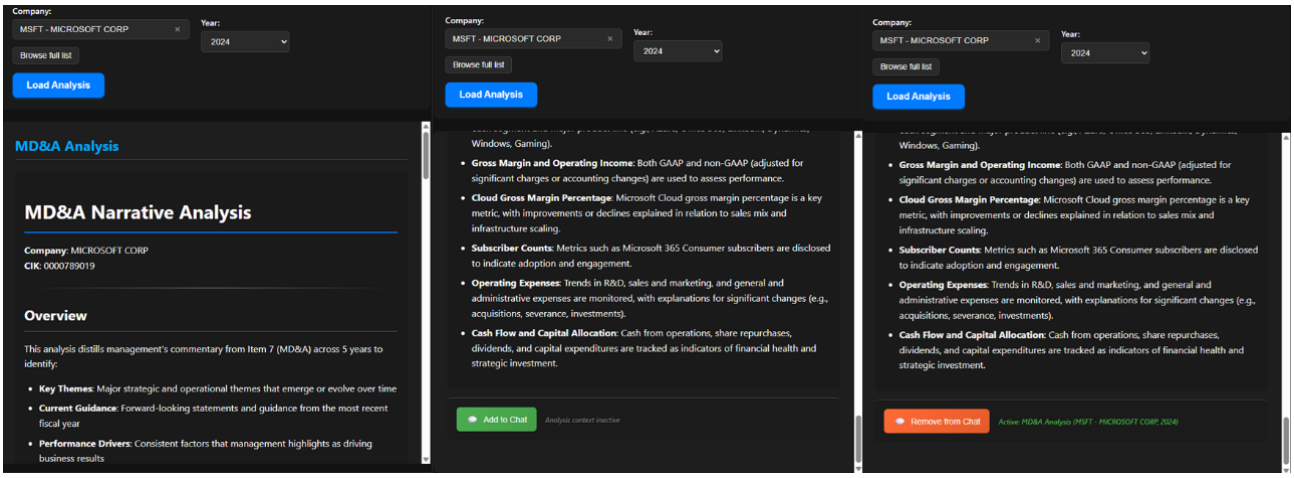
To illustrate how the agent handles a genuinely multi-step valuation task, consider the following query from our system validation logs: "Perform a dividend discount valuation (DDM) for Teradyne, Inc., using both the Gordon Growth and multi-stage DDM approaches." The router classifies this as a complex valuation request and produces a five-step plan involving (i) a Text-to-SQL call to retrieve Teradyne's recent dividends and earnings from Compustat, (ii) a RAG query over Item 7 and Item 8 to recover narrative guidance on payout policy, (iii)

Figure 6: System Homepage and User Interface Overview



Notes: This figure shows the initial interface of the Financial RAG platform. The left panel hosts the conversational workspace where users pose natural-language questions and receive streamed, citation-grounded responses. The right panel contains the structured control hub: analysts can search for a firm by name or ticker, select a fiscal year, and load a precomputed analysis module. Once loaded, modules such as Business Description, MD&A Narrative, DuPont, or Risk Taxonomy become available for inspection and can be injected directly into the chat as contextual guidance. The combined layout reflects the system's dual design philosophy—free-form conversational analysis supported by deterministic, module-based financial workflows, allowing users to transition seamlessly between open-ended queries and structured valuation tools.

Figure 7: Module Viewer and Context Injection Interface



Notes: This figure shows the workflow for inspecting and activating a precomputed financial analysis module. After selecting a company and fiscal year, the system loads the corresponding MD&A Narrative report, which is displayed in a structured markdown format summarizing key themes, guidance, and performance drivers extracted from Item 7 of the firm's 10-K filing. The bottom control panel includes the "Add to Chat" toggle: when activated, the displayed module becomes the active analysis context, enabling the agent to ground subsequent queries in the specific report the user is viewing. The toggle state is shown transitioning from inactive ("Add to Chat") to active ("Remove from Chat"), with metadata indicating the currently injected module (e.g., MD&A Analysis — MSFT, 2024). This interface allows users to fluidly combine free-form conversational analysis with deterministic, module-based financial workflows.

a macro-data lookup to obtain the current risk-free rate from FRED, (iv) a market-data call for the latest share price and shares outstanding, and (v) a Text-to-Python computation that implements single-stage and multi-stage DDM formulae.

Each tool call is logged with timestamps, inputs, and outputs, so the final answer—a 5,800-token narrative explaining assumptions, intermediate cash-flow forecasts, and implied values, which can be audited step by step from the original natural-language query to the numerical DDM estimates. A simplified machine-readable trace of this interaction is shown below; in the actual system, every block is timestamped and stored, allowing auditors to reconstruct the full chain of evidence from the original natural-language query to the final valuation output.

```
{
  "query": "Perform a dividend discount valuation (DDM) for Teradyne, Inc., using both the Gordon Growth and multi-stage DDM approaches.",
  "router_plan": ["text2sql", "rag", "macro_data", "market_data", "python"],
  "text2sql": {
    "sql": "SELECT fyear, dvpsx_f, epspx FROM fundamentals_annual WHERE cik = <TER_CIK> AND fyear >= 2019;",
```

```
      "rows_returned": 6
    },
    "rag": {
   "sections": ["Item 7 MD&A", "Item 8 Footnotes"],
   "key_snippet_topic": "Dividend policy; Share repurchase program authorization",
      "hits": 8
    },
    "macro_data": {
      "series": ["FRED:DGS10"],
      "risk_free_rate": 0.042
    },
    "market_data": {
      "ticker": "TER",
      "price": 123.45,
      "shares_out": 1.57e8
    },
    "python": {
      "task": "compute Gordon-growth and multi-stage DDM values",
      "formula_inputs": {"g": 0.05, "r": 0.095, "d0": 0.44},
      "result_values": {"ddm_single_stage": 108.50, "ddm_multi_stage": 112.30},
      "status": "success"
    },
    "latency_sec": 70.9,
    "answer_tokens": 5888
   }
```


B.7 System Validation and Quality Checks

To ensure that the AI valuation agent behaves as intended and that our experimental results are not driven by implementation artifacts, we conducted a structured validation exercise focused on numerical accuracy, citation fidelity, hallucination risk, and failure handling. We drew a stratified random sample of 100 user queries from pilot interactions with the system, covering 8 topical categories (business description, MD&A performance discussion, risk factors, DuPont and ratio diagnostics, earnings-quality modules, peer/multiples analysis, macro/market data, and valuation what-if questions) and 42 distinct firm–year observations. The sampling

frame included both simple fact-retrieval questions and more complex analytical prompts. All coauthors and four trained research assistants participated in the validation, with each query independently reviewed by at least two coders and disagreements adjudicated by a third coder when necessary.

For each sampled query, coders verified (i) *numerical accuracy*, by tracing every reported figure back to the underlying Compustat or SEC source (10-K/10-Q tables and notes) and checking equality within standard rounding tolerances; (ii) *citation accuracy*, by confirming that each citation in the agent's answer pointed to the correct filing, item, and table; and (iii) *toolpath integrity*, by inspecting the logged SQL and RAG calls to ensure that all facts in the answer were generated via the structured data warehouse or vector database rather than free-form language-model hallucination. Across the 100 queries, we audited 347 distinct numerical outputs and found no instances in which the agent reported a number that differed from the underlying data by more than a rounding unit (e.g., 0.01 in ratios or $1 million in scaled financials). For an additional 52 numerical expressions that involved simple arithmetic transformations (e.g., year-over-year growth rates, margins computed from retrieved components), all computations were consistent with spreadsheet recomputations. This yields an observed numerical error rate of 0% in this sample.

Citation checks likewise showed high fidelity. The 100 answers contained 387 explicit citations to filings, tables, and modules. Two citations (0.5%) were judged imperfect: in both cases, the agent cited the correct filing and item but pointed to a neighboring table rather than the exact table from which the number was drawn. These edge cases were traced to ambiguity in the underlying HTML structure and were corrected in the production system by tightening the chunking and metadata alignment routines. We found no instances in which the agent cited a non-existent filing, non-existent item, or table that was not present in the source documents.

To evaluate hallucination risk and failure handling, coders flagged any instances in which the agent referred to data or disclosures that did not appear in the tool logs. In the 100-query sample, we did not observe any cases in which the agent fabricated a financial ratio, line item, or filing outside of the SQL/RAG pipeline. We did, however, observe 7 queries (7% of the sample) in which the requested information was legitimately unavailable—for example, because a firm lacked a recent earnings-call transcript in the pilot ingestion, or because a requested macro series was not included in the curated FRED subset. In all such cases, the interface returned an explicit "not available" message or stated that the information could not be retrieved, rather

than silently fabricating an answer. These failure modes are therefore visible to users and can be distinguished from true zero values or negligible items.

Taken together, these checks indicate that the agent reliably retrieves and computes numerical information from the intended SEC/WRDS/market sources, maintains high citation fidelity, and handles missing data in a transparent manner. While no system can be guaranteed error-free, the absence of material numerical or citation errors in our validation sample, combined with explicit "not available" handling for missing inputs, supports our use of the agent as a research-grade tool for structuring the information environment in our field experiment.

Table 10: Glossary of System Components

Term	**Definition**
Router	The planning agent that interprets a user query, assembles context, and produces a machine-readable task plan.
Task Plan	A validated JSON structure specifying which tools to invoke (e.g., RAG, SQL, Python), in what order.
Task Decomposer	A reasoning agent that breaks multi-step queries into structured sub-queries and extracts implicit parameters.
Sub-Query	A smaller analytical instruction derived from a complex query (e.g., "retrieve MSFT revenue 2020–2023").
Context Snapshot	A structured record describing the active analysis module (company, year, module, markdown).
Analysis Context	User-selected module content ("Add to Chat") that guides routing and retrieval decisions.
Summary Node	The component that produces a compact JSON summary after each turn for future context-aware routing.
Dispatcher	The LangGraph node responsible for launching planned tasks in parallel.
Hybrid Search	A retrieval method combining dense embedding similarity with sparse lexical matching.
Collection	Separate Weaviate data stores representing narrative text, financial tables, and extracted footnotes.

Glossary of System Components (continued)

Term	Definition
Reranker	A cross-encoder model that re-scores retrieved passages to prioritize the most relevant evidence.
Retrieval Node	The agent component that issues hybrid searches and returns ranked disclosure snippets.
Filing Artifact	A JSONL-structured representation of a 10-K filing containing items, tables, and footnotes.
Metadata Alignment	The process of linking filing artifacts to structured data via CIK, accession number, and fiscal period.
Peer Cluster	A JSON-encoded list of comparable firms for a focal company, generated offline by GPT.
Vector Database	Weaviate v4 storage for embeddings and metadata used for hybrid retrieval.
Structured Warehouse	The MySQL repository holding Compustat fundamentals, ratios, and filing metadata.
Text-to-SQL	An LLM-based tool that converts natural-language numerical queries into validated SQL statements.
Text-to-Python	An LLM-based tool that generates restricted Python code for custom calculations.
Sandbox	A secure execution environment (RestrictedPython) that limits Python to safe operations.
RAG	Retrieval-Augmented Generation, a method where retrieved evidence is injected into the prompt before generation.
Context Assembly	The consolidation stage where results from SQL, Python, RAG, and other tools are merged into a unified evidence bundle.
Generator	The final-stage LLM that synthesizes a grounded, citation-complete answer from the assembled context.
Module Analyzer	Offline preprocessing tools (e.g., DuPont, Risk Taxonomy) that generate markdown reports per company-year.

Glossary of System Components (continued)

Term	Definition
Frontend Module	Precomputed financial analysis reports displayed to the user and optionally added to chat as context.
Add-to-Chat Toggle	UI mechanism allowing module markdown to be injected into backend routing for context-aware queries.
Turn Summary	A compact representation of the previous interaction used to maintain state without full transcripts.
Execution Engine	The LangGraph state machine orchestrating routing, task dispatch, validation, and generation.
Evidence Bundle	The combined output of all invoked tools, tables, text snippets, numerical frames, used for final synthesis.
Guardrails	Safety constraints ensuring all SQL is read-only, Python is sandboxed, and outputs are citation-verified.

Notes: This table shows key terms used throughout the system architecture. It provides standardized meanings for core components such as routing agents, retrieval mechanisms, execution modules, and data structures. By consolidating terminology across structured data, vector search, planning logic, and generative workflows, the glossary ensures that readers can interpret the agent's end-to-end pipeline consistently and without ambiguity.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Gao, L., Dai, Z., and Callan, J. (2021). Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S. H., Wu, L., Edunov, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented

generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lyle, M. R., Riedl, E. J., and Siano, F. (2023). Changes in risk factor disclosures and the variance risk premium. *The Accounting Review*, 98(6):327–352.

Muslu, V., Radhakrishnan, S., Subramanyam, K., and Lim, D. (2015). Forward-looking md&a disclosures and the information environment. *Management Science*, 61(5):931–948.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rouen, E., So, E. C., and Wang, C. C. (2021). Core earnings: New data and evidence. *Journal of Financial Economics*, 142(3):1068–1091.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., et al. (2023). Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551.

Scholak, T., Schucher, N., and Bahdanau, D. (2021). Picard: Parsing incrementally for constrained auto-regressive decoding from language models. *arXiv preprint arXiv:2109.05093*.

Sun, D., Guo, W., Liu, X., Zhang, Y., Hou, Z., and Li, Z. (2025). Zero-shot document retrieval with hybrid pseudo-document retriever. In *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing*.

Wang, J. and Duan, Z. (2024). Agent ai with langgraph: A modular framework for enhancing machine translation using large language models. *arXiv preprint arXiv:2412.03801*.

Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., and Ji, H. (2024). Executable code actions elicit better llm agents. In *Proceedings of the 41st International Conference on Machine Learning*.

Wu, Q., Liu, W., Luan, J., and Wang, B. (2024). Toolplanner: A tool augmented llm for multi granularity instructions with path planning and feedback. *arXiv preprint arXiv:2409.14826*.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Online Appendix C: Financial Statement Analysis Knowledge Quiz

This appendix presents the financial statement analysis (FSA) quiz used to measure participants' ex ante analytical knowledge. The quiz is administered before any interaction with the AI valuation agent. It consists of 40 multiple-choice questions and is designed to take approximately 40 minutes. Each correct answer is awarded one point; the FSA score is the sum of correct answers.

Instructions to Students

Please answer all questions without using notes, textbooks, or electronic resources. Select the best answer for each question. You will have 40 minutes to complete the quiz.

1. A firm's net profit margin rises from 6% to 8%, its asset turnover falls from 1.5 to 1.0, and its equity multiplier remains constant at 2.0. What happens to ROE?
   (a) It rises from 16% to 18%.
   (b) It falls from 18% to 16%.
   (c) It rises from 12% to 16%.
   (d) It remains unchanged at 18%.

2. A firm refinances a large portion of short-term debt with long-term debt one week before fiscal year-end. Which inference is most appropriate?
   (a) The quick ratio must decline because long-term debt increased.
   (b) Working capital necessarily declines.
   (c) The current ratio may improve even if underlying operating liquidity is unchanged.
   (d) The refinancing directly increases cash flow from operations.

3. A firm pays cash for software maintenance costs that it classifies as a capitalized software asset rather than expensing them immediately. Relative to proper expensing, which current-period pattern is most likely?
   (a) Higher net income, higher CFO, and higher total assets.
   (b) Higher net income, lower CFO, and lower total assets.
   (c) Lower net income, higher CFO, and higher total assets.
   (d) No effect on CFO because the classification affects only the income statement.

4. A company reports revenue growth of 20%, accounts receivable growth of 45%, a decline in deferred revenue, and CFO below net income. Which interpretation is most plausible?
   (a) Collections are improving and revenue recognition appears more conservative.
   (b) Customer prepayments are strengthening and cash conversion is improving.
   (c) Lower CFO is caused primarily by higher depreciation.
   (d) There may be a revenue-recognition or collection-quality concern.

5. A firm reports a beginning LIFO reserve of $40 million and an ending LIFO reserve of $70 million in a period of rising input prices. Relative to FIFO, current-year LIFO COGS is:

   (a) lower by $30 million.
   (b) lower by $70 million.
   (c) higher by $30 million.
   (d) higher by $70 million.

6. Inventory turnover improves from 4.0x to 6.0x, while gross margin declines from 35% to 29%. Which explanation is most consistent with this pattern?

   (a) The firm likely used price discounting or unfavorable mix changes to move product faster.
   (b) Purchasing efficiency improved materially while prices held constant.
   (c) Product scarcity allowed the firm to raise prices.
   (d) Working-capital needs necessarily increased because inventory turned faster.

7. Company X reports a current ratio of 2.4, a quick ratio of 0.6, and rising days inventory outstanding. Which inference is best?

   (a) The firm has strong near-cash liquidity.
   (b) Reported liquidity depends heavily on inventory that may be slow moving.
   (c) The company necessarily has negative working capital.
   (d) Accounts payable must be shrinking.

8. Near year-end, a firm factors a large block of receivables without recourse. All else equal, which pattern is most likely in that quarter?

   (a) Higher receivables, lower CFO, and no income effect.
   (b) Lower receivables, lower CFO, and higher net income.
   (c) Higher receivables, higher CFO, and lower net income.
   (d) Lower receivables, higher CFO, and little or no net-income effect.

9. Which situation is most consistent with negative working capital reflecting business strength rather than financial distress?

   (a) Chronic operating losses and covenant pressure.
   (b) An inability to refinance current obligations.
   (c) A business model with customer prepayments and strong supplier terms.
   (d) Inventories that are becoming obsolete.

10. A firm reports net income of $120 million, depreciation of $30 million, an increase in accounts receivable of $25 million, a decrease in inventory of $10 million, a decrease in accounts payable of $15 million, and an increase in deferred revenue of $12 million. Under the indirect method, CFO is:

    (a) $132 million.
    (b) $122 million.
    (c) $142 million.
    (d) $150 million.

11. A firm's allowance for doubtful accounts falls from 5% to 3% of gross receivables while DSO rises sharply and quarter-end sales accelerate. Which interpretation is most plausible?

    (a) Reserving is becoming more conservative and collections are improving.
    (b) Receivables are becoming shorter dated.
    (c) The pattern is normal seasonality with no earnings-quality implication.
    (d) The firm may be under-reserving while also recognizing revenue aggressively.

12. A company excludes restructuring charges from adjusted earnings in each of the last six years. Which treatment is most appropriate for core-performance analysis?

    (a) Exclude all such charges because management labels them nonrecurring.
    (b) Treat them as part of recurring economics unless strong evidence shows otherwise.
    (c) Ignore them because they are noncash by construction.
    (d) Capitalize them and amortize them over future periods.

13. A firm's ROE rises, but its ROIC falls and its net debt/EBITDA increases. Which interpretation is most consistent with this pattern?

    (a) Tax rates must have fallen enough to explain the entire change.
    (b) Asset turnover is the only possible driver.
    (c) Leverage is likely masking weaker operating performance.
    (d) Stronger operating economics are clearly driving equity returns.

14. A software firm reports revenue growth of 28%, billings growth of 12%, and a large increase in contract assets. Which concern is most plausible?

    (a) Revenue is being recognized faster than cash collections or billings would suggest.
    (b) The firm is necessarily becoming less profitable.
    (c) Gross margin must be understated.
    (d) The rise in contract assets proves demand quality is improving.

15. A retailer's DSO rises from 8 to 11 days, DIO falls from 82 to 74 days, and DPO falls from 40 to 38 days. The cash conversion cycle:

    (a) worsened by 3 days.
    (b) improved by 3 days.
    (c) was unchanged.
    (d) worsened by 5 days.

16. A firm's EBIT margin is unchanged, its pretax margin declines, its asset turnover improves modestly, and its equity multiplier is stable. Which factor most likely pressured ROE?

    (a) Lower DSO.
    (b) Lower leverage.
    (c) Higher tax burden.
    (d) Higher interest burden.

17. Two firms have the same debt/EBITDA ratio, but Firm A has much weaker EBITDA-to-CFO conversion because receivables and inventory are growing rapidly. Which statement is best?

(a) Firm A likely has lower credit risk because reported EBITDA is the same.
(b) Debt/EBITDA fully captures liquidity risk, so the firms are similarly risky.
(c) Firm A may be riskier because weak cash conversion makes the same EBITDA less protective.
(d) Firm A must have lower interest expense.

18. Two firms are expected to grow at 5% by reinvesting additional capital. Firm A earns a 14% ROIC on new investment and Firm B earns 8%; both have a 9% WACC. Which statement is correct?

(a) Growth creates value for Firm A but destroys value for Firm B.
(b) Growth creates value for both firms because growth is positive.
(c) Growth destroys value for both firms because reinvestment is required.
(d) Growth has no valuation effect because the firms share the same growth rate.

19. Two firms have identical EBITDA, growth, and risk, but Firm A requires much higher maintenance capex and working-capital reinvestment than Firm B. Relative to Firm B, Firm A should generally trade at:

(a) a higher EV/EBITDA multiple.
(b) a lower EV/EBITDA multiple.
(c) the same EV/EBITDA multiple.
(d) a higher price-to-book multiple and a lower EV/EBITDA multiple.

20. In a cyclical industry, a very low trailing P/E is often least informative when:

(a) book value is negative.
(b) debt is low.
(c) depreciation is high.
(d) earnings are near a cyclical peak.

21. When comparing EV/Sales across subscription software firms, which difference most strongly justifies a higher multiple, all else equal?

(a) Higher gross margins and lower customer churn.
(b) Higher depreciation expense.
(c) Lower book value of equity.
(d) Higher accounts payable days.

22. A target has lower margins, lower growth, and higher cyclicality than its peer group. Applying the peer median P/E to the target without adjustment would most likely:

(a) understate the target's value.
(b) be unbiased if capital structure is similar.
(c) overstate the target's value.
(d) be more reliable than EV-based multiples.

23. Holding near-term cash-flow forecasts fixed, which change most likely reduces intrinsic value in a DCF framework?

(a) Lengthening the period over which excess returns persist before fading.

(b) Shortening the period before excess returns fade toward the cost of capital.
(c) Replacing EV/EBITDA with P/E in a peer table.
(d) Increasing depreciation while holding capex unchanged.

24. Suppose EBIT = \$200 million, the tax rate = 25%, depreciation = \$40 million, capital expenditures = \$70 million, and the increase in net working capital = \$30 million. FCFF is:

(a) \$120 million.
(b) \$130 million.
(c) \$140 million.
(d) \$90 million.

25. A firm's expected FCFE next year is \$5 per share and is expected to grow at 4% in perpetuity. If the cost of equity is 9%, the equity value per share under a constant-growth FCFE model is:

(a) \$80.
(b) \$90.
(c) \$100.
(d) \$125.

26. Book value of equity today is \$40 per share. Expected ROE next year is 15%, the cost of equity is 10%, and residual income is expected to persist at the same dollar level indefinitely. Under a simple residual-income model, value per share is:

(a) \$60.
(b) \$48.
(c) \$52.
(d) \$68.

27. Invested capital is \$500 million, ROIC is 12%, and WACC is 9%. EVA is:

(a) \$45 million.
(b) \$15 million.
(c) \$60 million.
(d) \$10 million.

28. For a young firm with negative EBITDA but positive gross profit and rapidly scaling revenue, which multiple is generally the least useful?

(a) EV/Sales.
(b) EV/Gross Profit.
(c) Price-to-book.
(d) EV/EBITDA.

29. Price-to-book is often most informative for firms where:

(a) accounting book value is closely tied to deployable capital and ROE is central to valuation.
(b) most value resides in unrecognized intangible assets.

(c) leverage varies widely and equity is frequently negative.
(d) the balance sheet is light and earnings are highly cyclical.

30. If a firm repurchases shares at a price below its intrinsic value using excess cash, holding operating forecasts fixed, the repurchase will generally:

(a) reduce per-share intrinsic value for remaining shareholders.
(b) leave per-share intrinsic value unchanged by definition.
(c) increase per-share intrinsic value for remaining shareholders.
(d) increase enterprise value but decrease equity value.

31. A firm's ROE is flat year over year, but operating margin declines, asset turnover declines, and the equity multiplier rises. Which conclusion is best?

(a) Underlying operating performance improved.
(b) Stable ROE masks weaker operations offset by higher leverage.
(c) Declining turnover always implies lower leverage.
(d) The firm's cost of equity must have fallen.

32. A firm reports stable gross margin, lower operating margin, stable CFO, and sharply higher capex. Which interpretation is most reasonable?

(a) Free cash flow likely improved because CFO is stable.
(b) Capex is irrelevant because it is an investing cash flow rather than an operating cash flow.
(c) Operating margin can decline only if gross margin declines.
(d) Cost pressure below gross profit weakened operating performance, and free cash flow may still fall because capex is higher.

33. A serial acquirer reports strong adjusted EBITDA growth, but share count, debt, and acquired intangibles all rise sharply. Which valuation concern is most relevant?

(a) EBITDA growth may overstate value creation if growth is acquisition-driven and not per-share accretive or return-enhancing.
(b) Rising share count always lowers enterprise value.
(c) Acquired intangibles increase CFO one-for-one.
(d) Debt-financed acquisitions mechanically raise ROIC.

34. A target trades at 7x EV/EBITDA versus peers at 9x. Which additional fact most weakens the argument that the target is undervalued?

(a) The target has lower reinvestment needs than peers.
(b) The target has a net cash balance.
(c) The target has materially lower expected growth and ROIC, and higher cyclicality, than peers.
(d) Peers use different depreciation schedules.

35. A firm's CFO materially exceeds net income, but most of the gap is explained by large increases in accounts payable and accrued expenses while receivables and inventory also rise. Which inference is best?

(a) This is unambiguous evidence of high earnings quality.
(b) Strong cash flow may partly reflect temporary vendor financing rather than superior core economics.
(c) Rising receivables strengthen revenue quality.
(d) Working-capital expansion can never depress future CFO.

36. A firm reports a 300 bp decline in gross margin but only a 20 bp decline in operating margin because SG&A as a share of sales falls sharply. Which interpretation is most reasonable?

(a) Pricing power clearly improved.
(b) Inventory turnover must have improved.
(c) Net income should rise one-for-one with the gross-margin decline.
(d) Cost discipline below gross profit is offsetting weaker unit economics.

37. For banks and insurers, price-to-book often plays a larger role than EV/EBITDA because:

(a) debt-like liabilities are part of the operating model and book equity is more closely linked to deployable capital and ROE.
(b) EBITDA is always negative for financial firms.
(c) earnings for financial firms are largely irrelevant to valuation.
(d) enterprise value is undefined for regulated firms.

38. When using peer EV/EBITDA multiples, which issue can materially distort comparability if handled inconsistently across firms?

(a) Differences in fiscal year-end alone.
(b) Whether lease obligations and lease-related expenses are treated consistently in enterprise value and EBITDA.
(c) Whether firms report basic or diluted shares outstanding.
(d) Whether cash taxes differ across countries.

39. Which pattern is most consistent with possible cookie-jar reserve management?

(a) Inventory write-downs occur only when commodity prices collapse.
(b) Warranty expense rises when product failures increase.
(c) Large provisions are recorded in strong years and later reversed or released in weaker years to support earnings.
(d) Bad-debt expense tracks deteriorating customer credit quality.

40. A firm's dividend payout ratio rises even as FCFE weakens and the firm increasingly funds dividends with new borrowing. Which conclusion is most appropriate?

(a) The dividend is more sustainable because leverage substitutes for equity cash flow.
(b) FCFE is irrelevant as long as net income is positive.
(c) New borrowing reduces the cost of equity and therefore validates the dividend.
(d) Dividend sustainability is a concern because payouts are not supported by internally generated equity cash flow.

Answer Key (Not Shown to Participants)

Q1: B	Q11: D	Q21: A	Q31: B
Q2: C	Q12: B	Q22: C	Q32: D
Q3: A	Q13: C	Q23: B	Q33: A
Q4: D	Q14: A	Q24: D	Q34: C
Q5: C	Q15: B	Q25: C	Q35: B
Q6: A	Q16: D	Q26: A	Q36: D
Q7: B	Q17: C	Q27: B	Q37: A
Q8: D	Q18: A	Q28: D	Q38: B
Q9: C	Q19: B	Q29: A	Q39: C
Q10: A	Q20: D	Q30: C	Q40: D

Online Appendix D: Human Coding Manual and Coder Protocol

This appendix provides the detailed coding manual used for all human-coded measures. The goal of the protocol is to ensure that coders apply consistent standards across student–firm units while remaining blind to treatment assignment and the study hypotheses.

D.1 General Coding Principles

Blind coding. All conversations and memos must be coded independently by two trained coders who are blind to treatment assignment, outcome results, and the paper's hypotheses. Coders should not attempt to infer treatment arm from interface references or module names. If such references are visible, they should be ignored.

Unit of analysis. The research team supplies a *filtered conversation* for each student–firm unit. Coders should use the transcript *exactly as provided*. The relevant units are:

- **Conversation-level variables:** code once using the full filtered conversation.
- **Query-level variables:** code each student-authored query separately.
- **Memo-level variables:** code once using the final submitted memo and, when relevant, the concatenated AI responses from delegation turns.

What counts as student-authored reasoning. A *student-owned interpretive claim* is any substantive statement authored by the student about cause, mechanism, significance, comparability, valuation relevance, or likely implications. A query can contain a student claim even if phrased as a question. For example, "Could the margin decline reflect discounting rather than mix?" contains a student-authored explanatory frame and should be treated as the student's own reasoning.

Conservative coding rule. When a case is ambiguous, coders should choose the *more conservative* code:

- for binary measures, code 0 unless the positive condition is clearly satisfied;

- for ordinal 1–5 scales, choose the lower score unless the higher score is clearly warranted;
- for nominal classifications, choose the less interpretive category unless a stronger category is clearly supported by the student's language.

Coding sequence. Coders should follow the same sequence for every unit:

1. Read the full filtered conversation once without coding.
2. Code all query-level variables turn by turn.
3. Re-read the conversation and assign conversation-level scores.
4. Code memo dependence only after reviewing the student's final memo and the concatenated AI responses from delegation turns.
5. Record a brief note for any borderline case.

Training and calibration. Before production coding begins, coders should jointly review this manual and complete a pilot round on a small training set not used in the main analyses. After the pilot round, the research team should discuss disagreements, clarify edge cases, and lock the protocol. Production coding should begin only after this calibration round is complete.

D.2 Financial Metric Domains Used in Multi-Metric and Delegation Coding

For the *Multi-Metric Synthesis* and *Delegation* measures, coders must identify whether a query references multiple distinct financial metric domains. The domains are:

- **Revenue:** sales, top line, segment revenue, volume, pricing, mix.
- **EBIT / operating profit:** EBIT, operating income, EBITDA only when discussed as an operating-profit measure rather than as a valuation multiple denominator.
- **Net income / EPS:** net income, earnings, diluted EPS.
- **Free cash flow:** operating cash flow, FCF, cash generation.
- **Debt / leverage:** debt, net debt, leverage, interest coverage, debt maturity burden.

- **Margins / returns:** gross margin, EBITDA margin, operating margin, net margin, ROA, ROE, ROIC.
- **Working capital:** inventory, receivables, payables, cash conversion cycle, working-capital intensity.
- **Valuation multiples:** EV/EBITDA, EV/Sales, P/E, P/B, trading multiples.
- **Risk:** business risk, litigation, regulatory exposure, customer concentration, cyclicality, volatility.
- **Growth:** growth rates, acceleration/deceleration, organic growth, forward growth outlook.

Counting rule. A single phrase should be assigned to its *primary* domain only. Do not double count one expression across multiple domains. For example:

- "revenue growth" counts as **Growth**, not both Revenue and Growth;
- "EBITDA margin" counts as **Margins / returns**;
- "inventory build" counts as **Working capital**.

Named modules. For delegation coding, explicit references to two or more named interface modules or analyzers also qualify for screening into the delegation-coding universe, even if the query does not clearly list three metric domains.

D.3 Human-Coded Components of the Soft Information Index

D.3.1 Follow-Up Depth

Definition. *Follow-Up Depth* is a conversation-level 1–5 score capturing whether the student deepens inquiry across turns rather than issuing isolated or disconnected requests.

What to evaluate. Read the full conversation and assess whether later student queries build on earlier AI answers, narrow a specific puzzle, test a prior interpretation, or push the analysis into a more refined question. The score should reflect the *trajectory* of the conversation, not the number of turns.

Scoring anchors.

- **1 = No meaningful follow-up.** Queries are isolated data pulls or summaries. The student does not build on prior answers.
- **2 = Minimal follow-up.** There is occasional return to an earlier topic, but the conversation is mostly fragmented.
- **3 = Moderate follow-up.** The conversation contains at least one clear chain of follow-up, but the overall structure is uneven or broken into disconnected blocks.
- **4 = Strong follow-up.** The student repeatedly uses prior answers to drill into specific issues or puzzles.
- **5 = Sustained investigative follow-up.** The conversation shows a clear, coherent inquiry path in which the student consistently refines the question based on what the AI just returned.

Decision rules.

1. A long conversation can still score low if its turns are disconnected.
2. Merely moving across topics in a plausible order does not by itself imply deep follow-up.
3. Explicit language such as "based on that," "if so," "then," or "does that imply" is helpful but not required.
4. If only one short follow-up chain appears in an otherwise disconnected conversation, the score should generally not exceed 3.

Illustrative examples.

- *Score 1 example:* "What was revenue growth?" $\rightarrow$ "What was debt/EBITDA?" $\rightarrow$ "Summarize risk factors." These are isolated requests with no clear cumulative inquiry.
- *Score 5 example:* "Margins improved while inventory rose. Is that more consistent with mix shift or weakening demand? If mix shift, which segment disclosures support it? Does management commentary reconcile with that pattern?" This sequence clearly deepens a specific puzzle across turns.

D.3.2 Reasoning Depth

Definition. *Reasoning Depth* is coded at the query level on a 1–5 scale and later averaged within the conversation. It measures how much of the analytical thinking is authored by the student rather than delegated to the AI.

What to evaluate. Judge only the student's text in the focal query. Do not reward copied AI language or broad instructions that merely ask the AI to do the thinking.

Scoring anchors.

- **1 = Pure retrieval or instruction.** The student requests facts, summaries, or analysis with no student-authored reasoning.
- **2 = Directed interpretation request, no student claim.** The student points the AI to a topic or issue, but does not offer a substantive explanatory frame of their own.
- **3 = Partial framing.** The student supplies a tentative angle or partial interpretation, but the AI is still asked to do most of the analytical work.
- **4 = Student hypothesis, AI verification.** The student states a clear hypothesis or comparative claim and asks the AI to test, verify, or challenge it.
- **5 = Rich student-authored analytical claim.** The student integrates multiple facts into a substantive explanatory or valuation-relevant narrative and asks the AI only to extend, refine, or probe it further.

Decision rules.

1. A question can still score 4 or 5 if it embeds a student-authored hypothesis.
2. A list of finance terms without a claim is not reasoning.
3. Queries beginning with "analyze," "compare," or "draft" usually score 1 or 2 unless the student adds a genuine analytical frame.
4. If the student's contribution is merely to restate the AI's prior answer, do not score above 2.

Illustrative examples.

- *Score 1 example:* "Give me revenue, EBITDA, and debt for 2021–2023."
- *Score 5 example:* "Gross margin improved despite slower sales growth and higher inventory. My hypothesis is that temporary product mix, rather than durable pricing power, drove the improvement. Check whether the MD&A and segment data support that interpretation."

D.3.3 Cross-Period Synthesis

Definition. *Cross-Period Synthesis* is a query-level binary flag equal to one if the query:

1. references at least two distinct calendar years (or clearly labeled quarter-year periods), and
2. contains temporal comparison language such as "trend," "year-over-year," "historically," "over time," "improved," "deteriorated," or "compare."

Decision rules.

1. Both conditions must be satisfied.
2. Generic phrases such as "over the last few years" do *not* qualify unless the years are explicitly named.
3. A request for data from two years without comparison language does *not* qualify.
4. Quarter-year labels count if the periods are explicit (e.g., Q1 2022 and Q1 2023).

Illustrative examples.

- *Code 1 example:* "Compare inventory turns and gross margin from 2021 to 2023 and explain the trend."
- *Code 0 example:* "What were inventory turns in 2021 and 2023?" This names two years but does not ask for temporal synthesis.

D.3.4 Multi-Metric Synthesis Classification

Definition. Queries that reference at least three distinct financial metric domains are classified into one of four mutually exclusive categories:

`not_in_universe`, `synthesis`, `delegation`, or `retrieval`.

The conversation-level variable *Multi-Metric Synthesis* is the count of queries classified as `synthesis`.

Eligibility. If the query does not reference at least three distinct metric domains, code `not_in_universe`. The research team may pre-screen eligible queries, but coders should verify eligibility independently.

Classification rubric.

- `synthesis`. The student provides an interpretive claim, explanatory frame, or analytical tension linking at least two metrics, and asks the AI to verify, challenge, discriminate, or extend that reasoning.
- `delegation`. The student lists multiple metrics as ingredients and asks the AI to generate an integrated analytical output, with no substantive interpretive claim of the student's own.
- `retrieval`. The student requests multiple metrics or data points without interpretive assembly and without asking the AI to generate an integrated analytic product.

What counts as an integrated analytical output. Examples include a summary, diagnostic, write-up, ranking, investment thesis, valuation view, comparative report, or other output that combines multiple inputs into a single analytical product.

Decision rules.

1. If a query contains both task language and a student-authored claim, classify according to the *dominant* feature.
2. A substantive student claim should generally push the classification toward `synthesis` rather than `delegation`.

3. Generic prompts such as "what stands out?" or "pull it together" count as `delegation` when they ask the AI to do the synthesis and the student offers no interpretive claim.

4. Pure multi-metric data collection without interpretive assembly is `retrieval`.

Illustrative examples.

- *`synthesis` example:* "Revenue growth slowed even as operating margin improved and inventory rose. Is that more consistent with mix shift or with weakening sell-through?"
- *`delegation` example:* "Using revenue growth, margins, working capital, and leverage, draft a concise diagnostic of the firm's recent performance."
- *`retrieval` example:* "Give me revenue growth, operating margin, free cash flow, and net debt for 2021–2023."

D.3.5 Hypothesis Testing

Definition. *Hypothesis Testing* is a conversation-level 1–5 score capturing whether the student formulates testable explanations and uses the AI to evaluate or discriminate among them.

What counts as a hypothesis. A hypothesis is a specific explanation that could in principle be supported or weakened by additional evidence. High scores require more than asking "why"; they require one or more candidate explanations, and preferably some attempt to compare them.

Scoring anchors.

- **1 = No hypothesis formation.** The student requests data or summaries only.
- **2 = Minimal hypothesis language.** The student occasionally gestures at explanation, but no clear testable hypothesis is developed.
- **3 = Some hypothesis formation.** The student proposes at least one plausible explanatory frame, but testing is limited or sporadic.
- **4 = Active hypothesis testing.** The student advances multiple plausible explanations and asks the AI to evaluate or distinguish among them.
- **5 = Sustained hypothesis-driven inquiry.** The student repeatedly formulates, revises, and tests competing explanations across the conversation.

Decision rules.

1. Asking for a reason or driver, by itself, is not enough for a 4 or 5 unless the student articulates candidate explanations.
2. The highest scores require discrimination among plausible alternatives, not just a single conjecture.
3. A conversation can score 3 with one good hypothesis even if the rest is descriptive.

Illustrative examples.

- *Score 1 example:* "Summarize margin trends, debt, and risk factors."
- *Score 5 example:* "Could the margin improvement reflect mix shift, operating leverage, or reserve release? Check segment trends, SG&A behavior, and working-capital patterns to distinguish among them."

D.3.6 Mechanism-Seeking

Definition. *Mechanism-Seeking* is a query-level binary indicator equal to one if the student asks *why* or *how* an observed pattern arose, or asks the AI to identify the operative channel linking a cause to an outcome.

Code 1 if the query does any of the following:

- asks why something happened;
- asks how one factor affects another;
- asks whether a pattern is driven by one channel versus another;
- asks for the mechanism behind an observed financial change.

Code 0 if the query only:

- requests facts, values, or summaries;
- asks what happened but not why;
- asks for a conclusion without channel-level reasoning.

Illustrative examples.

- *Code 1 example:* "Why did free cash flow weaken even though EBIT improved? Was it inventory, receivables, or capex?"
- *Code 0 example:* "What happened to free cash flow and EBIT in 2023?"

D.4 Human-Coded Delegation Measures

D.4.1 Delegation Turn Indicator

Definition. A query receives `delegation = 1` if and only if *all three* of the following conditions are satisfied:

1. the student references at least two distinct financial metric domains as ingredients, or explicitly invokes at least two named modules;
2. the student asks the AI to produce an integrated analytical output;
3. the student contributes no substantive interpretive claim of their own.

If any one of these conditions fails, code `delegation = 0`.

Interpretive claim test. If the student offers even a tentative causal, comparative, or valuation-relevant interpretation, the query should generally be coded 0. For example, "Margins look temporarily inflated by mix; assess whether that is sustainable" contains a student-authored claim and is *not* delegation.

What counts as an integrated analytical output. Examples include a synthesized diagnostic, a comparative analysis, a recommendation, a ranked list of issues, a write-up, a valuation discussion, or a memo-like product. A simple request for raw numbers does not qualify.

Code 0 in the following cases.

- single-module invocation only;
- pure data retrieval or factual lookup;
- follow-up requests about prior evidence with no analytic handoff;

- fewer than two distinct domains and fewer than two named modules;
- procedural or social turns.

Decision rules.

1. Be conservative. If uncertain, code 0.
2. A request to "summarize" or "pull together" multiple inputs usually counts as delegation if no student interpretation is offered.
3. A query may be both multi-metric and non-delegating if the student provides a substantive analytical frame.
4. Within the eligible multi-metric universe, queries coded `delegation = 1` should ordinarily also receive the `delegation` label in the *Multi-Metric Synthesis* classification. Any discrepancy should be flagged for adjudication.

Illustrative examples.

- *Code 1 example:* "Using revenue growth, margin trends, leverage, and peer multiples, draft a short valuation-oriented performance assessment."
- *Code 0 example:* "Margins look temporarily supported by mix rather than demand strength. Use revenue, inventory, and segment data to test that." This is not delegation because the student supplies the analytical frame.

D.4.2 Delegation Any and Delegation Turns

Definition. These measures are derived mechanically from query-level delegation coding:

- **Delegation Any** equals one if the student has at least one query coded `delegation = 1` across all conversations.
- **Delegation Turns** equals the total number of queries coded `delegation = 1` across all conversations.

Coding note. Coders do not assign these student-level variables directly. They code query-level delegation turn indicators, from which the research team constructs the student-level outcomes.

D.4.3 Delegation Memo Score

Definition. *Delegation Memo Score* is a memo-level 1–5 score capturing the extent to which the student's final memo structurally depends on AI responses produced in the student's delegation turns.

Materials to compare. For each student, coders should compare:

1. the student's final memo; and
2. the concatenated AI responses from that student's delegation turns, in chronological order.

What counts as dependence. Dependence should be inferred from *clear structural import*, including:

- similar ordering of issues or sections;
- similar analytic skeleton or sequence of arguments;
- re-use of distinctive phrasing or framing;
- close tracking of AI-generated organization or emphasis.

Shared topics, generic finance language, or overlap in common facts are *not* sufficient by themselves.

Scoring anchors.

- **1 = No dependence.** The memo does not materially track the structure or framing of the AI outputs.
- **2 = Minor dependence.** The memo shows isolated borrowing of phrases or a small amount of structural borrowing.
- **3 = Moderate dependence.** One or more memo sections reflect the AI's organization or framing, but the memo still contains substantial independent framing.

- **4 = Heavy dependence.** The memo structure closely tracks the AI responses, with limited independent reorganization.
- **5 = Near-substitution.** The memo largely reorganizes or restates the AI responses with minimal independent analytical structure.

Decision rules.

1. Compare *structure and analytical framing*, not just wording.
2. If dependence is ambiguous, choose the lower score.
3. Do not infer high dependence solely because the memo and AI discuss the same broad topics.
4. A memo can score 4 or 5 even with paraphrasing if the organizational structure and argumentative sequence clearly track the AI outputs.

Illustrative examples.

- *Score 2 example:* The AI produced a peer-comparison paragraph and the memo borrows one sentence about a valuation discount, but the memo's overall structure and thesis are independently developed.
- *Score 5 example:* The AI responses are organized as business overview → margin drivers → peer multiples → valuation conclusion, and the student's memo follows the same sequence and content with only light rewriting.

D.5 Adjudication and Final Data Construction

All units are coded independently by two coders. Coders must submit their completed coding sheets before any disagreement discussion occurs. After independent coding is complete, the coauthor team reviews all disagreements and assigns final adjudicated codes.

D.6 Inter-Rater Reliability

All human-coded measures were double-coded by two independent coders. For binary and nominal variables, inter-rater reliability is summarized using Cohen's κ. For ordinal 1–5 measures, we report both quadratic-weighted Cohen's κ and the two-way random-effects intraclass correlation coefficient for absolute agreement, ICC(2,1). Reliability is high across all measures, with all coefficients in the strong range.

Table 11: Inter-rater reliability for human-coded measures

Variable	Unit	Reliability statistic	Agreement
Follow-Up Depth	Conversation	QWK = 0.81; ICC(2,1) = 0.86	72.4% exact; 97.8% within 1 point
Reasoning Depth	Query	QWK = 0.78; ICC(2,1) = 0.83	69.1% exact; 96.5% within 1 point
Cross-Period Synthesis	Query	Cohen's $\kappa = 0.93$	98.1%
Multi-Metric Classification	Query	Cohen's $\kappa = 0.84$	90.2%
Hypothesis Testing	Conversation	QWK = 0.79; ICC(2,1) = 0.84	68.7% exact; 97.1% within 1 point
Mechanism-Seeking	Query	Cohen's $\kappa = 0.87$	94.0%
Delegation Turn Indicator	Query	Cohen's $\kappa = 0.85$	92.8%
Delegation Memo Score	Memo	QWK = 0.80; ICC(2,1) = 0.85	70.3% exact; 98.0% within 1 point

Notes: QWK denotes quadratic-weighted Cohen's κ. ICC(2,1) denotes the two-way random-effects intraclass correlation coefficient for absolute agreement. Exact agreement is reported for all variables; for ordinal measures we also report the proportion of cases in which coders differed by no more than one scale point.

Online Appendix E: Instructions to Participants

Valuation Task. You will be assigned one public company and given 90 minutes to complete a valuation exercise using the AI valuation platform. Your objective is to produce your *own* well-reasoned investment view on the assigned firm. The platform is a decision-support tool: it can help you retrieve, organize, and interpret hard information, but the final judgment must be yours.

How to Use the Platform. You may interact freely with the chat interface and any analysis modules visible in your assigned interface. Use the system to investigate the firm's business model, recent performance, accounting quality, risk profile, and valuation. Strong analysis goes beyond one-shot fact gathering. It connects financial statements, disclosures, and other available information into an economic narrative about the firm's growth, profitability, reinvestment needs, risk, and value.

Required Deliverable. At the end of the exercise, submit a concise professional valuation memo that states and defends an investment thesis. Your memo should include: (i) a clear conclusion about the firm's valuation or investment attractiveness; (ii) the key evidence supporting that conclusion; (iii) discussion of the main value drivers and risks; and (iv) any important assumptions, caveats, or competing interpretations. The memo should read as a coherent piece of financial analysis, not as a transcript of the AI interaction.

Evaluation Criteria. Your work will be evaluated on the quality of your *independent* reasoning, not on whether you agree with the AI or arrive at any particular recommendation. Strong submissions will be rewarded for being coherent, evidence-based, valuation-relevant, and analytically disciplined. In particular, high-quality work will identify the most important facts, explain why they matter, weigh alternative explanations where appropriate, and translate those judgments into a defensible valuation view. Memos that merely restate AI outputs, reproduce dashboard language without interpretation, or rely on unexplained conclusions will receive lower evaluations.