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

Huan Liu[1] Miao Liu[2] Zhizhe Liu[3] Danqing Mei[4]

[1]Google

[2]Boston College

[3]University of Wisconsin-Madison

[4]CKGSB

May 2026

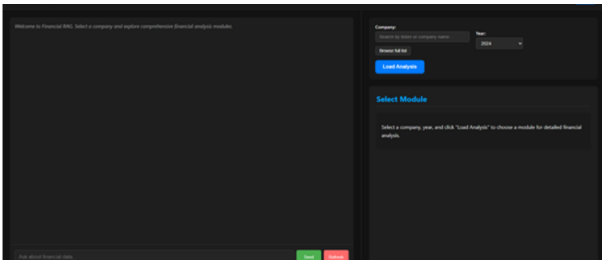
Research Question

- AI can now read filings, organize ratios, and benchmark peers—but the design question is **who produces what information** inside the human–AI team.
- In valuation, AI can lower search costs *and* anchor users, compressing independent reasoning.
- **When does AI-supplied structure complement rather than crowd out human reasoning?**
- We develop an AI valuation agent and run a field experiment varying the **amount and prescriptiveness** of AI-supplied hard information.

Experimental Design in One Slide

- **Task:** 126 advanced business students value real public firms using a custom AI valuation agent (90 minutes per firm).
- **Three treatment arms** (randomized, stratified on FSA knowledge):
  - **Low-Hard:** bare retrieval—agent answers questions, no dashboards
  - **Medium-Hard:** structured dashboards (MD&A, risk, DuPont, earnings quality) without valuation guidance
  - **High-Hard:** dashboards *plus* AI-generated peer diagnostics and valuation guidance
- **Key:** the underlying information set is held fixed; what changes is the **structure and prescriptiveness** of AI assistance.
- **Data:** full chat logs + final valuation memos $\rightarrow$ rich measures of human reasoning.

The AI Valuation Agent



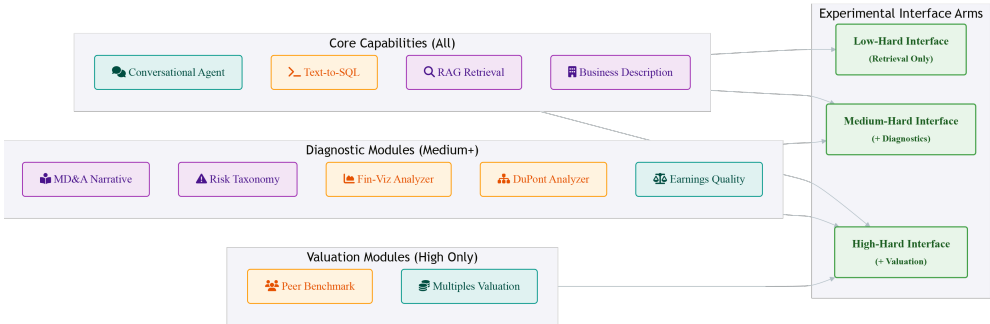
Data backbone

- Compustat fundamentals, ratio libraries, peer mappings
- 10-K text via RAG retrieval
- FRED macro, Yahoo Finance, SeekingAlpha

Interface

- Left: free-form chat
- Right: modular analytical reports
- Full interaction logging

Experimental Arms



- **Low-Hard**: bare retrieval + business description only.
- **Medium-Hard**: adds descriptive/diagnostic modules (no valuation assessment).
- **High-Hard**: adds peer-relative diagnostics and multiples-based valuation guidance.

How Prescriptive Is High-Hard, Really?

What High-Hard adds beyond Medium-Hard

- **Multiples Analyzer**: computes EV/EBITDA, P/E, etc. and benchmarks against an LLM-derived peer set with visual comparisons and textual interpretation.
- **Peer Benchmark**: compares operating and balance-sheet metrics across closest competitors, producing group medians, ranges, and relative positioning.

What it still does *not* do

- Does not determine which multiples deserve the most weight.
- Does not adjudicate peer comparability, transitory items, or cross-firm accounting distortions.
- Does not reconcile conflicting signals into a single justified target price.

Key point: High-Hard provides **structured comparison and directional framing**, not a complete valuation. The AI does not tell students what to conclude—so the crowding-out we observe is not because the AI handed them a final answer.

What the High-Hard Multiples Module Looks Like

Illustrative output:

	Focal Firm	Peer Median	Peer Min	Peer Max	Premium/Discount
EV / EBITDA	18.2x	14.7x	10.3x	22.1x	+23.8%
P / E (fwd)	25.4x	21.8x	15.6x	31.2x	+16.5%
EV / Revenue	4.1x	3.5x	1.8x	6.3x	+17.1%
P / FCF	30.7x	24.3x	16.9x	35.8x	+26.3%

AI-generated interpretation (paraphrased):

> *"The firm trades at a 24% EV/EBITDA premium to peers. This could reflect higher growth expectations or margin durability—but could also indicate overvaluation if peer comparability is imperfect."*

→ The module frames the question but **leaves the analyst to judge** whether the premium is justified—yet even this is enough to induce anchoring and delegation.

Measuring Human Reasoning

Multi-method measurement from chat traces and memos:

Human-coded (query level)

Soft Information Index =
sum of six standardized components:

- Follow-Up Depth
- Reasoning Depth
- Cross-Period Synthesis
- Multi-Metric Synthesis
- Hypothesis Testing
- Mechanism-Seeking

LLM-coded (conversation level)

- Hypothesis-Space Breadth
- Analytical Progression

Graph-based

- DAG Reasoning: validated evidence-to-interpretation edges

Cross-measure correlations:
0.74–0.76 (human vs. LLM coded)

Examples: What Do the Measures Capture?

Low-scoring query (retrieval request, no student reasoning)

"Show me Adobe's revenue and gross margin for the last three years."

High-scoring query (hypothesis testing + mechanism-seeking + cross-period synthesis)

"Gross margin improved from 2021 to 2023 despite flat revenue growth. Is that driven by product mix shifting toward subscriptions, or by cost discipline in COGS? Can you compare the trajectory of subscription revenue share with the margin trend to help me distinguish?"

- **Cross-Period Synthesis**: compares 2021 vs. 2023 trajectories.
- **Hypothesis Testing**: proposes two competing explanations (mix shift vs. cost discipline).
- **Mechanism-Seeking**: asks *how* margin improvement arose, not just *what* happened.
- **Follow-Up Depth**: builds on an observed pattern to pursue a refined question.

Example: Evidence-to-Interpretation Graph (DAG)

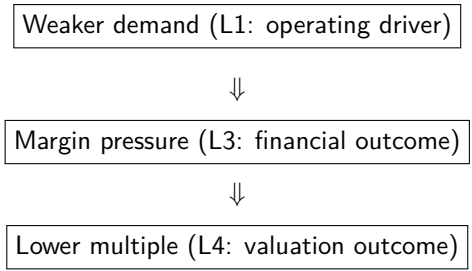
Student text

"Rising inventory days and management's discussion of weaker demand suggest future margin pressure, which supports a lower valuation multiple."

Evidence nodes

- Inventory days (Ratio)
- MD&A: weaker demand

Interpretation nodes



- **DAG Reasoning** counts validated edges from evidence to interpretation—measuring whether evidence is *converted into reasoning*, not just mentioned.

Main Results: Non-Monotonic Treatment Effects

	Soft Information Index	Hypothesis-Space Breadth	Analytical Progression	DAG Reasoning
Medium-Hard	3.227***	1.994***	0.368***	2.365**
	(0.693)	(0.411)	(0.126)	(1.058)
High-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)

Firm FE, SEs clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

- **Medium-Hard** substantially raises all four reasoning outcomes (0.43–0.88 SD).
- **High-Hard** delivers much smaller, mostly insignificant gains (0.02–0.44 SD).
- The pattern is **non-monotonic**: dashboards help; adding valuation guidance does not extend the gains.

Which Components of Reasoning Move?

	Follow-Up Depth	Reasoning Depth	Cross-Period Synthesis	Multi-Metric Synthesis	Hypothesis Testing	Mechanism-Seeking
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)

Firm FE, SEs clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

- **Medium-Hard** strengthens multiple margins: follow-up depth, cross-period synthesis, hypothesis testing, and mechanism-seeking.
- **High-Hard** significantly increases only reasoning depth—consistent with narrower engagement.

Expertise Amplifies, Not Equalizes

	Soft Information Index	Hypothesis-Space Breadth	Analytical Progression	DAG Reasoning
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)
Firm FE, SEs clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.				

- Implied FSA slope: 0.32 (Low-Hard) $\rightarrow$ 0.83 (Medium-Hard) $\rightarrow$ 1.74 (High-Hard).
- The most prescriptive interface is **expertise-amplifying**: strong users interrogate AI guidance; weaker users anchor on it.

High-Hard Induces Delegation

	Delegation Any	Delegation Turns	Delegation Memo Score
Medium-Hard	-0.077	0.235	-0.195
	(0.076)	(0.302)	(0.289)
High-Hard	0.317***	1.169***	1.103***
	(0.100)	(0.279)	(0.307)
Firm FE, SEs clustered by firm. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.			

- **Medium-Hard** does not increase delegation.
- **High-Hard** sharply increases analytical handoff to the AI across all three delegation measures.
- More valuation-oriented output, but partly AI-authored rather than human-authored.

Mechanism: Evidence-to-Interpretation Conversion

Support-edge coefficient	Medium-Hard	High-Hard
MD&A	0.616***	0.361*
Risk Factors	0.335**	0.210*
Ratio Patterns	0.660	-0.689
DuPont	0.849**	0.425
Earnings Quality	0.815**	0.070
$N = 756$ evidence-node obs. Firm FE, SEs clustered by firm.		

- **Medium-Hard** increases conversion of MD&A, risk factors, DuPont, and earnings-quality evidence into explicit reasoning.
- **High-Hard** preserves some narrative-evidence processing but converts less broadly.

Mechanism: Which AI Responses Stimulate Human Reasoning?

	Next-query Soft Info	Next-query Delegation
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)
$N = 1{,}177$ turn pairs.		

Key findings

- **Causal Gap** and **Mechanism Gap** strongly stimulate next-turn reasoning.
- **Valuation Stance** does *not* increase reasoning.
- AI is most complementary when it organizes evidence but **leaves inferential work to the human**.

Contributions

- **Hard vs. soft information:** first tightly controlled experiment directly measuring human soft-information production in response to AI-supplied hard information.
- **Human–AI collaboration:** moves beyond accept/reject framing to study *who produces what information* inside the team—showing non-monotonic effects of AI prescriptiveness.
- **Accounting and finance education:** FSA knowledge is complementary to AI, not substituted by it—foundational skills should be repositioned as complements in the LLM era.

Conclusion

- The best-performing interface is **not** the most prescriptive one.
- **Structured diagnostic support** (Medium-Hard) is more complementary to human reasoning than AI-generated valuation guidance (High-Hard).
- More prescriptive AI can improve output appearance but increases **delegation** and makes performance more **contingent on prior expertise**.
- **Design principle:** AI works best when it organizes the evidentiary base without collapsing the inferential task that remains for the human analyst.

Appendix: What Is on the FSA Quiz?

- DuPont decomposition and ratio analysis.
- Accruals, cash flows, working capital, and earnings quality.
- Growth, ROIC, WACC, DCF, FCFF, FCFE, and residual income.
- Peer multiples, comparability adjustments, and valuation interpretation.
- Several questions are designed to test whether students can distinguish operating performance from accounting presentation and financing effects.

The quiz is administered before the experiment and used to stratify randomization and as a continuous moderator.