Recovering Accounting Judgment with AI: Textual and Non-Textual Experiments

Whose judgment do AI participants reproduce?

Hunter Ng

November 2026

Can AI recover accounting judgement?

- Accounting AI research tests information extraction, disclosure analysis, and answers to accounting questions.
- As models improve, conclusions about these capabilities change.

Can AI recover how financial-statement users respond to an accounting choice?

Does describing their knowledge and characteristics improve recovery?

ROADMAP
Question
Increasing AI capability
Institutional details on judgment
Examples
Design
Tables

Whose judgement is the benchmark?

Judgement evaluates evidence and assigns weight to it in reaching a decision. Training, knowledge, objectives, and attitudes toward uncertainty affect those weights.

Question	Benchmark
Factual question	An identified correct answer
Judgement	The response of a specified group

Standard setters need timely evidence before adopting a rule. Published experiments provide human benchmarks for recognition, measurement, presentation, and disclosure choices.

(Bonner and Lewis, 1990; Libby and Luft, 1993; FASB, 2025; Koonce et al., 2024)

Preview of results

- **Textual versus visual presentation:** no statistically detectable difference in recovery: 40.5% versus 43.4% ($p = .804$). Overall, GPT-6 recovers 39.8% of 236 human directions.
- **What changes matters:** within presentation-based experiments, recovery is 64.7% for information changes and 31.0% for presentation-only changes.
- **Facts and judgement:** both models answer all 668 valid accounting-information questions correctly; GPT-6 recovers 37.9% of human directions across 195 findings where all available valid checks pass.
- **Participant descriptions:** knowledge raises recovery from 41.5% to 48.5% on 200 matched comparisons. Knowledge plus higher evidence thresholds adds 12.06 percentage points relative to neutral instructions.
- **Audience composition:** recovery is 41.9% under neutral instructions, 53.8% with equal risk-profile weights, and 55.4% with historical shares. Dated weights add 1.61 points beyond the equal mixture.
- **New tasks:** across 12 comparisons in two studies, recovery is 16.7% under neutral instructions, 25.0% with knowledge, and 33.3% with deliberate attention.

Accounting AI papers, 2024-2026



Koonce et al. (2024): papers by accounting standard

Conceptual framework

2006 Frederickson et al.
2009 Hewitt
2010 Hodge et al.
2010 Koonce et al.
2011 Koonce et al.
2012 Kadous et al.
2013 Libby & Brown
2014 Bonner et al.
2015 Anderson et al.
2015 Bloomfield et al.
2015 Hewitt et al.
2016 Kelton & Murthy
2016 Majors
2019 Cade et al.
2020 Doxey et al.
2021 Cascino et al.
2021 Young
2022 Barradale et al.
2022 Liu et al.
2023 Anderson

ASC 600: Revenue

2007 Clor-Proell & Nelson
2012 Bailey & Sawers
2017 Capps et al.
2021 Young

ASC 200: Presentation

1990 Marchant
1990 Mautz
1991 Cooper & Selto
1997 Maines et al.
1998 Hirst & Hopkins
2000 Maines & McDaniel
2001 Hronsky & Houghton
2004 Hirst et al.
2006 Bloomfield et al.
2008 Hodder et al.
2010 Elliott et al.
2010 Hodge et al.
2014 Clor-Proell et al.
2015 Elliott et al.
2015 Hewitt et al.
2015 Lachmann et al.
2017 Emett & Nelson
2019 Koonce et al.

ASC 700: Expenses

1991 Cooper & Selto
1992 Sami & Schwartz
1995 Cloyd
2001 Luft & Shields
2004 Hodge et al.
2006 Frederickson et al.
2009 Clor-Proell
2010 Seybert
2017 Emett & Nelson

ASC 300: Assets

1990 Marchant
1996 Cloyd et al.
1998 Hirst & Hopkins
2000 Hopkins et al.
2000 Maines & McDaniel
2001 Luft & Shields
2006 Bloomfield et al.
2008 Jackson
2009 Clor-Proell
2010 Elliott et al.
2010 Jackson et al.
2010 Seybert
2015 Elliott et al.
2015 Rennekamp et al.
2017 Asay et al.
2018 Tan & Trotman
2020 Kunz & Staehle

ASC 900: Industries

1997 Khumawala & Gordon
2001 Dietrich et al.
2003 Hirst et al.
2010 Koonce et al.
2021 Eilifsen et al.
2023 Anderson

ASC 400: Liabilities

1990 King et al.
1991 Davis et al.
1992 Reimers
1996 Hackenbrack & Nelson
1996 Hopkins
1998 Kennedy et al.
2006 Hodge et al.
2014 Clor-Proell & Maines
2015 Fanning et al.
2016 Clor-Proell et al.
2018 Kelton & Montague
2020 Du et al.

ASC 800: Transactions

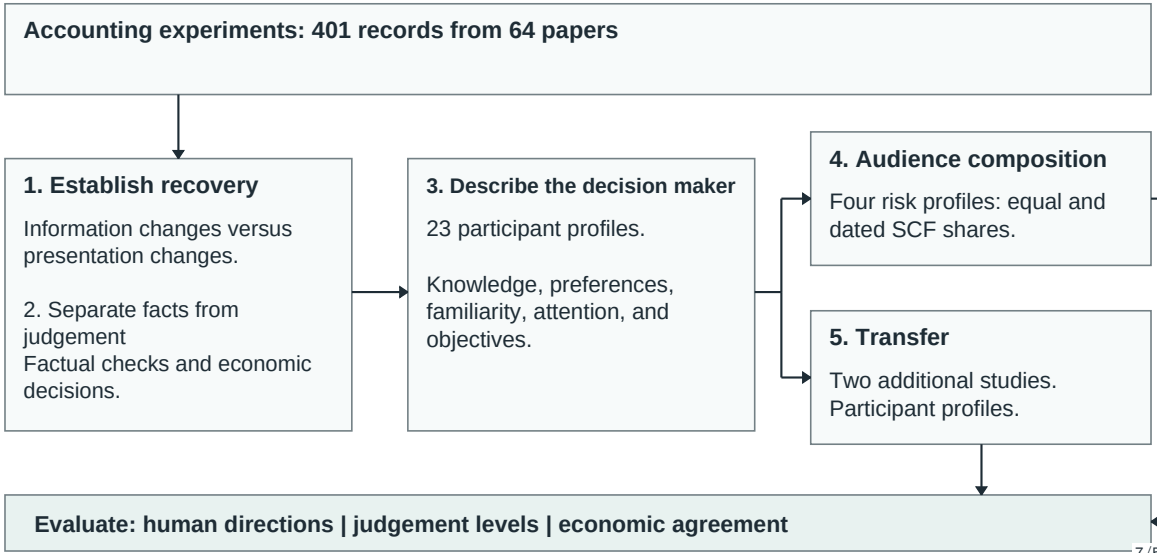
1995 Cloyd
1996 Hackenbrack & Nelson
2000 Hopkins et al.
2004 Hirst et al.
2005 Koonce et al.
2007 Nelson & Tayler
2010 Jackson et al.
2010 Jamal & Tan
2011 Agoglia et al.
2011 Gaynor et al.
2011 Koonce et al.
2012 Bailey & Sawers

ASC 800: Transactions (cont.)

2012 Hales et al.
2012 Kadous et al.
2013 Chen et al.
2013 Cohen et al.
2014 Bonner et al.
2014 Clor-Proell et al.
2014 Griffin
2015 Anderson et al.
2015 Lachmann et al.
2015 Nelson & Rupar
2016 Majors
2019 He et al.
2019 Koonce et al.
2020 Doxey et al.
2020 Tan & Mayorga
2021 Chen et al.
2021 Eilifsen et al.
2022 Barradale et al.
2022 Liu et al.

Red: papers in our sample. Papers can appear under more than one standard.

Paper setup


Accounting experiments: 401 records from 64 papers
1. Establish recovery
Information changes versus presentation changes.
2. Separate facts from judgement
Factual checks and economic decisions.
3. Describe the decision maker
23 participant profiles.
Knowledge, preferences, familiarity, attention, and objectives.
4. Audience composition
Four risk profiles: equal and dated SCF shares.
5. Transfer
Two additional studies.
Participant profiles.
Evaluate: human directions | judgement levels | economic agreement

What is a judgement?

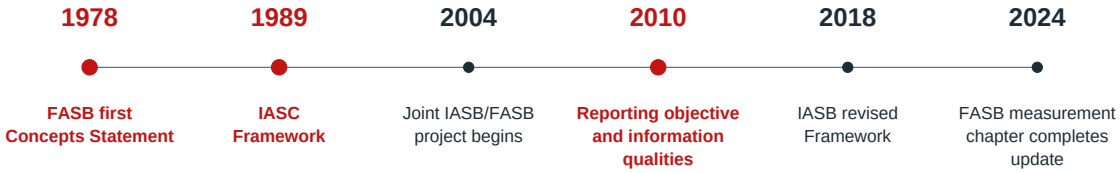
Judgement involves evaluating evidence and assigning weight to it in reaching an assessment or decision.

Evaluation	Meaning
Inaccuracy	An error about **facts**: an amount, classification, or calculation against a defined answer.
Inconsistency	A difference in **judgement** against a stated decision rule, economic criterion, or repeated-task comparison.
Human recovery	Reproducing the response of the intended person or group to the reporting choice.

(Bonner and Lewis, 1990; Libby and Luft, 1993)

A judgment comparison requires a stated criterion and a specified decision maker.

Pelger (2016): Conceptual frameworks



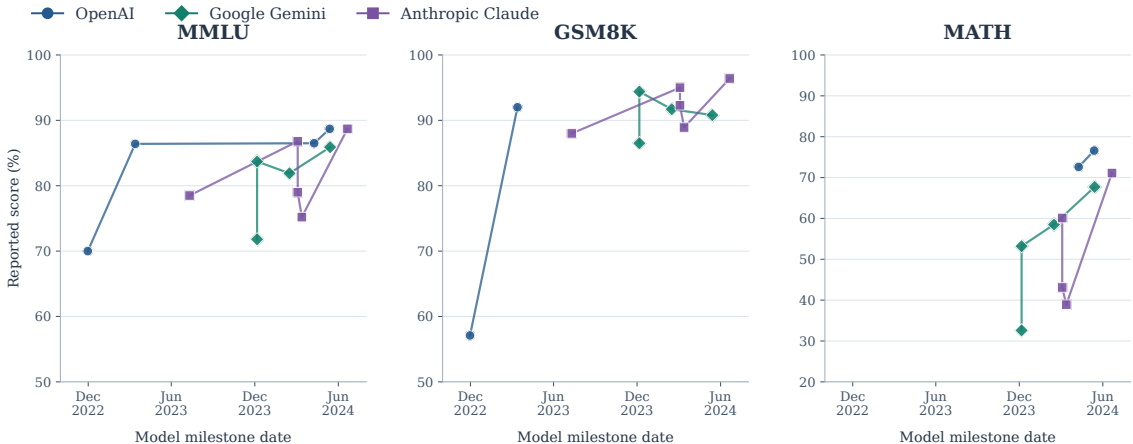
First statements: reporting objectives guide standard setting.

Conceptual framework: defines useful information and guides judgement about reporting choices and informational trade-offs.

Pelger: decision usefulness shaped debates over objectives and stewardship.

Sources: Pelger (2016); Erb and Pelger (2015); IASB (2018); FASB (2024).

Model performance over time: knowledge and mathematics

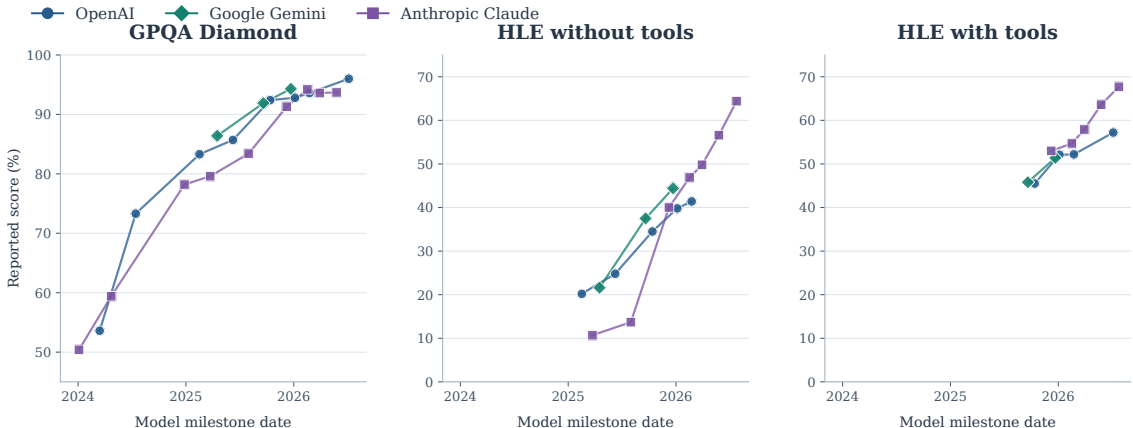


OpenAI: GPT-3.5 → GPT-4 → GPT-4 Turbo → GPT-4o

Google Gemini: 1.0 Pro / Ultra → 1.5 Pro (February preview → May update)

Anthropic Claude: 2 → 3 Opus / Sonnet → 3 Haiku → 3.5 Sonnet

Model performance over time: scientific and multidisciplinary reasoning



OpenAI: GPT-4o → o1-preview → o3 → GPT-5 → 5.2 → 5.4 → 5.5 → GPT-6 Astra

Google Gemini: 2.5 Pro → 3 Pro → 3.1 Pro

Anthropic Claude: 3 Opus → 3.5 Sonnet → 3.7 Sonnet → Opus 4 → Sonnet 4.5 → Opus 4.6 → 4.7 → 4.8 → 5 → 5.5

From academic questions to complete workflows

2019

- **Nov - ARC-AGI-1**
  Learning

2020

- **Sep - MMLU**
  Knowledge

2021

- **Mar - MATH**
  Mathematics
- **Jul - HumanEval**
  Software
- **Oct - GSM8K**
  Mathematics

2023

- **Jul - WebArena**
  Web / desktop
- **Oct - SWE-bench**
  Software
- **Nov - GPQA / Diamond**
  Knowledge
- **Nov - MMMU**
  Multimodal

2024

- **Apr - OSWorld**
  Web / desktop
- **Jun - MMLU-Pro**
  Knowledge
- **Jul - LAB-Bench**
  Science
- **Aug - SWE-bench Verified**
  Software
- **Sep - MMMU-Pro**
  Multimodal
- **Nov - FrontierMath**
  Mathematics

2025

- **Jan - Humanity's Last Exam**
  Knowledge
- **Mar - ARC-AGI-2**
  Learning
- **Apr - BrowseComp**
  Web / desktop
- **May - Terminal-Bench**
  Terminal
- **Sep - GDPval**
  Professional work
- **Nov - Terminal-Bench 2.0**
  Terminal

The widening benchmark landscape in 2026

Jan-Mar

- **Jan - FrontierMath Open Problems**
  Mathematics
- **Feb - LABBench2**
  Science
- **Mar - Riemann-bench**
  Mathematics
- **Mar - ARC-AGI-3**
  Learning

Apr-Jun

- **Apr - AutomationBench**
  Professional work
- **May - DeepSWE v1**
  Software
- **Jun - Agents' Last Exam**
  Professional work
- **Jun - FrontierMath Tier 4 v2**
  Mathematics
- **Jun - DeepSWE v1.1**
  Software
- **Jun - OSWorld 2.0**
  Web / desktop

Jul-Sep

- **Jul - AutomationBench v1.0.6**
  Professional work
- **Aug - Terminal-Bench-Science 0.1**
  Science
- **Aug - Terminal-Bench 4.0**
  Terminal
- **Sep - FrontierSWE v2**
  Software
- **Sep - OSWorld 2.1**
  Web / desktop
- **Sep - HLE-Diamond**
  Knowledge

October

- **Oct - Agents' Last Exam v1.1**
  Professional work
- **Oct - FrontierCode 2.0**
  Software
- **Oct - Artificial Analysis v5.0**
  Benchmark adoption

Increasing capability sharpens the judgement question

Reported benchmark score (%)	Earlier milestone	Later milestone
MMLU: GPT-3.5 to GPT-4o	70.0	**88.7**
GPQA Diamond: GPT-4o to GPT-6 Astra	53.6	**96.0**

These tests reward answers against a defined key. Accounting estimates and decisions also require a criterion for how evidence should be weighted.

Which criterion matters: the facts, an accounting rule, economic consistency, or the intended user's response?

ROADMAP
Question
Increasing AI capability
Institutional details on judgment
Examples
Design
Tables

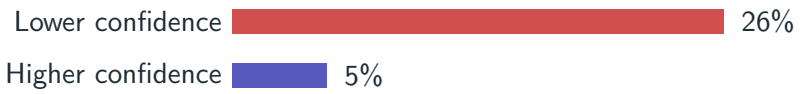
Koriat (2008): Answer switching

- Koriat (2008): 41 students answered 105 two-choice general-knowledge questions twice.
- After a short break, the same questions appeared in a new order.
- Participants repeated **83.4%** of their initial answers.

The implied average switching rate is **16.6%**.

Koriat (2012): Confidence and switching

Initial confidence	Repeat answer	Switch answer
Below own median	74%	**26%**
Above own median	95%	**5%**



Goldberg (1970): Modelling judgement

- Goldberg (1970) analyzed 29 clinical psychologists judging 861 MMPI profiles.
- For each psychologist, a linear model predicted that psychologist's rating from 11 MMPI scales.
- The fitted coefficients used the psychologist's judgments as the target.

Goldberg (1970): Criterion accuracy

Predictor	Correlation with recorded diagnosis
Psychologists	.28
Models of psychologists	**.31**

In split-sample cross-validation, models outperformed their corresponding psychologists in **86%** of comparisons.

Reproducing a person's judgement differs from predicting an external criterion.

Benchmarks for evaluating a judgment

Outcome	Benchmark
Factual accuracy	The stated amount, classification, or calculation
Rule compliance	The governing accounting requirement
Economic consistency	A specified payoff, ordering, or economic criterion
Human recovery	The response of the intended person or group

OpenAI interaction style and behavioral changes

25 Jan 2024

GPT-4 Turbo

Update targets incomplete tasks and code generation.

Apr–May 2025

Monday persona

An intentionally irreverent persona runs for one month.

24–29 Apr 2025

GPT-4o agreement

Excessive validation leads to mitigation and rollback.

15 Aug 2025

GPT-5 warmth

The default interaction becomes warmer and more familiar.

Nov 2025

GPT-5.1 language

Warmer conversation accompanies increased goblin and gremlin metaphors.

Mar 2026

Nerdy retired

The persona associated with creature-language rewards is retired.

29 Apr 2026

Goblin investigation

OpenAI links the language habit to personality-training rewards.

29 Oct 2026

GPT-6 flattery

A serving update makes specified traffic more concise and less flattering.

Useful information requires accounting judgement

Framework characteristic	Requirement
Relevance	Predictive or confirmatory value for decisions.
Faithful representation	Complete, neutral, and free from error.
Enhancing qualities	Comparability, verifiability, timeliness, understandability.
Cost constraint	Reporting benefits justify the costs.

(FASB, 2024), QC4–QC35.

Estimates incorporate future expectations and estimation error (Barth, 2006; Dechow and Dichev, 2002). Evaluating complex estimates requires integrating evidence (Griffith et al., 2015).

Accounting judgement applies reporting objectives to evidence, uncertainty, and informational trade-offs.

The development of benchmarks for tabular data



FinQA

Financial tables plus numerical programs

TAT-QA

Tables and narrative disclosures

TableBench preprint

General table reasoning and visualization

TableBench at AAAI

A broad table-reasoning benchmark

ToRR

Robustness across representations

TabArena

Prediction from structured features

ToRR revision

Updated representation tests

SpreadsheetBench 2

321 professional workbook tasks

Atomicity of tabular information

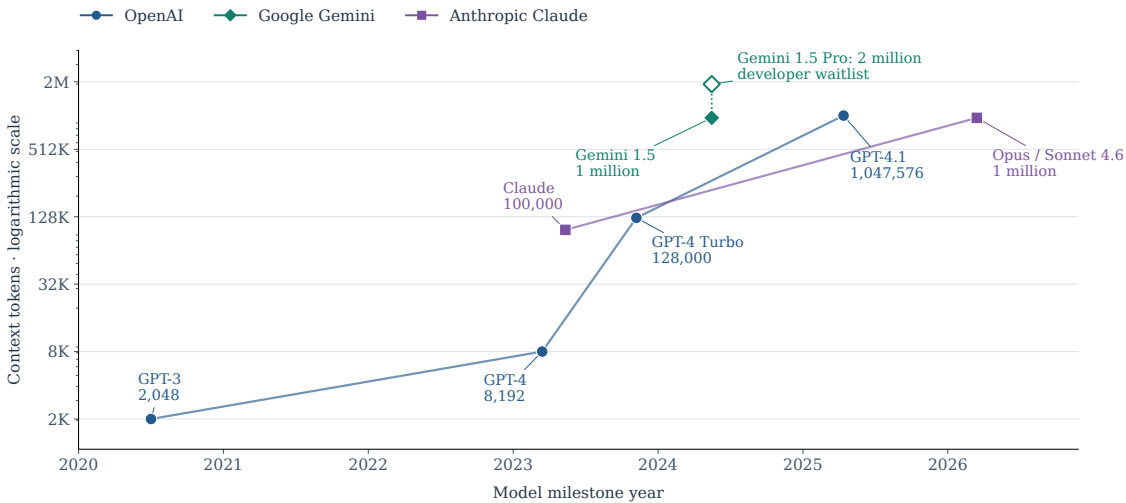
Kim, Kang, and Jung (2026) decompose table structure into six atomic features.

Atomic feature	Relationship conveyed
Grid formation	Which value occupies each row–column intersection.
Header designation	Which cells describe the data cells.
Cell merging	Which labels span rows or columns.
Header nesting	How categories form a hierarchy.
Section stacking	How sections share column structures.
Table composition	How different table structures fit together.

Source: (Kim et al., 2026).

An accounting amount must retain its relationships to periods, categories, and reporting sections.

The expansion of context-window capacity



Users bring knowledge and preferences to the same case

- Knowledge affects which classifications, relationships, and disclosures a user recognizes.
- Preferences affect how the user weighs uncertainty and decision consequences.
- A participant group combines people with different characteristics.

Predictions for accounting judgements

H1: What does the reporting choice change?

Within presentation-based experiments, AI recovers human responses more often when the treatment changes information than when it changes presentation while holding information constant.

H2: Is correct processing sufficient?

AI and human judgements can differ even when AI correctly answers separate questions about the relevant accounting information.

Hirst et al. (2004): The accounting task

56 bank-specialist analysts and portfolio managers assess exposed or hedged banks under full fair value (FFV) or piecemeal fair value (PFV).

Exposed bank (millions)	20X1		20X2		20X3	
	NI	CI	NI	CI	NI	CI
FFV	16.7	15.3	18.6	20.1	20.3	**12.1**
PFV	16.7	16.4	18.6	18.9	20.3	16.7

Actual questions (1–15):

"Assess the investment risk of the bank's common stock relative to that of an average bank of equivalent size."

"Assess the investment risk of the bank's common stock in the context of a diversified portfolio."

(Hirst et al., 2004), Figure 1 and Table 1.

Hirst et al.: Predictions and human results

Normative predictions: the same bank receives the same risk judgment under FFV and PFV; the exposed bank is riskier under both.

Risk composite (1–15)	FFV	PFV	FFV − PFV
Exposed	**10.10**	**8.11**	+1.99
Hedged	6.68	6.94	−0.26

(Hirst et al., 2004), Table 1.

The human response to economic exposure is stronger when fair-value changes are recognized in income.

Koonce et al. (2005): The accounting task

90 MBA students assess three descriptions of economically equivalent positions. Each position produces fixed net interest payments.

Actual question: "Overall, how risky do you consider the company's [instrument]?"

Instrument wording	Net payments
Fixed interest-rate debt	Fixed
Variable-rate debt combined with a variable-to-fixed-interest-rate swap	Fixed
Variable-rate debt combined with a variable-to-fixed-interest-rate swap used as a hedge	Fixed

Risk is rated from 0 to 100, then rated again after cash-flow and fair-value exposure information.

(Koonce et al., 2005), Figure 1 and Table 1.

Koonce et al.: Predictions and human results

Predictions: initially, swap > fixed-rate debt and swap > hedge; exposure information should eliminate labeling effects.

Risk rating (0–100)	Before information	After information
Fixed-rate debt	**35.43**	39.17
Debt with a swap	**44.43**	43.83
Debt with a swap used as a hedge	37.17	34.60

(Koonce et al., 2005), Table 1.

Added exposure information did not eliminate the overall labeling pattern.

Koonce et al. (2005): A calculation error

Separate legacy diagnostic, outside the common 668-question bank.

The instrument reports a $68,061,000 decrease, equal to 14.64% of next year's net income.

Saved diagnostic question:

"Treating the table's stated dollar decrease and percentage as exact inputs, what implied next year's net income, in dollars, does the table indicate?"

Calculation	Result
Correct: $68,061,000 ÷ 0.1464	**$464,897,540.98**
Saved GPT-6 response	**$465,170,081.97**
Absolute error	$272,540.99
Review result	Incorrect; retained as an arithmetic error

(Koonce et al., 2005), Experiment 3; saved case E3_LOSS_EQ_GAIN, stage E3_LOSS_ONLY.

Hold the accounting case fixed; describe the decision maker

- Fix the source-defined role, treatments, questions, and response scale.
- Vary assigned knowledge, preferences, and other participant characteristics.
- Compare the same finding under neutral and profile instructions.

The source catalogue and the analytical samples

Analytical population	Count	Papers
Source catalogue: 401 accepted records	395	64
Source-compatible nonzero directions	208	47
Controlled procedure patterns	45	28
Unprofiled model-comparison panel	236	46
Exact-source human condition-mean vectors	69	20

The inventory contains 65 studies. Paired tests use the same available findings across the compared profiles.

Participant profiles describe the decision maker

Profile family	Profiles	Description
Reference prompts	2	Original prompt; fresh neutral prompt
Assigned knowledge	1	Accounting expertise
Uncertainty and evidence	8	Cautious/open attitudes; higher/ordinary thresholds; with or without knowledge
Other characteristics	8	Horizons, familiarity, attention, objectives
Financial risk	4	SCF investment-risk categories

Primary profile model: GPT-6 Luna; temperature 0; reasoning disabled.

Four risk profiles define the audience mixture

Substantial risk; above-average risk; average risk; no financial risk.

$$\bar{y}_g(t) = \sum_{j=1}^{4} w_{gj} y_j(t), \qquad \sum_{j=1}^{4} w_{gj} = 1.$$

Audience mixture	Shares applied to the same four responses
Equal mixture	25% in each risk category
Dated audience	Closest preceding Survey of Consumer Finances wave
Fixed audience	2022 Survey of Consumer Finances distribution

Hold the four profile responses fixed; change only the audience shares.

Directions, judgement levels, and economic decisions

Outcome	Evaluation criterion
Human recovery	Same sign as the source-compatible nonzero human contrast
Controlled pattern agreement	Source-defined pattern under the adapted task procedure
Judgement-level agreement	Distance from published human condition means
Economic agreement	Source-supported accounting rule or economic decision

Separate factual checks evaluate amounts, classifications, and relationships.

Agreement in direction need not imply agreement in level or decision quality.

Two additional accounting studies test transfer

Study	Initial decision task	Conditions
Le (2026)	Accounting rules and accounting-career judgements	3
Cai et al. (2026)	Cryptocurrency classification, measurement, and valuation	2

Apply the 23 participant profiles to the five initial decision screens.

These applications use the profile descriptions without audience weights.

(Le, 2026; Cai et al., 2026)

The empirical sequence

1. Establish recovery on the accounting tasks.
2. Vary participant knowledge and characteristics.
3. Combine four risk-profile responses and vary their audience shares.
4. Assess judgement levels and the two additional accounting studies.
5. Examine robustness of the accounting comparisons.

Table 1: Panel A: Source populations

Sample Construction and Analytical Coverage

Source population	Mappings	Findings	Papers
	(1)	(2)	(3)
Panel A. Source universe and analytical populations			
All accepted source mappings	401	395	64
Unprofiled model-comparison panel	236	236	46
Source-eligible nonzero judgments	208	208	47
All controlled patterns	45	45	28
Combined source patterns	253	253	49

Table 1: Panel C: Analytical denominators

Sample Construction and Analytical Coverage

Analytical population	Findings	Papers
	(1)	(2)
Panel C. Analytical finding and paper denominators		
K versus N: source-eligible nonzero judgments	200	44
K versus N: controlled patterns	37	28
K versus N: combined source patterns	237	48
Four-arm mixture cohort: Human-direction agreement	186	39
Four-arm mixture cohort: Zero contrast	186	39
Four-arm mixture cohort: Reversed direction	186	39
Four-arm mixture cohort: Economic agreement	68	17
Exact-source complete human mean vectors: all sources	69	20
K versus N means: Normalized MAE	34	13
K versus N means: Like-unit improvement	60	19

Table 2: Textual versus visual presentation

Panel A: Source-eligible model contrasts

	(1) Textual	(2) Presentation-based	(3) Inseparable	(4) Indeterminate	(5) All
Eligible scoreable findings	163	53	19	1	236
Source papers	36	14	8	1	46
Same direction as humans	66	23	5	0	94
Zero contrast	61	21	9	1	92
Opposite direction	36	9	5	0	50
Same direction (%)	40.5	43.4	26.3	0.0	39.8

Panel B: Presentation-based minus textual regressions

Outcome	(1) Human direction recovered	(2) Zero contrast	(3) Reversed direction
Presentation-based treatment	2.91	2.20	−5.10
	(11.64)	(11.49)	(6.88)
Nominal p-value	0.804	0.849	0.462
Findings	216	216	216
Source papers	44	44	44
Intercept	X	X	X

No statistically detectable recovery difference (p = .804).

Panel C: Information and Presentation Changes within Presentation-Based Materials

	(1) GPT-4.1	(2) GPT-4.1 mini	(3) GPT-5	(4) GPT-6	(5) GPT-6.1 Sol
Information changes: recovered/findings	14/17	11/17	9/17	11/17	11/17
Recovery (%)	82.4	64.7	52.9	64.7	64.7
Presentation changes holding information constant: recovered/findings	18/29	18/29	14/29	9/29	11/29
Recovery (%)	62.1	62.1	48.3	31.0	37.9
Difference (percentage points)	20.28	2.64	4.67	33.67*	26.77
	(15.11)	(20.44)	(15.37)	(18.99)	(16.86)
Nominal p-value	0.202	0.899	0.766	0.100	0.136
Findings	46	46	46	46	46
Source papers	14	14	14	14	14

Table 3: Panel A: Accounting-information checks

Accounting Information and Human Judgment

Panel A. Accounting-Information Checks

Question type	(1) Valid	(2) GPT-6 correct (% of valid)	(3) GPT-6.1 Sol correct (% of valid)
Identify an amount	237	237 (100.0)	237 (100.0)
Classify accounting information	223	223 (100.0)	223 (100.0)
Combine related information	208	208 (100.0)	208 (100.0)
Total	668	668 (100.0)	668 (100.0)

Panel B. Human Recovery after Correct Information Responses

	(1)	(2)	(3)	(4)
	GPT-6		GPT-6.1 Sol	
	All available checks pass	Three-type coverage	All available checks pass	Three-type coverage
Distinct directional findings	195	137	125	108
Source papers	45	37	35	30
Human direction recovered	74	48	52	45
Zero contrast	77	60	55	48
Reversed direction	44	29	18	15
Recovery (%)	37.9	35.0	41.6	41.7
Textual (recovered / findings)	51/134	29/85	35/85	28/68
Presentation-based (recovered / findings)	18/44	14/35	11/23	11/23
Presentation-based minus textual (percentage points)	2.85	5.88	6.65	6.65
	(12.54)	(14.99)	(14.19)	(15.20)

Table 4: Panel B: Economic benchmarks

Human Responses and Economic Benchmarks

Panel B: Model agreement with source-supported benchmarks

	(1) GPT-4.1	(2) GPT-4.1 mini	(3) GPT-5	(4) GPT-6	(5) GPT-6.1 Sol
Eligible comparisons	45	45	45	45	45
Zero consistent with source invariance	15	13	20	23	23
Source-supported economic direction matched	0	0	0	1	5
Benchmark-consistent responses (%)	33.3	28.9	44.4	53.3	62.2
Human directions recovered	24	27	19	13	14
Human recovery (%)	53.3	60.0	42.2	28.9	31.1
Higher-confidence comparisons	28	28	28	28	28
Benchmark-consistent responses (%)	35.7	32.1	42.9	46.4	64.3
Equal-paper benchmark agreement (%)	38.7	32.1	32.0	53.1	52.9
Equal-paper human recovery (%)	54.2	62.2	54.6	34.6	29.8

Table 5: Knowledge versus neutral instructions

Accounting Knowledge and Human-Judgment Recovery

Source definition	Outcome	Source N	K minus N (pp)	Paired N	Papers	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Source-eligible nonzero judgments	Human direction	207	7.00 (6.76)	200	44	X
Source-correspondence sensitivity	Human direction	205	7.07 (6.83)	198	43	X
Controlled numeric reconstructions	Source pattern	31	-9.68 (10.09)	31	26	X
Controlled patterns: other adaptations	Source pattern	13	0.00 (0.00)	6	3	X
All controlled patterns	Source pattern	44	-8.11 (8.11)	37	28	X
Combined source patterns	Source pattern	251	4.64 (5.74)	237	48	X

Table 6: Evidence thresholds and horizon

Participant Profiles and Knowledge Interactions

Assigned description	Direction	Zero	Reversal	N	Papers	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Evidence threshold						
	9.05	-5.03	-4.02			
Higher evidence threshold	(6.45)	(7.14)	(6.01)	199	43	X
	3.00	-4.50	1.50			
Ordinary evidence threshold	(6.14)	(4.87)	(5.93)	200	44	X
	12.06	-8.54	-3.52			
Higher threshold + knowledge	(5.84)	(5.76)	(4.98)	199	43	X
	7.58	-8.08	0.51			
Ordinary threshold + knowledge	(7.10)	(5.66)	(5.73)	198	44	X
Investment horizon						
	4.46	-4.46	0.00			
Nearer investment horizon	(5.65)	(6.46)	(3.88)	202	44	X
	4.98	-4.48	-0.50			
Longer investment horizon	(4.80)	(5.37)	(4.67)	201	44	X

Table 7: Human recovery and zero contrasts

Recovery and Economic Agreement across Risk-Profile Mixtures

Arm	Events	Findings	Papers	Equal-finding %	Equal-paper %
	(1)	(2)	(3)	(4)	(5)
Human-direction agreement					
Neutral prompt (N)	78	186	39	41.94	44.97
Equal four-risk mixture	100	186	39	53.76	56.14
Historical population mixture	103	186	39	55.38	58.39
Fixed-2022 population mixture	105	186	39	56.45	59.52
Zero contrast					
Neutral prompt (N)	75	186	39	40.32	36.57
Equal four-risk mixture	47	186	39	25.27	23.34
Historical population mixture	44	186	39	23.66	21.26
Fixed-2022 population mixture	44	186	39	23.66	21.26

Table 8: Panel A: Equal finding weights

Risk-Profile Mixtures and Audience Composition

Signed comparison	Recovery (pp)	Zero (pp)	Reversal (pp)
	(1)	(2)	(3)
Historical population mixture minus N	13.44*	-16.67	3.23
	(5.43)	(6.95)	(4.32)
Equal four-risk mixture minus N	11.83	-15.05	3.23
	(5.48)	(7.22)	(4.32)
Historical population mixture minus Equal four-risk mixture	1.61	-1.61	0.00
	(1.69)	(1.30)	(1.54)

Table 9: Panel A: Normalized MAE

Agreement with Human Judgment Levels

Assigned description	Equal findings	Equal papers	N	Papers
	(1)	(2)	(3)	(4)
Strong assigned knowledge	0.021 (0.031)	0.020 (0.025)	34	13
Cautious preference	0.003 (0.027)	0.003 (0.030)	34	13
Opportunity-open preference	0.004 (0.018)	-0.004 (0.018)	34	13
Cautious + knowledge	0.056 (0.038)	0.050 (0.035)	30	12
Opportunity-open + knowledge	0.040 (0.023)	0.041 (0.021)	34	13
Higher evidence threshold	-0.025 (0.024)	-0.027 (0.025)	34	13
Ordinary evidence threshold	0.013 (0.032)	0.015 (0.030)	34	13

Table 9: Panel C: Equal finding weights

Agreement with Human Judgment Levels

Signed comparison	Normalized MAE	Like-unit improvement
	(1)	(2)
Equal finding weights		
Historical population mixture minus N	-0.051	0.302
	(0.054)	(0.224)
Equal four-risk mixture minus N	-0.044	0.208
	(0.056)	(0.214)
Historical population mixture minus Equal four-risk mixture	-0.007	0.113
	(0.006)	(0.177)

Table 10: Le: Reference prompts, knowledge, uncertainty

Participant Profiles in Two Additional Accounting Studies

Assigned description	Agreement / scored	Agreement %	Normalized effect error
	(1)	(2)	(3)
Original prompt			
Original prompt (B)	3/11	27.27	0.167
Neutral control			
Neutral prompt (N)	2/11	18.18	0.240
Assigned knowledge			
Strong assigned knowledge	3/11	27.27	0.208
Uncertainty preference			
Cautious preference†	2/11	18.18	0.240
Opportunity-open preference†	2/11	18.18	0.240
Cautious + knowledge†	2/11	18.18	0.240
Opportunity-open + knowledge†	2/11	18.18	0.240

References i

Mary E. Barth. Including estimates of the future in today's financial statements. *Accounting Horizons*, 20(3):271–285, 2006. doi: 10.2308/acch.2006.20.3.271.

Sarah E. Bonner and Barry L. Lewis. Determinants of auditor expertise. *Journal of Accounting Research*, 28(Supplement):1–20, 1990. doi: 10.2307/2491243. URL https://doi.org/10.2307/2491243.

Qingzhi Cai, Zheng J. Leitter, Long Niu, and Hun-Tong Tan. Accounting classification and measurement for cryptocurrencies: Impact on valuation judgments, 2026. URL https://ssrn.com/abstract=7011540. Available at SSRN.

Eddy Cardinaels. The interplay between cost accounting knowledge and presentation formats in cost-based decision-making. *Accounting, Organizations and Society*, 33(6):582–602, 2008. doi: 10.1016/j.aos.2007.06.003. URL https://doi.org/10.1016/j.aos.2007.06.003.

Patricia M. Dechow and Ilia D. Dichev. The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*, 77(Supplement):35–59, 2002. doi: 10.2308/accr.2002.77.s-1.35.

References ii

Carsten Erb and Christoph Pelger. "Twisting words"? A study of the construction and reconstruction of reliability in financial reporting standard-setting. *Accounting, Organizations and Society*, 40:13–40, January 2015. ISSN 0361-3682. doi: 10.1016/j.aos.2014.11.001. URL https://www.sciencedirect.com/science/article/pii/S0361368214001007.

Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A. Wood. Is it all hype? chatgpt's performance and disruptive potential in the accounting and auditing industries. *Review of Accounting Studies*, 29(3):2318–2349, 2024. doi: 10.1007/s11142-024-09833-9.

FASB. Conceptual framework for financial reporting. Technical report, Financial Accounting Standards Board, 2024. URL https://storage.fasb.org/Conceptual%20Framework%20for%20Financial%20Reporting%20%28September%202024%29.pdf. September 2024 compilation, Chapters 1–8.

FASB. Rules of procedure: Amended and restated through february 12, 2025, February 2025. URL https://storage.fasb.org/Rules%20of%20Procedure-Feb%202025.pdf. Amended and restated through February 12, 2025.

James R. Frederickson and Jeffrey S. Miller. The effects of pro forma earnings disclosures on analysts' and nonprofessional investors' equity valuation judgments. *The Accounting Review*, 79(3):667–686, 2004. doi: 10.2308/accr.2004.79.3.667. URL https://doi.org/10.2308/accr.2004.79.3.667.

References iii

Lewis R. Goldberg. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 73(6):422–432, 1970. ISSN 1939-1455. doi: 10.1037/h0029230.

E. E. Griffith, J. S. Hammersley, K. Kadous, and D. Young. Auditor mindsets and audits of complex estimates. *Journal of Accounting Research*, 53(1):49–77, 2015. doi: 10.1111/1475-679X.12066.

D. Eric Hirst, Patrick E. Hopkins, and James M. Wahlen. Fair Values, Income Measurement, and Bank Analysts' Risk and Valuation Judgments. *The Accounting Review*, 79(2):453–472, 2004. ISSN 0001-4826. URL https://www.jstor.org/stable/3203252. Publisher: American Accounting Association.

Eunbin Kim, Hyeonseok Kang, and Sangkeun Jung. LTEM: LLM-friendly table-enhanced markdown for table understanding and generation. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 43304–43324, 2026. doi: 10.18653/2026.emnlp-main.1900. URL https://aclanthology.org/2026.emnlp-main.1900/.

Lisa Koonce, Marlys Gascho Lipe, and Mary Lea McAnally. Judging the Risk of Financial Instruments: Problems and Potential Remedies. *The Accounting Review*, 80(3):871–895, 2005. ISSN 0001-4826. URL https://www.jstor.org/stable/4093180. Publisher: American Accounting Association.

References iv

Lisa Koonce, Cassie Mongold, Laura Quaid, and Brian J. White. Experimental research on standard-setting issues in financial reporting. *Accounting, Organizations and Society*, 112:101509, June 2024. ISSN 0361-3682. doi: 10.1016/j.aos.2023.101509. URL https://www.sciencedirect.com/science/article/pii/S0361368223000806.

Asher Koriat. Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4):945–959, 2008. doi: 10.1037/0278-7393.34.4.945.

Asher Koriat. The self-consistency model of subjective confidence. *Psychological Review*, 119(1): 80–113, 2012. doi: 10.1037/a0025648.

Anthony Le. Accounting rules and the labor market for accountants. *Journal of Accounting Research*, 2026. doi: 10.1111/1475-679X.70043. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4666335.

Bradford Levy. Caution ahead: Numerical reasoning and look-ahead bias in AI models. *Journal of Accounting Research*, 64(3):1139–1188, 2026. doi: 10.1111/1475-679x.70058. URL https://doi.org/10.1111/1475-679x.70058.

References v

Robert Libby and Joan Luft. Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment. *Accounting, Organizations and Society*, 18(5):425–450, 1993. doi: 10.1016/0361-3682(93)90040-D. URL https://doi.org/10.1016/0361-3682(93)90040-D.

Christoph Pelger. Practices of standard-setting – An analysis of the IASB's and FASB's process of identifying the objective of financial reporting. *Accounting, Organizations and Society*, 50:51–73, April 2016. ISSN 0361-3682. doi: 10.1016/j.aos.2015.10.001. URL https://www.sciencedirect.com/science/article/pii/S0361368215000896.