Recovering Accounting Judgment with AI — Textual and Non-Textual Experiments *

Hunter Ng[a]

[a] *Baruch College, City University of New York, United States*

November 3, 2026

Abstract

Financial-statement presentation affects human judgment even when the underlying information remains unchanged. I examine whether AI reproduces these responses using accounting experiments and 23 participant-profile arms. The source catalogue contains 401 accepted records and 395 distinct findings across 64 papers, with outcomes evaluated separately. The unprofiled model recovers 39.8% of 236 eligible human directions. Within presentation-based experiments, GPT-6 recovers 64.7% of information-changing findings and 31.0% of presentation-only findings. Correct factual answers do not ensure human recovery. On 45 source-supported economic comparisons, agreement rises from 33.3% under GPT-4.1 to 62.2% under GPT-6.1 Sol, while human recovery falls from 53.3% to 31.1%. The knowledge-minus-neutral change is 7.00 percentage points in source-compatible human recovery and -8.11 points in controlled pattern agreement. Combining the four risk-profile answers raises recovery by 11.83 percentage points relative to neutral instructions. Dated shares add 1.61 points relative to the equal mixture. Most of the observed recovery gain comes from combining risk-profile responses, with a smaller increment from replacing equal shares with dated shares. Condition means and transfer tasks show why direction recovery requires separate evaluation. These findings distinguish an appropriate accounting decision from reproduction of the intended users' judgments.

Keywords. Artificial intelligence, synthetic participants, accounting standard setting, FASB, behavioral accounting, experimental accounting, large language models

JEL Classification. M41, M48, C90, C91, C88, O33

*I am grateful to Edward Li, Hagit Levy, Lin Peng and two anonymous reviewers from the 2026 ABO Research Conference for their comments.

1 Introduction

Accounting research on artificial intelligence (AI) largely examines models' capabilities in different applications. Studies assess whether models extract information from financial reports, analyze disclosures, or answer accounting questions (Blankespoor et al., 2026, Huang et al., 2023). Financial benchmarks similarly test retrieval, calculation, and numerical reasoning (Chen et al., 2021, Zhu et al., 2021). As these capabilities improve, conclusions about what a model can do can change. In this paper, I examine judgment. The question is whether a model can recover how financial-statement users respond to an accounting choice and whether specifying those users' knowledge and characteristics improves recovery.

Judgment involves evaluating evidence and assigning weight to it in reaching an assessment or decision. An accounting user may assess whether earnings are sustainable, whether an estimate is reliable, or whether an investment is attractive. These assessments depend on the person's training, knowledge, objectives, and attitudes toward uncertainty (Bonner and Lewis, 1990, Libby and Luft, 1993). For a factual question, an identified correct answer provides a benchmark for accuracy. For a judgment, the relevant benchmark may instead be how a particular group responds. A model can calculate an amount correctly while giving its classification or presentation different weight from the human users it is intended to represent.

This distinction applies strongly to accounting standard setting. The FASB develops reporting requirements through research, consultation, and redeliberation, creating demand for timely evidence about users' responses before a rule is adopted (FASB, 2025). Published standard-setting experiments provide a useful setting for two reasons. First, their treatments and reported human findings provide publicly verifiable benchmarks. Second, they examine judgments about accounting policy, including judgements in recognition, measurement and presentation (Koonce et al., 2024). Using the Koonce et al. catalogue, I use all studies published between 1995 and 2025 and ending up with 401 records representing 395 distinct findings across 64 papers.

The users making these judgments differ in their knowledge and characteristics. Professional analysts and nonprofessional investors respond differently to pro forma earnings disclosures (Frederickson and Miller, 2004), and the usefulness of graphs and tables depends on accounting knowledge (Cardinaels, 2008). The FASB's conceptual framework also recognizes differences in users' information needs (FASB, 2021). Its qualitative characteristics can place competing demands on reporting choices, requiring judgment in how they are balanced.[1]

In my first set of tests, I direct models to recover the published human responses without the additional participant profiles. GPT-6 recovers 39.8% of 236 eligible human directions. Recovery is similar for textual and presentation-based findings, at 40.5% and 43.4%, respectively. The distinction between information and presentation changes is more informative. Within the same 46 presentation-based findings from 14 papers, GPT-6 recovers 64.7% of information changes and 31.0% of presentation changes holding information constant. All five tested configurations show the same ordering. These findings motivate examining

[1]The framework identifies relevance and faithful representation as fundamental qualities and comparability, verifiability, timeliness and understandability as enhancing qualities. Improving one may weaken another, requiring judgment in balancing users' needs (FASB, 2024, Chap. 3, paras. QC5, QC19, and QC34).

how the decision maker uses the accounting information, beyond whether the case contains text or a table.

I next ask whether the model correctly processes the facts needed for the judgment. GPT-6 correctly answers all 668 valid questions in the common accounting-information question bank. It nevertheless recovers only 37.9% of human directions in the subset where all available checks pass. The model can identify and combine the accounting information while responding differently to the reporting choice. This motivates the main tests of whether describing the participant's knowledge and characteristics brings the model's judgment closer to the reported human response.

In my second and main set of tests, I provide more detailed descriptions of the participant. I run the same accounting cases under 23 versions of the participant instructions. Two use the original and neutral prompts. The other 21 describe accounting knowledge, attitudes toward uncertainty, evidence thresholds, horizons, familiarity, attention, objectives and investment risk. The descriptions specify how a participant approaches a decision without telling the model which treatment response to produce. I hold the accounting case, source-defined role, questions, and scoring rules fixed. On 200 matched comparisons from 44 papers, recovery rises from 41.5% under neutral instructions to 48.5% with the knowledge description, a gain of 7.00 percentage points. Higher evidence thresholds combined with knowledge produce a gain of 12.06 points relative to neutral instructions. The profile tests show that specifying the decision maker changes the accounting judgments recovered.

In my last set of main tests, I examine the composition of the audience. I use the four mutually exclusive investment-risk categories in the Survey of Consumer Finances (SCF) to define participant profiles. I combine their responses using equal shares and historical shares from the closest preceding survey wave. The survey shares describe an external investment audience at the material time. They provide dated population weights for the profiles, rather than observations of the original laboratory participants. The comparisons hold the four profile responses fixed while changing their shares, separating the effect of combining profiles from the effect of audience composition.

On the same 186 findings from 39 papers, recovery is 41.9% under neutral instructions, 53.8% when the four risk profiles receive equal weight, and 55.4% with historical shares. Combining the profiles raises recovery by 11.83 percentage points. Replacing equal shares with historical shares adds 1.61 points. A fixed 2022 distribution gives recovery of 56.5%. Most of the observed gain therefore comes from combining different profile responses. Changing the audience weights produces a further, smaller change in the judgments recovered.

Collectively, my results identify the descriptions of the people represented and the composition of their audience as two components of reproducing judgements of accounting information. Accounting knowledge and participant characteristics affect the model's response to the same case. Population shares determine how those responses enter a group judgment. These components matter separately from the ability to extract information or calculate an answer. As models improve, reproducing these judgements for standard setting still requires an account of whose judgments the model is intended to recover.

This distinction also connects the paper to research on numerical reasoning. Levy (2026) shows that access to code execution improves performance on accounting calculations. Research on arithmetic heuristics examines how models produce numerical answers internally (Nikankin et al., 2025). Improvements

in computation can change the remedy needed for a capability problem. My tests address how to specify the knowledge and characteristics of a decision maker and combine responses for an intended audience. A streamlined procedure linking these inputs to the accounting task makes that requirement testable as models change.

I provide a reusable interface that supplies participant descriptions and compatible audience weights for an accounting task. I then test the participant descriptions on new accounting tasks. Across 12 directional comparisons in two additional studies, recovery is 16.7% under neutral instructions, 25.0% with the knowledge description and 33.3% with deliberate attention. These applications use participant descriptions without audience weights. They show how the procedure can be applied to new accounting tasks and how describing the decision maker can improve recovery.

In all tests, I keep the reported human findings and answer keys out of participant prompts. Robustness checks examine alternative encodings, accounting-attention instructions, table-cell indexing, context length, information position, entity masking, and response elicitation. Separate recognition and published-result recall questions assess familiarity with the source studies. I also compare information and presentation changes using source controls and within-paper and within-experiment specifications. The estimated information-change association remains positive across the reported specifications. These checks address how the accounting case and the model configuration affect recovery.

I separately examine whether agreement with humans also represents an economically appropriate decision. On the same 45 comparisons with source-supported economic or accounting-rule benchmarks, agreement rises from 33.3% under GPT-4.1 to 62.2% under GPT-6.1 Sol, while human recovery falls from 53.3% to 31.1%. I inspect the answers behind zero contrasts and evaluate agreement with published judgment levels separately from treatment directions. The knowledge description raises direction recovery but also raises estimated judgment-level error. The historical risk mixture lowers estimated judgment-level error. These tests preserve the distinction between making an appropriate accounting decision and representing the intended users' judgments.

My study makes three contributions. First, I systematically examine model performance on judgment tasks relevant to accounting standard setting. Research on AI capability establishes whether a model can obtain accounting information and answer questions about it. My tests examine whether it recovers how people use that information. The published experiments provide benchmarks tied to defined accounting choices and participant populations. This brings the intended financial-statement user into the evaluation of AI-generated participants.

Second, I link accounting judgment to the recovery of a group's responses. Knowledge, preferences, attention, and objectives affect how users evaluate accounting evidence. My design varies these descriptions while holding the accounting case fixed and then distinguishes combining profile responses from changing their population shares. The resulting gains connect silicon sampling to both participant descriptions and audience composition. They also explain why improved factual performance alone need not produce closer agreement with human judgments.

Third, I provide a reusable procedure linking experimental instruments, participant profiles, audience weights, and scoring rules. The locally tested interface supplies the profiles and compatible weights while

keeping published human outcomes out of participant prompts. Researchers can update these inputs as populations and models change and evaluate recovery against the accounting judgment of interest. Human evidence remains the benchmark for assessing whether the resulting responses represent the users a reporting requirement is intended to serve.

2 Institutional Details

2.1 Financial-Reporting Presentation and Human Judgment

Financial reporting communicates information through its arrangement as well as its amounts and wording. Experiments show that statement location, proximity, and format affect how users interpret and weight information. Clear presentation of comprehensive income helps professional analysts detect earnings management (Hirst and Hopkins, 1998), while the statement in which volatility appears affects nonprofessional investors' assessments (Maines and McDaniel, 2000). Placing related items closer together improves cash-flow forecasts (Hodge et al., 2010). Qualitative and quantitative hedge disclosures can also lead investors to weight offsetting exposures differently (He et al., 2019). These findings establish that unchanged economic information can produce different judgments when its presentation changes.

The response depends on the user making the judgment. Knowledge, objectives, and incentives affect how people interpret accounting evidence (Bonner and Lewis, 1990, Libby and Luft, 1993). Pro forma earnings disclosures influence nonprofessional investors' valuations differently from those of professional analysts (Frederickson and Miller, 2004). Accounting knowledge also affects the usefulness of graphs and tables in a cost-accounting task (Cardinaels, 2008). Research therefore needs a reason to expect one participant group to represent another (Elliott et al., 2007). The same requirement applies when a model is asked to represent financial-statement users.

These differences matter for standard setting because reporting requirements serve defined information needs. The FASB develops requirements through research, public deliberation, proposals for comment, and redeliberation before issuing an Accounting Standards Update (FASB, 2025). Experiments inform this process by holding features of the decision setting fixed while varying a reporting choice (Kagel and Roth, 2016, Libby and Lewis, 1982, Smith, 1994). The review by Koonce et al. (2024) organizes accounting experiments around these issues. Their reported treatment effects provide evidence about users' responses rather than a universal key for the correct judgment.

2.2 AI-Generated Participants and Accounting Decisions

Using a computational model to represent judgment does not require it to reproduce every feature of human cognition. Goldberg (1970) shows that a *model of man* can capture systematic judgment patterns and apply them more consistently than the person. Behavioral decision research similarly examines how people combine evidence and how those patterns affect accounting decisions (Einhorn and Hogarth, 1981, Joyce and Libby, 1981). Language models learn from extensive human-generated material about decisions, institutions, and experience. Evidence from surveys and cognitive tasks shows that they reproduce some

human response patterns (Argyle et al., 2023, Binz and Schulz, 2023, MacKenzie et al., 2024). Whether those patterns extend to accounting choices is an empirical question.

AI use in financial analysis gives this question practical consequences. Research examines generative AI in financial reporting and investors' interactions with AI tools (Blankespoor et al., 2026, Croom, 2026). It has also examined model-generated financial-statement analysis and earnings forecasts (Kim et al., 2024). The draft summarized in the appendix was subsequently withdrawn. Evidence about assistance must be assessed separately from evidence about substitution for a user. For example, Croom (2026) finds that greater interactivity raises investors' perceived abilities and willingness to invest despite impairing information processing. A tool can change the user's judgment without accurately representing that judgment on its own.

The form of the accounting case presents a further issue. Converting a statement into text can preserve amounts and labels while changing their proximity or prominence. Research on document and table understanding shows that layout and ordering affect model performance (Sui et al., 2024, Xu et al., 2020). The position of evidence within a long input also affects its use (Bai et al., 2024, Liu et al., 2024). A model may identify the correct amount and statement without giving their placement the importance a human investor assigns to it. I therefore distinguish how information is supplied, whether it is processed correctly, and how it affects the judgment.

2.3 Representation, Processing, and Judgment

I organize the tests around three possible sources of divergence. **Representation differences** arise when reconstruction changes the information or its presentation. **Processing differences** arise when a model fails to identify or combine the information supplied. **Judgment differences** arise when a model answers separate questions about the relevant facts correctly but responds differently to the reporting choice. The distinctions identify questions the evidence can address. They do not identify three mutually exclusive causal mechanisms.

The first question concerns what the treatment changes. An amount can change within a fixed table, while a presentation treatment can move an unchanged amount between a statement and a note. A table therefore does not by itself establish that the experiment tests presentation holding information constant. If AI responds to economic information without sharing human sensitivity to its presentation, it should recover information-changing findings more often within presentation-based experiments. This gives the first hypothesis.

H1. *Within presentation-based experiments, AI recovers human treatment effects more often for information changes than for presentation changes holding information constant.*

The second question concerns whether correct processing is sufficient for human recovery. Limited attention affects the weight investors assign to prominent or readily accessible information (Blankespoor et al., 2020, Hirshleifer and Teoh, 2003). A model can identify an amount and its classification yet assign them different weight. I use separate accounting-information questions to assess processing without directing attention before the focal judgment. This gives the second hypothesis.

H2. *AI and human judgments differ even when AI correctly answers separate questions about the relevant*

accounting information.

Human recovery and decision quality require separate benchmarks. A presentation effect need not be an error, and a subjective rating need not have a uniquely correct level. Where the original source documents a bias or supports an economic or accounting-rule benchmark, however, nonrecovery can accompany either an appropriate decision or a mistake. I inspect the underlying answers and apply these benchmarks only where the source supports them (Heath and Alexander, 2026, Jang et al., 2026). This assessment determines what disagreement with the human finding means for the accounting decision.

3 Data and Design

3.1 Primary Experimental Data

My primary data are experimental findings about how financial-reporting choices affect human judgment. I begin with the studies compiled by Koonce et al. (2024) and extend coverage through 2025. The source inventory contains 65 studies, with journal coverage between 1995 and 2025. The accepted coding contains 401 records from 64 papers. The articles, appendices, and available experimental materials provide public sources for the accounting treatments and reported human findings.

I reconstruct the accounting decision from these sources. For each study, I record the setting, participant characteristics, sample sizes, treatments, instructions, accounting amounts and disclosures, judgment questions, response scales, and reported outcomes and tests. Statement records retain classifications, subtotals, and information placement where available. Structured JSON records link each condition and judgment to source references and supporting excerpts. Where the original instrument is unavailable, I reconstruct the task to the extent supported by the source and identify changes to the procedure. Unreported amounts remain missing.

The unit of observation is an experimental comparison. A study that varies recognition versus disclosure and the size of a loss can contribute several effects. I record each effect with its treatment conditions, participant group, outcome, and reported human finding. Main effects, simple effects, and specified interactions remain separate. Comparisons can share participants and conditions, so inference accounts for dependence within source papers.

I manually code the studies and compare this coding with an independent extraction by full GPT-5.4. The extraction has no access to the manual coding or behavioral model responses. I match records by outcome, participant group, experimental phase, treatment conditions, and effect type. A separate GPT-5.4 context checks the proposed matches. This procedure yields 401 accepted records. Six records repeat a finding under another hypothesis, leaving 395 distinct findings from 64 papers. Acceptance concerns correspondence between the coded findings. Task reconstruction separately addresses the information, questions, and procedures needed to administer each comparison.

I complete source-supported task definitions for 95 comparisons and repair material or scoring gaps in 33 other comparisons. The full frame retains all 401 records. Each task links its accounting case, questions, response format, and scoring rule to the source. I distinguish source-compatible treatment comparisons from controlled tasks that change a memory, acquisition, market, or stimulus procedure. The source-

compatible nonzero directional frame contains 208 distinct comparisons from 47 papers. The controlled pattern-agreement frame contains 45 comparisons. Unsigned, joint, distributional, associative, and coding-dependent findings retain their respective targets and component responses. Table 1 reports these source classes and the matched samples used for each outcome.

Human-direction recovery requires an identified nonzero treatment contrast and responses to its required conditions. Numerical scoring definitions cover 207 of the 208 source-compatible targets and 44 of the 45 controlled targets. The remaining target in each class requires source-rubric adjudication of a written response. A measured employer category or a knowledge subgroup is not replaced by an assigned participant description. A null human finding or inactive initial-phase treatment is not treated as a direction to recover. Controlled tasks are evaluated against their stated direction patterns and reported separately. I also compare model judgments with source-exact human condition means wherever available. This condition-mean frame contains 69 distinct human condition-mean vectors from 20 papers. Its eligibility is separate from the directional frame. Missing or ambiguous model responses remain unavailable rather than being coded as zero treatment effects.

The unprofiled comparisons provide a common accounting benchmark across model configurations and include controlled reconstructed tasks. The profile comparisons report source-compatible targets and procedure adaptations separately. The unprofiled panel contains 236 directional findings from 46 papers, comprising 163 textual, 53 presentation-based, 19 inseparable content and presentation comparisons, and one indeterminate comparison. The broader source-composition figure includes 239 distinct findings before the null, placebo, and subgroup restrictions. Its counts describe this unprofiled subset, while Table 1 describes the full source frame and new profile comparisons.[2]

I classify the treatment that changes in each comparison. Textual treatments change wording, accounting labels, instructions, or explicitly stated information, including amounts when their organization remains fixed. Presentation-based treatments change financial-statement placement, the arrangement of amounts, tables, graphs, or other display features. Inseparable and indeterminate findings enter overall coverage and remain separate in pure-channel comparisons. I further distinguish changes in economic conditions or decision-relevant disclosures from changes in classification, location, aggregation, and emphasis that hold information constant. The classification applies to comparisons rather than entire papers. Authors' expectations and economic or accounting-rule benchmarks are coded from the original sources.

[2]The source inventory includes a dissertation instrument labeled `2023 Anderson`. Its correspondence to the related journal experiment is unverified. I identify its two findings and report a source-correspondence sensitivity that excludes them.

Table 1. 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
Panel B. All source benchmark classes			
Scalar treatment comparisons	234	228	48
Joint or multivariate comparisons	22	22	11
Other scalar comparisons	68	68	27
Reconstructed numeric comparisons	32	32	27
Association comparisons	32	32	14
Distributional comparisons	8	8	4
Comparisons without a signed benchmark	3	3	2
Written-response comparisons	2	2	2

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

This table reports source coverage and analytical sample sizes. The 401 mappings identify 395 distinct findings from 64 papers. Source-eligible nonzero judgments and controlled patterns are separate populations. The combined population measures source-pattern agreement. The unprofiled panel is a separate model-comparison sample. Papers may occur in several source classes. Panel C distinguishes paired analytical samples from complete human condition-mean vectors. The four-arm mixture cohorts use source-eligible nonzero judgments, separately for each endpoint. Paired mean counts use the all-source-class exact-source condition-mean benchmark. Counts identify findings and papers. Normalized MAE additionally requires explicit source scale endpoints. Sources are the published studies, supplements, source-supported task reconstructions and saved model responses.

3.2 Standard-Setting and Additional Data

I collect FASB Accounting Standards Updates (ASUs), issuance dates, and identifiable exposure-draft dates to describe the standard-setting context. Figure 2, Panel (a), reports 214 ASUs issued between 2010 and 2025, averaging approximately 13 annually. Continued issuance creates a recurring need for evidence on how reporting changes affect users. Panel (b) reports the mean months from the first identifiable exposure draft to final issuance for 189 ASUs. This interval captures the elapsed time during which a reporting proposal is developed, consulted on, and finalized. It varies across issuance years. The mean for updates issued in 2025 is 22.1 months.

Panels (c) and (d) describe the experimental benchmark and the reporting changes addressed by ASUs. Panel (c) uses the 239 distinct coded source findings from 46 studies. Panel (d) reports my manual classification of the 214 ASUs by their dominant amendment purposes, using FASB summaries and main provisions (FASB, 2026). Of these, 176 concern textual/content changes, 32 concern visual/presentation changes, and six combine both. Excluding 41 updates primarily concerning effective dates, transition, Codification maintenance, or SEC guidance increases the presentation-related share from 17.8% to 21.4%. The experiment and ASU percentages use different units and coding definitions. They describe the range of reporting choices covered, not proportional representativeness.

I use two accounting experiments from 2026 to assess responses outside the main benchmark. Their initial decision tasks require limited adaptations to the available source materials. Separate accounting-information questions, study-recognition questions, and prompt variants assess processing and robustness. Each supplementary test retains its stated sample and source-defined outcome.

To assess whether the distinction between accuracy and human agreement extends beyond accounting, I reanalyze three public human-judgment benchmarks. This use of complementary evidence follows Bloomfield et al. (2016). Related accounting applications combine surveys with interviews or field experiments (Gassen and Muhn, 2025, Graham et al., 2005). For Hagendorff et al. (2023), I use released human response codes and published model category counts for 150 cognitive-reflection and 50 semantic-illusion tasks. BiasMonkey provides saved responses from nine primary configurations across five survey response biases (Tjuatja et al., 2024). WildFrame provides 5,000 human votes on 1,000 framed texts and saved model sentiment labels (Lior et al., 2026). Each reanalysis uses the source's own tasks and scoring rules, keeping correctness, response to wording, and agreement with humans distinct.

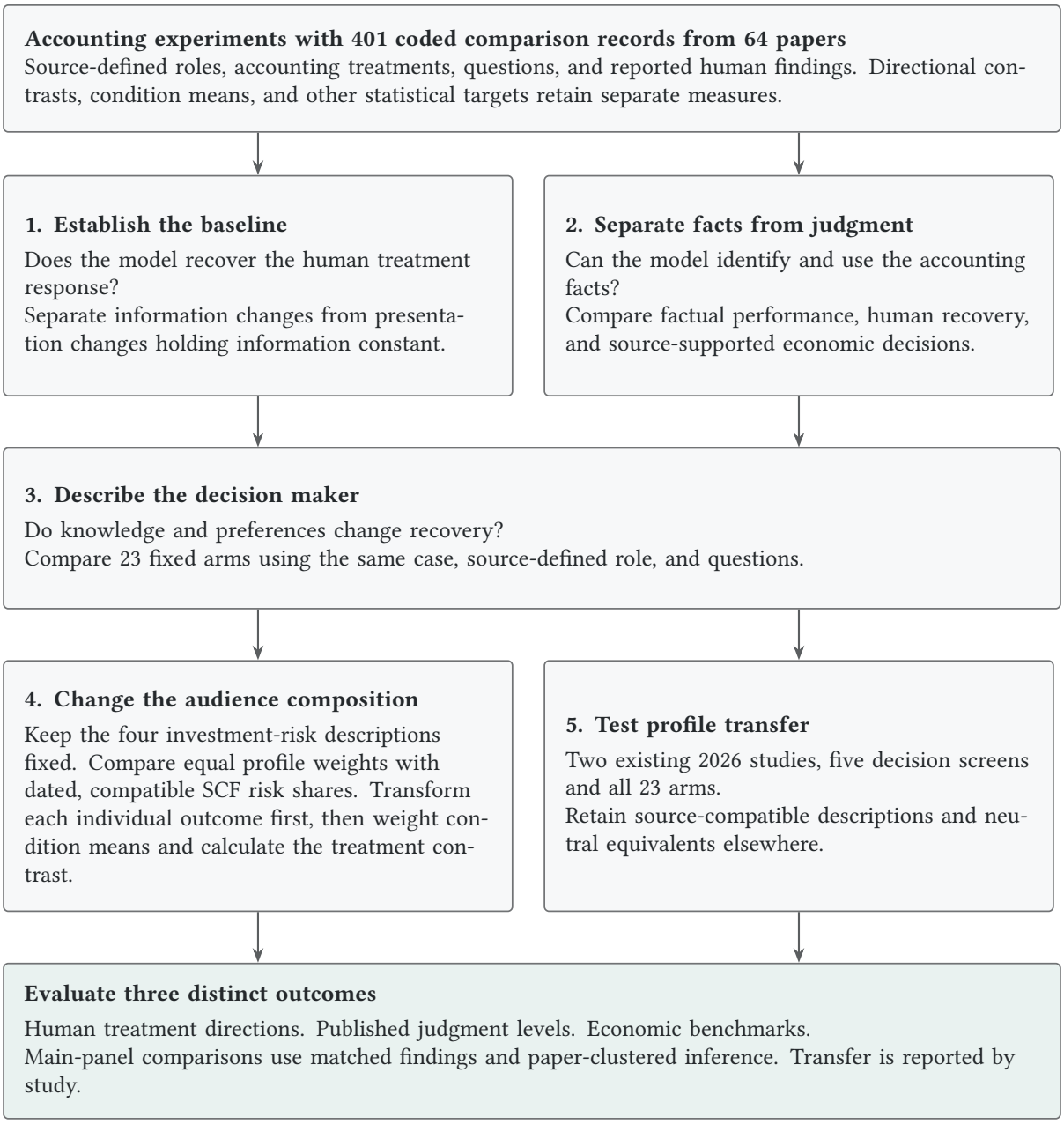
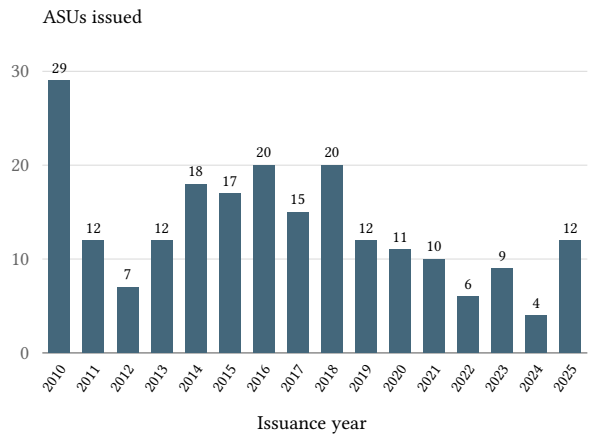
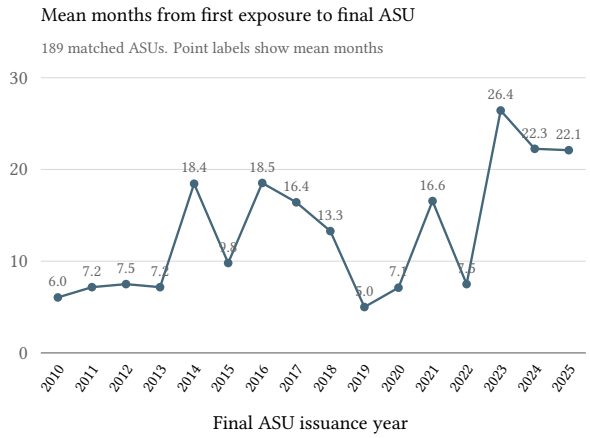


Figure 1. Accounting Judgment and Audience Composition

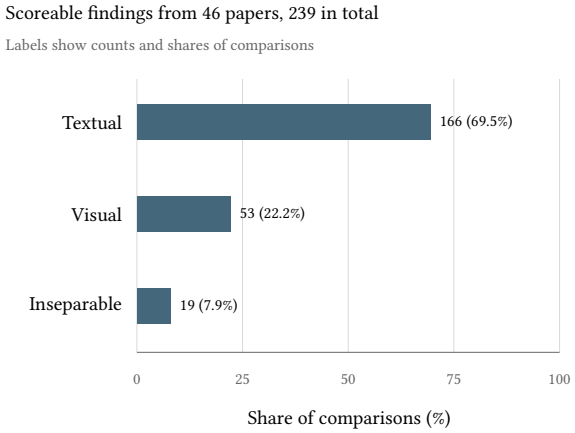
This figure describes the empirical design. The source frame contains 401 accepted comparison records from 64 papers. The first tests compare human-direction recovery with factual performance and source-supported economic decisions. The profile tests hold the accounting case fixed while varying knowledge, preferences, familiarity, attention, objectives, and investment-risk descriptions. The audience tests hold the four risk-profile responses fixed while changing their weights, using equal shares, dated Survey of Consumer Finances (SCF) shares, and fixed 2022 shares. Each individual outcome is calculated before weighting condition means and constructing the treatment contrast. The dated shares describe an external investment audience. The two 2026 applications separately evaluate fixed descriptions on source-adapted initial questions, without audience weights. Directional recovery, judgment-level agreement, and economic agreement are evaluated separately.



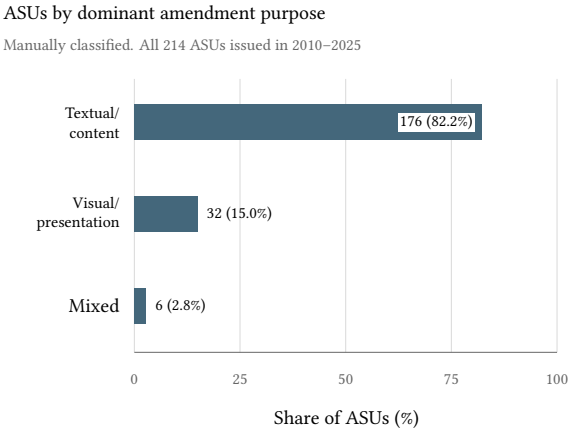
(a) Annual Number of ASUs Issued, 2010–2025



(b) Mean Months from Exposure Draft to Final ASU



(c) Treatment Channels across 46 Source Studies



(d) ASU Amendment Purposes, 2010–2025

Figure 2. Standard-Setting Activity and the Coded Experimental Benchmark

In the figure, panel (a) counts final ASUs by calendar year of issuance, not by effective date. Bar labels report the number issued. The 214 updates include changes to accounting requirements, effective dates, and Codification maintenance. Panel (b) measures elapsed calendar months from the first exposure draft explicitly linked to an amendment to its final ASU issuance. Each point is the equally weighted mean for matched ASUs finalized in the year on the horizontal axis, irrespective of when their exposure drafts appeared. Point labels report mean months. For example, the 2025 mean is 22.1 months across 10 matched updates. The series covers 189 ASUs. 25 without an identified applicable exposure date are omitted, not assigned zero duration. Research before exposure and implementation after issuance are outside this measure. Panel (c) reports 239 distinct scoreable findings across 46 experimental papers within the 50-paper administration panel. The 19 inseparable findings combine content and presentation changes and do not enter pure textual-versus-visual regressions. One indeterminate comparison (0.4%) enters the denominator but is not plotted. Each comparison receives one channel classification. A paper may contribute to several categories. Panel (d) reports our manual classification of all 214 ASUs by dominant amendment purpose. Bar labels in panels (c) and (d) give counts and percentages. Textual/content refers to amendment purpose, not the absence of numerical information. The panels use different units and coding definitions and are not a test of proportional representativeness.

3.3 Empirical Design

Figure 1 summarizes the design. The primary profile model is GPT-6 Luna with temperature zero and reasoning disabled. I obtain one answer per distinct condition and stage, retaining sequential information and the model's own prior answers where required by the instrument. Identical requests reuse the saved

response. I impose no additional output-token cap. The profiles are alternative descriptions of a decision maker, not independent draws of human participants.

I apply the contrast supported by each experiment to its condition-level answers. Main effects, simple effects, interactions, and planned contrasts retain their source-defined forms (Guggenmos et al., 2018). Positive and negative weights in zero-sum contrasts are normalized to one and minus one. Exact evaluation of source-defined fractional weights preserves zero contrasts when equal judgments cancel. Midpoint comparisons use the source-defined reference. Unknown pooling weights produce bounds, with a point score only when the bounds collapse. Recovery equals one when an identified model contrast has the same sign as the nonzero human contrast. Zero contrasts and reversals are separate outcomes. Controlled procedure comparisons use pattern agreement as a separate endpoint.

I estimate linear probability regressions to examine recovery by treatment channel and information content. The specifications include the stated comparison and reconstruction controls, with paper or source-experiment fixed effects where applicable. Profile and input-format comparisons match the same finding across conditions and use finding fixed effects. Coefficients are reported in percentage points. Standard errors are clustered by source paper using the stated CR1 finite-sample correction and $t(G-1)$ inference. Equal-paper estimates give each source paper the same total weight. The table descriptions identify each sample, adjustment family, and specification.

I evaluate agreement in judgment levels using exact-source human condition means. Normalized mean absolute error expresses the discrepancy relative to the supported response scale. Like-unit improvement compares errors within a judgment measure before aggregation. I also examine effect-magnitude endpoints where both the human effect and the model contrast are identified. These measures preserve the difference between reproducing a treatment direction and reproducing the level of a valuation or rating.

I assess representation and processing using alternative text formats, history ordering, accounting-attention instructions, and table-cell indexing. Separate accounting-information questions test processing in conversations that do not precede the focal judgment. Source-supported economic benchmarks and the answers behind zero contrasts assess the decisions themselves. Additional checks vary paper weights, source support, measurement scales, study periods, information position, context length, incidental company names, and response elicitation. Separate questions assess familiarity with published findings.

I grade separate accounting-information questions using complete responses and use the same eligibility rules for both models. Questions with ambiguous referents are omitted from factual scoring. The model comparison holds reconstructed tasks and scoring rules fixed. GPT-4.1 and GPT-4.1 mini use temperature zero without a reasoning parameter. GPT-5 uses minimal reasoning with temperature omitted. GPT-6.1 Sol uses high reasoning with temperature omitted. The common unprofiled sample contains 229 findings from 45 papers under all five configurations. Smaller encoding and table-indexing comparisons also include high-reasoning GPT-6, non-reasoning GPT-6 Sol, and GPT-5.4 mini. GPT-6 Sol and GPT-6.1 Sol are distinct configurations. Each supplementary comparison retains its own matched sample and stated multiplicity adjustment.

Using source-defined eligibility, I bound profile recovery differences by allowing each unknown binary outcome to equal zero or one while retaining every known arm value, including findings with only one

observed arm.

3.4 Participant Profiles and Audience Composition

The profile design asks whether describing the decision maker improves recovery beyond supplying the accounting case. The design uses 23 fixed arms across all 64 papers and 401 accepted records. The primary model is GPT-6 Luna with reasoning disabled, temperature zero, and no additional output-token cap. I distinguish the original baseline prompt, B, from a fresh neutral prompt, N, that uses the same instruction format as the new profiles. Verified identical completed B requests can reuse historical responses. The main comparisons use contemporaneous N responses matched by model configuration, case, and stage. A B–N comparison combines changes in instruction format and collection time and cannot isolate either one.

In addition to B and N, the arms vary assigned accounting knowledge, cautious versus open attitudes toward uncertainty, with and without added knowledge, higher versus ordinary evidence thresholds, with and without added knowledge, near versus long decision horizons, routine versus extensive familiarity, deliberate versus efficient attention, and stewardship versus valuation objectives. Four further arms use the Survey of Consumer Finances investment-risk categories. These dimensions distinguish what a decision maker knows, how the decision maker evaluates evidence, and what the decision maker seeks to achieve. They do not prescribe the treatment direction to be recovered.

I hold the source-defined role, treatments, questions, response scales, incentives, and available information fixed across applicable profiles. Assigned knowledge describes expertise rather than teaching new accounting content, and it is not a measure of the historical participants' knowledge. Source-defined knowledge or trait factors remain part of the original experiment. I identify explicit mismatches rather than treating student or nonprofessional status as evidence of low knowledge. Horizon, objective, and risk instructions apply only where the task permits discretion. When they do not apply, an instruction can be equivalent to N rather than a distinct substantive condition.

The profile comparisons use the same eligible findings to assess whether a participant description changes recovery, zero contrasts, reversals, and source-supported economic-benchmark agreement. The specified analysis uses finding fixed effects and paper-clustered inference, with equal-paper sensitivity checks and Holm adjustments across the specified profile-family comparisons. Directional findings and other source-defined outcomes remain separate. Profile definitions and applicability are based on source characteristics rather than model recovery.

Audience weighting asks whether changing the profile mixture changes recovery of the reported human responses. The audience register records original recruitment groups, experiment-level source evidence, and collection dates where supported. The register preserves an unknown collection year when the source does not report it. For external audience calibration, a publication-year anchor is separately identified as a period proxy. I use independently dated external survey shares where the population and measured characteristic are compatible. The equal mixture assigns a quarter of the weight to each investment-risk category. The 23 arms are alternative instructions that can overlap. The four survey risk categories define the population mixture.

For the four investment-risk arms, I use public microdata from the Federal Reserve's Survey of Consumer Finances for 1995, 1998, 2001, 2004, 2007, 2010, 2013, 2016, 2019, and 2022. The survey's four reported levels of willingness to take financial risk define the categories. The weighting procedure divides each survey weight by five when pooling the five imputations and distinguishes primary economic units with directly held stocks from broader age and college-education groups. Review of source dates, publication timing, and country compatibility determines which external audience a set of weights can describe. These historical shares represent transfer to an external audience. They do not reproduce the original professional, MBA, or other recruited sample merely because the survey year is close.

The composition tests hold the four risk-profile responses fixed while changing their shares. I compare the equal mixture, the closest preceding SCF wave, and a fixed 2022 distribution on the same eligible findings. Comparisons with the neutral prompt measure the combined effect of using the risk descriptions and combining their responses. Comparisons among the mixtures measure the incremental effect of their weights. For outcomes such as forecast error, I calculate each profile's error before forming the weighted condition mean. I then calculate the accounting treatment contrast and score its agreement with the reported human direction.

The implementation makes instruments, participant descriptions, and scoring rules reusable while withholding human findings and answer keys from participant prompts. A local read-only interface returns documented profiles and compatible audience weights for an accounting task. It does not submit model requests or disclose the human findings. The two 2026 studies supply the separate accounting tasks for transfer assessment. Their existing adaptations and completed responses are retained. I apply all 23 arms to the three Le conditions and the two Cai conditions using fixed descriptions, prompts, and scoring rules. These applications retain the initial career questions and first valuation tasks available in the source materials. The interface has been tested locally and has not been publicly deployed.

4 Findings

I examine whether AI responds to reporting choices as the human participants do. I first distinguish changes in accounting information from changes in its presentation. I then assess factual performance and the decisions behind agreement or disagreement with humans. The profile tests ask whether specifying the decision maker changes recovery. The audience tests separate combining risk-profile responses from weighting that mixture to dated external shares. These comparisons evaluate different requirements for representing financial-statement users.

4.1 Accounting Information and Financial-Statement Presentation

I begin by asking whether the broad form of the materials predicts recovery. Table 2 reports the unprofiled GPT-6 results. The model recovers 94 of 236 eligible human directions (39.8%). Recovery is 66/163 (40.5%) for textual findings and 23/53 (43.4%) for presentation-based findings. It gives 92 zero contrasts and 50 reversals. Recovery of accounting judgments therefore extends to both textual and presentation-based experiments.

The broad channel leaves open what the treatment changes. I distinguish information changes from classification, aggregation, location, or emphasis changes that leave the underlying information available. An amount can change within a fixed table. An unchanged amount can also move between a statement and a note. In the selected comparisons from Bonner et al. (2014), participants evaluate a single gain of $70,000 or a gain of $1,000,000 alongside a loss of $930,000. By contrast, Experiment 3 of Koonce et al. (2005) adds potential-gain information to a disclosure that covered losses. The accounting treatment, rather than the presence of a table, determines whether information changes.

I hold the reporting channel fixed using 46 presentation-based findings from 14 papers. These contain 17 information changes and 29 presentation changes holding information constant. GPT-6 recovers 64.7% of information changes and 31.0% of presentation-only changes. All five configurations show the same ordering. Table 2 also reports the adjusted content associations and their stated controls. The association remains positive across the reported specifications. This provides evidence that responding to new accounting information and reproducing human sensitivity to its presentation are different tasks.

Table 2. Baseline Recovery of Accounting Judgments

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

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

Panel D. Information changes minus presentation changes holding information constant

Specification	(1) Unadjusted	(2) Channel and source controls	(3) Within paper	(4) Within source experiment
Information change	16.90**	12.46	5.66	15.62
	(8.23)	(10.09)	(13.45)	(17.62)
Holm-adjusted p-value	0.190	0.676	0.881	0.881
Findings	163	148	41	23
Source papers	36	35	6	4
Channel controls	–	X	–	–
Source and reconstruction controls	–	X	–	–
Paper FE	–	–	X	–
Source-experiment FE	–	–	–	X

Panel E. Same-sample channel and information-content regressions

	(1)	(2)	(3)
Information change	19.98**	18.68*	13.87
	(8.83)	(9.78)	(10.40)
Presentation-based treatment	8.86	8.88	9.34
	(10.47)	(10.25)	(9.02)
Additional source-finding indicator	–	X	X
Reconstructed or corrected-material indicator	–	–	X
Findings	154	154	154
Papers	34	34	34

This table reports baseline recovery of human directions and associations with accounting information and presentation. The first panels report GPT-6 recovery and channel regressions. The dependent variables in Panel B equal one for recovery of the nonzero human direction in column (1), a zero model contrast in column (2), and a reversed direction in column (3). Panel A reports counts and recovery percentages for 236 eligible findings from 46 papers. Panel B uses 216 textual or presentation-based findings from 44 papers. Each regression includes a presentation-based indicator and intercept, with textual findings as the reference and no other controls or FE. Coefficients and standard errors are in percentage points. Sources are the published human findings, reconstructed instruments, and condition-level responses. Standard errors are paper-clustered CR1, with $t(43)$ inference. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$. Panel C reports recovery for information changes and presentation changes holding information constant within presentation-based materials. The dependent variable equals one when the model contrast matches the reported nonzero human direction. Columns identify five configurations using the same 46 findings from 14 papers, comprising 17 information changes from six papers and 29 presentation changes from 11 papers. Paper counts overlap. Each regression includes an information-change indicator and intercept. Presentation changes form the reference category. Coefficients and standard errors are in percentage points. Sources are published human findings, coded treatments, reconstructed instruments, and model responses. Standard errors are paper-clustered CR1, with $t(13)$ inference. Stars use nominal p-values. The broader specification-family Holm adjustment covers GPT-4.1, mini, GPT-6, and Sol. GPT-5 statistics use nominal inference. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$. Panels D and E report associations between information content, treatment channel, and GPT-6 recovery. The dependent variable equals one when the model recovers the reported nonzero human direction. Coefficients and standard errors are in percentage points. Panel D compares information changes with presentation changes holding information constant in four specifications with 163, 148, 41, and 23 findings from 36, 35, six, and four papers. Column (2) includes channel and reconstruction-tier controls. Columns (3) and (4) include Paper FE and Source-experiment FE. Its adjusted p-values use a four-test Holm family. Panel E holds the sample fixed at 154 pure-channel findings from 34 papers and includes both treatment indicators. Columns (2) and (3) add the additional source-finding indicator. Column (3) adds the reconstructed or corrected-material indicator. The reference category is textual presentation changes holding information constant. Sources are published findings, source coding, reconstructed instruments, and condition-level responses. These supplementary specifications use the directional administration panel before subgroup eligibility. Standard errors are paper-clustered CR1, with $t(G-1)$ inference. Stars use nominal p-values. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

4.2 Accounting Facts and the Judgment They Support

One explanation for nonrecovery is that the model does not identify or combine the information needed for the decision. I assess this explanation using separate questions about amounts, classifications, and relationships. These questions are asked in separate conversations, so they do not direct attention before the focal judgment. Both GPT-6 and GPT-6.1 Sol correctly answer all 668 valid questions in the common question bank. Twenty-five questions are excluded because they do not identify which of several displayed securities is being assessed. Table 3 reports this factual performance alongside human recovery. The model usually identifies the tested accounting facts correctly.

I next ask whether correct factual answers accompany the human response to the reporting choice. Using all source-matched accounting checks, GPT-6 recovers 74/195 human directions (37.9%) where all available checks pass. Requiring amount, classification, and integration checks at each referenced judgment stage leaves recovery of 48/137 (35.0%). These diagnostic subsets use the directional administration

panel before the subgroup restriction. This provides evidence that accurate answers about accounting facts do not ensure that a model reproduces how users respond to reporting choices.

Table 3. 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) GPT-6	(2)	(3) GPT-6.1 Sol	(4)
	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)

This table reports accounting-information accuracy and recovery of the corresponding human judgments. Panel A reports correct responses to separate accounting-information questions. Entries are counts and percentages of the 668 valid questions from a common bank of 693 questions concerning 39 task instruments from 38 papers. Rows identify amount, classification, and integration questions. Columns compare GPT-6 and GPT-6.1 Sol. Both models correctly answer all 668 valid questions. The same 25 questions are excluded for both models because the question does not identify which displayed security is being assessed. Percentages divide correct answers by valid questions. Both models receive the same question bodies in separate conversations, without human results or answer keys. Grading uses complete responses, accepting equivalent stated units, signed amounts when the question asks for magnitude, and conventional rounding. Sources are the accounting instruments, source-supported answer keys, and saved model responses. Panel B reports human recovery conditional on correct accounting-information responses. The dependent variable in the channel regressions equals one for recovery of the nonzero human direction. Columns (1) and (3) require all available valid checks to pass. Columns (2) and (4) additionally require amount, classification, and integration checks at each referenced judgment stage. GPT-6 uses the 237-finding directional administration panel. Sol uses its 236 available judgment pairs before subgroup eligibility. The model-specific samples contain 195, 137, 125, and 108 findings from 45, 37, 35, and 30 papers. Each model uses its own source-matched factual responses. These samples combine the common question bank with other available source-matched accounting checks. The channel regressions include a presentation-based indicator and intercept, with textual findings as the reference. Coefficients and standard errors are percentage points. Missing checks remain ungraded. Sources are the accounting instruments, approved keys, and condition-level responses. Standard errors are paper-clustered CR1, with $t(G-1)$ inference. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

4.3 Economic Decisions and Recovery of Human Responses

Disagreement with humans does not by itself establish that an accounting decision is wrong. I compare recovery with economic or accounting-rule benchmarks supported by the original sources. Table 4 holds the same 45 comparisons from 13 papers fixed across five configurations. Economic-benchmark agreement is 33.3% for GPT-4.1, 28.9% for GPT-4.1 mini, 44.4% for GPT-5, 53.3% for GPT-6, and 62.2% for Sol. Human-direction recovery on those cases is 53.3%, 60.0%, 42.2%, 28.9%, and 31.1%, respectively. Greater agreement

with the source-supported economics can accompany lower agreement with the human treatment effect.

A zero contrast also leaves open what decisions the model made. I examine the constituent answers to distinguish correct actions from repeated mistakes and ratings without an identified correct level. Twenty-five model-comparisons from two papers show correct economic actions in every relevant condition. These involve efficient hedging and machinery replacement. Four GPT-4.1 comparisons instead repeat the same incorrect keep decision. Other cases use subjective ratings for which the source does not identify a uniquely correct numerical level. The same zero contrast can therefore represent different accounting decisions.

The machinery task illustrates why the answers matter. GPT-4.1 keeps the machinery in both conditions, referring to the disposal loss and remaining asset value. GPT-5 and Sol choose efficient replacement in both. Each model produces a zero contrast, but the economic choices differ. A model can make an appropriate replacement decision without reproducing the human depreciation effect.

Table 4. Human Responses and Economic Benchmarks

Panel A. Human findings and the authors' directional expectations

	(1) Textual	(2) Presentation-based	(3) Other	(4) All
Findings reviewed	164	53	20	237
Codable directional expectations	134	36	16	186
Human finding in expected direction	128	33	16	177
Agreement (%)	95.5	91.7	100.0	95.2
Null predictions	10	3	0	13
Other or unresolved expectations	20	14	4	38

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

Panel C. Answers Underlying Zero Treatment Contrasts

	(1) GPT-4.1	(2) GPT-4.1 mini	(3) GPT-6	(4) GPT-6.1 Sol
Recorded benchmark-consistent zeros reviewed	16	13	23	24
Correct economic action in relevant conditions	4	4	7	10
Same wrong substantive choice	4	0	0	0
Missing expressed economic response	0	0	2	0
Limited direction or rule correctness	2	0	3	3
Correct answer level not identified	5	9	11	10
Conditional endpoint ineligible or unverified	1	0	0	1
Diagnostic of different answers collapsed to a binary zero				
Source-defined binary collapse	3	0	1	0

This table reports agreement with human expectations and economic benchmarks, together with the answers underlying zero contrasts. The first panels report human agreement with authors' predictions and model agreement with economic benchmarks. Panel A classifies 237 directional human findings by channel and expected direction. Agreement is a count or percentage of findings with a directional expectation. Panel B compares the same 45 findings from 13 papers across five configurations, with 28 findings from nine papers in the higher-confidence subset. Equal-paper rows average within-paper agreement and recovery rates across the 13 papers. The other rows weight findings equally. The human-recovery rows use the same 45 findings as the benchmark rows. An economic benchmark is a source-supported direction, ordering, invariance condition, or accounting rule. It is distinct from an author's behavioral prediction and from correctness of every numerical answer. The benchmark panel excludes operational reference conditions, unresolved economic targets, and the ineligible subgroup endpoint. Sources are the original hypotheses, reported human findings, coded economic benchmarks, and condition-level model responses. Panel C reports the economic content of answers underlying 76 zero model-comparisons across four configurations. The sample covers 33 distinct findings from 11 papers and 174 condition-level answer components. Counts classify correct actions in all relevant conditions, repeated wrong choices, missing economic responses, limited correctness, ratings without an identified correct level, and unresolved subgroup eligibility. The separate diagnostic row identifies different answers mapped to the same source-defined binary score and overlaps the classifications above. Sources are the original accounting tasks, source-supported economic benchmarks, and condition-level model responses.

4.4 Knowledge and the Description of the Decision Maker

The preceding evidence leaves a question about the participant rather than the accounting facts. I ask whether assigned accounting knowledge changes recovery relative to the fresh neutral description, N. The source-compatible directional sample contains 208 comparisons from 47 papers. The matched sample is reported separately. Table 5 contains 200 paired findings from 44 papers. Recovery is 41.5% under N and 48.5% under K. The estimated K-minus-N change is 7.00 percentage points. The paper-clustered standard error is 6.76 percentage points. Recovery is higher under the knowledge description on these matched accounting tasks.

I separately examine controlled comparisons with reconstructed or changed procedures. Their endpoint is agreement with the source-defined direction pattern. The paired sample contains 37 comparisons from 28 papers. The estimated K-minus-N change in pattern agreement is -8.11 percentage points. The paper-clustered standard error is 8.11 points. The profile estimates differ across source-compatible accounting decisions and controlled task procedures.

Table 5. 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

This table reports assigned knowledge minus the fresh neutral prompt with equal finding weights. Human-direction agreement and controlled or combined source-pattern agreement are distinct outcomes. The source-correspondence sensitivity excludes source-paper records with unverified correspondence. Controlled patterns separate numeric reconstructions from other procedural or stimulus adaptations. Source N counts findings with a supported point definition. Paired N is the complete comparison sample. Neither is an observed human sample size. Changes and their standard errors are percentage points. X indicates a finding fixed-effect estimator. No additional controls are asserted. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. An em dash denotes unavailable estimation, not a zero effect. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The combined sensitivity retains the source-compatible and controlled comparison classes. Its endpoint is pattern agreement rather than original human-direction recovery. The accounting instrument and procedure determine source compatibility. The combined comparison evaluates how the profile works across these distinct tasks.

Knowledge is one part of a participant description. I also compare uncertainty attitudes, evidence thresholds, horizons, familiarity, attention, objectives, and investment-risk descriptions with N. Table 6 reports all 21 profile-arm comparisons and the two specified knowledge-cross contrasts. Higher evidence thresholds combined with knowledge have an estimated change of 12.06 percentage points, while the no-financial-risk description has a change of -5.50 points. The other profile estimates range from 0.50 to 9.05 points. Each comparison retains its source-defined applicability and matched support. This provides evidence that risk attitudes and evidence thresholds produce different accounting-treatment responses.

The crossed descriptions ask whether knowledge has the same association with recovery under different attitudes or evidence thresholds. The cautious-minus-open difference in the knowledge effect is -0.52 percentage points (standard error 5.47). The higher-minus-ordinary threshold difference is -2.02 points

(standard error 6.88). The underlying role, available information, questions, and response scales remain fixed within each applicable comparison. These tests connect expertise to the particular assessment the participant is asked to make.

Table 6. Participant Profiles and Knowledge Interactions

Assigned description	Direction	Zero	Reversal	N	Papers	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Assigned knowledge						
Strong assigned knowledge	7.00 (6.76)	-6.00 (5.19)	-1.00 (5.36)	200	44	X
Uncertainty preference						
Cautious preference	3.52 (6.46)	-7.04 (6.42)	3.52 (6.20)	199	43	X
Opportunity-open preference	4.50 (5.69)	-6.00 (5.62)	1.50 (4.07)	200	43	X
Cautious + knowledge	1.04 (7.12)	-7.25 (6.89)	6.22 (6.78)	193	43	X
Opportunity-open + knowledge	3.98 (5.46)	-6.97 (6.57)	2.99 (5.56)	201	44	X
Evidence threshold						
Higher evidence threshold	9.05 (6.45)	-5.03 (7.14)	-4.02 (6.01)	199	43	X
Ordinary evidence threshold	3.00 (6.14)	-4.50 (4.87)	1.50 (5.93)	200	44	X
Higher threshold + knowledge	12.06 (5.84)	-8.54 (5.76)	-3.52 (4.98)	199	43	X
Ordinary threshold + knowledge	7.58 (7.10)	-8.08 (5.66)	0.51 (5.73)	198	44	X
Investment horizon						
Nearer investment horizon	4.46 (5.65)	-4.46 (6.46)	0.00 (3.88)	202	44	X
Longer investment horizon	4.98 (4.80)	-4.48 (5.37)	-0.50 (4.67)	201	44	X
Task familiarity						
Routine task familiarity	4.48 (5.88)	-3.98 (5.08)	-0.50 (4.89)	201	44	X
Extensive task familiarity	0.50 (5.12)	-4.50 (5.18)	4.00 (5.42)	200	44	X
Processing attention						
Deliberative attention	4.00 (5.81)	-6.00 (5.04)	2.00 (4.17)	200	44	X
Efficient attention	3.50 (5.15)	-5.50 (5.00)	2.00 (4.73)	200	44	X
Decision objective						
Stewardship objective	2.50 (3.19)	0.50 (3.13)	-3.00 (2.98)	200	43	X
Valuation objective	0.52 (2.88)	-0.52 (3.07)	0.00 (3.17)	192	43	X
Financial risk willingness						
Substantial financial risk	2.50 (5.03)	-2.50 (5.13)	0.00 (2.86)	200	44	X
Above-average financial risk	4.50 (5.73)	-2.00 (5.30)	-2.50 (3.98)	200	44	X
Average financial risk	2.97 (5.33)	0.50 (5.98)	-3.47 (4.94)	202	44	X

Continued on next page

Table 6. Participant Profiles and Knowledge Interactions (continued)

Assigned description	Direction	Zero	Reversal	N	Papers	FE
	(1)	(2)	(3)	(4)	(5)	(6)
No financial risk	-5.50 (5.77)	7.00 (6.32)	-1.50 (4.25)	200	43	X
Planned balanced knowledge interactions						
Knowledge, cautious minus opportunity-open	-0.52 (5.47)	0.00 (4.63)	0.52 (4.09)	193	43	—
Knowledge, higher minus ordinary threshold	-2.02 (6.88)	1.01 (6.39)	1.01 (5.17)	198	43	—

This table reports every current assigned-profile contrast with fresh N and the two planned balanced knowledge interactions on source-eligible nonzero judgments, with equal finding weights. Direction, zero and reversal denote human-direction agreement, an exactly zero model contrast and the opposite human direction. Interactions are the cautious-minus-open and higher-minus-ordinary differences in the knowledge effect. All changes are percentage points. Shared N and paper counts apply to all three endpoints within each row. X denotes the finding fixed-effect estimator. The planned interactions use within-finding linear changes without a separately estimated fixed effect. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. In the FE column, an em dash indicates that no fixed effect is separately estimated. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.5 Combining Risk Profiles and Weighting an Audience

A participant description does not determine the composition of a group. I examine this distinction using the four mutually exclusive SCF investment-risk categories. I apply the source-defined outcome rule to each profile's answer, combine the resulting condition outcomes with equal weights, dated historical shares, and fixed 2022 shares, and then calculate the treatment contrast. The 23 arms overlap in knowledge, attention, objectives, and crossed instructions. They are not 23 mutually exclusive population types. Only the four risk categories define the mixture evaluated here.

Table 7 compares N and the three mixtures on the same 186 findings from 39 papers. Recovery is 78/186 (41.9%) under N, 100/186 (53.8%) under the equal four-risk mixture, 103/186 (55.4%) under historical shares, and 105/186 (56.5%) under fixed 2022 shares. The common cohort makes the mixture comparison about the same accounting choices and human directions.

Table 7. 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
Reversed direction					
Neutral prompt (N)	33	186	39	17.74	18.47
Equal four-risk mixture	39	186	39	20.97	20.52
Historical population mixture	39	186	39	20.97	20.35
Fixed-2022 population mixture	37	186	39	19.89	19.21
Economic agreement					
Neutral prompt (N)	44	68	17	64.71	60.85
Equal four-risk mixture	42	68	17	61.76	58.80
Historical population mixture	41	68	17	60.29	52.91
Fixed-2022 population mixture	40	68	17	58.82	51.93

This table reports paired descriptive outcomes for the neutral prompt (N), the equal four-risk mixture, historical population weights and fixed-2022 weights on source-eligible nonzero judgments. All four arms use identical findings within each endpoint, matching the composition-contrast support. Events and finding denominators are unweighted counts. Equal-finding rates equal events divided by that denominator. Economic agreement uses source-supported accounting or decision benchmarks and its own matched sample. Equal-paper rates average within-paper rates. The equal mixture assigns one-quarter to each risk arm. The fixed-2022 distribution provides the common later-period weighting comparison. Population weights use public Survey of Consumer Finances microdata for US primary economic units with directly held stocks. Accounting data come from the source studies, reconstructed tasks and saved model responses.

I next separate combining profile answers from changing their weights. Table 8 reports all six pairwise comparisons, using both equal-finding and equal-paper weights. The estimated equal-mixture-minus-N change is 11.83 percentage points. The historical-minus-N change is 13.44 points, while the historical-minus-equal increment is 1.61 points. Fixed 2022 shares give a change of 14.52 points relative to N. The paper-clustered standard errors are 5.48 and 5.43 points for the equal and historical mixtures relative to N. The six-comparison Holm-adjusted probabilities are 0.149 and 0.090, respectively. Most of the observed recovery gain comes from combining risk-profile responses, with a smaller increment from replacing equal

shares with dated shares.

The equal mixture reduces zero treatment contrasts from 40.3% under N to 25.3%. Historical shares reduce them to 23.7%. Reversals are 17.7% under N and 21.0% under historical shares. The gain in recovery accompanies a larger reduction in zero responses and a smaller increase in reversals. This provides evidence that combining risk-profile responses mainly changes whether the model responds to the reporting treatment.

I next ask how these mixtures affect economic agreement. On the same 68 comparisons from 17 papers with supported benchmarks, agreement is 64.7% under N, 61.8% under equal risk-profile weights, 60.3% under historical shares, and 58.8% under fixed 2022 shares. Economic agreement moves down while human recovery moves up in their respective matched samples. This provides evidence that representing an audience's reporting responses and giving economically appropriate answers are different objectives.

The historical and fixed-2022 shares describe external investment audiences. They do not identify the risk distribution of an original professional, MBA, or other recruited sample. Fixed 2022 shares also provide a common later-wave sensitivity for studies anchored to earlier periods. These comparisons assess how a stated external composition relates to the model's accounting judgments.

Table 8. Risk-Profile Mixtures and Audience Composition. Panel A. Equal finding weights

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)
Fixed-2022 population mixture minus N	14.52*	-16.67	2.15
	(5.58)	(6.95)	(4.66)
Fixed-2022 population mixture minus Historical population mixture	1.08	0.00	-1.08
	(1.06)	(0.00)	(1.06)
Fixed-2022 population mixture minus Equal four-risk mixture	2.69	-1.61	-1.08
	(2.23)	(1.30)	(2.16)
Identical six-comparison support, separately by endpoint			
Paired findings	186	186	186
Papers	39	39	39
Holm-family size	6	6	6

This table reports all six predefined composition and timing contrasts on source-eligible nonzero judgments. Changes are percentage points. Each endpoint uses the same four-arm findings for all six comparisons. Its finding and paper counts are shown separately. Weights use public Survey of Consumer Finances microdata. Fixed-2022 is a later-wave sensitivity for earlier anchors. Population condition means follow the source individual-outcome transformation. Binary recovery indicators are not mixed to create a population outcome. Each endpoint and weighting retains its fixed Holm family of six, including unavailable comparisons. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. An em dash denotes unavailable estimation, not a zero effect. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 8. Risk-Profile Mixtures and Audience Composition. Panel B. Equal paper weights

Signed comparison	Recovery (pp)	Zero (pp)	Reversal (pp)
	(1)	(2)	(3)
Historical population mixture minus N	13.42*	-15.31*	1.89
	(5.01)	(5.96)	(4.33)
Equal four-risk mixture minus N	11.18*	-13.23	2.06
	(4.76)	(5.99)	(3.93)
Historical population mixture minus Equal four-risk mixture	2.24	-2.08	-0.17
	(2.46)	(1.95)	(1.68)
Fixed-2022 population mixture minus N	14.56*	-15.31*	0.75
	(5.58)	(5.96)	(5.13)
Fixed-2022 population mixture minus Historical population mixture	1.14	0.00	-1.14
	(1.27)	(0.00)	(1.27)
Fixed-2022 population mixture minus Equal four-risk mixture	3.38	-2.08	-1.31
	(3.22)	(1.95)	(2.72)
Identical six-comparison support, separately by endpoint			
Paired findings	186	186	186
Papers	39	39	39
Holm-family size	6	6	6

This table reports all six predefined composition and timing contrasts on source-eligible nonzero judgments. Changes are percentage points. Each endpoint uses the same four-arm findings for all six comparisons. Its finding and paper counts are shown separately. Weights use public Survey of Consumer Finances microdata. Fixed-2022 is a later-wave sensitivity for earlier anchors. Population condition means follow the source individual-outcome transformation. Binary recovery indicators are not mixed to create a population outcome. Each endpoint and weighting retains its fixed Holm family of six, including unavailable comparisons. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. An em dash denotes unavailable estimation, not a zero effect. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

4.6 Agreement in the Level of Accounting Judgments

The preceding tests measure treatment directions. I next ask whether the model gives the same valuations, risk ratings, and other judgment levels as the human participants. I assess levels using the source-exact condition-mean frame across source classes, which contains 69 distinct human condition-mean vectors from 20 papers. Each vector is linked to the exact source version. Table 9 reports the judgment-level endpoints on their own matched samples. K versus N uses 34 findings from 13 papers for the normalized mean absolute error comparison. The estimated change is 0.0209 in normalized error units. The historical-risk-mixture versus N comparison uses 32 findings from 12 papers, with an estimated change of -0.0506. Lower normalized mean absolute error indicates closer agreement in level. Estimated judgment-level error is higher under the knowledge description and lower under the historical risk mixture. These estimates

show that the descriptions associated with closer judgment levels can differ from those associated with closer treatment directions.

Table 9. Agreement with Human Judgment Levels. Panel A. Normalized MAE

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
Higher threshold + knowledge	-0.015 (0.019)	-0.010 (0.019)	34	13
Ordinary threshold + knowledge	-0.001 (0.030)	0.005 (0.026)	34	13
Nearer investment horizon	0.011 (0.029)	0.008 (0.027)	34	13
Longer investment horizon	0.008 (0.020)	0.010 (0.018)	34	13
Routine task familiarity	0.015 (0.025)	0.015 (0.023)	34	13
Extensive task familiarity	-0.007 (0.026)	-0.006 (0.024)	34	13
Deliberative attention	-9.51×10^{-5} (0.023)	-0.001 (0.022)	34	13
Efficient attention	-0.027 (0.041)	-0.027 (0.035)	34	13
Stewardship objective	0.009 (0.024)	0.016 (0.026)	34	13
Valuation objective	0.010 (0.022)	0.005 (0.020)	34	13
Substantial financial risk	-0.007 (0.031)	-0.001 (0.030)	34	13
Above-average financial risk	-0.019 (0.062)	-0.015 (0.054)	34	13
Average financial risk	-0.046 (0.053)	-0.041 (0.047)	34	13
No financial risk	0.043 (0.033)	0.052 (0.044)	34	13

This table reports all 21 current assigned-profile contrasts with fresh N for the all-source-class exact-source condition-mean benchmark. Normalized MAE is profile loss minus N loss using explicit source scale endpoints. Smaller loss is closer and signs are not reversed. Both weights use the same paired sample within a row. Only complete exact-source vectors enter inference. These condition-mean measures do not recover individual human errors. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. An em dash denotes unavailable estimation, not a zero effect. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9. Agreement with Human Judgment Levels. Panel B. Like-unit improvement

Assigned description	Equal findings	Equal papers	N	Papers
	(1)	(2)	(3)	(4)
Strong assigned knowledge	0.017	0.003		
	(0.208)	(0.211)	60	19
Cautious preference	0.000	0.125		
	(0.255)	(0.253)	59	18
Opportunity-open preference	0.186	0.216		
	(0.186)	(0.194)	59	18
Cautious + knowledge	-0.309	-0.231		
	(0.245)	(0.250)	55	17
Opportunity-open + knowledge	-0.283	-0.274		
	(0.201)	(0.230)	60	19
Higher evidence threshold	0.276	0.270		
	(0.208)	(0.242)	58	18
Ordinary evidence threshold	-0.067	-0.127		
	(0.194)	(0.228)	60	19
Higher threshold + knowledge	0.071	0.037		
	(0.189)	(0.244)	56	17
Ordinary threshold + knowledge	-0.033	0.003		
	(0.227)	(0.243)	60	19
Nearer investment horizon	0.017	0.070		
	(0.167)	(0.196)	60	19
Longer investment horizon	0.150	0.132		
	(0.144)	(0.178)	60	19
Routine task familiarity	-0.017	-0.015		
	(0.186)	(0.201)	60	19
Extensive task familiarity	0.083	0.089		
	(0.231)	(0.254)	60	19
Deliberative attention	0.083	0.161		
	(0.198)	(0.214)	60	19
Efficient attention	0.200	0.275		
	(0.173)	(0.184)	60	19
Stewardship objective	-0.034	-0.059		
	(0.152)	(0.169)	59	18
Valuation objective	-0.102	-0.004		
	(0.130)	(0.176)	59	18
Substantial financial risk	-0.050	-0.091		
	(0.134)	(0.182)	60	19
Above-average financial risk	-0.067	0.014		
	(0.140)	(0.184)	60	19
Average financial risk	0.283	0.309		
	(0.161)	(0.183)	60	19
No financial risk	-0.136	-0.154		
	(0.129)	(0.177)	59	18

This table reports all 21 current assigned-profile contrasts with fresh N for the all-source-class exact-source condition-mean benchmark. Like-unit improvement averages +1 for smaller raw MAE, zero for a tie and -1 for larger raw MAE within matched unit groups relative to N. It is not a binary probability or an average of errors in unlike units. Both weights use the same paired sample within a row. Only complete exact-source vectors enter inference. These condition-mean measures do not recover individual human errors. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. An em dash denotes unavailable estimation, not a zero effect. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9. Agreement with Human Judgment Levels. Panel C. Six mixture mean contrasts

Signed comparison	Normalized MAE	Like-unit improvement
	(1)	(2)
Equal finding weights		
Historical population mixture minus N	-0.051 (0.054)	0.302 (0.224)
Equal four-risk mixture minus N	-0.044 (0.056)	0.208 (0.214)
Historical population mixture minus Equal four-risk mixture	-0.007 (0.006)	0.113 (0.177)
Fixed-2022 population mixture minus N	-0.050 (0.056)	0.264 (0.220)
Fixed-2022 population mixture minus Historical population mixture	0.001 (0.005)	0.000 (0.251)
Fixed-2022 population mixture minus Equal four-risk mixture	-0.006 (0.009)	0.151 (0.182)
Identical six-comparison support, separately by endpoint		
Paired findings	32	53
Papers	12	16
Holm-family size	6	6
Equal paper weights		
Historical population mixture minus N	-0.053 (0.048)	0.358 (0.232)
Equal four-risk mixture minus N	-0.043 (0.049)	0.244 (0.222)
Historical population mixture minus Equal four-risk mixture	-0.009 (0.007)	0.204 (0.239)
Fixed-2022 population mixture minus N	-0.051 (0.049)	0.296 (0.226)
Fixed-2022 population mixture minus Historical population mixture	0.001 (0.005)	0.037 (0.265)
Fixed-2022 population mixture minus Equal four-risk mixture	-0.008 (0.010)	0.225 (0.235)
Identical six-comparison support, separately by endpoint		
Paired findings	32	53
Papers	12	16
Holm-family size	6	6

This table reports all six population-mixture contrasts for the all-source-class exact-source condition-mean benchmark, separately from the assigned-profile comparisons. Normalized-MAE changes retain their original loss signs. The like-unit improvement outcome retains its score scale, not percentage points. Historical, equal-four-risk, fixed-2022 and N outcomes use common four-arm support within each endpoint. Support may differ between normalized loss and like-unit improvement. The four investment-risk categories and dated shares come from public Survey of Consumer Finances microdata for US primary economic units with directly held stocks. All six comparisons remain in each exported Holm family. Fixed-2022 weights are a future-weight sensitivity for older source anchors and do not replace historical primary weights. N counts paired distinct findings. Papers are inference clusters. The accounting data are the published experimental findings, source-supported instruments and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. An em dash denotes unavailable estimation, not a zero effect. Parentheses contain paper-cluster CR1 standard errors. Stars use Holm-adjusted probabilities within the predefined families, including unavailable hypotheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.7 Transfer to Additional Accounting Tasks

I apply the fixed descriptions to the source-adapted initial questions in the two 2026 accounting studies. Table 10 reports all 23 arms separately for each study. The applications use the initial accounting-rule judgments of Le (2026) and the first valuation task of Cai et al. (2026). In Le, recovery changes from 2/11 under N to 3/11 under K and 4/11 under deliberate attention. In Cai, every profile gives zero recovery on the one directional comparison. This evaluates the participant description against the particular accounting judgment elicited in each study.

The applications retain the initial decision screens supported by the source materials and use fixed participant descriptions without audience weights. This provides evidence on how the descriptions transfer to two additional accounting tasks.

Table 10. Participant Profiles in Two Additional Accounting Studies. Le

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
Evidence threshold			
Higher evidence threshold†	2/11	18.18	0.240
Ordinary evidence threshold†	2/11	18.18	0.240
Higher threshold + knowledge†	2/11	18.18	0.240
Ordinary threshold + knowledge†	2/11	18.18	0.240
Investment horizon			
Nearer investment horizon†	2/11	18.18	0.240
Longer investment horizon†	2/11	18.18	0.240
Task familiarity			
Routine task familiarity	2/11	18.18	0.058
Extensive task familiarity	2/11	18.18	0.195
Processing attention			
Deliberative attention	4/11	36.36	0.163
Efficient attention	3/11	27.27	0.216
Decision objective			
Stewardship objective†	2/11	18.18	0.240
Valuation objective†	2/11	18.18	0.240
Financial risk willingness			
Substantial financial risk†	2/11	18.18	0.240
Above-average financial risk†	2/11	18.18	0.240
Average financial risk†	2/11	18.18	0.240
No financial risk†	2/11	18.18	0.240

This table reports every frozen arm separately for the Le external study. Numerators count directionally agreeing contrasts. Denominators count human-scored contrasts, not participants or independent studies. Le has 11 scored contrasts and one contrast without a human direction. Sources are the initial accounting-rule career questions in Le (2026) and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. Normalized effect error is the within-study mean absolute effect error divided, contrast by contrast, by the fixed source question range times the positive contrast-weight mass. It is not re-normalized using observed model responses. Contrasts share source cases, so no pooled main-finding inference, standard errors or significance stars are assigned. The dagger denotes N-equivalent handling when a descriptor is unsupported by that study task. Historical B differs from current N in serving conditions, time and prompt.

Table 10. Participant Profiles in Two Additional Accounting Studies. Cai

Assigned description	Agreement / scored	Agreement %	Normalized effect error
	(1)	(2)	(3)
Original prompt			
Original prompt (B)	0/1	0.00	0.023
Neutral control			
Neutral prompt (N)	0/1	0.00	0.123
Assigned knowledge			
Strong assigned knowledge	0/1	0.00	0.023
Uncertainty preference			
Cautious preference	0/1	0.00	0.023
Opportunity-open preference	0/1	0.00	0.123
Cautious + knowledge	0/1	0.00	0.023
Opportunity-open + knowledge	0/1	0.00	0.023
Evidence threshold			
Higher evidence threshold	0/1	0.00	0.023
Ordinary evidence threshold	0/1	0.00	0.023
Higher threshold + knowledge	0/1	0.00	0.023
Ordinary threshold + knowledge	0/1	0.00	0.023
Investment horizon			
Nearer investment horizon†	0/1	0.00	0.123
Longer investment horizon†	0/1	0.00	0.123
Task familiarity			
Routine task familiarity	0/1	0.00	0.223
Extensive task familiarity	0/1	0.00	0.023
Processing attention			
Deliberative attention	0/1	0.00	0.023
Efficient attention	0/1	0.00	0.023

Continued on next page

Table 10. Participant Profiles in Two Additional Accounting Studies. Cai (continued)

Assigned description	Agreement / scored	Agreement %	Normalized effect error
	(1)	(2)	(3)
Decision objective			
Stewardship objective†	0/1	0.00	0.123
Valuation objective†	0/1	0.00	0.123
Financial risk willingness			
Substantial financial risk	0/1	0.00	0.123
Above-average financial risk	0/1	0.00	0.123
Average financial risk	0/1	0.00	0.023
No financial risk	0/1	0.00	0.023

This table reports every frozen arm separately for the Cai external study. Numerators count directionally agreeing contrasts. Denominators count human-scored contrasts, not participants or independent studies. Cai has one scored marginal A/B contrast, distinct from its original measured moderation. Sources are the initial cryptocurrency-valuation questions in Cai et al. (2026) and saved model responses. Responses use GPT-6 at temperature zero with reasoning disabled. Normalized effect error is the within-study mean absolute effect error divided, contrast by contrast, by the fixed source question range times the positive contrast-weight mass. It is not re-normalized using observed model responses. Contrasts share source cases, so no pooled main-finding inference, standard errors or significance stars are assigned. The dagger denotes N-equivalent handling when a descriptor is unsupported by that study task. Historical B differs from current N in serving conditions, time and prompt.

4.8 Unavailable Responses

To assess how unavailable responses affect the profile comparisons, I allow each unknown binary result to be a recovery or a failure and retain every observed arm value. The knowledge description and the equal, historical, and fixed 2022 risk mixtures retain positive gains relative to neutral instructions under both weighting rules. With equal-finding weights, the knowledge difference ranges from 4.34 to 9.67 percentage points, and the historical-mixture difference ranges from 9.23 to 16.42 points. The historical-versus-equal bounds include zero under both weighting rules, so dated shares need not improve on the equal mixture. Table A1 reports all seven comparisons. These are logical bounds, not sampling intervals.

4.9 Representation, Context, and Model Configuration

I assess whether the way the case is supplied changes recovery. Continuous text increases recovery from 82 to 100 of 208 findings. Ordered lines increase it from 93 to 105 of 235, and accounting attention increases it from 94 to 104 of 236. Chronological ordering changes recovery from 38 to 37 of 92. Table A8 reports the matched input comparisons in the appendix. Changes in representation can recover some reporting responses and lose others.

I also compare JSON requests with a reversible labeled transcript that retains the same accounting content. The test uses 24 comparisons from 24 papers. For GPT-6 Sol, presentation-based recovery changes

from four to seven of 12 and textual recovery from six to five. For GPT-5.4 mini, presentation-based recovery changes from seven to four while textual recovery remains four of 12. Changing GPT-6 reasoning from disabled to high changes recovery from nine to ten of 24 under JSON and from ten to 12 under the transcript. The reported Holm-adjusted p-values are 0.982. The effect of an encoding depends on the model evaluating the accounting information.

To assess whether explicit table structure adds to the content, I compare row-and-column indexing with repetition of the same table text. The test covers nine comparisons from nine papers. Four configurations have lower recovery with indexing, and two have unchanged recovery. Making cell addresses explicit does not by itself reproduce how users respond to a financial table.

Context can also affect a judgment without changing the focal accounting information. In the goodwill-allocation task, changing information position, surrounding context, or question wording changes some judgments. I vary incidental company names and response elicitation as additional features of the supplied case. These comparisons identify features of the setting that must be held fixed when interpreting a model's response to a reporting choice.

I assess study composition using paper weights, effect types, source support, and publication periods. In 2005–2014, GPT-6 recovers 31/77 textual findings (40.3%) and 3/25 presentation-based findings (12.0%). Presentation-only treatments account for 20 of those 25 presentation-based comparisons. Omitting each paper in that period leaves GPT-6's textual-minus-presentation-based gap between 25.0 and 32.8 points. The period association accompanies a different mix of reporting treatments and extends beyond one source paper.

I compare model configurations on the same reconstructed tasks to separate model choice from task coverage. Table A11 reports the full comparisons in the appendix. On the same 229 findings from 45 papers, recovery is 57.6% for GPT-4.1, 56.8% for mini, 48.9% for GPT-5, 39.3% for GPT-6, and 45.4% for Sol. The greater economic-benchmark agreement under GPT-6 and GPT-6.1 Sol therefore does not imply closer reproduction of human judgments.

I examine familiarity with published results separately from participant-task responses. Across 144 numerical recall items, none exactly matches the published result. Nine of 72 direction responses are correct, with eight from GPT-4.1 and one from high-reasoning GPT-6. In a separate 16-paper diagnostic, recovery is 13/37 when the model's independently elicited prediction agrees with the human direction and 4/22 when it conflicts. The expected treatment direction is associated with recovery even though exact numerical recall is absent in these requests.

4.10 Evidence from Public Human-Judgment Benchmarks

I use three public benchmarks to distinguish correctness, response to wording, and agreement with people outside the accounting panel. On the same 150 cognitive-reflection tasks in Hagendorff et al. (2023), correct answers increase from 4.7% for GPT-3 davinci-003 to 60.0% for ChatGPT-3.5 and 95.3% for ChatGPT-4. Characteristic intuitive mistakes fall from 80.0% to 14.7% and then to zero. The semantic-illusion tasks show correct-answer rates of 18.0%, 74.0%, and 88.0%, respectively. More accurate answers can move a model away from common human errors.

BiasMonkey asks a different question because the survey responses have no single correct opinion (Tjuatja et al., 2024). I use saved responses for nine primary configurations and 739 eligible question-level effects per configuration. For GPT-3.5 turbo, acquiescence, allow/forbid wording, and a midpoint option follow the human response-bias direction. Response order and an explicit don't-know option run in the opposite direction. The mean absolute acquiescence effect is 22.9 percentage points, compared with 13.3 under key typos and 28.3 under randomized interior letters. Departing from a human response bias does not eliminate sensitivity to how a question is written.

WildFrame compares judgments under positive or negative framing (Lior et al., 2026). I reconstruct the human majority from 5,000 votes on 1,000 texts and compare it with the saved model labels. Agreement is 62.8% for GPT-3.5, 70.7% for GPT-4o, and 77.6% for GPT-5 mini. Here the newer configuration more closely reproduces the human sentiment judgment. Improvement in human resemblance therefore depends on the judgment elicited.

The accounting economic-benchmark pattern also remains when papers receive equal weight. Agreement is 52.9% for Sol and 38.7% for GPT-4.1, while human recovery is 29.8% and 54.2%, respectively. Each exclusion of one paper preserves that ordering. The public benchmarks retain their own models and tasks. Together, these comparisons show why an accounting decision aid and a substitute for financial-statement users require separate measures of performance.

5 Conclusion

I examine whether AI reproduces users' judgments about accounting reporting choices. GPT-6 recovers 39.8% of 236 eligible directions. Within presentation-based experiments, recovery is 64.7% for information changes and 31.0% for presentation changes holding information constant. All five configurations show the same ordering. Correct factual answers also coexist with different judgments, separating identification of information from its influence on a user.

Economic-benchmark agreement and human recovery can move in different directions. On the same 45 comparisons, economic agreement rises from 33.3% under GPT-4.1 to 62.2% under Sol, while human recovery falls from 53.3% to 31.1%. Correct actions and repeated wrong choices can both produce zero contrasts. The accounting decision must therefore be examined alongside the treatment response.

Across 200 matched comparisons, the knowledge description changes human recovery by 7.00 percentage points. Its change in controlled pattern agreement is -8.11 points across 37 comparisons. The profile estimates differ across source-compatible accounting decisions and controlled task procedures. The profile-family tests distinguish knowledge from uncertainty attitudes, evidence thresholds, attention, and objectives. Estimated judgment-level error is higher under the knowledge description and lower under the historical risk mixture. Agreement in direction and agreement in judgment level remain different outcomes.

Combining the four risk-profile answers differs from N by 11.83 percentage points. Historical shares add 1.61 points relative to the equal mixture, while fixed 2022 shares differ from N by 14.52 points. Most of the observed recovery gain comes from combining risk-profile responses, with a smaller increment from

replacing equal shares with dated shares. The survey shares describe an external investment audience.

In Le, recovery changes from 2/11 under N to 3/11 under K and 4/11 under deliberate attention. In Cai, every profile gives zero recovery on the one directional comparison. These applications concern the source-adapted initial questions and use no audience weights. For standard setting, a model's suitability depends on both the accounting decision and the users it is intended to represent.

References

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023, July). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis 31*(3), 337–351.

Bai, Y., X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, and J. Li (2024). Longbench: A bilingual, multitask benchmark for long context understanding. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. arXiv:2308.14508.

Binz, M. and E. Schulz (2023). Using cognitive psychology to understand GPT-3. *PNAS Proceedings of the National Academy of Sciences of the United States of America 120*(6), 1–10. Place: US Publisher: National Academy of Sciences.

Blankespoor, E., J. Croom, and S. M. Grant (2026). Generative ai and investor processing of financial information. *Journal of Accounting and Economics 82*(2), 101908.

Blankespoor, E., E. deHaan, and I. Marinovic (2020, November). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics 70*(2-3), 101344.

Bloomfield, R., M. W. Nelson, and E. Soltes (2016). Gathering data for archival, field, survey, and experimental accounting research. *Journal of Accounting Research 54*(2), 341–395.

Bonner, S. E., S. M. Clor-Proell, and L. Koonce (2014, November). Mental Accounting and Disaggregation Based on the Sign and Relative Magnitude of Income Statement Items. *The Accounting Review 89*(6), 2087–2114.

Bonner, S. E. and B. L. Lewis (1990). Determinants of auditor expertise. *Journal of Accounting Research 28*(Supplement), 1–20.

Cai, Q., Z. J. Leitter, L. Niu, and H.-T. Tan (2026). Accounting classification and measurement for cryptocurrencies: Impact on valuation judgments. Available at SSRN.

Cardinaels, E. (2008). The interplay between cost accounting knowledge and presentation formats in cost-based decision-making. *Accounting, Organizations and Society 33*(6), 582–602.

Chen, Z., W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang (2021). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics.

Croom, J. (2026). Interactivity and illusions of ability: How using generative AI affects investor judgments. *Journal of Accounting Research 64*(2), 681–719.

Einhorn, H. J. and R. M. Hogarth (1981). Behavioral Decision Theory: Processes of Judgment and Choice. *Journal of Accounting Research 19*(1), 1–31.

Elliott, W. B., F. D. Hodge, J. J. Kennedy, and M. Pronk (2007). Are M.B.A. students a good proxy for

nonprofessional investors? *The Accounting Review 82*(1), 139–168.

FASB (2021, December). Conceptual framework for financial reporting: Chapter 7, presentation. Statement of financial accounting concepts no. 8, Financial Accounting Foundation.

FASB (2024, September). Conceptual Framework for Financial Reporting. Technical report, Financial Accounting Standards Board. Statement of Financial Accounting Concepts No. 8, Chapters 1–8.

FASB (2025, February). Rules of procedure: Amended and restated through february 12, 2025. Amended and restated through February 12, 2025.

FASB (2026). Accounting standards updates issued.

Frederickson, J. R. and J. S. Miller (2004). The effects of pro forma earnings disclosures on analysts' and nonprofessional investors' equity valuation judgments. *The Accounting Review 79*(3), 667–686.

Gassen, J. and M. Muhn (2025). Financial transparency of private firms: Evidence from a randomized field experiment. *Journal of Accounting Research 63*(1), 413–460.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin 73*(6), 422–432.

Graham, J. R., C. R. Harvey, and S. Rajgopal (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics 40*(1–3), 3–73.

Guggenmos, R. D., M. D. Piercey, and C. P. Agoglia (2018). Custom contrast testing: Current trends and a new approach. *The Accounting Review 93*(5), 223–244.

Hagendorff, T., S. Fabi, and M. Kosinski (2023, October). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science 3*(10), 833–838. Publisher: Nature Publishing Group.

He, Y., H.-T. Tan, F. Yeo, and J. Zhang (2019). When Do Qualitative Risk Disclosures Backfire? The Effects of a Mismatch in Hedge Disclosure Formats on Investors' Judgments. *Contemporary Accounting Research 36*(4), 2093–2112. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1911-3846.12518.

Heath, O. and R. Alexander (2026). Correcting mode collapse in silicon sampling with semantic similarity rating. arXiv preprint arXiv:2607.28550.

Hirshleifer, D. and S. H. Teoh (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics 36*(1–3), 337–386.

Hirst, D. E. and P. E. Hopkins (1998). Comprehensive Income Reporting and Analysts' Valuation Judgments. *Journal of Accounting Research 36*, 47–75. Publisher: [Accounting Research Center, Booth School of Business, University of Chicago, Wiley].

Hodge, F. D., P. E. Hopkins, and D. A. Wood (2010). The Effects of Financial Statement Information Proximity and Feedback on Cash Flow Forecasts. *Contemporary Accounting Research 27*(1), 101–133.

Huang, A. H., H. Wang, and Y. Yang (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research 40*(2), 806–841.

Jang, C., D. Lee, and J. Kim (2026). Instruction-tuned language models cannot sample from distributions they can describe. arXiv preprint arXiv:2607.25292. Findings of EMNLP 2026.

Joyce, E. J. and R. Libby (1981). Some Accounting Implications of "Behavioral Decision Theory: Processes of Judgment and Choice". *Journal of Accounting Research 19*(2), 544–550.

Kagel, J. H. and A. E. Roth (2016). *The Handbook of Experimental Economics, Volume 2*. Princeton university press.

Kim, A., M. Muhn, and V. Nikolaev (2024). Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866. Working paper withdrawn February 20, 2025; the authors report inconsistencies in the data and analyses.

Koonce, L., M. G. Lipe, and M. L. McAnally (2005). Judging the Risk of Financial Instruments: Problems and Potential Remedies. *The Accounting Review 80*(3), 871–895. Publisher: American Accounting Association.

Koonce, L., C. Mongold, L. Quaid, and B. J. White (2024, June). Experimental research on standard-setting issues in financial reporting. *Accounting, Organizations and Society 112*, 101509.

Le, A. (2026). Accounting rules and the labor market for accountants. *Journal of Accounting Research*.

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

Libby, R. and B. L. Lewis (1982). Human information processing research in accounting: The state of the art in 1982. *Accounting, Organizations and Society 7*(3), 231–285.

Libby, R. and J. Luft (1993). Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment. *Accounting, Organizations and Society 18*(5), 425–450.

Lior, G., L. Naccache, and G. Stanovsky (2026, October). Comparing the framing effect in humans and LLMs on naturally occurring texts. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, Budapest, Hungary, pp. 825–841. Association for Computational Linguistics.

Liu, N. F., K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics 12*, 157–173.

MacKenzie, N. L., J. Moon, S. Rykowski, Q. T. Swanquist, and R. L. Whited (2024, November). A Comparison of Artificial Intelligence and Human Responses in Audit Experiments.

Maines, L. A. and L. S. McDaniel (2000, April). Effects of Comprehensive-Income Characteristics on Nonprofessional Investors' Judgments: The Role of Financial-Statement Presentation Format. *The Accounting Review 75*(2), 179–207.

Nikankin, Y., A. Reusch, A. Mueller, and Y. Belinkov (2025). Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations*.

Smith, V. L. (1994, March). Economics in the Laboratory. *Journal of Economic Perspectives 8*(1), 113–131.

Sui, Y., M. Zhou, M. Zhou, S. Han, and D. Zhang (2024). Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654. Association for Computing Machinery.

Tjuatja, L., V. Chen, T. Wu, A. Talwalkar, and G. Neubig (2024). Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics 12*, 1011–1026.

Xu, Y., M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou (2020). LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200. Association for Computing Machinery.

Zhu, F., W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua (2021). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287. Association for Computational Linguistics.

A Logical Bounds for Unavailable Responses

Table A1. Logical Bounds on Profile Recovery Differences

			Missing binary outcomes			Equal findings		Equal papers	
Comparison	Findings	Papers	Left only	Right only	Both	Lower	Upper	Lower	Upper
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A. Added accounting knowledge									
K versus N	207	46	2	1	4	4.34	9.67	1.62	12.77
Panel B. Risk-profile mixtures and dated shares									
Equal risk mixture versus N	195	42	4	0	5	7.69	14.88	2.37	17.28
Historical mixture versus N	195	42	4	0	5	9.23	16.42	4.45	19.36
Historical versus equal mixture	195	42	0	0	9	-3.08	6.16	-6.87	11.03
Fixed 2022 mixture versus N	195	42	4	0	5	10.25	17.44	5.51	20.42
Fixed 2022 versus historical mixture	195	42	0	0	9	-3.59	5.65	-7.89	10.01
Fixed 2022 versus equal mixture	195	42	0	0	9	-2.06	7.18	-5.81	12.09

This table reports logical lower and upper bounds on the left-minus-right difference in human-direction recovery. Panel A uses the 207-comparison primary numeric source frame. Panel B uses the 195-comparison primary source frame with an identified pre-response operator. Eligibility follows the source-defined accounting targets and operators. N denotes the fresh neutral description. K adds accounting knowledge. The equal, historical, and fixed 2022 mixtures combine the four mutually exclusive SCF risk categories. Findings and papers are full-frame denominators rather than complete-pair counts. Left and right refer to the arms in column (1). Left only and Right only count findings missing just that arm. Both counts findings missing both arms. These three categories are disjoint. Each unknown binary recovery outcome may equal zero or one. Every known arm value constrains the bounds, including findings with only one observed arm. Equal-finding weights give each finding equal weight. Equal-paper weights give each paper equal weight and then average its findings. Limits are percentage points and reflect logical possibilities rather than sampling uncertainty. Lower limits round down and upper limits round up to two decimals. Signs use the exact rational limits.

B Supplementary Unprofiled Comparisons

The following exhibits retain the unprofiled sample construction, model comparisons, input-format tests, and economic benchmarks. Their samples correspond to the analyses stated in each description. The full source frame and participant-profile samples are reported in Table 1.

Table A2. Sample Construction and Treatment-Channel Composition

	(1) Comparisons	(2) Papers
Panel A. Sample construction		
Accepted matches between manual and independent coding	401	64
Selected source-linked comparisons	269	50
Less. responses or classifications without a point score	(12)	–
Less. comparisons with nondegenerate pooling bounds	(4)	–
Less. changed-estimand or memory-procedure descriptors	(8)	–
Scoreable comparison records	245	46
Less. duplicate mappings of the same source finding	(6)	–
Distinct scoreable findings	**239**	**46**
Less. human-zero benchmark	(1)	–
Less. reconstructed initial-phase placebo	(1)	–
Distinct directional human benchmarks	**237**	**46**
Less. source-ineligible conditional model endpoint	(1)	–
Source-eligible model contrasts	**236**	**46**
Panel B. Treatment channels of distinct scoreable findings		
	Comparisons	Sample (ASU) share (%)
Textual	166	69.46 (82.24)
Visual	53	22.18 (14.95)
Inseparable content and presentation changes	19	7.95 (2.80)
Indeterminate	1	0.42 (–)
Total	**239**	**100.00 (100.00)**

This table reports sample construction and treatment-channel composition. Panel A counts coding matches, administered comparisons, numerical records, distinct findings, and eligible model contrasts. The 401 accepted matches cover 64 papers. 269 administered comparisons cover 50 papers. The 245 numerical records from 46 papers map to 239 distinct findings. Excluding one human-zero benchmark, one structural placebo, and one ineligible subgroup yields 236 directional model contrasts. Panel B reports counts and shares of the 239 findings. Parentheses give shares of 214 ASUs classified by dominant amendment purpose. Experimental findings and ASUs are separate units. Sources are the published studies, supplements, structured coding, and FASB updates issued in 2010–2025.

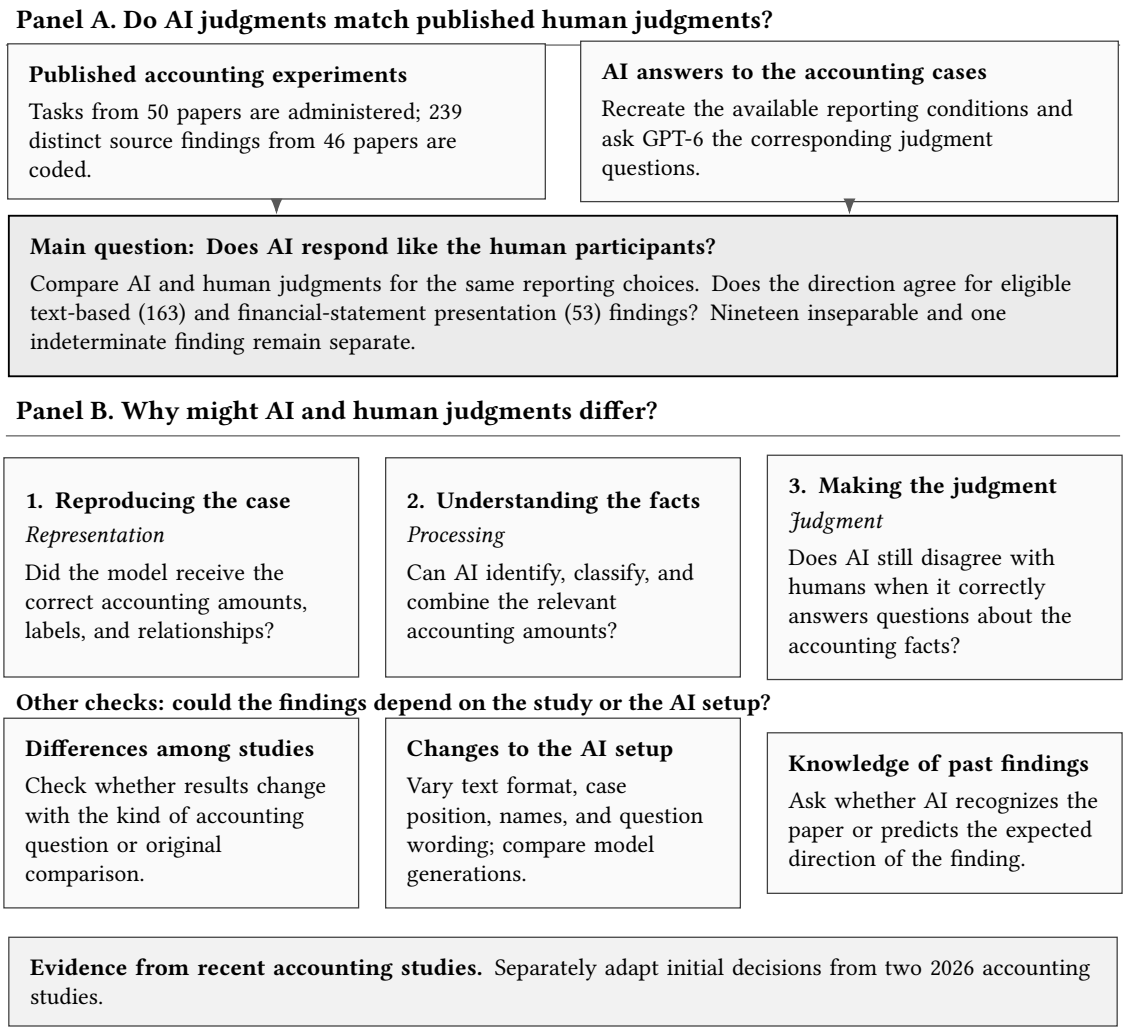


Figure A1. Overview of the empirical research design

This figure shows how the tests are conducted. 401 accepted coding matches are mapped to 245 point-scoreable mappings and 239 distinct findings. Of the 239 findings, 237 have nonzero human directions and are not structural placebos. The design distinguishes reconstruction tiers, pooling bounds, and source-defined outcomes. Some comparisons involve more than two conditions. Channel labels describe the original human treatments, while AI receives reconstructed text rather than verified copies of the original visual displays. We use one saved response for each required condition and stage. The information checks and economic benchmarks assess processing accuracy and decision quality alongside human recovery.

Table A3. Recovery of Reported Human Accounting Judgments

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

This table reports GPT-6 recovery and channel regressions. The dependent variables in Panel B equal one for recovery of the nonzero human direction in column (1), a zero model contrast in column (2), and a reversed direction in column (3). Panel A reports counts and recovery percentages for 236 eligible findings from 46 papers. Panel B uses 216 textual or presentation-based findings from 44 papers. Each regression includes a presentation-based indicator and intercept, with textual findings as the reference and no other controls or FE. Coefficients and standard errors are in percentage points. Sources are the published human findings, reconstructed instruments, and condition-level responses. Standard errors are paper-clustered CR1, with $t(43)$ inference. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A4. Information Changes 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

This table reports recovery for information changes and presentation changes holding information constant within presentation-based materials. The dependent variable equals one when the model contrast matches the reported nonzero human direction. Columns identify five configurations using the same 46 findings from 14 papers, comprising 17 information changes from six papers and 29 presentation changes from 11 papers. Paper counts overlap. Each regression includes an information-change indicator and intercept. Presentation changes form the reference category. Coefficients and standard errors are in percentage points. Sources are published human findings, coded treatments, reconstructed instruments, and model responses. Standard errors are paper-clustered CR1, with $t(13)$ inference. Stars use nominal p-values. The broader specification-family Holm adjustment covers GPT-4.1, mini, GPT-6, and Sol. GPT-5 statistics use nominal inference. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A5. Information Changes, Presentation, and Recovery. Regression Evidence

Panel A. Information changes minus presentation changes holding information constant

Specification	(1) Unadjusted	(2) Channel and source controls	(3) Within paper	(4) Within source experiment
Information change	16.90**	12.46	5.66	15.62
	(8.23)	(10.09)	(13.45)	(17.62)
Holm-adjusted p-value	0.190	0.676	0.881	0.881
Findings	163	148	41	23
Source papers	36	35	6	4
Channel controls	–	X	–	–
Source and reconstruction controls	–	X	–	–
Paper FE	–	–	X	–
Source-experiment FE	–	–	–	X

Panel B. Same-sample channel and information-content regressions

	(1)	(2)	(3)
Information change	19.98**	18.68*	13.87
	(8.83)	(9.78)	(10.40)
Presentation-based treatment	8.86	8.88	9.34
	(10.47)	(10.25)	(9.02)
Additional source-finding indicator	–	X	X
Reconstructed or corrected-material indicator	–	–	X
Findings	154	154	154
Papers	34	34	34

This table reports associations between information content, treatment channel, and GPT-6 recovery. The dependent variable equals one when the model recovers the reported nonzero human direction. Coefficients and standard errors are in percentage points. Panel A compares information changes with presentation changes holding information constant in four specifications with 163, 148, 41, and 23 findings from 36, 35, six, and four papers. Column (2) includes channel and reconstruction-tier controls. Columns (3) and (4) include Paper FE and Source-experiment FE. Its adjusted p-values use a four-test Holm family. Panel B holds the sample fixed at 154 pure-channel findings from 34 papers and includes both treatment indicators. Columns (2) and (3) add the additional source-finding indicator. Column (3) adds the reconstructed or corrected-material indicator. The reference category is textual presentation changes holding information constant. Sources are published findings, source coding, reconstructed instruments, and condition-level responses. These supplementary specifications use the directional administration panel before subgroup eligibility. Standard errors are paper-clustered CR1, with $t(G-1)$ inference. Stars use nominal p-values. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A6. Human Predictions and Source-Supported Economic Benchmarks

Panel A. Human findings and the authors' directional expectations

	(1) Textual	(2) Presentation-based	(3) Other	(4) All
Findings reviewed	164	53	20	237
Codable directional expectations	134	36	16	186
Human finding in expected direction	128	33	16	177
Agreement (%)	95.5	91.7	100.0	95.2
Null predictions	10	3	0	13
Other or unresolved expectations	20	14	4	38

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

This table reports human agreement with authors' predictions and model agreement with economic benchmarks. Panel A classifies 237 directional human findings by channel and expected direction. Agreement is a count or percentage of findings with a directional expectation. Panel B compares the same 45 findings from 13 papers across five configurations, with 28 findings from nine papers in the higher-confidence subset. Equal-paper rows average within-paper agreement and recovery rates across the 13 papers. The other rows weight findings equally. The human-recovery rows use the same 45 findings as the benchmark rows. An economic benchmark is a source-supported direction, ordering, invariance condition, or accounting rule. It is distinct from an author's behavioral prediction and from correctness of every numerical answer. The benchmark panel excludes operational reference conditions, unresolved economic targets, and the ineligible subgroup endpoint. Sources are the original hypotheses, reported human findings, coded economic benchmarks, and condition-level model responses.

Table A7. Actual Answers Underlying Zero Treatment Contrasts

	(1) GPT-4.1	(2) GPT-4.1 mini	(3) GPT-6	(4) GPT-6.1 Sol
Recorded benchmark-consistent zeros reviewed	16	13	23	24
Correct economic action in relevant conditions	4	4	7	10
Same wrong substantive choice	4	0	0	0
Missing expressed economic response	0	0	2	0
Limited direction or rule correctness	2	0	3	3
Correct answer level not identified	5	9	11	10
Conditional endpoint ineligible or unverified	1	0	0	1
Diagnostic of different answers collapsed to a binary zero				
Source-defined binary collapse	3	0	1	0

This table reports the economic content of answers underlying 76 zero model-comparisons across four configurations. The sample covers 33 distinct findings from 11 papers and 174 condition-level answer components. Counts classify correct actions in all relevant conditions, repeated wrong choices, missing economic responses, limited correctness, ratings without an identified correct level, and unresolved subgroup eligibility. The separate diagnostic row identifies different answers mapped to the same source-defined binary score and overlaps the classifications above. Sources are the original accounting tasks, source-supported economic benchmarks, and condition-level model responses.

Table A8. Paired Changes in Recovery under Alternative Accounting Inputs

	(1) Continuous text	(2) Ordered lines	(3) Accounting attention	(4) Chronological ordering
Matched distinct findings	208	235	236	92
Source papers	44	46	46	26
Recovered. baseline	82	93	94	38
Recovered. alternative input	100	105	104	37
Baseline recovery (%)	39.4	39.6	39.8	41.3
Alternative-input recovery (%)	48.1	44.7	44.1	40.2
ΔRecovery (percentage points)	8.65*	5.11	4.24	−1.09
	(4.38)	(3.71)	(3.00)	(4.28)
Holm-adjusted p-value	0.601	1.000	1.000	1.000
Additional presentation-based ΔRecovery	−2.88	5.73	−3.68	20.44
	(11.08)	(11.62)	(8.21)	(12.35)

This table reports paired recovery differences under four alternative accounting inputs. The dependent variable is ΔRecovery. alternative-input recovery minus baseline recovery for the same finding, multiplied by 100. Columns identify continuous text, ordered lines, accounting attention, and chronological ordering. Matched samples contain 208, 235, 236, and 92 findings from 44, 46, 46, and 26 papers. Average changes are intercept-only regressions. Additional presentation-based changes add a channel indicator on 188, 215, 216, and 79 pure-channel findings. Coefficients and standard errors are percentage points. The chronology test changes the order of prior material and responses. Sources are reconstructed tasks and paired model responses from the directional administration panel. Standard errors are paper-clustered CR1 with $t(G-1)$ inference. Holm-adjusted p-values use 12 tests. four inputs crossed with recovery, zero contrasts, and reversals. Stars use nominal p-values. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A9. Separate Checks of Accounting Information in Reconstructed Materials

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)

This table reports correct responses to separate accounting-information questions. Entries are counts and percentages of the 668 valid questions from a common bank of 693 questions concerning 39 task instruments from 38 papers. Rows identify amount, classification, and integration questions. Columns compare GPT-6 and GPT-6.1 Sol. Both models correctly answer all 668 valid questions. The same 25 questions are excluded for both models because the question does not identify which displayed security is being assessed. Percentages divide correct answers by valid questions. Both models receive the same question bodies in separate conversations, without human results or answer keys. Grading uses complete responses, accepting equivalent stated units, signed amounts when the question asks for magnitude, and conventional rounding. Sources are the accounting instruments, source-supported answer keys, and saved model responses.

Table A10. Recovery after Correct Answers to Accounting-Information Questions

	(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)

This table reports human recovery conditional on correct accounting-information responses. The dependent variable in the channel regressions equals one for recovery of the nonzero human direction. Columns (1) and (3) require all available valid checks to pass. Columns (2) and (4) additionally require amount, classification, and integration checks at each referenced judgment stage. GPT-6 uses the 237-finding directional administration panel. Sol uses its 236 available judgment pairs before subgroup eligibility. The model-specific samples contain 195, 137, 125, and 108 findings from 45, 37, 35, and 30 papers. Each model uses its own source-matched factual responses. These samples combine the common question bank with other available source-matched accounting checks. The channel regressions include a presentation-based indicator and intercept, with textual findings as the reference. Coefficients and standard errors are percentage points. Missing checks remain ungraded. Sources are the accounting instruments, approved keys, and condition-level responses. Standard errors are paper-clustered CR1, with $t(G-1)$ inference. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

Table A11. Recovery across Model Generations and Configurations

	(1) GPT-4.1	(2) GPT-4.1 mini	(3) GPT-5	(4) GPT-6	(5) GPT-6.1 Sol
Panel A. Source-eligible directional recovery					
Scoreable findings	235	231	232	236	235
Source papers	46	46	45	46	46
Recovered findings	135	130	115	94	107
Zero contrasts	63	62	66	92	98
Reversed contrasts	37	39	51	50	30
Recovery (%)	57.4	56.3	49.6	39.8	45.5
Textual. recovered/findings	92/162	88/158	77/159	66/163	75/162
Presentation-based. recovered/findings	35/53	33/53	27/53	23/53	26/53
Panel B. Identical findings under all five configurations					
Findings	229	229	229	229	229
Source papers	45	45	45	45	45
Recovered findings	132	130	112	90	104
Recovery (%)	57.6	56.8	48.9	39.3	45.4
Panel C. Paired recovery difference relative to GPT-6					
ΔRecovery (percentage points)	17.45***	16.88***	9.48**	–	5.53
	(5.61)	(6.03)	(4.51)	–	(5.42)
Nominal p-value	0.003	0.007	0.041	–	0.313
Matched findings	235	231	232	–	235
Source papers	46	46	45	–	46
Intercept	X	X	X	–	X

This table reports directional recovery and paired model comparisons. The dependent variable in Panel C is ΔRecovery. comparison-model recovery minus GPT-6 recovery for the same finding, multiplied by 100. Panel A reports each configuration's source-eligible point-scoreable sample. Panel B uses the same 229 findings from 45 papers. Panel C compares each alternative with GPT-6 on its matched findings, using an intercept and no other controls or FE. GPT-4.1 and mini use temperature zero without a reasoning parameter. GPT-5 uses minimal reasoning with temperature omitted. GPT-6 uses temperature zero with reasoning disabled. Sol uses high reasoning with temperature omitted. All models receive the same reconstructed tasks and source scoring rules, with their own permitted prior answers and no additional output cap. GPT-5 has four outcomes without a unique point score. two mixed recommendations and two pooling-bound contrasts. Sources are the accounting studies and condition-level model responses. Coefficients and standard errors are percentage points. Standard errors are paper-clustered CR1, with $t(G-1)$ inference. $G = 45$ for GPT-5 and 46 for the other paired comparisons. Stars use nominal p-values. Significance thresholds are $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

C Recent AI Experiments in Accounting

Table A12 traces AI experiments in accounting from November 2023 to November 2026, covering The Accounting Review, Journal of Accounting Research, Journal of Accounting and Economics, Contemporary Accounting Research, Review of Accounting Studies, and Accounting, Organizations and Society, together with SSRN and arXiv working papers. The literature develops along three lines of AI task performance, AI judgement, and human responses to AI.

I distinguish three research questions. AI capability concerns accuracy, forecasting, and information processing. AI judgment concerns model decisions in behavioral and subjective accounting tasks. Human response concerns people's reliance, perceptions, and decisions when AI is available or used by others. A paper can address more than one question. The distinction matters because evidence that a model completes a task does not by itself show that it represents the person making the decision.

Table A12. Recent AI Experiments and Capability Tests in Accounting, in Chronological Order.

This table reports accounting studies of AI capability, AI judgement, and human responses to AI. Rows follow the displayed source date and identify publication status, methods, and findings. The coverage includes the six accounting journals named above and SSRN or arXiv papers. Highlighted labels identify studies of AI judgement.

Source date / study focus	Paper and publication status	Methods	Findings
2023-11-20 AI capability	Islam, Kannappan, Kiela, Qian, Scherrer and Vidgen FinanceBench. A New Benchmark for Financial Question Answering arXiv 2311.11944. Repository paper.	Open-book questions with answers and source evidence. 16 model/retrieval/context configurations. Manual scoring of 2,400 answers on 150 cases.	The tested systems often fail even straightforward financial-document questions. In the reported retrieval configuration, GPT-4-Turbo answers incorrectly or abstains on 81% of questions. Results are configuration-specific.
2024-02-26 Human response	Libby and Witz Can artificial intelligence reduce the effect of independence conflicts on audit firm liability? CAR (2024), 41(2). 1346-1375. Published. Repository version.	Two human-participant experiments vary human versus AI-performed audit procedures and the appearance of auditor independence conflicts. Measure juror liability judgments.	Described AI use attenuates the liability penalty associated with independence concerns, through perceived auditor objectivity and trust.
2024-03-11 AI judgement	Isack Replicating Reason. The Advent of Human-like Audit Judgment by Generative AI SSRN 4716860. Working paper.	Administers five existing audit experiments to GPT-4 and compares its task responses with human-auditor benchmarks.	Reports some human-like audit judgments but context-dependent deviations.
2024-06-24 AI capability	Kirkos, Boskou, Chatzipetrou, Tiakas and Spathis Exploring the Boundaries of Financial Statement Fraud Detection with Large Language Models SSRN 4842962. Working paper. Duplicate SSRN versions consolidated.	Tests ChatGPT-4 fraud classification using CEO letters and risk disclosures. Varies prompting and incorporates human feedback, without complex fine-tuning.	Reports approximately 67% sensitivity, specificity and F-measure with human feedback.

Continued on next page

Table A12 continued. Recent AI Experiments in Accounting

Source date / study focus	Paper and publication status	Methods	Findings
2024-06-27 AI capability	Eulerich, Sanatizadeh, Vakilzadeh and Wood Is it all hype? ChatGPT's performance and disruptive potential in the accounting and auditing industries RAST (2024), 29(3). 2318-2349. Published.	Benchmarks CPA, CMA, CIA and EA exam questions. Compares GPT-3.5/GPT-4, 10-shot examples and reasoning/acting with external tools.	Average scores rise from 53.1% for GPT-3.5 to 85.1% with the combined enhancements, passing the tested sections.
2024-07-25 AI capability	Kim, Muhn and Nikolaev Financial Statement Analysis with Large Language Models arXiv 2407.17866. WITHDRAWN. Withdrawn working paper.	Standardized, anonymized statements. GPT-4 earnings-direction forecasts with reasoning prompts. Comparisons with analysts and statistical/ML forecasts.	The earlier draft claimed superior predictive performance. Withdrawn on 2025-02-20 after a coauthor identified data/analysis inconsistencies.
2024-09 Human response	Estep, Griffith and MacKenzie How do financial executives respond to the use of artificial intelligence in financial reporting and auditing? RAST (2024), 29(3). 2798-2831. Published.	Financial-executive survey plus randomized experiment varying company AI use and auditor AI use in a complex accounting-estimate adjustment.	Managers at companies described as using AI accept larger adjustments when the auditor also uses AI. Auditor AI use has no detected effect without company AI. First published online in May 2023. Issued in 2024.
2024-09-12 Human response	Commerford, Eilifsen, Hatfield, Holmstrom and Kinserdal Control issues. How providing input affects auditors' reliance on artificial intelligence CAR (2024), 41(4). 2134-2162. Published.	Experienced-auditor experiment compares AI evidence with human-specialist evidence and varies opportunity to provide input. Measures locus of control. Corroborating interviews.	Auditors rely less on AI evidence. Providing input mitigates aversion for auditors with external locus of control, with little benefit for those with internal locus of control.
2025-01-06 AI judgement	MacKenzie, Moon, Rykowski, Swanquist and Whited A Comparison of Artificial Intelligence and Human Responses in Audit Experiments SSRN 5036687. Working paper. Revised 2026-07-08.	Practitioner interviews and published audit experiments used to compare ChatGPT response patterns with auditors. Reflects the revised 2026 paper.	ChatGPT can complete the tasks, but recovery depends on the underlying bias. motivational response patterns are not reproduced, while some information-weighting patterns are.
2025-06-14 AI capability	Wang, Liu, Zhao, Li and Zhang Automating Financial Statement Audits with Large Language Models arXiv 2506.17282. Repository paper.	Real financial tables combined with synthetic transactions. Five-stage audit evaluation tests error detection, explanations, standards mapping and statement revision.	Models can locate errors when transaction evidence is supplied, but struggle with explanations, appropriate accounting-standard citations, complete audits and corrective revisions.
2025-06-25 AI capability	Cao, Chen, Tucker and Wan Can generative AI help identify peer firms? RAST (2025), 30(4). 3344-3386. Published. Repository version.	Prompts generative AI (Bard) for product-market peers. Compares with expert peers, established systems, business-description similarity and SIC classifications. Validates economic comovement.	Generated peers overlap substantially with expert choices and exhibit useful subsequent economic comovement, particularly for large firms.

Continued on next page

Table A12 continued. Recent AI Experiments in Accounting

Source date / study focus	Paper and publication status	Methods	Findings
2025-08 Human response	Bertomeu, Lin, Liu and Ni The impact of generative AI on information processing. Evidence from the ban of ChatGPT in Italy JAE (2025), 80(1). 101782. Published.	Natural experiment from the unexpected Italian ChatGPT ban. Compares domestic and foreign analysts covering the same firms, alongside AI-text measures and market outcomes.	The ban coincides with fewer and less accurate domestic analyst forecasts, more industry-level information use, and weaker market efficiency.
2025-09-11 Human response	Croom Interactivity and Illusions of Ability. How Using Generative AI Affects Investor Judgments JAR (2026), 64(2). 681-719. Published.	Three laboratory experiments manipulate FinGPT interface interactivity while controlling provided information. Distinguish asking questions from broader interactivity. Vary task complexity.	Greater interactivity inflates self-assessed processing/investing ability and willingness to invest without improving actual information processing. Some processing outcomes worsen. The mechanism is misattribution of tool ability.
2025-10-10 AI capability	Wang et al. FinAuditing. A Financial Taxonomy-Structured Multi-Document Benchmark for Evaluating LLMs arXiv 2510.08886. SIGIR 2026 accepted. Conference-accepted repository version, revised 2026-05-17.	Evaluates 13 LLMs on 1,102 XBRL-based instances averaging over 33,000 tokens. Semantic matching, relation extraction and mathematical reasoning.	Finds substantial weaknesses in concept retrieval, taxonomy relations and consistent cross-document reasoning.
2025-10-17 Human response	Ulla and Commerford Reliance on Algorithmic Estimates. The Joint Influence of Algorithm Adaptability and Estimation Uncertainty TAR (2025), 100(6). 285-308. Published.	Two experiments vary learning/adaptive versus static algorithms and estimation uncertainty. Auditor setting followed by an accounting-estimate preparation setting.	Participants rely more on learning algorithms when uncertainty is high. Reliance depends jointly on algorithm adaptability and the estimation environment.
2025-11 Human response	Cheng, Lin and Zhao Does generative AI facilitate investor Trading? Early evidence from ChatGPT outages JAE (2025), 80(2-3). 101821. Published.	Uses ChatGPT service outages as interruptions to GenAI availability. Examines intraday trading, news timing, ownership and price informativeness.	Trading volume, short-run price impact and return variance fall during outages, consistent with less informed trading.
2026-01-14 AI capability	Bernard, Blankespoor, de Kok and Toynbee Using GPT to Measure Business Complexity TAR (2026), 101(3). 67-102. Published.	Fine-tunes an autoregressive model on narrative disclosures and iXBRL tags. Uses tag-prediction confidence as complexity. Validation includes rewritten disclosure inputs and archival applications.	The measure captures complexity beyond simple readability and is associated with slower price formation. Complex debt features also correlate with persistent interest expense and resilience when lending conditions deteriorate.
2026-03-11 AI capability	Malarkkan et al. FinRule-Bench. A Benchmark for Joint Reasoning over Financial Tables and Principles arXiv 2603.11339. Repository paper.	Pairs real financial statements with curated accounting principles. Zero-/few-shot tests of rule verification, violated-rule selection and joint diagnosis. Counterfactual-consistency protocol.	Strong isolated verification does not carry over to distinguishing rules or detecting and localizing multiple simultaneous violations. Accuracy deteriorates as diagnostic demands increase.

Continued on next page

Table A12 continued. Recent AI Experiments in Accounting

Source date / study focus	Paper and publication status	Methods	Findings
2026-03-11 Human response	Peecher, Pietsch, Stirnkorb and Yamoah Coping With Changing Skill Requirements. Does Disaffirmation Versus Affirmation Affect Auditors' Reliance on AI -Supported Advice From Specialists? CAR (2026), 43(2). 659-679. Published.	Two auditor experiments manipulate specialist reliance on AI and self-affirmation/disaffirmation of professional skills. Measure reliance on advice in a fair-value case.	Disaffirmed auditors discount advice more when specialists rely heavily on AI. Affirming traditional audit skills mitigates the defensive response and associated denigration of specialist competence.
2026-04-07 AI capability	Blankespoor, de Haan and Li Generative AI in Financial Reporting JAR (2026), 64(3). 1189-1232. Published.	Injects known amounts of GPT-3.5-rewritten text at document/sentence levels. Tests GPTZero detection power and errors. Then applies the validated measure to corporate disclosures.	Detection works for small amounts of AI text in the studied settings. Estimated AI writing appears in five disclosure classes, reaching up to 4.5% of new text in 2024. These prevalence estimates depend on detector validity.
2026-04-10 AI capability	Dou and Fu Supervised Fine-tuning of Large Language Models on Financial Statements to Predict Future Earnings Changes SSRN 6484841. Working paper. Revised 2026-10-17.	Supervised fine-tuning on anonymized statements and realized next-year earnings directions. Evaluates after model release and compares with three tabular ML models. Probes temporal and accounting-label knowledge.	Reports out-of-sample AUC of 67.88%, above the tabular benchmarks. Authors attribute much of the gain to prior knowledge of temporal structure and accounting terminology.
2026-04-16 Human response	Choi and Xie Human + AI in Accounting. Early Evidence from the Field JAR (2026), 64(3). 1333-1373. Published.	Survey of 277 accountants and platform field records. Framed field experiment randomizes 99 accountants to classify 43 transactions with or without AI recommendations.	AI assistance increases classification accuracy by about 17.5 percentage points. Field adoption correlates with productivity and reporting improvements. Less accurate non-consensus recommendations can propagate errors.
2026-04-16 Human response	Lin, Cheng and Humphreys Tough Ratings, Tougher Sell. How Different Types of Adjustment Affect Managers' Asymmetric Algorithm Use in Performance Evaluation Judgments TAR (2026), 101(3). 413-440. Published.	Human experiments vary high/low algorithm-advised performance ratings and decision rights to adjust the computation, the final rating, or neither.	Managers avoid algorithmic advice more when ratings are low. Allowing changes to how the rating is computed increases use of low-rating advice, consistent with improved perceived understanding.
2026-04-17 AI judgement Human response	Bertomeu, Cheynel, Lunawat and Milone On Humans and AI. A Financial Reporting Dilemma SSRN 6590198. Working paper.	Humans and LLMs act as CFOs choosing whether to reverse a biased but understood reporting policy. Varies ethical guidance, management pressure and advice explanations.	Models prefer truthful reporting more consistently than humans and show greater judgment-policy coherence. Humans are more receptive to explained model advice and may discount unexplained advice.
2026-04-22 Human response	Chang, Dong, Martin and Zhou AI Democratization and Trading Inequality JAR (2026), 64(3). 1287-1331. Published.	Extracts AI sentiment from earnings calls. Compares retail/short-sale alignment around ChatGPT deployment and exploits exogenous service outages.	Retail trading becomes more aligned with AI sentiment, especially with high processing costs. Alignment falls during outages. Higher alignment accompanies lower asymmetry and improved retail profitability.

Continued on next page

Table A12 continued. Recent AI Experiments in Accounting

Source date / study focus	Paper and publication status	Methods	Findings
2026-05-06 AI capability	Levy Caution Ahead. Numerical Reasoning and Look-Ahead Bias in AI Models JAR (2026), 64(3). 1139-1188. Published. Repository version.	Accounting-arithmetic and ratio tasks. Code-interpreter intervention. LLMs trained from scratch with yearly knowledge cutoffs. Least-significant-digit perturbations. Gradient-boosted-tree benchmarks.	Numerical accuracy deteriorates with task length/complexity and improves when calculation is delegated to tools. Timing-controlled models and perturbations reveal memorization/look-ahead problems in earnings prediction.
2026-06-04 AI capability	Ciconte, Rozario and Urcan Using AI to identify exogenous shocks and conduct archival accounting research RAST (2026), 31(3). 1638-1676. Published.	Prompts Claude 4-Sonnet to identify securities-law shocks, analyze supplied data and generate papers. Contrasts U.S./non-U.S. laws. Compares models and human/AI referee assessments.	AI generates polished studies with spurious causal claims, including for implausible non-U.S. shocks. Human and AI reviewers also have difficulty detecting these problems.
2026-06-14 AI capability	Tumpati et al. FinBalance. A Multi-Document Accounting Reconciliation Benchmark arXiv 2606.15949. Repository paper.	Generated source-document bundles with exact ledger-derived entries, balance sheets and 23 inconsistency labels. Six LLMs on 710 evaluation records. Citation-pressure and ledger-feedback ablations.	Best exact balance-sheet accuracy is 46%. Reported balance sheets often conflict with models' own entries. Ledger feedback helps more than citation pressure. Separates plausible-looking accounting from consistent reconciliation.
2026-07-22 AI capability	Tong et al. Are the Financial Reasoning from LLMs Credible? A Real World Test over Long-Horizon Statements arXiv 2607.28661 (FinIndices). Repository paper.	Uncropped statements up to 32,000 tokens. Single-index versus full-table computations. Adversarial traps, removal of formula hints and supervised fine-tuning.	Removing hints and increasing tabulation demands expose temporal, column-selection and stock-flow errors. Fine-tuning partly improves zero-hint performance. Formula recall is insufficient for reliable statement reasoning.
2026-07-30 Human response	Samet, Williamson and Yip AI-Augmented Design and the Expertise Bias in Subjective Evaluations of Creative Output TAR (2026), 101(5). 445-465. Published. Repository version.	Multiple human experiments hold creative output constant while varying designer expertise and awareness/availability/use of AI augmentation.	AI availability reduces expertise-based favoritism by reducing perceived exclusivity of expertise. Designers can restore this bias by declining AI use.
2026-08-11 AI capability	Johnson, Jiang, Chaudhuri, Chen, Falvey and O'Cofaigh Long-Horizon Forecasting of Complete Financial Statements with Forma arXiv 2608.11327. Repository paper.	ProForma-20Q benchmark predicts 78 line items over 1–20 quarters. Specialized tuple-based transformer versus ML, time-series and frontier-LLM baselines. Accounting-identity checks and scenario interventions.	The specialized model outperforms the tested baselines, with larger gains at long horizons. Accounting coherence can be enforced without a statistically significant accuracy loss.

D Model Releases and Cognitive Benchmarks

The following timelines show how reported model performance has evolved across academic knowledge, mathematics, and scientific reasoning.

D.1 Timeline 1. Academic Knowledge and Mathematical Reasoning

MMLU tests knowledge and problem solving across 57 subjects. GSM8K tests grade-school mathematical word problems. And MATH tests competition-style mathematical problem solving.

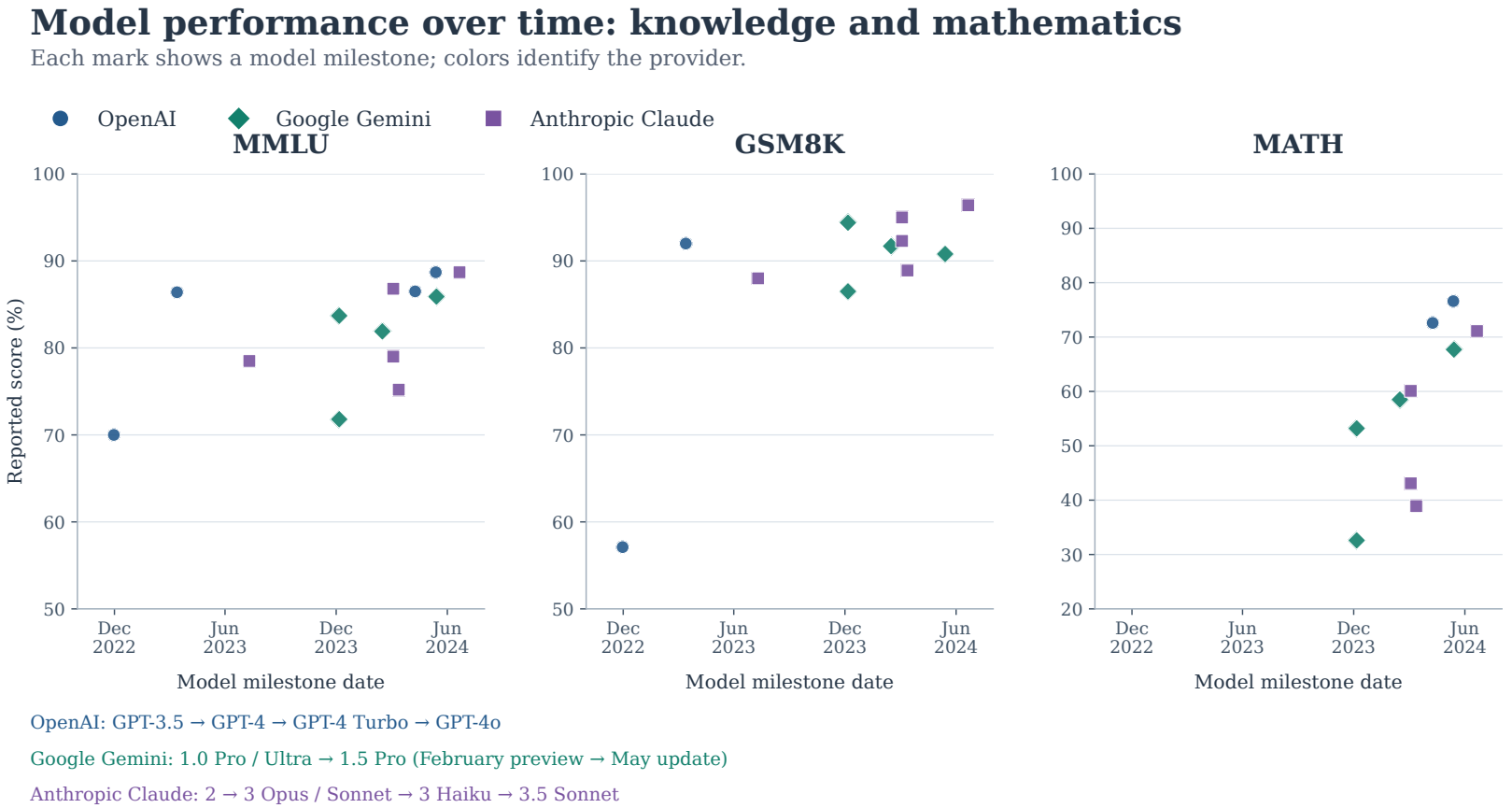


Figure A2. This figure reports MMLU, GSM8K, and MATH scores over model milestones. Scores are percentages. Colors and symbols identify providers, and the sequences below the panels identify the models. Table A13 gives evaluation settings and sources.

Table A13. Timeline 1. Academic Knowledge and Mathematical Reasoning.

This table reports model milestones and reported MMLU, GSM8K, and MATH scores. Scores are percentages, with higher values indicating better performance. The date identifies a release, preview, or update. The last column gives the evaluation settings and source. Shot counts refer to prompt examples. CoT denotes chain-of-thought prompting. Dashes denote unreported scores.

Milestone date	Provider / model	MMLU	GSM8K	MATH	Evaluation settings and source
2022-11-30	OpenAI GPT-3.5 (report baseline)	70.0	57.1	—	Family launch date. Benchmark reference published March 2023. MMLU and GSM8K. 5-shot. Evaluation source.
2023-03-14	OpenAI GPT-4 (original report)	86.4	92.0	—	MMLU. 5-shot. GSM8K. 5-shot CoT. Report notes training on part of the GSM8K training set. Academic evaluations use the base model. Evaluation source.
2023-07-11	Anthropic Claude 2	78.5	88.0	—	MMLU. 5-shot CoT, as reproduced in Google's report. GSM8K. 0-shot. Also reported by Anthropic at launch. Evaluation source.
2023-12-06	Google Gemini 1.0 Pro	71.8	86.5	32.6	MMLU. 5-shot. GSM8K. majority vote over 32 attempts. MATH. 4-shot. Original technical-report configuration. Evaluation source.
2023-12-06	Google Gemini 1.0 Ultra	83.7	94.4	53.2	Announcement, before broad February 2024 availability. MMLU. 5-shot. GSM8K. majority vote over 32. MATH. 4-shot. Headline MMLU 90.04 uses a different voting protocol. Evaluation source.
2024-02-15	Google Gemini 1.5 Pro (February preview)	81.9	91.7	58.5	February preview snapshot. MMLU. 5-shot. GSM8K. 11-shot. MATH. 4-shot. Original figures reproduced in Claude 3's card. Evaluation source.
2024-03-04	Anthropic Claude 3 Opus	86.8	95.0	60.1	MMLU. 5-shot. GSM8K and MATH. 0-shot CoT. MMLU with 5-shot CoT is separately reported as 88.2. Evaluation source.
2024-03-04	Anthropic Claude 3 Sonnet	79.0	92.3	43.1	Original March card. MMLU 5-shot. GSM8K and MATH 0-shot CoT. Evaluation source.
2024-03-13	Anthropic Claude 3 Haiku	75.2	88.9	38.9	MMLU. 5-shot. GSM8K and MATH. 0-shot CoT. Evaluation source.
2024-04-09	OpenAI GPT-4 Turbo (April snapshot)	86.5	—	72.6	MMLU and MATH. 0-shot CoT as reproduced in Anthropic's June comparison. April API snapshot. Evaluation source.
2024-05-13	OpenAI GPT-4o (May release)	88.7	—	76.6	MMLU and MATH. 0-shot CoT in the June comparison. Evaluation source.
2024-05-14	Google Gemini 1.5 Pro (May update)	85.9	90.8	67.7	Updated model. MMLU 5-shot. GSM8K 11-shot. MATH 4-shot Minerva prompt. Evaluation source.
2024-06-21	Anthropic Claude 3.5 Sonnet (initial)	88.7	96.4	71.1	Date shown on the current announcement page. Snapshot ID is dated June 20. MMLU 5-shot. Math 0-shot CoT. MMLU 90.4 with CoT is a different protocol. Evaluation source.

D.2 Timeline 2. Scientific and Multidisciplinary Reasoning

GPQA Diamond is the 198-question expert-vetted subset of GPQA's difficult science questions. Humanity's Last Exam (HLE) is a harder multidisciplinary benchmark that includes text and multimodal questions. HLE was introduced in January 2025. The timeline follows the rise of reasoning models and the addition of search and code tools.

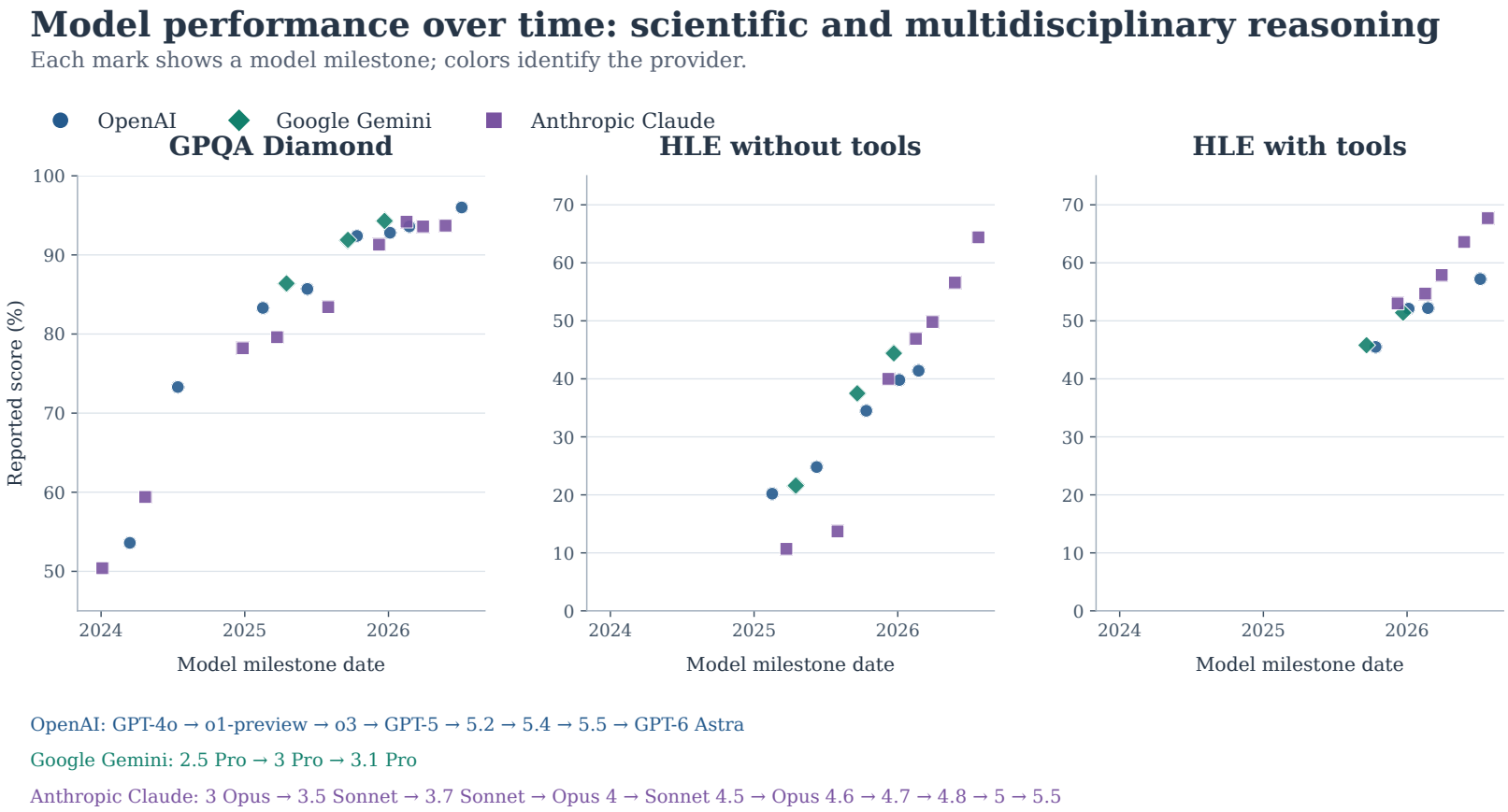


Figure A3. This figure reports GPQA Diamond and HLE scores over model milestones. Scores are percentages. The HLE panels distinguish evaluations with and without tools. Table A14 gives model identities, settings, and sources.

Table A14. Timeline 2. Scientific and Multidisciplinary Reasoning.

This table reports model milestones and reported GPQA Diamond and Humanity's Last Exam scores. Scores are percentages. The HLE columns separate evaluations without tools and with search or code tools. The last column identifies reasoning settings and sources. Dashes denote unreported scores.

Milestone date	Provider / model	GPQA Diamond	HLE no tools	HLE with tools	Evaluation settings and source
2024-03-04	Anthropic Claude 3 Opus	50.4	—	—	GPQA Diamond. 0-shot CoT, mean of 10 rollouts. Evaluation source.
2024-05-13	OpenAI GPT-4o (May release)	53.6	—	—	GPQA Diamond. 0-shot CoT in Anthropic's June comparison. Evaluation source.
2024-06-21	Anthropic Claude 3.5 Sonnet (initial)	59.4	—	—	GPQA Diamond. 0-shot CoT. Announcement page date. June 20 snapshot. Evaluation source.
2024-09-12	OpenAI o1-preview	73.3	—	—	Released preview, pass@1. Evaluation source.
2025-02-24	Anthropic Claude 3.7 Sonnet (thinking)	78.2	—	—	64K extended thinking. Pass@1 averaged over trials. Evaluation source.
2025-04-16	OpenAI o3 (high)	83.3	20.2	—	No-tool figures in OpenAI's later GPT-5 developer comparison. Evaluation source.
2025-05-22	Anthropic Claude Opus 4 (thinking)	79.6	10.7	—	32K thinking. Figures reproduced in Google's June 2.5 Pro card. HLE without tools. Evaluation source.
2025-06	Google Gemini 2.5 Pro (June/GA)	86.4	21.6	—	June/GA snapshot. GPQA pass@1. The HLE no-tool designation is explicit in Google's later Gemini 3 comparison. Evaluation source.
2025-08-07	OpenAI GPT-5 (high reasoning)	85.7	24.8	—	Developer report, high reasoning. HLE follows the report's revised test version. Evaluation source.
2025-09-29	Anthropic Claude Sonnet 4.5 (thinking)	83.4	13.7	—	Google's November comparison. Evaluation source.
2025-11-18	Google Gemini 3 Pro (thinking)	91.9	37.5	45.8	Full HLE set, including multimodal items. The tool-assisted result uses search and code. Evaluation source.
2025-12-11	OpenAI GPT-5.2 Thinking	92.4	34.5	45.5	Reported high-end reasoning configuration. HLE with tools uses search and Python. Evaluation source.
2026-02-05	Anthropic Claude Opus 4.6	91.3	40.0	53.0	Own system card. adaptive thinking, max effort. HLE tool run uses search/fetch/code and compaction. Evaluation source.
2026-02-19	Google Gemini 3.1 Pro (thinking)	94.3	44.4	51.4	Official model card. HLE full text-plus-multimodal set. No tools versus search with a blocklist plus code. Evaluation source.
2026-03-05	OpenAI GPT-5.4	92.8	39.8	52.1	xhigh reasoning in the reported research environment. Tool and no-tool HLE remain separate. Evaluation source.
2026-04-16	Anthropic Claude Opus 4.7	94.2	46.9	54.7	GPQA averaged over 10 trials. HLE at max effort, auto thinking, 1M total-token cap. No context compaction. Evaluation source.
2026-04-23	OpenAI GPT-5.5	93.6	41.4	52.2	xhigh reasoning, research environment. Evaluation source.

Continued on next page

Table A14 continued. Timeline 2. Scientific and Multidisciplinary Reasoning

Milestone date	Provider / model	GPQA Diamond	HLE no tools	HLE with tools	Evaluation settings and source
2026-05-28	Anthropic Claude Opus 4.8	93.6	49.8	57.9	GPQA average over 25 trials. HLE figures from the system-card summary. Evaluation source.
2026-07-24	Anthropic Claude Opus 5	93.7	56.6	63.6	HLE values in the later Opus 5.5 system card. GPQA in OpenAI's Astra comparison. Evaluation source. Additional score source.
2026-09-03	OpenAI GPT-6 Astra	96.0	—	57.2	Academic table reports GPQA and tool-assisted HLE. Evaluation source.
2026-09-22	Anthropic Claude Opus 5.5	—	64.4	67.7	Own system card. adaptive thinking at max effort, average of five trials. Tool HLE uses search/fetch/code with a 1M-token cap. Evaluation source.
2026-09-29	OpenAI GPT-6.1 Sol	—	—	—	Launch comparisons emphasize professional work and agentic tasks. Evaluation source.
2026-10-13	Google Gemini 4 Argon	—	—	—	Broad-availability milestone. The evaluation suite emphasizes professional and agentic tasks. Evaluation source.
2026-10-28	Google Gemini 4 Carbon	—	—	—	Announcement/limited early access. The evaluation report introduces a new benchmark suite. Evaluation source.

E Emergence and Scope of AI Benchmarks

The visual timelines trace the emergence of harder questions and complete-workflow evaluations. The accompanying tables identify each benchmark's tasks, release dates, and sources.

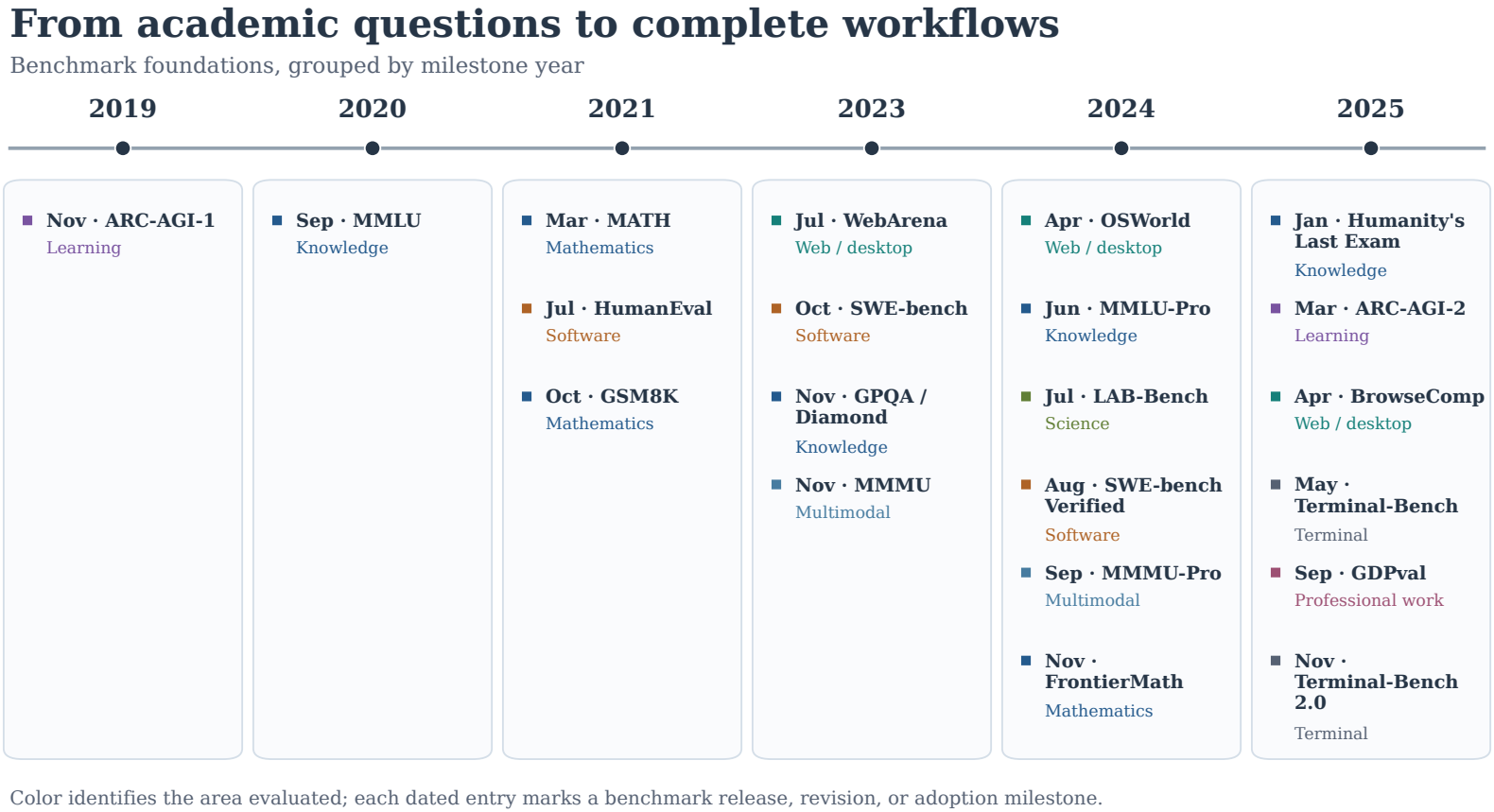


Figure A4. This figure reports the emergence of benchmark families from 2019 through 2025. Entries identify the milestone month and benchmark. Colors identify the capability area. Table A15 gives the tasks and sources.

Table A15. Benchmark Timeline 1. Selected Foundations, 2019–2025.

This table reports benchmark foundations from 2019 through 2025. Each row identifies the dated release or publication, the capability area, the task evaluated, and the benchmark's development. Source links accompany the entries.

Date / event	Benchmark	Primary area	What it tests	Development
2019-11-05 Paper	ARC (now ARC-AGI-1)	Adaptive learning	Infer a new grid transformation from a few examples, then solve unseen cases.	Measures skill acquisition through unfamiliar static input/output puzzles. Source.
2020-09-07 Paper	MMLU	Academic knowledge	Multiple-choice knowledge and reasoning across 57 subjects.	Broad subject coverage becomes a reference point for language-model evaluation. Original English MMLU is distinct from later Pro variants. Source.
2021-03-05 Paper	MATH	Mathematics	Solve held-out competition mathematics problems. The complete dataset contains 12,500 problems and worked solutions.	Moves beyond simple arithmetic toward multistep competition reasoning. Source.
2021-07-07 Paper	HumanEval	Software engineering	Generate Python functions from docstrings. Execute hidden tests for 164 problems.	Introduces execution-based functional correctness at function scale. Pass@k depends on the allowed number of samples. Source.
2021-10-27 Paper	GSM8K	Mathematics	Solve grade-school word problems requiring several arithmetic reasoning steps.	A reproducible mathematical reasoning target below competition/research level. Its original dataset contains 8,500 problems. Source.
2023-07-25 Paper	WebArena	Web interaction	Complete user goals on realistic, stateful websites using browser actions.	Extends assessment from answering a question to changing an environment. Browser workflow completion differs from fact retrieval. Source.
2023-10-10 Paper	SWE-bench	Software engineering	Modify real repositories to resolve GitHub issues. Verify the resulting patch with tests.	Moves from isolated functions to repository-level debugging and engineering. Source.
2023-11-20 Paper	GPQA family / Diamond subset	Academic knowledge	Graduate-level biology, chemistry, and physics questions. Diamond is the 198-question expert-vetted subset.	Raises the difficulty of closed-ended scientific reasoning. Source.
2023-11-27 Paper	MMMU	Multimodal understanding	Answer college-level questions that combine text with images, charts, diagrams, and other visual material.	Adds visual perception plus subject-specific reasoning. Understanding an image is distinct from operating a desktop interface. Source.
2024-04-11 Paper	OSWorld	Desktop applications	Carry out open-ended tasks in real computer environments spanning applications and operating systems.	Tests GUI actions and final execution state across 369 tasks. Source.
2024-06-03 Paper	MMLU-Pro	Academic knowledge	Answer a more difficult, reasoning-focused revision of broad academic questions, with ten answer options.	Improves discrimination and robustness relative to original MMLU. Source.

Continued on next page

Table A15 continued. Benchmark Timeline 1. Selected Foundations, 2019–2025

Date / event	Benchmark	Primary area	What it tests	Development
2024-07-14 Paper	LAB-Bench	Scientific work	Answer practical biology research questions on literature, figures, databases, and biological sequences.	Over 2,400 multiple-choice items target research-related skills. Source.
2024-08-13 Validated subset	SWE-bench Verified	Software engineering	Resolve 500 human-validated software issues selected from SWE-bench.	Improves task specification and evaluation reliability. Source.
2024-09-04 Paper	MMMU-Pro	Multimodal understanding	A more robust visual-understanding evaluation, including vision-only inputs and harder answer selection.	Reduces text-only shortcuts and answer guessing. Its scoring and task format differ from original MMMU. Source.
2024-11-07 Paper	FrontierMath	Mathematics	Solve original, expert-authored research-level mathematics problems with verifiable answers.	Extends beyond school/competition mathematics. Source.
2025-01-24 First preprint	Humanity's Last Exam (HLE)	Academic knowledge	Difficult, closed-ended multidisciplinary questions, including images. The Nature version describes 2,500 questions.	Adds harder expert-level coverage as older exams saturate. Nature publication on 2026-01-28 is a later publication event. Sources. [1], [2].
2025-03-24 Launch	ARC-AGI-2	Adaptive learning	Infer deeper, compositional rules in unfamiliar static grid puzzles.	Hardens the ARC task family while retaining the static input/output format. ARC-AGI-3 later adds interaction. Source.
2025-04-10 Launch	BrowseComp	Web research	Find hard-to-locate facts across the open web. 1,266 questions with short, checkable answers.	Measures persistent browsing and evidence discovery. Source.
2025-05-19 Launch	Terminal-Bench (initial release)	Terminal workflows	Complete realistic command-line tasks inside executable environments. The launch set contained 80 tasks.	Evaluates multistep execution and tests the resulting state. Sources. [1], [2].
2025-09-25 Launch	GDPval	Professional work	Produce economically valuable work products across 44 occupations and nine industries.	Moves toward documents, spreadsheets, and other professional deliverables, graded against expert work. GDPval-AA is a separate implementation. Source.
2025-11-07 New version	Terminal-Bench 2.0	Terminal workflows	A harder terminal task set with revised evaluation infrastructure and the Harbor framework.	Shows that environment orchestration and task maintenance evolve alongside benchmark content. Sources. [1], [2].

E.1 Benchmark Development in 2026

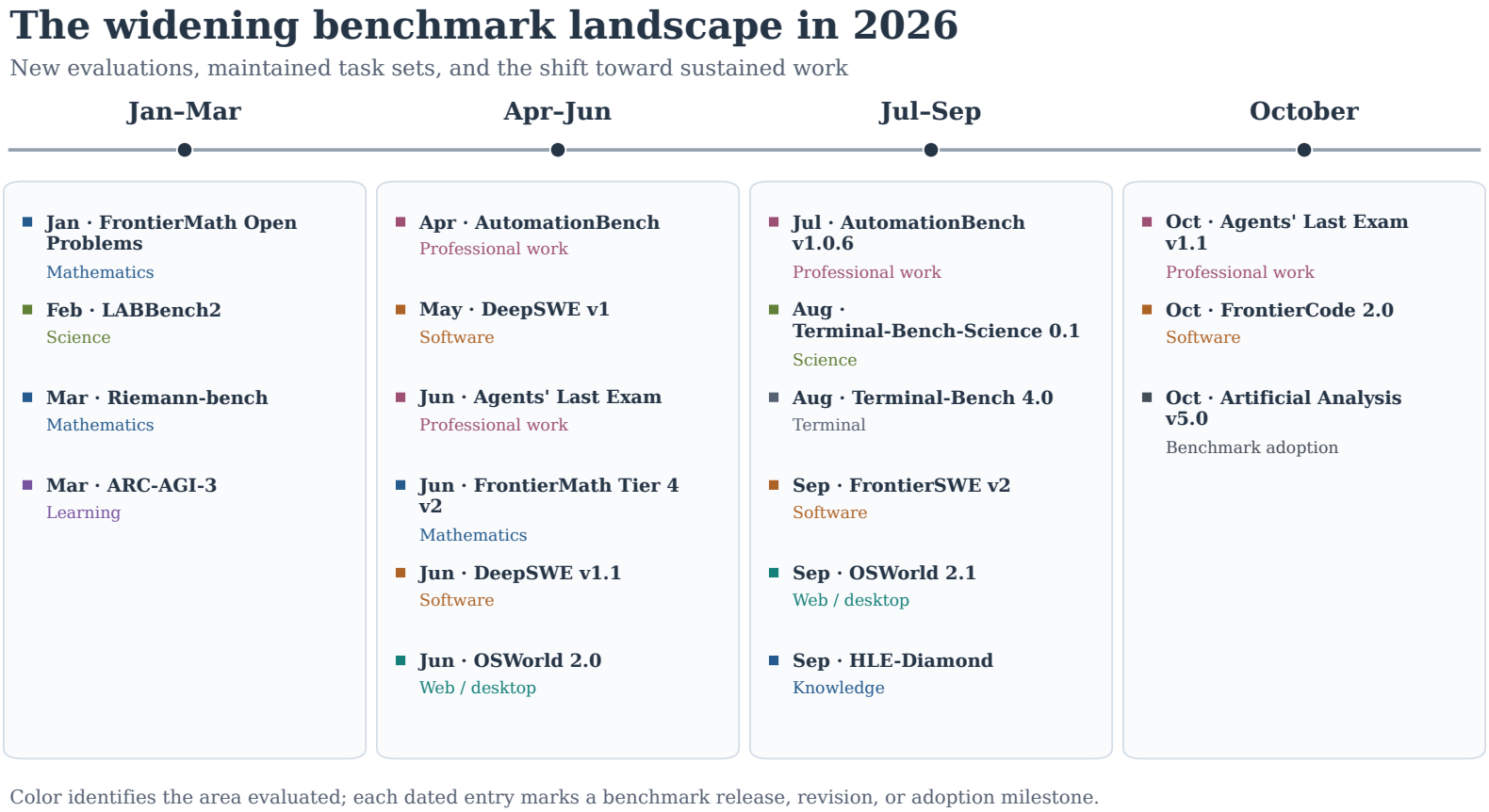


Figure A5. This figure reports benchmark introductions, revisions, and adoption milestones in 2026. The columns group events by period, and colors identify the evaluated area. Table A16 gives event dates, tasks, and sources.

Table A16. Benchmark Timeline 2. New Benchmarks, Revisions, and Adoption in 2026.

This table reports benchmark introductions, revisions, and adoption milestones in 2026. Rows identify the dated event, evaluated area, task, and development. The event labels distinguish first releases, revised task sets, grading changes, and index adoption.

Date / event	Benchmark	Primary area	What it tests	Development
2026-01-27 Pilot launch	FrontierMath. Open Problems	Mathematics	Seek solutions to mathematical problems unsolved at release, with programmatically verifiable candidate answers.	Extends from known-answer research problems to potential new mathematical results. A distinct benchmark from FrontierMath Tiers 1-4. Source.
2026-02-04 First preprint	LABBench2	Scientific work	Perform useful biology research tasks. Expands and revises the earlier LAB-Bench design.	Targets practical scientific abilities beyond textbook recall. Source.
2026-03-24 Dated launch post	Riemann-bench	Mathematics	Solve private research mathematics problems written and independently checked by domain experts.	Surge's post carries March 24 plus an October 21 update date. The first preprint is April 8. Sources. [1], [2].
2026-03-25 Competition launch	ARC-AGI-3	Adaptive learning	Explore unfamiliar interactive environments, discover goals and dynamics, and solve them efficiently.	Tests active exploration and learning from feedback. Date marks the full competition launch, after previews. The technical report is April 22. Sources. [1], [2].
2026-04-20 Launch	AutomationBench	Business workflows	Orchestrate cross-application business processes through tools/APIs across sales, marketing, operations, support, finance, and HR.	Scores final environment state against deterministic criteria. Source.
2026-05-26 Launch	DeepSWE v1	Software engineering	Complete original, long-horizon software engineering tasks in real codebases.	Broadens beyond short fixes using newly authored tasks and task-specific verification. Date is the launch post, before the July research paper. Source.
2026-06-03 First preprint	Agents' Last Exam (ALE)	Professional work	Complete sustained, economically useful professional workflows with verifiable outcomes.	Adds end-to-end performance across diverse professions. This paper date is separate from the October v1.1 release. Source.
2026-06-12 Corrected dataset	FrontierMath Tier 4 (v2)	Mathematics	An updated set of 43 exceptionally difficult problems, separate from the 295 Tiers 1-3 problems.	Corrects problems in the broader dataset. The Tier 4 private/public split matters. Source.
2026-06-14 Grading revision	DeepSWE v1.1	Software engineering	Retains the long-horizon engineering tasks but evaluates committed patches in a separate, clean verifier environment.	Improves reproducibility, test reporting, and resistance to shortcuts. Source.
2026-06-28 First preprint	OSWorld 2.0	Desktop applications	Complete 108 longer workflows involving websites and professional desktop applications.	Adds realistic state, cross-source constraints, and longer horizons. Source.
2026-07-31 Task/verifier revision	AutomationBench v1.0.6	Business workflows	Refines task discoverability, API fidelity, constraints, and success checks in the workflow environments.	Repairs unfair or overly strict tasks and adjusts private-set difficulty. Source.

Continued on next page

Table A16 continued. Benchmark Timeline 2. New Benchmarks, Revisions, and Adoption in 2026

Date / event	Benchmark	Primary area	What it tests	Development
2026-08-27 Launch	Terminal-Bench-Science 0.1	Scientific work	Execute 70 scientific workflows across life, physical, Earth, mathematical, and engineering sciences.	Grades artifacts such as code, analyses, simulations, and data products with task-specific tests. Extends terminal evaluation into research practice. Sources. [1], [2].
2026-08-28 Major revision	Terminal-Bench 4.0	Terminal workflows	A maintained terminal task set with recalibrated resources, task fixes, and saturated-task removals.	The authors set an eight-hour task timeout and require re-running trials after breaking changes. Sources. [1], [2].
2026-09-02 New version	FrontierSWE v2	Software engineering	Complete 34 demanding implementation, optimization, scientific-computing, and AI-research tasks.	Uses a 20-hour agent budget with a revised harness. Expands scope and retires some earlier tasks. Distinct from Cognition's FrontierCode. Source.
2026-09-16 Release	OSWorld 2.1	Desktop applications	Updates the OSWorld 2.0 workflows, agents, assets, mocked websites, and evaluation pipeline.	A reproducibility/maintenance release, recommended by the maintainers. Source.
2026-09-22 Refined subset release	HLE-Diamond	Academic knowledge	A cleaned 1,000-question HLE subset. 500 reasoning and 500 expert-knowledge questions.	Introduces a focused reasoning and expert-knowledge subset within the HLE family. Source.
2026-10-26 Version release	Agents' Last Exam v1.1	Professional work	122 public professional-workflow tasks spanning 55 sub-industries.	One of the newest dated releases in this selection. Source.
2026-10-27 New version	FrontierCode 2.0	Software engineering	Assess whether repository maintainers would accept an agent's code, including correctness, design, and inferred user intent.	Expands task scope and grading criteria. Uses weighted rubric scores and blocking criteria. Source.
2026-10 Index update	Artificial Analysis Index v5.0	Cross-area adoption	Adds HLE-Diamond, Terminal-Bench-Science, and FrontierCode 2.0 to the aggregate evaluation suite.	Shows the shift in benchmark adoption. Removes original HLE and GDPval-AA from the index. Standalone evaluations can remain available. Source.

Table A17. Capability Areas and Recent Benchmark Developments.

This table reports the capability areas covered by the benchmark timelines. It identifies the work required, benchmark families, recent milestones, and the evaluated outcomes, including answer accuracy, task completion, rubric scores, and comparison-based ratings.

Capability area	What the tasks require	Benchmark families	Recent benchmark milestone	Evaluated outcome
Academic knowledge and reasoning	Answer difficult closed-ended questions across science and other academic subjects.	MMLU / MMLU-Pro. GPQA Diamond. HLE	HLE-Diamond (2026-09-22). a refined 1,000-question set.	Answer accuracy across academic subjects. Sources. [1], [2], [3].
Advanced mathematics	Solve competition, research-level, or genuinely open mathematical problems.	GSM8K. MATH. FrontierMath. Riemann-bench	Riemann-bench. March 24 post, updated October 21, 2026. FrontierMath Tier 4 v2. June 12. Open Problems pilot. January 27, 2026.	Correct answers and verified mathematical solutions. Sources. [1], [2], [3], [4].
Learning unfamiliar tasks	Infer new rules and goals, adapt from examples or feedback, and act efficiently.	ARC-AGI-1. ARC-AGI-2. ARC-AGI-3	ARC-AGI-3 full competition (2026-03-25), following previews.	Grid-solution accuracy and interactive action efficiency. Sources. [1], [2], [3].
Multimodal understanding	Interpret images, charts, diagrams, and text jointly using domain knowledge.	MMMU. MMMU-Pro	MMMU-Pro (2024-09-04).	Answer accuracy for joint visual and textual inputs. Sources. [1], [2].
Software engineering	Debug or extend codebases, implement complex changes, and produce acceptable code.	HumanEval. SWE-bench / Verified. DeepSWE. FrontierSWE. FrontierCode	FrontierCode 2.0 (2026-10-27). FrontierSWE v2 (2026-09-02). DeepSWE v1.1 (2026-06-14).	Unit-test pass rates, performance targets, and maintainers' weighted rubrics. Sources. [1], [2], [3], [4], [5].
Terminal workflows	Use command-line tools to configure systems, transform data, and finish executable tasks.	Terminal-Bench and its versioned task sets	Terminal-Bench 4.0 (2026-08-28).	Task resolution under a fixed environment and budget. Sources. [1], [2].

Continued on next page

Table A17 continued. Capability Areas and Recent Benchmark Developments

Capability area	What the tasks require	Benchmark families	Recent benchmark milestone	Evaluated outcome
Professional deliverables	Produce work products or complete sustained occupational workflows.	GDPval. GDPval-AA. Agents' Last Exam	ALE v1.1 (2026-10-26). GDPval-AA v2.1.	GDPval-AA uses blind pairwise comparison and Elo. ALE reports pass rate plus partial credit. Sources. [1], [2], [3].
Cross-application business workflows	Coordinate CRM, email, calendars, spreadsheets, and other enterprise systems through APIs/tools.	AutomationBench	AutomationBench v1.0.6 (2026-07-31), following its April 20 launch.	Deterministic final-state checks. Sources. [1], [2].
Desktop applications	Operate graphical interfaces across websites and professional desktop applications.	OSWorld. OSWorld 2.0 / 2.1	OSWorld 2.1 (2026-09-16), maintaining the longer 2.0 workflows.	Full task completion and partial-credit scores. Sources. [1], [2], [3].
Scientific workflows	Carry out research-related literature, data, coding, analysis, and specialized-tool tasks.	LAB-Bench. LABBench2. Terminal-Bench-Science	Terminal-Bench-Science 0.1 (2026-08-27). LABBench2 first preprint (2026-02-04).	Research-question accuracy and verified workflow or artifact completion. Sources. [1], [2], [3].
Web research	Locate difficult factual evidence across the web and synthesize a checkable answer.	BrowseComp. WebArena for action-oriented browser tasks	BrowseComp (2025-04-10).	Factual-answer correctness and browser-task completion. Sources. [1], [2].

F OpenAI Interaction Style and Behavioral Quirks

Table A18 traces changes in OpenAI models' interaction style and behaviour, including task persistence, designed personas, warmth, excessive agreement, and unusual language habits. The episodes show how model behaviour changes through training, product design, and subsequent updates.

For experiments using AI participants, model version, collection date, product settings, and prompt design are part of the decision setting. Changes in interaction style can alter a response even when the accounting instrument remains fixed. Recording these features helps distinguish a change in the reporting choice from a change in the model asked to evaluate it.



Figure A6. This figure reports eight OpenAI interaction-style and behavioral episodes in chronological order. The path connects changes in task completion, personas, warmth, agreement, and language habits. Table A18 gives the dated provider accounts.

Table A18. Timeline of Selected OpenAI Interaction-Style and Behavioral Episodes.

This table reports dated OpenAI episodes involving task completion, personas, conversational warmth, excessive agreement, and unusual language. It distinguishes the event, the provider's account, and the behavioral change. Dates refer to the deployment period, retirement, or publication identified in the row.

Date / period	Example / event type	What happened	Behavioural change
2024-01-25	GPT-4 Turbo. incomplete tasks Task completion	OpenAI released gpt-4-0125-preview to improve completion of tasks such as code generation and reduce the behavior it described as laziness. Source.	Improves task persistence and code completion.
2025-04-01 to 2025-05-01	Monday persona Deliberate persona	The irreverent Monday personality was available in text and voice for a planned one-month run. The May 1 release note records its retirement. Source.	Introduces a deliberately irreverent conversational style for a one-month run.
2025-04-24 to 2025-04-29	GPT-4o. excessive agreement Unintended behavior	An April 24-25 update produced excessive agreement and inappropriate validation. OpenAI applied a prompt mitigation, began rollback on April 28, and completed it about a day later. Source.	An update increased agreement and validation, leading to mitigation and rollback.
2025-08-15	GPT-5. warmer default Style adjustment	Following feedback that GPT-5 felt too reserved, OpenAI made its default style warmer and more familiar. It reported no increase in sycophancy on its internal evaluations. Source.	Warmer and more familiar default interaction style.
2025-11-12. November observations	GPT-5.1. warmth and creature metaphors Launch / later analysis	GPT-5.1 launched with a warmer conversational style. An April 2026 account places the first clear pattern in November. goblin use rose 175% and gremlin use 52% after launch. Sources. [1], [2].	Warmer conversational style accompanied by increased creature metaphors.
2026-03 (mid-month)	Nerdy personality retired Persona retirement	OpenAI retired Nerdy after GPT-5.4 launched. Its April account dates this to mid-March and associates Nerdy training rewards with the creature-language habit. Source.	Retires the persona associated with the creature-language reward signal.
2026-04-29	Goblin investigation published Retrospective / mitigation	OpenAI linked the habit to personality-training rewards and transfer beyond Nerdy. Nerdy accounted for 2.5% of responses but 66.7% of goblin mentions. GPT-5.5 training predated the diagnosis. Codex received a suppressing instruction. Source.	Links reward design to the spread of a lexical habit across model responses.
2026-10-29	GPT-6. reduced flattery Targeted serving update	OpenAI announced shorter responses and less flattering language for GPT-6 serving Free, Go, and logged-out users. On Plus/Pro, the update applied only to fallback traffic after GPT-6 message limits. Source.	More concise responses and reduced flattery for the specified serving traffic.

G Accounting Judgment, Trade-Offs, and the Formation of Reporting Concepts

Accounting judgment involves interpreting evidence and applying reporting objectives and standards to recognition, measurement, presentation, disclosure, and estimates. It concerns amounts, classifications, assumptions, and explanations. Table A19 links this literature to the FASB Framework and the institutional formation of reporting concepts.

The FASB Framework (2024, QC4–QC19, QC31–QC35) identifies relevance and faithful representation as fundamental qualities. Relevance concerns predictive or confirmatory value; faithful representation concerns completeness, neutrality, and freedom from error. An estimate can faithfully represent an uncertain amount without being perfectly accurate if its process and explanation are appropriate (QC15–QC16). Enhancing qualities can also conflict: omitting complexity may impair completeness, and a reporting improvement may temporarily reduce comparability. The cost constraint requires the benefits of information to justify its costs. These considerations make accounting judgment depend on the evidence, reporting objective, and uncertainty of the decision.

How concepts shape rules. Pelger (2016) studies the joint IASB/FASB choice of reporting objective. Erb and Pelger (2015) study the reconstruction of reliability. Their analyses show how the vocabulary and purposes used to assess reporting are themselves institutionally formed. Table A20 traces this development from the early decision-usefulness programme through the joint IASB/FASB project and subsequent framework revisions.

The FASB Rules of Procedure connect reporting concepts to the development of requirements through agenda decisions, research, consultation, deliberation, exposure for comment, and a Board vote. Amendments are communicated through an Accounting Standards Update. Applying the resulting requirements still calls for judgment about evidence, assumptions, and reporting objectives. A benchmark for an AI accounting decision must specify which of these judgments the task asks the model to make.

Table A19. Accounting Judgment, Estimation, and Conceptual Trade-Offs.

This table reports literature on accounting judgement, estimates, evidence processing, and the formation of reporting concepts. It identifies the judgement considered, the method or conceptual contribution, and the trade-off or implication for evaluating accounting decisions.

Paper / conceptual source	What the judgment concerns	Method and contribution	Trade-off or implication
FASB (2024) Conceptual Framework for Financial Reporting. Concepts Statement No. 8, Chapters 1-8.	Choose what to recognize, measure, present, and disclose so that financial information supports economic decisions.	Normative framework: relevance and faithful representation are fundamental; comparability, verifiability, timeliness, and understandability enhance usefulness. Estimation requires an appropriate process and explanation of uncertainty.	Both fundamentals are required. Enhancing qualities can conflict; complexity, comparability, and reporting costs require judgment. The Framework guides standard setting and is nonauthoritative US GAAP guidance. Source.
Dechow, P. M., and Dichev, I. D. (2002) The Quality of Accruals and Earnings. The Role of Accrual Estimation Errors. The Accounting Review 77(Supplement). 35-59.	Estimate accruals that align cash flows with the periods whose performance is being reported.	Model and archival analysis link working-capital accruals to past, current, and future cash flows. Larger residual estimation noise indicates lower accrual quality.	Accruals reduce cash-flow timing mismatches while introducing assumptions and estimation error. Sources. [1], [2].

Continued on next page

Table A19 continued. Accounting Judgment, Estimation, and Conceptual Trade-Offs

Paper / conceptual source	What the judgment concerns	Method and contribution	Trade-off or implication
Richardson, S. A., Sloan, R. G., Soliman, M. T., and Tuna, I. (2005) Accrual Reliability, Earnings Persistence and Stock Prices. Journal of Accounting and Economics 39(3). 437-485.	Assess how measurement subjectivity affects the interpretation of accrual components and future earnings.	A model and archival tests classify accruals by reliability. Less reliable categories have lower earnings persistence, which investors do not fully anticipate.	Measurement quality has economic consequences. Sources. [1], [2].
Barth, M. E. (2006) Including Estimates of the Future in Today's Financial Statements. Accounting Horizons 20(3). 271-285.	Select and interpret measurement attributes that incorporate expectations about assets, liabilities, and future economic benefits.	Conceptual analysis explains why estimates of the future are already integral to reporting. Recognition, measurement attributes, and disclosures determine how they enter statements.	The issue is how expectations are incorporated under the reporting objective. More current information can bring estimation demands. Sources. [1], [2].
Griffith, E. E., Hammersley, J. S., Kadous, K., and Young, D. (2015) Auditor Mindsets and Audits of Complex Estimates. Journal of Accounting Research 53(1). 49-77.	Evaluate the reasonableness of an estimate by integrating evidence, including information inconsistent with management's position.	An experiment finds that a deliberative-mindset intervention improves identification of unreasonable estimates through better recognition and use of contradictory evidence, rather than simply greater effort.	Judgment quality depends on evidence processing. Evaluating estimates requires integrating conflicting evidence and assessing assumptions. Sources. [1], [2].
Griffith, E. E., Hammersley, J. S., and Kadous, K. (2015) Audits of Complex Estimates as Verification of Management Numbers. How Institutional Pressures Shape Practice. Contemporary Accounting Research 32(3). 833-863.	Decide how to challenge management's model and assumptions and combine auditors' and specialists' knowledge.	Interviews with 24 experienced auditors identify piecemeal verification of management's estimates, missed inconsistencies, and reliance on specialists. The authors explain these patterns institutionally.	Procedural verification and an overall skeptical evaluation can diverge. Institutional pressures shape how evidence and specialist knowledge are combined. Source.
Ulla, J. W., and Commerford, B. P. (2025) Reliance on Algorithmic Estimates. The Joint Influence of Algorithm Adaptability and Estimation Uncertainty. The Accounting Review 100(6). 285-308.	Decide how much weight to assign algorithmic advice while evaluating or preparing an uncertain accounting estimate.	Two experiments examine auditor advice use and estimate preparation. Under greater uncertainty, participants rely more on learning algorithms than on static algorithms.	Reliance depends on uncertainty and the system's described adaptability. Source.
Pelger, C. (2016) 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.	Determine which users and purposes define useful financial reporting, including the place of stewardship.	Due-process documents and interviews show how the decision-usefulness programme bounded debate over a separate stewardship objective, with staff organization influencing board deliberations.	The criteria used to assess reporting are themselves formed through institutional choices. Sources. [1], [2].

Continued on next page

Table A19 continued. Accounting Judgment, Estimation, and Conceptual Trade-Offs

Paper / conceptual source	What the judgment concerns	Method and contribution	Trade-off or implication
Erb, C., and Pelger, C. (2015) Twisting Words? A Study of the Construction and Reconstruction of Reliability in Financial Reporting Standard-Setting. Accounting, Organizations and Society 40. 13-40.	Define an acceptable representation and resolve competing meanings of reliability and verification.	Historical analysis traces the replacement of reliability with faithful representation and conflicts between practitioner understandings and the boards' conceptual vocabulary.	Even quality labels embody contested interpretations. The authors connect the change to wider acceptance of current/fair values. Sources. [1], [2].

G.1 Formation of Reporting Concepts

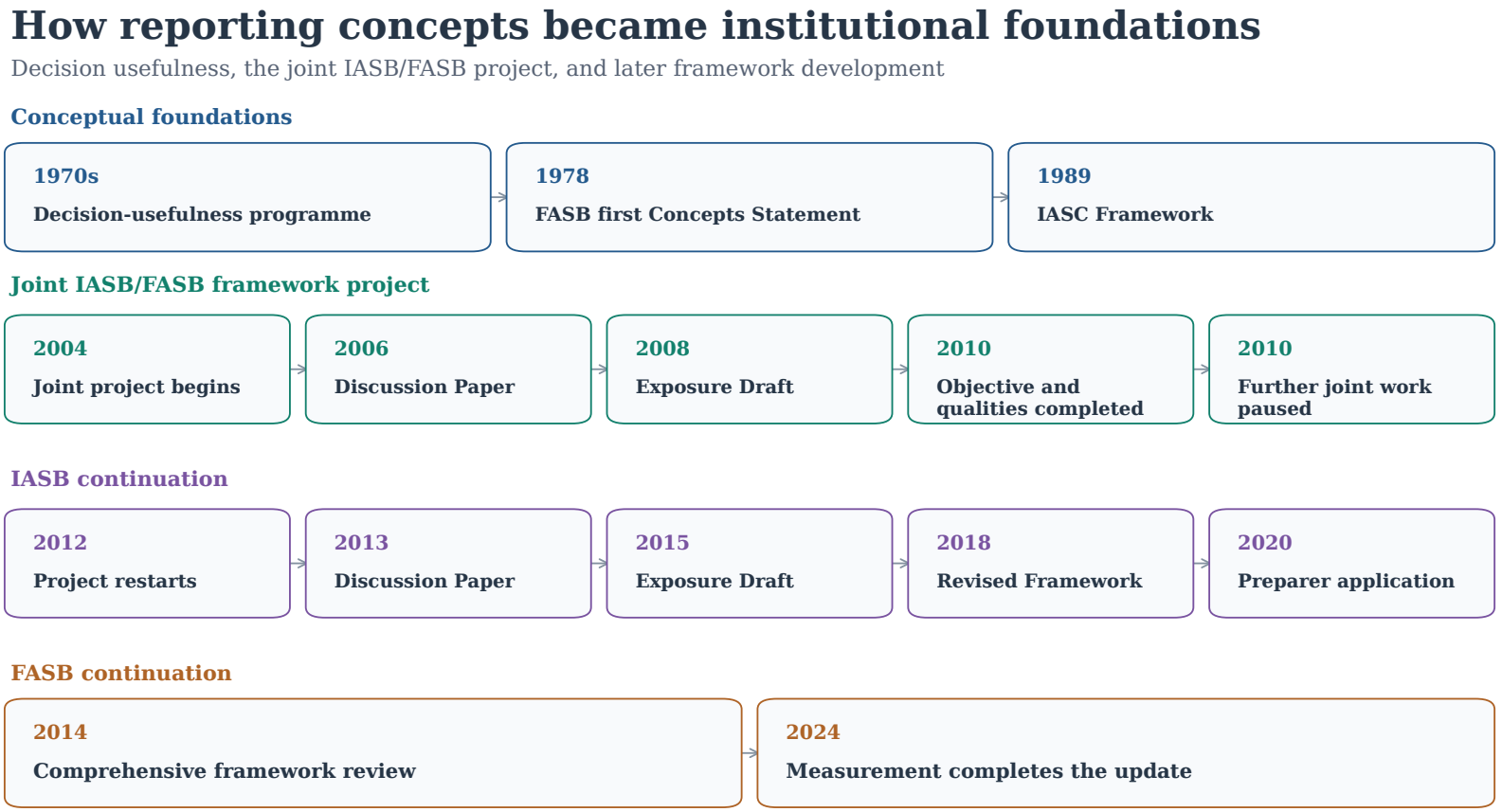


Figure A7. This figure reports the formation of financial-reporting concepts. The lanes identify early foundations, the joint IASB/FASB project, and the institutions' later continuations. Table A20 gives the chronology and its relation to Pelger's analysis.

Table A20. Formation of Reporting Concepts and Their Role in Accounting Rule-Making.

This table reports milestones in the formation of financial-reporting concepts. It follows early conceptual foundations, the joint IASB/FASB project, and later developments within each institution. The final column explains their significance for judgement and standard setting.

Date / period	Institution / setting	Milestone	Significance for judgment and standard setting
1970s	US intellectual context	The decision-usefulness programme develops as a framing for financial reporting.	Pelger identifies this inherited programme as shaping later debates. Source.
1978	FASB	FASB issues its first financial-reporting Concepts Statement.	Formal concepts articulate reporting objectives as a foundation for subsequent standard-setting choices. Source.
1989	IASC	The international Framework is issued. It is later partly revised in 2010.	Provides the foundation for the later IASB framework project. Source.
2004-10	IASB / FASB	The boards add a joint project to their agendas to converge and improve their conceptual frameworks.	Defines the project's remit and sequencing. Staff proposals and agenda organization are part of how the conceptual questions reach the boards. Sources. [1], [2].
2006-07-06	IASB / FASB	A Discussion Paper exposes preliminary views on the reporting objective and qualitative characteristics for comment.	Makes alternative understandings of usefulness available for debate. Source.
2008-05-29	IASB / FASB	An Exposure Draft proposes revised objective and qualitative-characteristic chapters.	The move from preliminary views to draft wording narrows the concepts to be adopted while retaining a formal opportunity for constituent response. Source.
2010-09-28	IASB / FASB Phase A	The objective and qualitative-characteristic chapters are completed. Relevance and faithful representation become the fundamental qualities.	Pelger analyzes stewardship's place within decision usefulness. Erb and Pelger analyze the reliability terminology change. Their accounts explain this outcome retrospectively. Sources. [1], [2], [3], [4].
2010 (later)	Joint project	Further joint framework work is suspended to concentrate on other agenda projects.	Agenda priorities affect the pace and scope of conceptual development. Framework formation is an interrupted institutional process. Source.

Continued on next page

Table A20 continued. Formation of Reporting Concepts and Their Role in Accounting Rule-Making

Date / period	Institution / setting	Milestone	Significance for judgment and standard setting
2012	IASB	Following its 2011 agenda consultation, the IASB restarts its Framework project.	Begins a later IASB phase. Source.
2013-07	IASB	A Discussion Paper proposes updates, clarification, and completion of the Framework.	Reopens consultation, including unresolved recognition and measurement questions. Later feedback changes the treatment of stewardship. Sources. [1], [2].
2014	FASB	FASB decides to revisit its own Framework more comprehensively in response to stakeholder feedback.	The US continuation is distinct from the IASB restart and eventual 2018 Framework. Source.
2015-05-28	IASB	An Exposure Draft proposes a more complete Framework and gives stewardship greater prominence.	Consultation can revise objectives and explanations. This proposal belongs to the later IASB phase, beyond the joint 2010 settlement examined by Pelger. Source.
2018-03-29	IASB	A revised Framework is issued, explicitly discussing stewardship and clarifying measurement uncertainty and prudence.	The IASB's concepts evolve after the 2010 settlement. Sources. [1], [2], [3].
2020-01-01	IFRS preparers	The 2018 Framework becomes applicable for annual periods beginning on or after this date when preparers develop policies by reference to it.	Distinguishes the issue date and immediate use by the IASB from the later preparer transition date. Source.
2024-07-12	FASB	Issuance of the Measurement chapter completes FASB's comprehensive Framework update.	The September 2024 compilation supplies the US conceptual anchor used here. Sources. [1], [2].

H Benchmarks for Tabular Data and Spreadsheet Workflows

Table A21 groups benchmarks into three uses of tabular data, covering reading and reasoning over table content, operating spreadsheet workbooks, and predicting outcomes from structured features. The sequence runs from financial-report question answering to representation robustness and executable business workflows.

FinQA and TAT-QA connect financial tables with narrative disclosures and arithmetic. TableBench broadens table reasoning to analysis and visualization. ToRR tests how answers respond to changes in table representation. SpreadsheetBench 2 moves to workbook deliverables, while TabArena evaluates prediction from structured data.

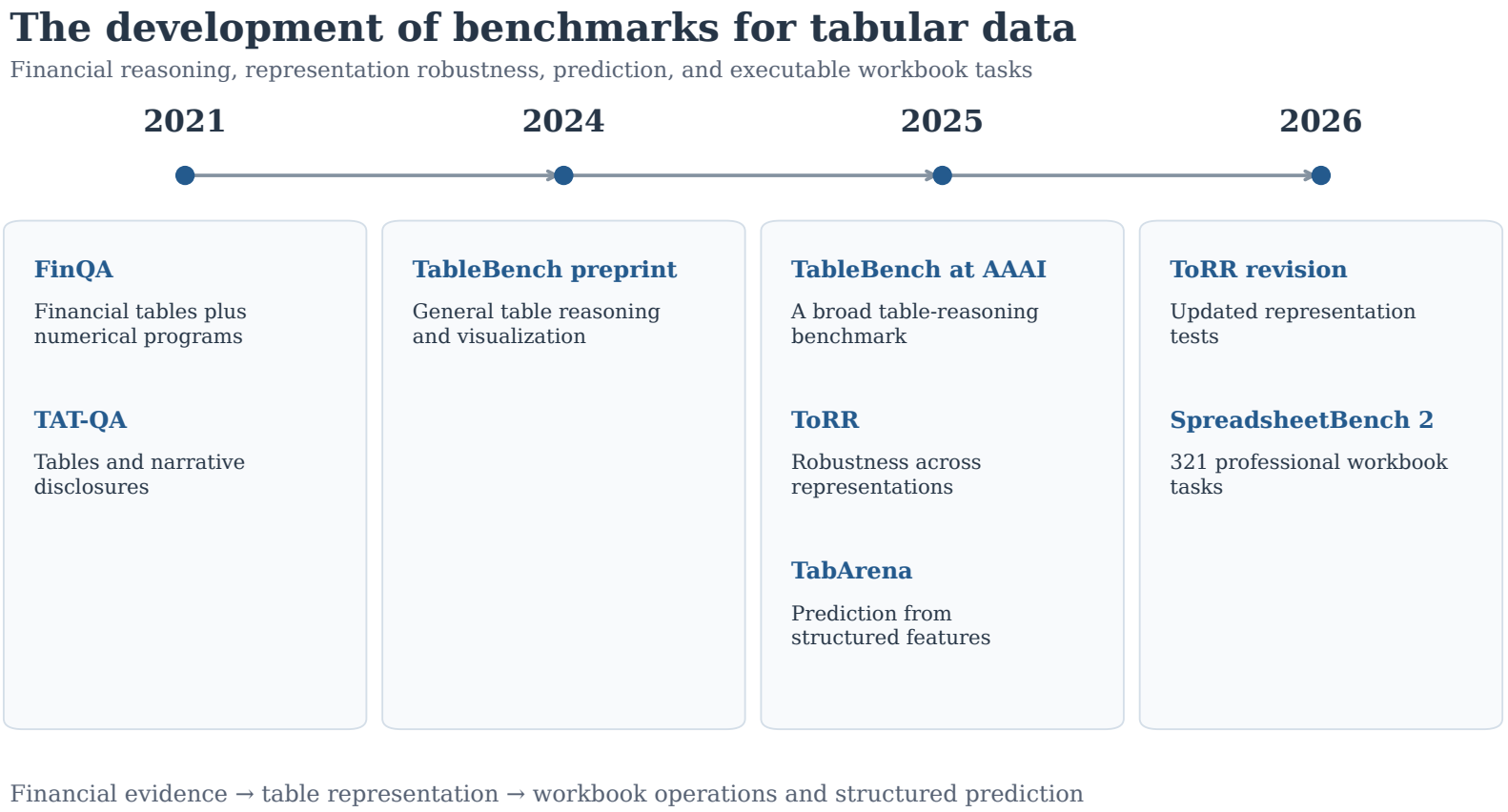


Figure A8. This figure reports milestones in financial table reasoning, representation robustness, predictive modelling, and spreadsheet workflows. The columns group releases and revisions by year. Table A21 identifies the papers and evaluation tasks.

Table A21. Selected Benchmarks for Table Reasoning, Spreadsheet Workflows, and Tabular Prediction.

This table reports benchmarks for financial and general table reasoning, representation robustness, spreadsheet workflows, and predictive tabular modelling. It identifies the published version, inputs, task counts and units, scoring approach, and accounting relevance.

Benchmark / publication / version	Primary area	Inputs and tasks	Evaluated output / scoring	Accounting relevance
FinQA Chen et al. EMNLP 2021 Original published dataset.	Financial table/text reasoning	8,281 question-answer pairs over financial-report tables and text, with annotated numerical reasoning programs. The count includes training, validation, and test data.	Execution accuracy. Mathematically equivalent program accuracy.	Tests retrieval and multistep financial arithmetic. Program checking adds information beyond final-answer agreement. Sources. [1], [2].
TAT-QA Zhu et al. ACL-IJCNLP 2021 Original published dataset.	Financial table/text reasoning	16,552 questions associated with 2,757 hybrid table/text contexts from financial reports. Tasks combine extraction with arithmetic, counting, and comparison.	Exact match. Numeracy-focused F1, including sign/scale handling.	Tests integration of table cells and narrative disclosures. Sources. [1], [2].
TableBench Wu et al. AAAI 2025. Preprint August 2024 Project data/metric update April 18, 2025.	General table reasoning	886 test cases spanning 18 subcategories within fact checking, numerical reasoning, data analysis, and visualization.	Task-category and overall scores. Revised evaluator in April 2025.	Covers broad table reasoning, including chart generation. Sources. [1], [2].
ToRR (The Mighty ToRR) Ashury-Tahan et al. First preprint February 26, 2025 v3, February 17, 2026.	Representation robustness	10 datasets tested across seven serializations and four structural perturbations, including row/column swaps, transposition, and empty rows. Includes FinQA.	Mean dataset scores across formats. Robustness from within-item score ranges.	Directly tests robustness when content is preserved but representation changes. Sources. [1], [2].
SpreadsheetBench 2 Zhu et al. Preprint June 29, 2026 Version 2. 321 tasks.	Workbook workflows	321 business spreadsheet workflows involving financial modeling, template completion, debugging, and visualization, with large multi-sheet workbooks and cross-sheet dependencies.	Cell modification and task accuracy. Vision-language-model chart grading.	Tests workbook changes and deliverables. Version 2 differs from the original 912-task SpreadsheetBench. Sources. [1], [2], [3].
TabArena Erickson et al. NeurIPS 2025, Datasets and Benchmarks Published v0.1 study. Living benchmark.	Predictive tabular ML	The original study uses 51 datasets for supervised classification and regression on structured features. Compares trees, neural methods, foundation models, and ensembles.	AUROC, log-loss, RMSE. Aggregate rankings including Elo.	Tests prediction from structured features, such as credit-risk modeling. Sources. [1], [2].

I Context Windows and the Use of Accounting Information

A context window holds the instructions, question, documents, conversation history, and generation budget for the current response. Larger windows allow complete statements, notes, exhibits, and earlier experimental stages to enter the same decision setting. The visual timeline shows their expansion. The table gives model-specific capacities and availability dates.

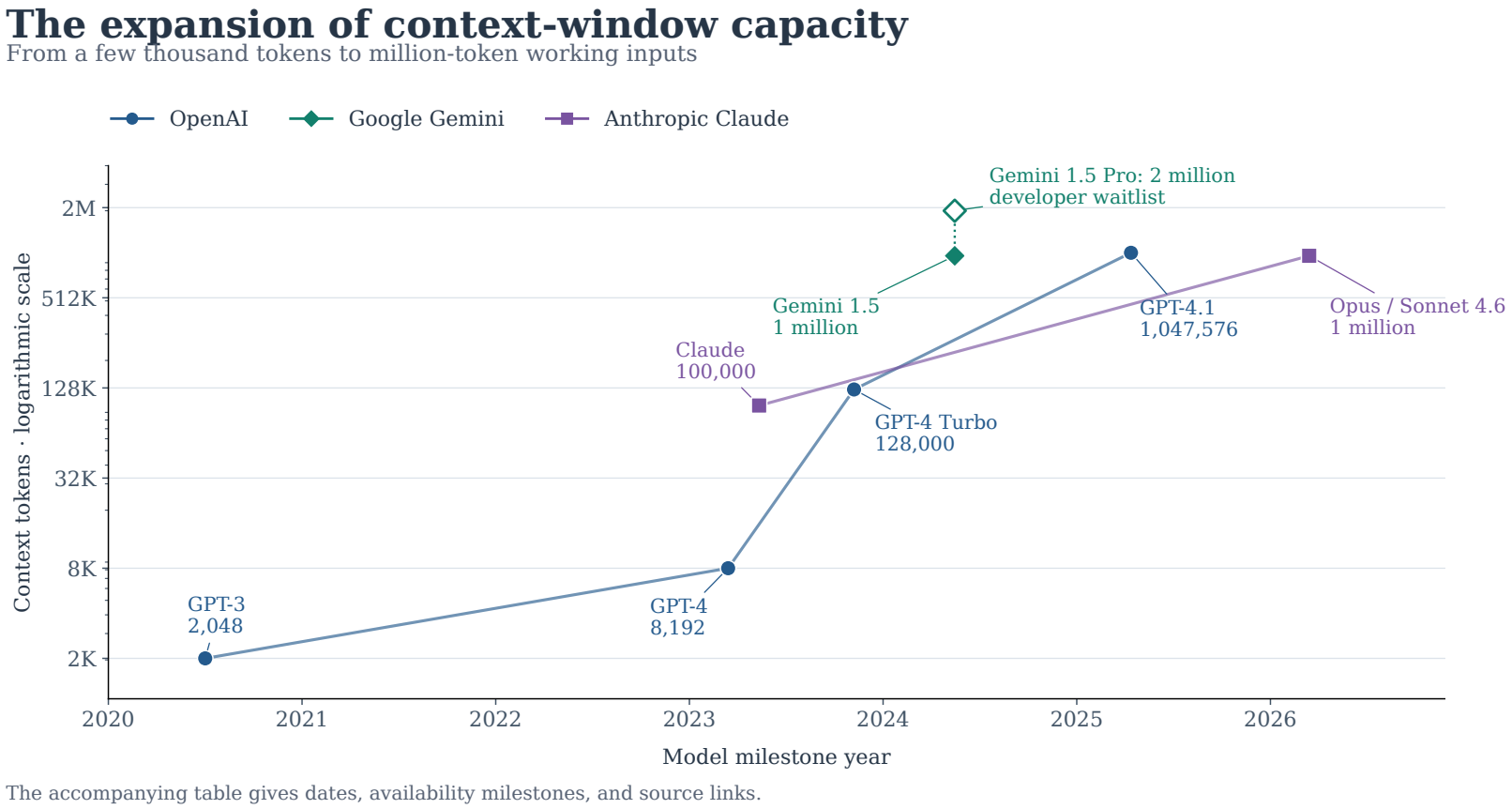


Figure A9. This figure reports context-window capacity over model milestones. The vertical axis measures tokens on a logarithmic scale. The open Gemini marker identifies the two-million-token developer waitlist milestone. Table A22 gives dates and availability.

Table A22. Expansion of Model Context Windows.
This table reports the expansion of model context windows. Capacities are measured in tokens. Dates identify the model or availability milestone. The final column distinguishes API access, public preview, waitlist access, and general availability.

Date	Model	Context window	Development
2020	GPT-3	**2,048**	Few-shot examples and task instructions share a short input span. Source.
2023-03	GPT-4	**8,192**	A longer working input for instructions and documents. Source.
2023-05-11	Claude	**100,000**	API access expands from 9,000 to 100,000 tokens. Source.
2023-11	GPT-4 Turbo	**128,000**	Brings substantially longer documents into a single request. Source.
2024-05-14	Gemini 1.5 Pro / Flash	**1 million. Pro. 2 million via waitlist**	One million tokens in public preview. Two million for Pro through a developer waitlist. Source.
2025-04-14	GPT-4.1	**1,047,576**	Approximately one million tokens in the API. Source.
2026-03-13	Claude Opus 4.6 / Sonnet 4.6	**1 million**	One-million-token context becomes generally available at standard pricing. Source.

I.1 How Larger Windows Became Practical

A longer input requires computational capacity as well as training for distant positions. Conventional full attention compares token pairs, so doubling sequence length roughly quadruples the attention arithmetic. FlashAttention reduces transfers between GPU memory levels by organizing exact attention into blocks. RingAttention distributes sequence blocks across devices, while Longformer limits direct attention to local and selected global connections. These approaches help make complete accounting cases feasible inputs. They do not establish how a model will use a disclosure once it is supplied.

Training affects the model's use of the extended input. YaRN rescales position representations and extends them through further training. Information-intensive training (IN2) trains retrieval across positions and integration across passages. Retrieval-augmented generation instead supplies selected passages from an external collection. For an accounting experiment, input capacity, evidence selection, and the model's use of evidence are therefore separate features of the design.

I.2 From Available Information to Accounting Judgment

Context research identifies three demands relevant to accounting judgment. First, evidence position and surrounding narrative affect its use. Liu et al. move relevant information within a context. Shi et al. add irrelevant material while holding the calculation fixed. These designs motivate my information-position, narrative-context, and accounting-attention tests.

Second, accounting evidence requires integration across locations. RULER tests tracing and aggregation. NoLiMa tests recognition with little literal overlap. LongBench v2 includes structured data and cross-document reasoning. Accounting tasks similarly link amounts to classifications, assumptions in notes, and other statement items.

Third, representation and earlier responses shape the decision setting. ToRR varies table formats. Laban et al. examine prior conversational turns. My representation tests rearrange accounting content, while chronological ordering changes the placement of prior responses. Separate factual checks distinguish information processing from its influence on judgment.

Table A23. Context-Window Research and Its Connection to Accounting Experiments.

This table reports research on context-window capacity and the use of information within long inputs. Panel A covers attention algorithms, position extension, and document retrieval. Panel B covers evidence position, integration, table representation, and prior-response history. Each row identifies methods, findings, and the connection to the accounting experiments.

Paper / publication	Methods	Findings	Connection to this study
Panel A. Expanding capacity and document access			
Beltagy, Peters and Cohan (2020) Longformer. The Long-Document Transformer. arXiv 2004.05150.	Combines local sliding-window attention with selected global connections. Evaluates language modelling, long-document question answering, and summarization.	Attention scales linearly with sequence length under the sparse pattern. The pretrained model improves long-document tasks relative to RoBERTa.	Shows how an attention structure can connect a focal question with distant evidence in a long reporting document.
Lewis et al. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.	Combines a neural retriever over a document index with a pretrained sequence-to-sequence generator. Evaluates knowledge-intensive tasks.	Improves open-domain question answering and generates more specific and factual text than the parametric-only baseline.	Retrieval determines which disclosures reach the model. Supplying the full experimental case holds this selection step fixed.
Dao et al. (2022) FlashAttention. Fast and Memory-Efficient Exact Attention with IO-Awareness. NeurIPS 2022.	Tiles exact attention into GPU-memory blocks, reducing transfers between high-bandwidth memory and on-chip memory.	Reduces memory traffic and speeds up training while computing exact attention. Supports longer-sequence experiments.	Explains how larger collections of statements and notes become computationally practical inputs.
Peng, Quesnelle, Fan and Shippole (2024) YaRN. Efficient Context Window Extension of Large Language Models. ICLR 2024.	Rescales rotary position embeddings and continues training to extend the positions over which a model can operate.	Extends LLaMA-family models to 128,000-token contexts with substantially fewer training tokens and steps than the compared extension methods.	Separates the engineering of a longer input span from the subsequent processing of accounting evidence within that span.
Liu, Zaharia and Abbeel (2024) RingAttention with Blockwise Transformers for Near-Infinite Context. ICLR 2024.	Distributes sequence blocks across devices and overlaps attention computation with communication of key-value blocks.	Demonstrates contexts of millions of tokens, with supported sequence length scaling with the number of devices.	Provides an infrastructure route to keeping complete reports, exhibits, and prior task stages in the same input.

Continued on next page

Table A23 continued. Context-Window Research and Accounting Experiments

Paper / publication	Methods	Findings	Connection to this study
Panel B. Using evidence across positions, formats, and task stages			
Shi et al. (2023) Large Language Models Can Be Easily Distracted by Irrelevant Context. ICML 2023.	Adds irrelevant sentences to arithmetic problems while preserving the required calculation. Compares prompting and decoding strategies on GSM-IC.	Irrelevant context changes answers and lowers accuracy. Self-consistency and instructions to ignore irrelevant information improve robustness.	Closely parallels changes in surrounding narrative and accounting-attention instructions while the focal economic information is held fixed.
Liu et al. (2024) Lost in the Middle. How Language Models Use Long Contexts. TACL 12, 157-173.	Varies the location of relevant evidence in multi-document question answering and key-value retrieval, holding the target information fixed.	Performance is often highest when relevant information appears at the beginning or end and lower when it appears in the middle.	Directly motivates the information-position test and the distinction between preserving an accounting amount and preserving its prominence.
Hsieh et al. (2024) RULER. What's the Real Context Size of Your Long-Context Language Models? COLM 2024.	Evaluates 17 models on 13 synthetic tasks, varying sequence length, the number of relevant facts, multi-hop tracing, and aggregation.	Accuracy declines as length and task complexity increase, including for models that perform near perfectly on a simple needle-retrieval task.	Connects locating a disclosed number to the harder task of combining several amounts, classifications, and relationships.
Bai et al. (2024) LongBench. A Bilingual, Multitask Benchmark for Long Context Understanding. ACL 2024, 3119-3137.	Evaluates 21 datasets across six categories, including single- and multi-document questions, summarization, and code completion.	Longer-sequence fine-tuning and scaled position embeddings improve context use. Retrieval-based compression helps models with weaker long-context performance.	Supplies a framework for separating input length, cross-document processing, and the task asked of the model.
An et al. (2024) Make Your LLM Fully Utilize the Context. NeurIPS 2024.	Information-intensive training requires retrieval from different positions and integration across segments in 4,000-32,000-token inputs. Produces FILM-7B.	Improves retrieval across positions and raises NarrativeQA F1 from 23.5 to 26.9 while keeping short-context MMLU performance similar.	Shows that training can change how position and distributed evidence affect processing, motivating comparisons across model generations.

Continued on next page

Table A23 continued. Context-Window Research and Accounting Experiments

Paper / publication	Methods	Findings	Connection to this study
Panel B continued. Meaning, structured data, and prior responses			
Bai et al. (2025) LongBench v2. Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. ACL 2025, 3639-3664.	Uses 503 multiple-choice questions with 8,000-2 million words of context, including long structured data, multiple documents, and dialogue histories.	Direct-answer model accuracy reaches 50.1%. O1-preview with extended reasoning reaches 57.7%, against 53.7% for humans under a 15-minute limit.	Adds structured-data and sustained-reasoning tasks close to working across statements, notes, and experimental stages.
Modarressi et al. (2025) NoLiMa. Long-Context Evaluation Beyond Literal Matching. ICML 2025.	Tests 13 models with questions that require latent associations to locate evidence, minimizing literal word overlap between the question and the relevant passage.	At 32,000 tokens, 11 models fall below half their short-context baseline. GPT-4o falls from 99.3% to 69.7%.	Connects accounting labels and paraphrased disclosures to the need to recognize their economic meaning across a long case.
Ashury-Tahan et al. (2025) The Mighty ToRR. A Benchmark for Table Reasoning and Robustness. arXiv 2502.19412. Revised 2026.	Tests ten table datasets across alternative serializations and structural transformations, including row and column rearrangements.	Answers and model rankings vary with table representation. Evaluating several formats yields more reliable performance estimates.	Directly connects to continuous-text, ordered-line, JSON/transcript, and table-indexing tests that rearrange the same accounting content.
Laban, Hayashi, Zhou and Neville (2025) LLMs Get Lost In Multi-Turn Conversation. arXiv 2505.06120. ICLR 2026.	Compares fully specified single-turn instructions with incrementally disclosed multi-turn instructions across six generation tasks and over 200,000 simulated conversations.	Reports an average 39% performance decline in multi-turn settings, associated with early assumptions and reliance on prior generated solutions.	Connects prior-response history and task-stage ordering to path dependence when models make sequential accounting decisions.