Synthetic data can beat human data. Where it has, and where it hasn't yet.

AI-made training data has beaten crowdworkers' demonstrations and, built from clinicians' vetted answers, matched physicians' rubrics; AI-made research has beaten lawyers on bounded legal research, but not yet professionals in finance.

Aperture

Published 19 November 2026 · Research cutoff 18 November 2026

The question. Can synthetic data, made by AI from a real corpus with verification and structure, be better than human data, as material for training models and as finished research and due diligence?

Executive summary. Synthetic data can beat human data.

As training data it has beaten crowd and benchmark data when it fits how the model learns, and tied expert data when expert-vetted material goes back in.

As finished work, AI has beaten lawyers on bounded legal research; in finance it has not yet won, and verification and structure are the conditions to test.

Whether an expert checking AI's work closes that gap is untested in finance: the test needs three arms, expert alone, AI alone and expert checking AI.[1]

Scope: This essay reviews published studies, benchmarks and industry reports from 2025–26 that set AI-made work against human work as training data, or AI research against professionals' work; the evidence is mostly small and partly vendor-run, and sources read only in part are cited only for what was read.[2] Aperture produces research for buy-side and sell-side

[1] Training data: section 02 (Gupta et al.; Gunjal et al., Scale AI, vendor-run; Chandra et al.; Harsha et al.; Huang and Simpson; Ioanovici et al.; Himi et al., abstract only; Su et al.; Barlow et al.; Mendes et al.; Akshansh et al., Pareto AI, vendor-run; Lu and Smith; Banerjee et al.). The human data beaten are crowdworker trajectories (Gupta et al.), textbook-style benchmark solutions (Chandra et al.) and, only on another dataset's test, 2021–22 expert-built finance datasets (Harsha et al.). Bounded work: section 03 (Vals AI; Krumdick et al.; Lu, Jiacheng, et al.). Full research: section 04 (Lau et al., BankerToolBench; Haque et al.). Checking: section 05 (Autor et al., patent law, Google-funded, read in part; Vaccaro et al.). No study found tests an expert checking AI against both the expert alone and the AI alone on full research deliverables; the Federal Reserve study (Botti et al.) is a bounded task with three analysts, and Tseng and Chang has no human-alone arm.

[2] The search behind this essay covered 2025–26 studies with a measured human comparison on research, extraction, diligence or forecasting tasks, and studies comparing synthetic with

institutions, so it has an interest in the answer. No Aperture system, data or result is examined or reported here. Proprietary deployments and vendor claims without a published protocol were not examined in this review, and no adverse inference follows.

01 The claim, and what should make it hold

Synthetic work can be better than human work, as training data and as finished research, and in some published studies it already is. Trained on 2,700 web-browsing trajectories from language-model agents, a web agent completed 53.0% of WebVoyager tasks, against 35.4% when trained on 2,700 crowdworkers' trajectories for the same task instructions (Figure 1). Against physicians the result is a tie: in a vendor-run study, rubrics a model wrote from clinician-vetted reference answers trained a medical model as well as physicians' own rubrics did (35.9% against 34.8% on HealthBench, comparable in the authors' word), and fell behind without them (32.0%; section 02). As finished work, four AI products outscored lawyers at a US firm on 200 legal research questions graded blind (section 03).

This essay calls synthetic data any work an AI system produces in place of a person's: solutions, trajectories, rubrics, question-answer pairs, research notes, diligence reports. It calls such work grounded when it is made from a defined, real source (a data room, a set of filings, a dataset's own problems and answers, the websites a task names) rather than from tasks or sources the model invents, and checked when its claims or answers are tested against that source. Three properties should make grounded finished work better than a person's. Scale: a model can read every page of a data room in the time an analyst reads a few. Consistency: it applies the same procedure to every document. Verification: each claim can carry a pointer to its source, so it can be checked. As training data, section 02's studies credit other things: data that suits how the student model learns, information the human demonstrators lacked, and expert material put back into the synthetic set. None isolates grounding or verification as the cause of a win.

Scale is untested: EPSBench varied which sources an agent had, not how much it read, and news alone did no detectably worse than all of them (section 04). The need for verification is best supported: a model judged 11% to 34% of research agents' claims not found in the sources they cited, so each claim needs checking (section 04). Consistency is open: no study isolates it, and it can repeat one error everywhere.

Each property can fail. Reading everything does not mean judging well, and a claim that can be checked still has to be checked by someone.

---

human training data in knowledge work, including a final pass for comparisons with paid expert data; it is not a systematic review. Items read only as abstracts, excerpts or secondary reports are marked as such in their notes and cited only for what was read. Vendor-run studies are labelled as such.

02 Where it has beaten human data as training data, and against whom

As training data, synthetic work has beaten human data made by crowdworkers or written for textbooks and benchmarks, when it suits how the model learns; against expert data it mostly still loses, and it ties when expert-vetted material goes back in. No 2025–26 study we found shows synthetic data beating paid expert data on that data's own test at matched size; in finance, the only wins come on another dataset's test (Harsha and colleagues, below). Two confounds run through these studies: the synthetic data was usually made or judged by a stronger model than the student, and some was shaped by the test (Banerjee's synthetic set was modelled on its test data's source; MolmoWeb's training filter used the WebVoyager judge).

The clearest win is on agent tasks. Ai2 trained MolmoWeb, the web agent above, on 2,700 trajectories from each source: language-model agents' trajectories, kept when a judge rated the task done, against crowdworkers' trajectories, each reviewed. The synthetic set won on all three benchmarks (Figure 1). The authors suggest two causes: humans' longer, exploratory trajectories, and the agents' view of each page's accessibility tree.[3]

[3] Gupta et al., Ai2, arXiv:2604.08516v1 (Apr. 2026), secs. 2, 4.1 and 4.4, table 6: DeepShop 24.4 against 19.8, WebVoyager 53.0 against 35.4, Online-Mind2Web 16.8 against 9.0. Synthetic trajectories were kept when the WebVoyager language-model judge rated the task complete, so the training filter used the WebVoyager judge. The human trajectories came from crowdworkers, with annotation support from Snorkel AI, not domain experts. The authors hypothesise two causes: humans' longer, more exploratory trajectories, and the agents' access to each page's accessibility tree. The paper's main evaluations average three to five runs per benchmark; it does not say how many runs or seeds lie behind table 6, and reports no uncertainty for it. Read in full. Paraphrased.

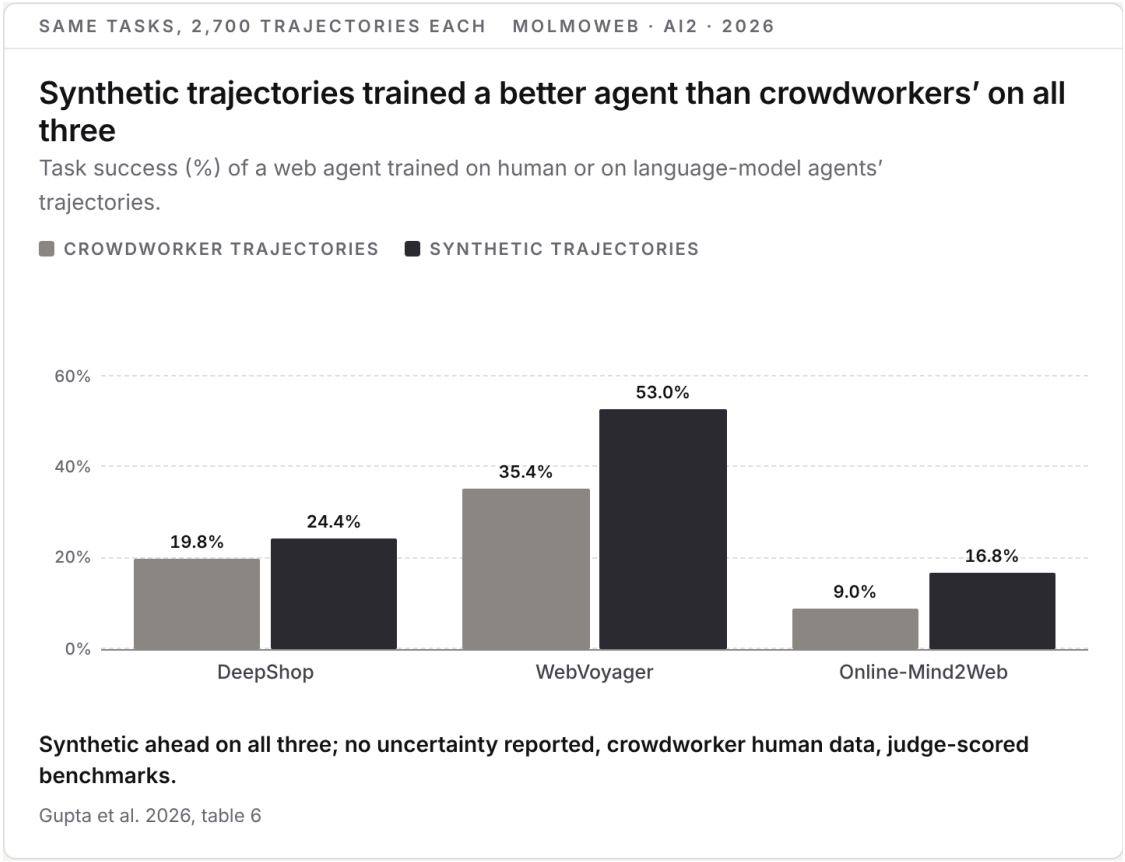

Figure 1. On the same task instructions, a web agent trained on 2,700 synthetic trajectories beat one trained on 2,700 crowdworker ones on all three benchmarks. Task success (%) for MolmoWeb trained on language-model agents' trajectories, kept when a judge rated the task done, or on crowdworkers' reviewed trajectories, each mixed with the same perception data. No uncertainty reported for this comparison; benchmarks scored by a language-model judge; the human trajectories had thoughts written afterwards and a partly different action space, and the agents saw each page's accessibility tree.

In maths, Chandra and colleagues fine-tuned Gemma-2-2B on MATH's training problems, once on the dataset's human-written solutions and once on solutions Gemma-2-27B wrote: at each run's best checkpoint, chosen on the test set, it solved 22.3% of MATH500 against 16.2%. Four-shot prompting alone gives the base model 17%, so the human solutions left it slightly below that. Solutions that reached wrong answers also beat the human ones (18.3%), and the human solutions, rewritten by the larger model, rose from 17% to 20% at best. The authors credit closeness to the student's style, not correctness. Llama-3.1-8B gained less (22% against 19% at

best); Qwen2.5-1.5B, already at 53%, did worse.[4] A similar result by Hou and colleagues rests largely on two near-zero scores and is not relied on.[5]

Finance question answering splits by test. Harsha and colleagues had GPT-4o write question-answer pairs from the report pages behind two expert-built datasets, FinQA and TAT-QA, kept when their code ran, without checking the answers. Trained on them, Phi-3-mini lost on each dataset's own test (68.43% against 73.49% on FinQA; 88.99% against 90.94% on TAT-QA) and won on the other dataset's test both ways (84.26% against 81.19%; 67.82% against 63.81%). No repeats are reported; the authors conclude real data did better.[6]

Against expert data the record is thin and mostly goes the other way. In two small clinical extraction studies, synthetic text lost: F1 0.33 against 0.90 for expert-annotated real colonoscopy reports, tested on real reports, and 0.682 against 0.767 for manually annotated,

---

[4] Chandra et al., arXiv:2512.22255v2 (Jan. 2026), secs. 4–5, tables 2–6 and app. tables 7, 9, 10 and 11: Gemma-2-2B on MATH500, the mean over five seeds of each run's best checkpoint, 0.223 against 0.162 (95% intervals ±0.006), selected on the test set; the base model scores 12% zero-shot and 17% four-shot (sec. 5.1); the teacher reached a correct answer on 5,809 of 7,500 problems. Solutions that reached wrong answers also beat the human-written ones (0.183). Human solutions paraphrased by Gemma-2-27B: best 20% against 17% (sec. 5.3, one run's maximum). Llama-3.1-8B on the same teacher's solutions: 0.22 against 0.19 at best, 0.18 against 0.17 on average (app. table 7). Qwen2.5-1.5B: 0.48 against 0.52 on average (app. table 7); the authors put its result down to its high starting accuracy and a possibly unsuitable teacher. MATH's solutions are competition solutions written for human readers, not annotation made for training. Not peer reviewed. Read in full. Paraphrased.

[5] Hou et al., arXiv:2611.14912v1 (Nov. 2026), sec. 3.1 and app. B.5, table 5: Qwen2.5-7B trained on one verified solution per maths prompt from its own reinforcement-learned version averaged 38.7 on twelve general-reasoning tasks, against 34.9 when trained on solutions from Qwen3-14B. In the Qwen3-14B arm, MMLU-Pro and MMLU-Pro+ score 1.1 and 0.8, below the untrained base model's 9.2 and 21.0, which suggests a formatting or scoring failure rather than reasoning; over the other ten tasks the averages are 43.2 against 41.7 (Aperture's calculation from table 5), with no intervals. No human-written data. An arXiv preprint. Read in full apart from the proofs. Paraphrased.

[6] Harsha et al., FinNLP 2025 workshop, secs. 1, 4–5, tables 1–3 and 5 (zero-shot prompt; Phi-3-mini). FinQA was annotated by experts and TAT-QA by annotators with financial expertise (sec. 4.1). Own tests: FinQA 68.43 against 73.49, TAT-QA 88.99 against 90.94 (table 1). Cross-dataset: trained on FinQA-style data and tested on TAT-QA, 84.26 against 81.19, a win at all four training-set sizes (table 2: 82.17/81.19, 83.56/79.66, 82.31/81.75, 84.26/81.19); trained on TAT-QA-style data and tested on FinQA, 67.82 against 63.81, also a win at all four sizes (table 3: 65.47/64.86, 65.47/64.16, 67.56/63.46, 67.82/63.81). The authors read the cross-dataset results as a hint of better generalisation from broader question variety, but conclude that for both datasets models trained on real data perform better (sec. 5.1). The filter removed pairs whose code did not run or returned the wrong type; it did not check whether answers were correct (sec. 1). Across three models synthetic data lost on FinQA's own test and lost on ConvFinQA by about 8 to 14 points, depending on the synthetic set (table 5). No repeats are reported; TAT-QA and ConvFinQA were scored on their dev splits. Read in full. Paraphrased.

physician-written simulated lung-cancer notes, tested by cross-validation on those notes.[7] Clinical letters are the exception: letters GPT-5 wrote from templates, not real records, then re-labelled, tied expert-labelled real letters (F1 0.88, or 0.896 with chain-of-thought, against 0.874).[8] Two other results are not synthetic-against-expert tests. In finance sentiment, a small model trained on 105 human seed sentences expanded ninefold by GPT-4o scored 77.14% on financial tweets against 86.60% for the full human-labelled set, about ten times larger and of undocumented provenance; on expert-authored statements the gap was 1.3 points. Seed-matched, the expansion helped: accuracy rose from 74.2% to 77.1% (clustered seeds) and from 59.9% to 74.3% (random seeds).[9] Su and colleagues' dementia-prediction study has no human-written arm: rationales a clinical expert rated accurate and faithful to each patient's record, added to the training targets, cut mean ROC-AUC from 0.734 with labels alone to 0.604.[10]

---

[7] Ioanovici et al., *Journal of Personalized Medicine*, July 2025: 100 real Romanian colonoscopy reports against 100 GPT-4o-generated ones, both annotated by two gastroenterologists, so only the text differs; tested on real reports, F1 0.90 against 0.33; a 1:1 mix scored best on a mixed test set (0.94). Read in part (abstract and methods excerpts; result tables not read). Himi et al., JMIR preprint, June 2026: 96 manually annotated, physician-written simulated progress notes (cross-validated) against 4,366 synthetic document–event pairs generated from PubMed case reports, micro-F1 0.767 against 0.682; synthetic then manual, 0.800. Not size-matched; the manual arm is evaluated on held-out folds of its own notes. Read at abstract depth only. Paraphrased.

[8] Barlow et al., BioNLP 2026 workshop, secs. 3–5 and table 3 (Qwen-2.5-14B, three seeds; a gold set of 300 letters double-annotated by epileptologists, 21 positive). GPT-5 wrote 15,099 letters from 10 manually composed letters and 74 templates of letter characteristics, then re-labelled each and discarded mislabelled ones; no real patient letters were used, so the synthetic set is not grounded in a real corpus. Training sets: 908 balanced synthetic letters against 1,814 real letters annotated by five epilepsy experts (about 6% positive). F1: synthetic 0.88 ± 0.007, synthetic with chain-of-thought 0.896 ± 0.023, real 0.874 ± 0.049, within seed variation of each other. Read in full; saved as verbatim excerpts. Paraphrased.

[9] Huang and Simpson (University of Bristol), arXiv:2606.18875 (2026), secs. 4–5 and tables 1–2: ModernBERT on Twitter Financial News Sentiment, full training set (9,544 sentences) 86.60% accuracy against 77.14% for 105 clustered human seed sentences expanded ninefold by GPT-4o; labels assigned by prompt, not verified. On Financial PhraseBank, expert-authored statements, the gap was smaller: 96.48% against 95.15% with under 6% of the human labels. Not size-matched; the Twitter set's annotation source is not documented. Seed-matched (table 2, ModernBERT): 105 clustered seeds 74.20% alone and 77.14% expanded (IDs 12 and 14); 105 random seeds, mean of three runs, 59.90% and 74.34% (IDs 11 and 13). Read in full. Paraphrased.

[10] Su et al. (University of Pennsylvania and others), arXiv:2606.10279 (2026), secs. 3.1 and 4.1: five-year dementia prediction from longitudinal health records; 504 configurations on Qwen2.5-7B-Instruct and Qwen3-8B; mean ROC-AUC 0.734 for label-only fine-tuning against 0.604 with free-form rationales and 0.592 with stepwise rationales (paired t-test, $P \approx 7 \times 10^{-52}$). One clinical expert rated a stratified sample of the rationales medically accurate and faithful to each record. There is no human-written training arm: the comparison is labels alone against labels plus model-written rationales. Read in full. Paraphrased.

Quality work narrows the gap; for grounded data, it closed only when expert signal went back in. In Scale AI's vendor-run Rubrics as Rewards, model-written rubrics tied physicians' as rewards only with clinician-vetted reference answers (35.9% against 34.8% on HealthBench, graded on physician rubrics; 32.0% without).[11] Expert olympiad solutions rewritten into the model's own style improved models that the raw expert solutions made worse.[12]

[11] Gunjal et al., Scale AI, "Rubrics as Rewards", ICLR 2026, secs. 4.5 and 6, table 2 and app.: Qwen2.5-7B trained with GRPO on 3,500 HealthBench prompts, evaluated on 1,000 held-out prompts graded with HealthBench's physician-authored rubrics by a language-model grader (GPT-4o-mini). Rewards from o3-mini rubrics written with the reference answers 35.9%, from the physicians' rubrics 34.8%, from o3-mini rubrics without reference answers 32.0%; fine-tuning directly on the expert answers 20.4%. The reference answers are HealthBench's reference completions, vetted by licensed clinicians. The authors call the first two comparable. One policy model; no intervals or seed counts. Vendor-run: Scale AI sells human and synthetic training data. Main text read in full. Paraphrased.

[12] Mendes, Park and Ritter, "Didactic to Constructive", ICML 2026 (arXiv:2602.02405), abstract and secs. 1 and 3: fewer than 1,000 expert solutions (AIME community solutions; olympiad proofs from Evan Chen's site, used with permission) on Qwen2.5-Instruct and Qwen3 models. Training directly on the expert solutions, and on the model's own rationalisations, both made the model worse than untrained; rewriting the expert solutions in the model's distribution and training with a contrastive loss gave 10–25% pass@k gains (abstract). Results are in figures, so no per-arm numbers are given here. Main text read in full. Paraphrased.

A human base with generated variants held its own: in a vendor-run study, 10 hand-authored agent tasks plus 319 gated variants matched or slightly beat 97 hand-authored tasks (55.14 against 54.18; one training seed).[13] Replacing flagged model labels with gold ones recovered only 60–75% of the gap to gold labels in a classification study, by the authors' summary; gold here means crowd or benchmark labels, not expert ones.[14] Cost changes the comparison: per dollar, synthetic preference data matched or beat every human source at every shared budget, including HelpSteer3, a 2025 set from paid vendor annotators; at more than ten times the budget, HelpSteer3 won back helpfulness for two of three models. That synthetic set was not built from a real corpus. Human data varied widely: SHP left every model family below its untrained baseline, so the authors advise vetting human data, not assuming it is gold.[15]

---

[13] Akshansh et al., Pareto AI, arXiv:2606.03800v2 (2026), secs. 4–6 and table 5: Qwen3.5-27B trained with reinforcement learning on agent tasks; ten-benchmark mean 53.61 with 10 hand-authored tasks, 54.18 with 97, 54.38 with 10 plus 80 gated variants and 55.14 with 10 plus 319 (8 of 10 benchmarks at or above the 97-task arm). The 97-task arm trained to matched compute scores 54.72, a 0.42-point gap. One training seed; intervals reflect decoding seeds only; the authors claim parity, not superiority. The 87 extra hand-authored tasks added only 0.57 points. Vendor-run: Pareto AI sells expert-authored data. Read in full. Paraphrased.

[14] Lu and Smith, arXiv:2504.15432 (Apr. 2025): RoBERTa trained on gold labels against Qwen2.5 3B and 7B labels for the same texts, five runs each; an entropy filter and a consistency ensemble flagged doubtful model labels and replaced them with gold labels, and recovered only 60–75% of the gap, in the authors' summary; on Manifestos macro-F1 the best mitigation recovers more (gold 79.38, model labels 60.62, best mitigation 77.97). The gold labels are crowd or benchmark labels, not expert ones. Read in full (short paper); saved as excerpts.

[15] Banerjee et al. (Georgia Tech), arXiv:2611.16895 (Nov. 2026), secs. 2–3 and app. A: preference data for helpfulness and summarisation, dollar-matched from $100 to $1,000, with a $13,000 tier for human data only; Olmo-3-7B, Llama-3.1-8B and Qwen-2.5-7B; two seeds; win rate judged by a language model. HelpSteer3 overtook the synthetic peak at $13,000 for Llama and Qwen, not Olmo; on summarisation human data never caught up. Human cost is priced at $1.30 a sample from crowd work, which likely understates HelpSteer3's cost and so favours the human data. Mixing the two sources did not beat the better pure source. The synthetic helpfulness set was generated by Llama-3.1-Tulu-3-8B-SFT from Persona Hub personas, labelled by a Qwen-3-32B judge and modelled structurally on Anthropic's Helpful-Harmless data, whose test set is the evaluation (secs. 2.2, 2.3 and 2.5); it is not built from a real corpus. Human sources diverged: SHP, preferences inferred from Reddit scores, fell below the untrained baseline in every model family, and the authors conclude that a human source should be vetted rather than assumed to be the standard (sec. 3.5); Anthropic's Helpful-Harmless data, drawn from the test distribution, still trailed HelpSteer3. HelpSteer3's annotators were contractors recruited through Scale AI and Translated, per its dataset documentation (read in search excerpts); the study used its general domain only. Paper read in full. Paraphrased.

Grounding may matter: in Agent-R^2, one model trained on tasks built around real files scored 43.3% on GDPval against 41.4% with generated files, with no interval, and its real-file teacher trajectories used 2.54 times as many tool calls.[16] And verification is not judgment: GAUGE's authors fine-tuned an agent on 146 of its own verified valuation runs; valuation scores rose 8.4, assumption scores fell 4.9, and the overall score did not move (−0.8, 95% interval −3.6 to 2.0).[17]

No study we found trains a model on AI-made analyst or diligence work and compares it with training on professionals' work.

03 Where it already beats human work: bounded legal tasks

On bounded tasks with checkable answers, every study we found in which AI beat professionals overall comes from law. In finance, AI lost (Kensho, GAUGE, Vals's SEC-filing task against lawyers), had no human arm (Banco de España, CLAIR-Fin) or did not detectably beat them (Federal Reserve, section 05). AI also trails on redlining, and experienced professionals still lead where judgment enters. In Vals AI's October 2025 legal research study, four AI products (three legal research tools and ChatGPT, a general-purpose assistant; the page does not say what each searched) and lawyers at a US firm answered the same 200 legal research questions, graded blind by lawyers and law librarians. The products scored 74% to 78% on a weighted measure of accuracy, sourcing and clarity, against 69% for the lawyers, who had two weeks. The products scored higher on three-quarters of the questions; the lawyers led on four of ten question types, where, the authors suggest, context and judgment counted most. The largest legal research vendors did not take part.[18]

---

[16] Pan et al., arXiv:2611.08219 (Nov. 2026), sec. 4.5 and table 4: 1,000 matched tasks per side, one supervised fine-tuning run per side under one recipe, no interval or seed count; trajectories came from Claude Opus 4.5, checked against the task rubric. With real files the teacher trajectories used 20.4 tool calls against 8.0, so longer trajectories may account for part of the gap. No human-demonstration arm; GDPval scored by a language-model judge. Read in full apart from appendix prompts. Paraphrased.

[17] Lu, Jiacheng, et al., GAUGE, arXiv:2607.24889 (version of Sept. 2026), sec. 5.4. The trajectories were rejection-sampled from the training split; there is no human-trajectory arm. The authors conclude that corpus data teaches an agent where professional numbers live, not the judgment around them. StepFun, an author affiliation, evaluated its own model. Main text read in full; appendices not read. Paraphrased.

[18] Vals AI, "VLAIR – Legal Research", 14 Oct. 2025, executive summary, key findings 4 and 5, and methodology. 200 US legal research questions written with a consortium of law firms (one was later disregarded for an error in its wording); responses collected in July 2025; scored blind by lawyers and law librarians, at least two per response, on accuracy (50%), authoritativeness (40%) and appropriateness (10%). The products were Alexi, Counsel Stack, Midpage and ChatGPT; Vals discloses a customer relationship with one or more of them. On accuracy alone the AI products averaged 80% where they gave a response; a secondary report gives the lawyers 71% (Ambrogi), a figure not in the report's text. The same secondary report names Thomson Reuters, LexisNexis and vLex as the large vendors that did not take part. Vals's page text was read in full; its charts were not readable. Paraphrased.

A February 2025 Vals AI study set legal AI products against lawyers on the same document tasks, scored by a language model against correctness checks. On extracting data from documents such as credit agreements, the best product scored 75.1% against the lawyers' 71.1%, a gap within the study's standard errors, and on questions about documents 94.8% against 70.1%; on open-ended research in SEC filings the lawyers led, 70.1% against 55.2%, and on redlining contracts they led every product, 79.7% against 65.0%.[19]

[19] Vals AI, "Vals Legal AI Report", 27 Feb. 2025, results table and methodology. Lawyers were "average" independent lawyers sourced by an alternative legal services provider, with the same instructions as the tools; scoring was by a language-model judge against correctness checks. Data extraction: Harvey Assistant 75.1 ± 3.2 and CoCounsel 73.2 ± 3.1 against 71.1 ± 3.2; document Q&A: Harvey 94.8 ± 2.5 against 70.1 ± 5.2; redlining: Harvey 65.0 ± 5.0 and Vincent AI 53.6 ± 6.0 against 79.7 ± 4.8 (the report's text gives Harvey 59.4%; the results-table value is used); EDGAR research: Oliver 55.2 ± 3.3 against 70.1 ± 3.0, with Harvey withdrawn from that task. The ± values are standard errors. Of the other tasks, chronology tied at 80.2 and summarisation and transcript analysis favoured the products. Vendors could withdraw from any task; LexisNexis withdrew from these. Read in full. Paraphrased.

Figure 2 sets these beside finance studies that score professionals and AI on one scale, best AI system against professionals' average, which flatters the AI.

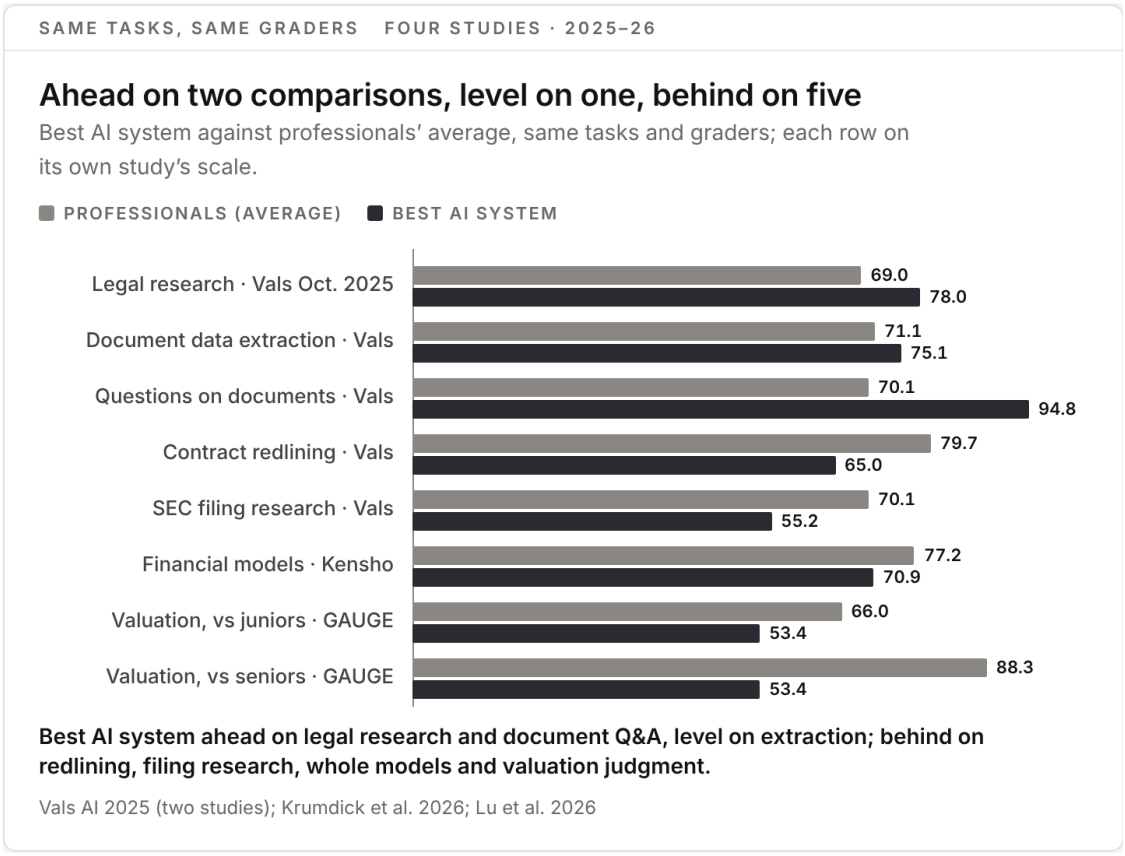

Figure 2. The best AI system scored above professionals' average on legal research and document questions, level on extraction, and below on redlining, filing research, whole models and valuation judgment. Best AI system against professionals' average, each pair on its own study's scale and grading: compare within a row. Every legal research product (74–78) and every document Q&A product (72.7–94.8) beat the lawyers; two of four extraction products did. Four of seven Feb. 2025 Vals tasks are shown. No intervals are drawn: Vals Feb. 2025 standard errors are 2.5 to 5.2 points, and the extraction gap is within them; Kensho and Vals Oct. 2025 report none; GAUGE's group means beat the agent in every bootstrap resample.

Finance's closed-corpus studies score AI against an answer key, not professionals. Banco de España researchers had models pre-fill a 30-question supervisory template from 20 bank audit reports; against the templates experienced supervisors had completed, the best scored 0.87, high on factual questions and well below on questions that need supervisory judgment.[20] CLAIR-Fin, which keeps a ledger of claims, has agents debate contested ones and audits each

[20] Alonso-Robisco et al., Banco de España Documentos Ocasionales 2613, July 2026, abstract, secs. 3–5 and tables 5–6: with semantic retrieval, Kimi 0.873 and Llama 70B 0.870; Claude Sonnet 4.6 0.795. Scored by a language-model judge (Kimi, also one of the models tested) with expert back-checks. No human-alone arm. Read in excerpts of about 40% of the paper (search-result full text); the annex was not read. Paraphrased.

citation against the source, raised faithfulness on 500 questions about a central bank's annual report from 0.780 to 0.889, scored by a language model, with no human arm.[21]

Speed is where the gap is widest. Kensho's human experts took 18.3 hours on average to build a financial model; GPT-5.4 took under an hour.[22] In a venture fund's trial of DealsFlow, a system that gates each claim on evidence, analysts' time per application fell from 44.7 to 18.9 minutes even after they verified and corrected the reports; 84% of 240 cited claims were fully supported, but 70% of reports needed at least one correction.[23]

[21] Faria et al., CLAIR-Fin, arXiv:2608.13706 (Sept. 2026), secs. 4–5 and tables 2–3. Faithfulness is a RAGAS score judged by GPT-4.1-mini; the paper does not report exact correctness for the plain retrieval baseline, and two banking experts rated only the CLAIR-Fin answers. Read in full. Paraphrased.

[22] Krumdick et al., Kensho and S&P Global, sec. 4.1 and table 3 (0.97 hours for GPT-5.4, 0.71 for Opus 4.6). See the note on the same study below.

[23] Tang et al., DealsFlow, arXiv:2610.32380 (Oct. 2026), secs. on RQ4–RQ5 and limitations. 20 inbound applications at one venture fund; two reviewers who work with the fund manager, not blind to the system's labels; a heading in the paper reads "The analyst study is small and not independent." Read in full. Paraphrased apart from that heading.

The limits are where the task becomes a whole model or asks for judgment. On Kensho's ten financial models the human experts averaged 77.2% against GPT-5.4's 70.9%, though GPT-5.4 scored higher on three.[24] GAUGE scores valuation models against the ranges in 1,001 professional workbooks: 12 senior analysts averaged 88.3, 18 juniors 66.0, and the best agent 53.4, above finance students but below every senior. It passed 93% of mechanical checks and 78% of judgment checks.[25] And AI in a professional's hands does not always help: after FactSet added generative AI to its platform, sell-side forecasts associated with it became less accurate when analysts covered complex portfolios, worked in small teams or revised several firms' forecasts on one day, though a machine-learning model given the same inputs did not deteriorate.[26]

04 Where it hasn't yet, and why

On full research and diligence deliverables, AI still loses to professionals, and the failures point to two gaps: claims nobody checked, and the missing structure of professional work. Grounding alone closes neither, and the model itself still matters.

BankerToolBench, built with 502 investment bankers, asks agents for junior-banker deliverables from a data room: models, decks and memos that took bankers five hours on average. Bankers rated none of the AI deliverables sendable as is; the best model, GPT-5.4, passed 16% of tasks at a bar of light edits.[27] On a simplified v1.1 of the benchmark with a stricter

[24] Krumdick et al., Kensho and S&P Global, arXiv:2604.05912 (Apr. 2026), secs. 3.2 and 4.1, table 3. Ten of 25 tasks (three-statement, LBO, DCF, merger and lender models from public filings) were built by human experts and by two models in a minimal computer-use harness, and all were graded by domain experts against the rubric: humans 77.2%, GPT-5.4 70.9%, Opus 4.6 61.8%; time 18.3 hours against 0.97 and 0.71. No intervals. Main text read in full; appendices not read. Paraphrased.

[25] Lu, Jiacheng, et al., GAUGE, arXiv:2607.24889 (version of Sept. 2026), secs. 1, 3–5 and table 1. 55 people made 165 attempts under the agents' conditions (no network, two tools, inputs to an as-of date, consensus withheld); judgment facets are scored against ranges observed across professional workbooks. The best agent, Claude Fable 5, was below 15 of 18 juniors; the gap between each agent's mechanical and judgment pass rates had a median of 26 points. StepFun, an author affiliation, evaluated its own model. Main text read in full; appendices not read. Paraphrased.

[26] Xue, Zhang and Zhu, arXiv:2512.19705v2 (Sept. 2026), abstract and sec. 1: 46,853 report–firm pairs, 413 analyst teams, January 2022 to October 2024; the treatment is citing FactSet after its December 2023 integration, not observed use. A secondary summary of the result tables gives a fall in relative accuracy of 0.157 standard deviations; those tables were not read. Read in part (introduction and design). Paraphrased.

[27] Lau et al., BankerToolBench, arXiv:2604.11304 (Apr. 2026), secs. 3–5. 100 tasks; bankers took five hours on average and up to 21. Bankers rated AI deliverables on a five-point scale from unusable to sendable as is; deliverables rated as needing light edits or better set the pass threshold. GPT-5.4's pass rate falls to 13% when all three runs must pass. The automated verifier agreed with bankers 88.2% of the time against 84.6% between bankers. Handshake AI,

pass rule, aggregators report a later model passing about a third; that is a secondary report of a different version, not a trend.[28]

In J.P. Morgan's Deep FinResearch Bench, one firm's equity research scored 2.84 on a four-point quality scale against 2.02 to 2.31 for four commercial deep-research agents, scored by a language model checked against experts on a sample; the best agent beat the analysts only on stating assumptions, 2.00 against 1.91. A model judged 11.2% to 34.1% of the agents' claims hallucinated, not found in the sources they cited. The professional reports, which cite little, were checked differently, by another model searching the web; on the share of claims supported, which both report, agents (41% to 86%) and professionals (51% and 76%) overlap. In the authors' diagnosis, the agents' reports lacked sector KPIs, justified assumptions and scenario analysis. On forecasts the analysts were more accurate on most line items (Figure 3).[29]

---

the authors' employer, built and runs the benchmark. Read in full. Paraphrased.

[28] BankerToolBench v1.1 README (Handshake AI, GitHub), read in full: the same 100 tasks, simplified, with more objective rubrics. The pass rates (Claude Opus 5 32.6%, GPT-5.4 13.7%) come from two aggregator sites quoting Handshake's leaderboard, which is published as an image that could not be read; secondary reports only.

[29] Haque et al., J.P. Morgan AI Research, arXiv:2604.21006 (Apr. 2026), secs. 4–6, tables 3–4, 7, 11 and 14, and apps. A.4.1, A.5 and A.7.4. 50 reports per firm on 25 randomly sampled S&P 500 companies, fiscal 2025 Q1 and Q2; the agents searched the open web and wrote their reports after the professional reports, sometimes after the earnings were released, with the cut-off set in the prompt, so look-ahead would favour the agents. The quality judge was GPT-5, checked against eight domain experts on eight companies. Agents' claims were checked against their cited URLs by GPT-4.1, with GPT-4o web search as a fallback: hallucination rates 11.2% to 34.1% for reports in Firm A's structure (table 7; 53.2% to 86.0% supported) and 10.8% to 45.7% in Firm B's (table 14). The professional reports cite little, so their claims were extracted and checked by GPT-5.2 searching the web: error rates 4.72% (Firm A) and 2.32% (Firm B), supported 75.93% and 51.08% (app. A.7.4). The authors note that claims a model flags as errors were generally rounding or conflicting evidence and may not be wrong. The two procedures differ, so the hallucination and error rates are not like-for-like; both report the share of claims supported, and there the agents and the professionals overlap within each firm: 53.2% to 86.0% against 75.93% in Firm A's format, 40.6% to 83.0% against 51.08% in Firm B's (tables 7 and 14). On buy and sell calls the main and appendix tables disagree and are not used. No uncertainty reported. Read in full. Paraphrased.

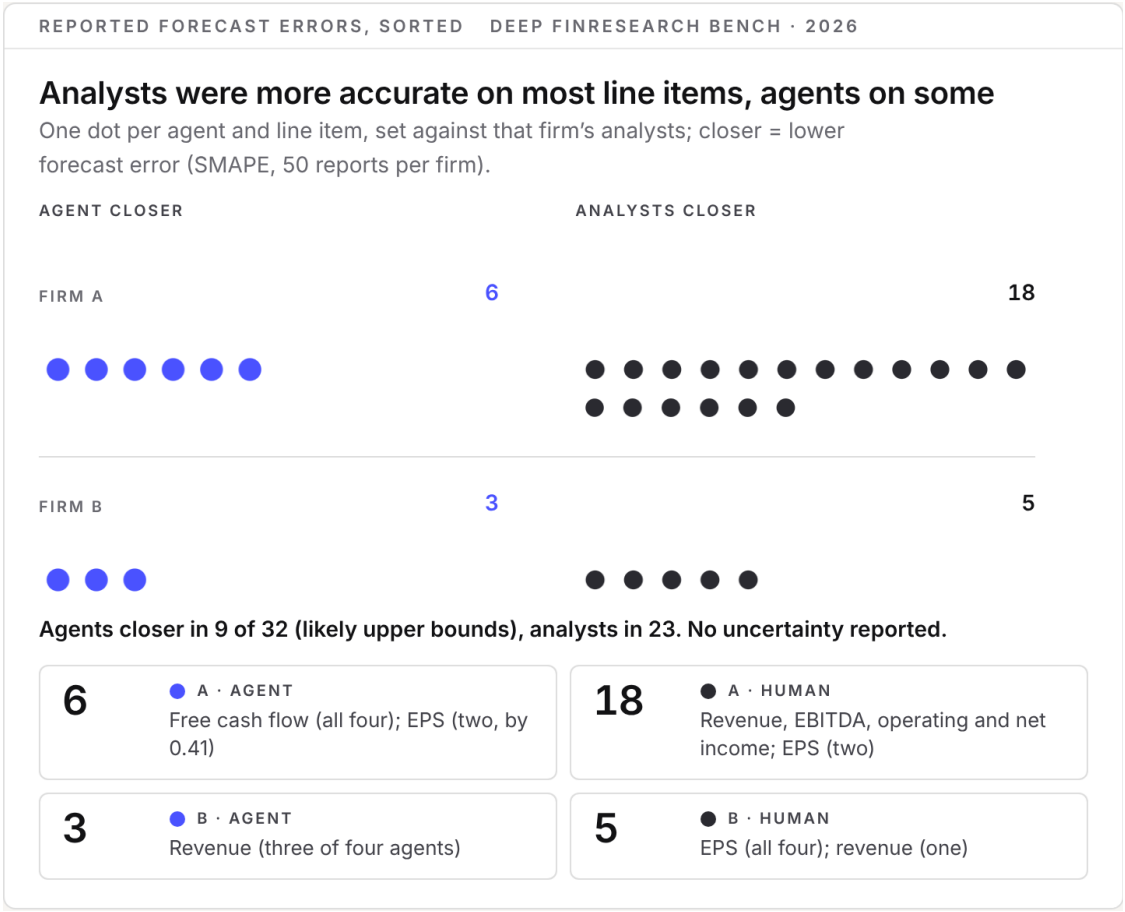

Figure 3. Analysts were more accurate on most forecasts; agents on free cash flow and, for two of them, EPS against Firm A, and three on revenue against Firm B. Each dot is one of four commercial web-search agents against one firm's analysts on one line item, by percentage forecast error (SMAPE) over 50 reports per firm; 9 of 32 favour the agent. The agent wins are likely upper bounds, as reports were written after the fact under a prompted cut-off; two are by 0.41 points; no uncertainty is reported. Call tables are omitted: the paper's main and appendix versions disagree. The authors conclude that professionals still do better research.

Forecasting is closer. In EPSBench, agents forecast quarterly earnings per share from filings, call transcripts and news dated before each announcement; the three strongest showed no detectable difference in error from analyst consensus, each closer on 50% of 149 firm-quarters. Removing both kinds of source cut one model's win rate to 31%.[30] Pre-cutoff results claiming

[30] Yin, Yanbin, et al., arXiv:2610.38293 (Oct. 2026), secs. 3–5, tables 1–3. 149 firm-quarters from 86 US companies, announced 2 Dec. 2025 to 24 June 2026; four of six models' stated cut-offs precede every announcement. The three strongest models' median price-deflated errors were 0.113% to 0.131% against consensus 0.111%, with paired 90% intervals that include zero; the authors note this does not establish equivalence. The source ablation is one model (GPT-5.5): with news alone its error did not detectably change (paired change −0.000 points, 90% interval −0.010 to +0.009; win rate 40% against 50%) while its input use fell to about a quarter, with filings and transcripts alone it rose slightly (+0.009, +0.002 to +0.016), and with neither it rose by 0.033 (+0.019 to +0.047), its win rate falling from 50% to 31% (table 3, sec. 5.4). With all

more are suspect: Levy found GPT-4's 60% accuracy at predicting earnings direction fell to chance when only the least significant digit of each financial statement item was randomised, the signature of memorisation.[31]

Two gaps explain most of these failures. Factuality is a verification gap: claims were wrong or unsupported, and nothing checked them. Quality is a structure gap: the reports lacked what professionals build in, and BankerToolBench's authors found that adding banker domain context and formatting guidance to prompts significantly improved scores.[32] The model matters too. In FinanceHarness, a finance harness lifted one open model from 25.3% to 32.4%, but the authors found that changing the model spread results more than changing the scaffold.[33] Westworld's vendor-run diligence benchmark attributes 74% of score variation to the model and 10% to the harness,[34] while Samaya's FrontierFinance finds the harness dominant, from 33.0% with web search alone to 56% with its own retrieval.[35]

---

sources it tied consensus; no condition beat it. The ablation removes sources rather than varying how much of a fixed set the agent read, and the agents forecast through a supplied schema and accounting calculator. The weaker three models lost to consensus. Calibration was poor: surprise-direction accuracy was at most 63% (GPT-5.5), and the models' 80% prediction intervals covered 37% to 70% of outcomes, with 20% to 38% landing above the upper bound (secs. 5.1–5.2). The comparison is with consensus, not individual analysts, and is not controlled for information access. Read in full. Paraphrased.

[31] Levy, *Journal of Accounting Research*, 2026, introduction and secs. 3–4: a re-creation of an earlier study's earnings-direction test on anonymised statements; the author also finds that net income and sales figures together identify over 99.9% of earnings announcements. Read in full. Paraphrased.

[32] Lau et al., BankerToolBench, sec. 5 and fig. 9. The authors' wording is that domain and formatting context "significantly improve performance"; no test statistic is printed beside it.

[33] Xiao et al., FinanceHarness, arXiv:2607.27853 (July 2026), abstract, secs. 4–6 and table 4 (Qwen3.6-27B: plain search 25.3, a naive harness 29.6, the full harness 32.4); scored by a language-model judge; questions generated by a model and validated by experts; no human analyst score. Main text read in full; tables and appendix read through a summarising fetch, so only the printed totals are used. Paraphrased.

[34] Halluminate, "Westworld Finance Diligence Bench", 11 Aug. 2026, section "Model or harness": a two-way decomposition of 20 model-by-harness mean scores over 88 problems from one anonymised acquisition; the best configuration earned 0.61 of available credit. A vendor-run benchmark published as a blog post; no human score. Read in full. Paraphrased.

[35] Zhang, Yuhao, et al., Samaya AI, arXiv:2608.11683 (Aug. 2026), secs. 5–6 and table 4: Claude Opus 4.8 with a web-search harness 33.0%; the best system under a finance tool harness 49.2%; Samaya's own system 56%. Rubrics written by former buy-side, sell-side and banking staff; vendor-authored; no human score. Read in full. Paraphrased.

Comparisons with analysts also need look-ahead controls. In Investor's Last Exam, 57% of one agent's public-web runs retrieved sources dated after their cut-off, some containing the answer.[36] Qian and colleagues found that telling GPT-5 to answer as of a past date cut its claims to know the answer from 39% to 1%, while its accuracy stayed at about 0.76.[37] Only time-frozen corpora or post-cutoff tests settle such comparisons.

05 Who has to check

No finance study has yet tested whether an expert checking AI's research beats both the expert and the AI alone on full deliverables; the nearest are small, bounded or outside finance, and show that who checks, and how, matters. In the one finance experiment with all three arms, a Federal Reserve study of a bounded extraction task, three analysts answered 30 yes/no, categorical or exact-figure questions about each of nine European banks from their public disclosures, against an answer key vetted by an earlier research team. Working alone with a PDF reader they were right 52.9% of the time, at four minutes a question. Recording an AI tool's first answer, they were right 58.9% at 0.3 minutes, a gain the authors could not distinguish from chance. Free to question the tool and open the documents (they opened them on a quarter to two-fifths of questions), they were right 65.9% at 1.2 minutes, a significant gain (Figure 4).[38] The gain was not an expert's: the analyst with the most subject

---

[36] Jiang, Huangfei, et al., "Investor's Last Exam", arXiv:2610.38342 (Oct. 2026), sec. 5 and app. I: 49 of the agent's runs on the public web, with a dated query, received sources dated after their cut-off; its score was about the same as on the time-gated corpus (68.3 against 68.1). No human comparator. Read in full. Paraphrased.

[37] Qian et al., arXiv:2611.06808 (Nov. 2026), abstract, sec. 1 and table 4: 2,226 dated US facts (first-release macro data, month-end prices, Federal Reserve decisions), chance accuracy about 0.27; accuracy 0.76 before and 0.75 after the dated instruction; one run per prompt. Read in part (about 60%, including the main results table). Paraphrased.

[38] Botti et al., IFC Bulletin 67 (Bank for International Settlements, July 2026; prepared for a February 2025 workshop), secs. 3.1–3.6 and 5.1, tables 5, 9 and 14. Questions were yes/no, categorical or exact figures, answered from the banks' public disclosures; the answer key is the vetted dataset of Beltran et al. (2023). The tool was Claude 3.5 Sonnet in a basic retrieval set-up the authors call commonly accessible but not state of the art. The 52.9% is 81 of 153 answers; 58.9%, 99 of 168; 65.9%, 114 of 173. In the interactive condition analysts could ask follow-up questions and consult the PDFs; they opened them on 24% of low-complexity and 38% and 39% of medium- and high-complexity questions, and neither opening them nor the number of messages was significantly associated with accuracy (sec. 5.1). Significance is from the authors' logistic regression with annotator and complexity controls ($p = 0.014$ for the interactive condition); the two AI conditions are each compared with the control, not with each other, and whether the standard errors are clustered by analyst cannot be seen in the part read. Questions were randomly assigned between the control and first-answer conditions, but the interactive phase always ran at least two weeks later, so order and familiarity are not randomised away (sec. 3.5). The authors note that every condition stays below 80% of what a full team review achieves (conclusion, read in a search excerpt). Read through sec. 5.1 in the bulletin text; sec. 5.2 and the conclusion in part. Paraphrased.

expertise scored 72.9% alone, 60.0% recording the tool's first answer and 66.7% working with it, while the least accurate analyst rose from 33.3% to 75.0%.[39] This is evidence about AI assistance on a bounded task, with a 2024 model, not about expert verification.

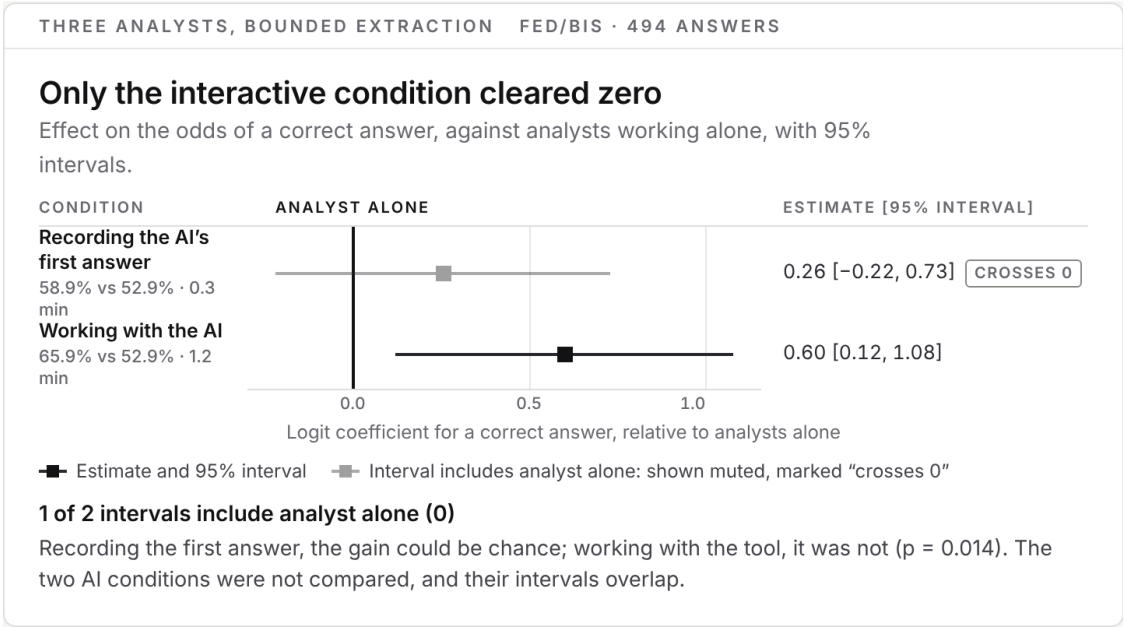

Figure 4. Only working with the AI interactively gave a significant gain over analysts alone; recording its first answer did not, and the two AI conditions were not compared. Logit coefficients for a correct answer against analysts alone, with 95% intervals (coefficient plus or minus 1.96 printed standard errors) and analyst and complexity controls; 494 answers by three analysts. Accuracy: 52.9% alone, 58.9% recording the first answer, 65.9% working with the AI (follow-up questions, documents available; opened on 24–39% of questions, not significantly linked to accuracy). The most expert analyst scored lower with the AI than alone; the least accurate gained most. A 2024-era model; a convenience sample.

The checker's expertise matters. In Tseng and Chang's study of debt contracts, AI without a verification loop classified 8 of 10 contracts correctly, an accountant-guided loop 9 of 10 and a loop guided by a non-accountant 5 of 10, each with one verifier, no human-alone arm and no significance test.[40] In Autor and colleagues' Google-funded trial with 133 patent lawyers, three

[39] Botti et al., table 13: analyst C, the most domain-expert and least AI-skilled of the three, 72.9%, 60.0% and 66.7% across the three conditions; analyst B 33.3%, 68.3% and 75.0%; analyst A 48.6%, 53.2% and 60.3%. Each cell is about 50 answers. Accuracy gains were concentrated in low-complexity questions (66.7% alone, 90.0% interactive) against high-complexity ones (34.5% and 40.7%) (table 11).

[40] Tseng and Chang, *Accounting Horizons*, published online 3 Nov. 2026. The article is paywalled and was not read; the figures come from the journal's synopsis, a blog post by the first author (10 Nov. 2026) and a co-author's description of the method in another paper. Each test sample held ten contracts, so 80%, 90% and 50% are 8, 9 and 5 of 10; the accountant's later rounds reached 10 of 10 on new samples. The synopsis states that a non-accountant loop dropped accuracy below the no-loop baseline. Read in part (synopsis and secondary descriptions only).

months of AI drafting improved seniors' unaided review of an AI-written draft by 0.45 standard deviations; juniors showed no average gain. While using the AI, treated lawyers' drafts scored 0.34 to 0.38 standard deviations higher, with larger gains among juniors.[41]

Checking has to be claim by claim, and by someone other than the model. Luo and colleagues verified 77,272 claims in 1,440 AI-written investment memos against each memo's own filings: only 1.17% to 1.77% were refuted, yet 36.9% to 41.1% of memos contained at least one.[42] In the vendor-run ATLAS-Finance, one agent's workbook passed all 13 of its own checks, and a second implementation the agent built agreed with it, because both used the same wrong deduction; financial logic, not data handling, caused 53.8% of failures.[43]

Adding a human is not enough on its own. Vaccaro and colleagues' meta-analysis of 54 experiments found human-AI teams worse on average than the better of human or AI alone ($g = -0.35$); in an exploratory moderator analysis, teams that divided the work did relatively better ($p = 0.031$).[44] And expertise is not what catches obvious errors: in Jin and colleagues'

Paraphrased.

[41] Autor et al., NBER Working Paper 35720 (Sept. 2026, revised Oct. 2026), abstract and secs. 1–3. A pre-registered trial at eleven US IP firms; 91 of 133 lawyers completed it; blinded expert graders. The study was run and funded by Google, which employs six of the seven authors, and the tool was unreleased. 15 of 91 review submissions may have used AI; excluding them, the senior effect stays significant at the 10% level. The outcome is the lawyers' own review skill after AI use, not a test of AI output checked by them against AI output unchecked. Read in part (abstract, introduction, design; read as excerpts); the result tables were not read, and the drafting gains (0.34 SD at 10 days, 0.38 SD at 90 days) are from the abstract excerpt. Paraphrased.

[42] Luo et al., arXiv:2611.04375v1 (Nov. 2026), abstract, secs. 3–6, tables 2–3: 30 companies over 12 quarters; four agents (Codex with GPT-5.5, Claude Code with Sonnet 4.6, DeepSeek-V4-Pro, Qwen3.8-flash); human review confirmed 188 of 200 sampled refutations. The authors did not test whether a verify-and-revise loop reduces the share of memos with an error. No human-written memos were compared. Main text read in full. Paraphrased.

[43] Lau et al., ATLAS-Finance, arXiv:2611.22085 (Nov. 2026), secs. 5.3–5.4: 100 investment-banking tasks in simulated workplaces, built by bankers; the best model passed 10.7% at the first attempt. Handshake AI built and ran the benchmark. No human score. Main text read in full. Paraphrased.

[44] Vaccaro, Almaatouq and Malone, arXiv:2610.42001 (Oct. 2026), main text: 171 effect sizes from 54 peer-reviewed experiments that also tested AI alone, 95% interval −0.52 to −0.18; against humans alone, teams did better ($g = 0.53$). In an exploratory moderator analysis, some division of labour was associated with more augmentation ($p = 0.005$) and more synergy ($p = 0.031$), not with net gains over the better solo performer. The $g = 0.53$ is from the review's full set of 1,955 effect sizes, not only the 54 experiments. The review was run by AI agents with small human audits; heterogeneity is high (I^2 about 93%); not specific to finance. Main text read in full. Paraphrased.

experiment, legal professionals rejected conspicuously wrong AI advice no more reliably than laypeople did.[45] The value of an expert lies in subtle errors, which none of these studies isolates.

06 Conditions to test, and implications

The evidence points to conditions that can be tested before anyone relies on grounded AI research; the second is the least tested. Five follow:

- **Claims checked against the source.** Every claim carries a pointer to the passage that supports it and is checked there (Luo; CLAIR-Fin; DealsFlow).
- **An expert checker with time.** The verifier knows the domain and has the time and standing to reject the draft. Non-expert and rushed checking did worse in the studies above (Tseng and Chang; Autor; Xue).
- **Domain structure.** The work follows the profession's own frame: KPIs, assumptions, scenarios, house formats (Deep FinResearch Bench; BankerToolBench; GAUGE).[46]
- **Look-ahead controlled.** Evidence is frozen at the decision date, and results are tested after the model's training cut-off (Investor's Last Exam; Qian; Levy).
- **Independent grading.** The grader is not the system, and its agreement with experts is measured (BankerToolBench; Kensho; GAUGE).

For funds, the evidence supports using grounded AI for reading, extraction and first drafts now, and treating expert review as a cost to be budgeted and measured, not a formality whose value is assumed. A team that accepts output unchecked, or hands the checking to whoever has time, should expect the results in sections 04 and 05.

For labs, the open problems are the structure and judgment gaps, not retrieval: the agents in these studies found the right documents and still missed the analysis a professional builds on them. Benchmarks that hold evidence to a date, grade against professional practice and report a human score are what separate those gaps.

Two questions remain open. The first is whether an expert checking AI beats both the expert and the AI alone on full research and diligence deliverables: a finance trial with those three arms, on work seeded with subtle errors, scored blind. The second is whether synthetic analyst work, with experts' material put back in, trains a better model than professionals' own work, tested on real outcomes.

Aperture produces research for buy-side and sell-side institutions; it would benefit commercially if these claims held, and nothing in this essay tests Aperture's own research. Until such tests are published, AI has beaten professionals here only on bounded legal work; in finance, verification and structure remain conditions to test.

---

[45] Jin, Sun and Zhang, *Legal and Criminological Psychology*, 2026: 222 participants (62 legal professionals, 62 law students, 98 laypeople), Chinese law; expertise had no significant effect on adopting flawed advice ($p = 0.067$). The authors note the errors were deliberately conspicuous and that real errors tend to be subtle. Read in part (about half, including all main results). Paraphrased.

[46] Lu, Jiacheng, et al., sec. 5.4: industry assumption tables in the prompt, with no figures from the test company, raised the valuation-judgment subscore by 4.0 (95% interval 1.3 to 6.7).

Publication scope

A research commentary on published work, revision r8.9, published 19 November 2026. It is not a systematic review, a replication or an independent audit. Figure 1 restates published task-success rates; Figure 2 restates published scores side by side, the best AI system against professionals' average; Figure 3 restates published forecast errors; Figure 4 restates a published regression, with intervals computed by Aperture from its printed standard errors. Several sources are vendor-run or read only in part, as their notes state. No Aperture system, data or result is reported.

Works cited

Published sources

Akshansh, et al. "Trading Human Curation for Synthetic Augmentation in RLVR." *arXiv*, version 2, 2026, arxiv.org/abs/2606.03800. Accessed 18 Nov. 2026.

Alonso-Robisco, Andrés, et al. *Retriever or Reasoner? Decomposing Retrieval-Augmented Generation Performance in External Audit Supervision*. Documentos Ocasionales 2613, Banco de España, July 2026, repositorio.bde.es/handle/123456789/43986. Accessed 18 Nov. 2026.

Ambrogi, Bob. "Vals AI's Latest Benchmark Finds Legal and General AI Now Outperform Lawyers in Legal Research Accuracy." *LawSites*, 23 Oct. 2025, www.lawnext.com/2025/10/vals-ais-latest-benchmark-finds-legal-and-general-ai-now-outperform-lawyers-in-legal-research-accuracy.html. Accessed 18 Nov. 2026.

Autor, David, et al. "Does AI Assistance Enhance or Erode Expertise? Evidence from a Three-Month Field Experiment in Patent Drafting." NBER Working Paper 35720, National Bureau of Economic Research, Sept. 2026, www.nber.org/papers/w35720. Accessed 18 Nov. 2026.

Banerjee, Shourik, et al. "Human or Synthetic? What a Fixed Budget Buys for Post-Training." *arXiv*, 12 Nov. 2026, arxiv.org/abs/2611.16895. Accessed 18 Nov. 2026.

Barlow, Stephen H., et al. "Using Synthetic Records to Improve Automated Identification of Seizure Freedom in Clinical Text about People with Epilepsy." *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, 2026, pp. 20–30, aclanthology.org/2026.bionlp-1.3. Accessed 18 Nov. 2026.

Botti, Nicholas, et al. "Efficacy of AI RAG Tools for Complex Information Extraction and Data Annotation Tasks: A Case Study Using Banks' Public Disclosures." *IFC Bulletin*, no. 67, Bank for International Settlements, July 2026, www.bis.org/2026-07/ifcb67_10.pdf. Accessed 18 Nov. 2026.

Chandra, Abhranil, et al. "Shape of Thought: When Distribution Matters More than Correctness in Reasoning Tasks." *arXiv*, version 2, 22 Jan. 2026, arxiv.org/abs/2512.22255. Accessed 18 Nov. 2026.

Faria, Fatema Tuj Johora, et al. "CLAIR-Fin: An Adversarial Multi-Agent Framework for Claim-Level Verification and Adaptive Debate in Cross-Modal Financial QA." *arXiv*, 7 Sept. 2026, arxiv.org/abs/2608.13706. Accessed 18 Nov. 2026.

Gunjal, Anisha, et al. “Rubrics as Rewards: Reinforcement Learning beyond Verifiable Domains.” *International Conference on Learning Representations (ICLR 2026)*, 2026, proceedings.iclr.cc/paper_files/paper/2026/file/cfd7bee7a651ee9af525098ef67a9e45-Paper-Conference.pdf. Accessed 18 Nov. 2026.

Gupta, Tanmay, et al. “MolmoWeb: Open Visual Web Agent and Open Data for the Open Web.” *arXiv*, 9 Apr. 2026, arxiv.org/abs/2604.08516. Accessed 18 Nov. 2026.

Halluminate. “Westworld Finance Diligence Bench.” *Halluminate*, 11 Aug. 2026, www.halluminate.ai/blog/due-diligence-bench. Accessed 18 Nov. 2026.

Haque, Mirazul, et al. “Deep FinResearch Bench: Evaluating AI’s Ability to Conduct Professional Financial Investment Research.” *arXiv*, 22 Apr. 2026, arxiv.org/abs/2604.21006. Accessed 17 Nov. 2026.

Harsha, Chetan, et al. “Synthetic Data Generation Using Large Language Models for Financial Question Answering.” *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, 2025, pp. 76–95, aclanthology.org/2025.finnlp-1.7. Accessed 18 Nov. 2026.

Himi, Y., et al. “Clinical Event Extraction in Lung Cancer Progress Notes Using Synthetic Data Derived from Case Reports: Development and Evaluation Study.” *JMIR Preprints*, 26 June 2026, doi.org/10.2196/preprints.105600. Accessed 18 Nov. 2026.

Hou, Yunlong, et al. “Rethinking ‘RL Generalizes, SFT Memorizes’: The Role of SFT Data.” *arXiv*, version 1, 11 Nov. 2026, arxiv.org/abs/2611.14912. Accessed 18 Nov. 2026.

Huang, Wen-Fong, and Edwin Simpson. “Efficient Financial Language Understanding via Distillation with Synthetic Data.” *arXiv*, 2026, arxiv.org/abs/2606.18875. Accessed 18 Nov. 2026.

Ioanovici, A.-C., et al. “Enhancing and Not Replacing Clinical Expertise: Improving Named-Entity Recognition in Colonoscopy Reports through Mixed Real-Synthetic Training Sources.” *Journal of Personalized Medicine*, vol. 15, no. 8, 2025, p. 334, www.mdpi.com/2075-4426/15/8/334. Accessed 18 Nov. 2026.

Jiang, Huangfei, et al. “Investor’s Last Exam: Can AI Agents Do Equity Investment Without Seeing the Future?” *arXiv*, Oct. 2026, arxiv.org/abs/2610.38342. Accessed 18 Nov. 2026.

Jin, Yaohua, et al. “The Trojan Horse in the Courtroom: The Impact of Flawed AI Advice on Legal Discretionary Decision-Making.” *Legal and Criminological Psychology*, 23 Oct. 2026, doi.org/10.1111/lcrp.70071. Accessed 18 Nov. 2026.

Krumdick, Michael, et al. “FrontierFinance: A Long-Horizon Computer-Use Benchmark of Real-World Financial Tasks.” *arXiv*, 7 Apr. 2026, arxiv.org/abs/2604.05912. Accessed 18 Nov. 2026. Authors at Kensho and S&P Global.

Lau, Elaine, et al. “BankerToolBench: Evaluating AI Agents in End-to-End Investment Banking Workflows.” *arXiv*, 13 Apr. 2026, arxiv.org/abs/2604.11304. Accessed 18 Nov. 2026.

Lau, Elaine, et al. “ATLAS-Finance: Benchmarking AI Agents in Realistic Banking Environments.” *arXiv*, 17 Nov. 2026, arxiv.org/abs/2611.22085. Accessed 18 Nov. 2026.

Levy, Bradford. “Caution Ahead: Numerical Reasoning and Look-Ahead Bias in AI Models.” *Journal of Accounting Research*, 6 May 2026, doi.org/10.1111/1475-679x.70058. Accessed 18 Nov. 2026.

Lu, Jiacheng, et al. “One Analyst Is Not Ground Truth: Grading Agent-Built Financial Models Against Observed Professional Practice.” *arXiv*, Sept. 2026, arxiv.org/abs/2607.24889. Accessed 18 Nov. 2026.

Lu, Y., and K. Smith. “Feeding LLM Annotations to BERT Classifiers at Your Own Risk.” *arXiv*, 21 Apr. 2025, arxiv.org/abs/2504.15432. Accessed 18 Nov. 2026.

Luo, Chenri, et al. “LLM-Generated Investment Research Memos Need Validation.” *arXiv*, version 1, 3 Nov. 2026, arxiv.org/abs/2611.04375. Accessed 18 Nov. 2026.

Mendes, Ethan, et al. “Didactic to Constructive: Turning Expert Solutions into Learnable Reasoning.” *International Conference on Machine Learning (ICML 2026)*, 2026, arxiv.org/abs/2602.02405. Accessed 18 Nov. 2026.

Pan, Teng, et al. “Agent-R^2: Scaling Agents for Realistic Productivity Work via Real-Source Data Synthesis.” *arXiv*, 5 Nov. 2026, arxiv.org/abs/2611.08219. Accessed 18 Nov. 2026.

Qian, Lingfei, et al. “LLMs Don’t Know What They Know: Benchmarking Temporal Self-Knowledge for Financial Backtesting.” *arXiv*, 4 Nov. 2026, arxiv.org/abs/2611.06808. Accessed 18 Nov. 2026.

Su, Buxin, et al. “Supervised Fine-Tuning with Synthetic Rationale Data Hurts Real-World Disease Prediction.” *arXiv*, 2026, arxiv.org/abs/2606.10279. Accessed 18 Nov. 2026.

Tang, Ming Ze, et al. “DealsFlow: An Evidence-Gated Agentic Harness for Venture Capital Deal Triage.” *arXiv*, 23 Oct. 2026, arxiv.org/abs/2610.32380. Accessed 18 Nov. 2026.

Tseng, Ayung, and Chengyee Janie Chang. “Domain Knowledge and Human-AI Collaboration: Extracting Accounting Information from Contracts.” *Accounting Horizons*, published online 3 Nov. 2026, doi.org/10.2308/HORIZONS-2026-019. Accessed 18 Nov. 2026.

Vaccaro, Michelle, et al. “AutoAnalysis: An Automated, Self-Updating Meta-Analysis of Human-AI Collaboration.” *arXiv*, 30 Oct. 2026, arxiv.org/abs/2610.42001. Accessed 18 Nov. 2026.

Vals AI. “Vals Legal AI Report.” *Vals AI*, 27 Feb. 2025, www.vals.ai/industry-reports/vlair-2-27-25. Accessed 18 Nov. 2026.

Vals AI. “VLAIR – Legal Research.” *Vals AI*, 14 Oct. 2025, www.vals.ai/industry-reports/vlair-10-14-25. Accessed 18 Nov. 2026.

Xiao, Yijia, et al. “FinanceHarness: Autonomous Financial Deep Research Framework.” *arXiv*, 30 July 2026, arxiv.org/abs/2607.27853. Accessed 18 Nov. 2026.

Xue, Jian, et al. “Generative AI for Analysts.” *arXiv*, version 2, Sept. 2026, arxiv.org/abs/2512.19705. Accessed 18 Nov. 2026.

Yin, Yanbin, et al. “EPSBench: Benchmarking LLM Agents for Earnings Forecasting.” *arXiv*, 28 Oct. 2026, arxiv.org/abs/2610.38293. Accessed 17 Nov. 2026.

Zhang, Yuhao, et al. “FrontierFinance: A Challenging Benchmark for Measuring Frontier Intelligence of Finance Agents.” *arXiv*, 12 Aug. 2026, arxiv.org/abs/2608.11683. Accessed 18 Nov. 2026. Authors at Samaya AI.