

FINANCEHARNESS: Autonomous Financial Deep Research Framework

Yijia Xiao^{1,2*} Rujun Han¹ Yanfei Chen¹ Zifeng Wang¹ Ke Jiang¹
Zhongying Cui¹ Wei Wang² Tomas Pfister¹ Chen-Yu Lee¹

¹Google Cloud AI Research ²University of California, Los Angeles

Abstract

Powered by advances in LLMs and autonomous agents, deep research has become one of the most widely adopted agentic products. However, most deep research systems write general-purpose reports, which are inadequate for financial deep research. Financial research demands specialized knowledge to analyze historical patterns and forecast upcoming events. Automating financial deep research therefore requires both a layered harness to drive the research agent and a verifiable, point-in-time benchmark that prevents leakage of future information. We present FINANCEHARNESS, a harness that runs finance-oriented tools and practitioner-guided workflows, automating financial deep research end to end: environment and data construction, the agent execution loop, and reward modeling. We further propose FINANCEGYM, comprising thesis-driven research questions and rubrics that combine pre-cutoff and post-cutoff criteria. Professional expert validation yields an 82% pass rate. With the same open-weight backbone, FINANCEHARNESS improves the overall rubric score from 25.3% to 32.4%, demonstrating the effectiveness of our specialized harness design. However, even pairing FINANCEHARNESS with the most cutting edge LLM (e.g. Opus-5), the FINANCEGYM score is below 45%, showing that it is a challenging benchmark for financial deep research. FINANCEGYM¹ and FINANCEHARNESS² code are publicly available.

1 Introduction

Deep research is now among the most widely adopted agentic products in the industry (OpenAI, 2025; Perplexity, 2025; Team et al., 2025b;

Google, 2026). Equipped with leading LLMs and a carefully designed agent harness, deep research conducts comprehensive searches that can answer challenging multi-hop reasoning questions and synthesize insightful long-form reports from multiple sources and topics (Han et al., 2025; Team et al., 2025b,a, 2026).

Despite the impressive progress, a general deep research harness falls short for highly specialized domains (Wu et al., 2023; Xie et al., 2024, 2023). In this work, we target *financial deep research*, which requires capabilities such as evidence gathering, cross-source validation, and report synthesis that go beyond a generic deep research skillset. It further requires the system to grasp the relationships among financial entities and events, and anticipate their evolution over time (Ding et al., 2015; Xu and Cohen, 2018). For example, an equity analyst investigating a semiconductor company needs to examine historical earnings reports, analyze current competitor dynamics, and predict future demand and macroeconomic conditions to derive a reasonable rating for the stock. A general deep research system without financial tools and expertise to analyze earnings reports and economic events may fail to identify specific information for such complex analysis.

Moreover, despite the abundance of recent deep research benchmarks (Coelho et al., 2025; Du et al., 2025; Li et al., 2026a; Sharma et al., 2025; Wang et al., 2025), none of them provide scalable evaluation data for financial research. Existing deep research benchmarks focus mostly on testing past or current conditions. Professional financial research reports typically need to reason about future events (post-cutoff reasoning), which requires point-in-time (PIT) evaluation to avoid leakage of future information. On the other hand, existing financial NLP benchmarks target isolated capabilities (sentiment, entity recognition, filing QA) (Shah et al., 2022; Islam et al., 2023; Yang et al., 2023), which

*This work was done while Yijia Xiao was a Student Researcher at Google Cloud AI Research. Corresponding author(s): {yijia.xiao, weiwang}@cs.ucla.edu, {rujunh, chenyulee}@google.com.

¹<https://financegym.github.io/>

²<https://github.com/Yijia-Xiao/FinanceHarness>

are insufficient to measure the quality of comprehensive, temporally constrained financial reports.

We address these gaps with a financial deep research suite built on a PIT search sandbox. We construct a large-scale web corpus with real publication dates from the public web, and build a dense retrieval system with cutoff-date access control. This ensures a financial deep research agent can locate finance-related knowledge effectively, while information published beyond each question’s cutoff date is excluded from the research reports.

Leveraging this sandbox, we construct FINANCEGYM, a large-scale, expert-validated financial deep research benchmark. We build an entity graph from the corpus and sample financial situations from this graph to ground each research question and its rubric. Drawing on input from financial practitioners, we generate an investment thesis and paired pre-cutoff and post-cutoff rubrics for every question. A multi-stage filtering pipeline then removes low-quality records, and expert annotators validate the rest to form the final benchmark.

Building on this new environment, we propose FINANCEHARNESS, an expert-knowledge-guided agent harness for financial deep research (Section 4). Following the common definition in the community, FINANCEHARNESS implements a harness as layered services of LLM-facing tools, APIs, execution workflows, and evaluation rubrics (Lee et al., 2026; Ning et al., 2026). It runs agents against the PIT sandbox and scores them with the FINANCEGYM rubrics, so that evaluation and in-environment optimization share one contract.

We summarize our contributions. (1) We build a point-in-time financial search sandbox to serve as the foundation for FINANCEGYM: an expert-validated financial deep research benchmark of 400 high-quality research questions and rubrics that separate pre-cutoff evidence retrieval from post-cutoff reasoning. (2) We propose FINANCEHARNESS, an expert-knowledge-guided harness that automates environment and data construction, agent execution, and reward modeling for financial deep research under a strict point-in-time contract. (3) We benchmark a wide range of leading LLMs and agents and show that FINANCEGYM is challenging, with every system scoring below 45%, making it a useful target for advancing financial deep research.

2 Related Work

Financial benchmarks. Financial NLP benchmarks have traditionally targeted isolated skills such as sentiment analysis, named entity recognition (Shah et al., 2022), factual question answering over SEC filings (Islam et al., 2023), or broad evaluation of finance-specialized LLMs (Yang et al., 2023). These tasks are valuable, but they do not capture analyst-style financial deep research: long-form reports that connect entities, sectors, events, and time. FINANCEGYM instead builds questions from a finance entity graph and evaluates reports with a two-tier rubric that separates pre-cutoff-date evidence retrieval from post-cutoff-date outcome anticipation.

Research agents. Modern research agents build on retrieval-augmented generation (Lewis et al., 2020) and ReAct-style tool use (Yao et al., 2022), but differ in where the research policy lives. Scaffolded agents encode orchestration in code, including Test-Time Diffusion for Deep Research (TTD-DR) (Han et al., 2025), GPT-Researcher,³ STORM (Shao et al., 2024), and subagent-based frameworks. Specialized deep-research models instead train the backbone on agentic-search trajectories with a fixed tool stack, as in Tongyi-DeepResearch-30B-A3B (Tongyi-DR) (Team et al., 2025b) and MiroThinker-1.7-mini (Team et al., 2025a, 2026). Our evaluation includes trained models, a fixed ReAct search wrapper with multiple backbones, and engineered harnesses on the same PIT corpus retriever, which lets us separate backbone capability, scaffold contribution, and tool-distribution shift.

Deep-research evaluation. Recent deep-research benchmarks emphasize long-form, citation-grounded answers and richer rubrics (Coelho et al., 2025; Du et al., 2025; Li et al., 2026a; Sharma et al., 2025; Wang et al., 2025), but most rely on live-web retrieval or a single global snapshot rather than a per-question publication-date cutoff. LLM-as-judge methods offer scalable evaluation (Zheng et al., 2023), yet finance requires criterion-level attribution: a report may retrieve the right historical facts while failing to anticipate the later outcome, or reason plausibly while omitting source-grounded evidence. We therefore use evidence-aware rubric

³<https://github.com/assafelovic/gpt-researcher>

judging with explicit hindsight and foresight criteria; Appendix A.2 gives a compact positioning table against representative benchmarks.

3 FINANCEGYM Construction

As mentioned in Section 1, our benchmark infrastructure consists of a point-in-time (PIT) financial search sandbox and FINANCEGYM, the benchmark constructed on top of it (Figure 1). In this section, we first describe the corpus, retrieval, and temporal access contract that define the search sandbox (§3.1). We then describe how FINANCEGYM is generated from a finance entity graph, filtered, balanced, and expert-annotated (§3.2). Finally, we define the scoring contract used by the benchmark and by downstream harness training (§3.3).

3.1 Point-in-Time Financial Search Sandbox

Source corpus. We built the search sandbox from a large-scale web corpus that we collected from thousands of public web domains. We built the corpus so that all articles carry reliably extracted publication dates: each article’s date was extracted with `htmldate` (Barbaresi, 2020), allowing the environment to enforce PIT access. We extracted clean text with `trafilatura` (Barbaresi, 2021) with normalized metadata. The resulting corpus contains 100+ million articles, deliberately preserving the noise and breadth of web-scale financial search rather than reducing the task to a curated filing set. The corpus serves only as the internal testbed for building and evaluating the harness. Users can input any public corpus in our pipeline to produce research questions and rubrics. The released benchmark consists of standalone research questions that are decoupled from the underlying corpus (§3.2).

Embedding and storage. Articles are embedded with Qwen3-Embedding-4B (Zhang et al., 2025), producing normalized dense vectors for retrieval. Full texts are stored separately from the vector index so agents can retrieve candidate documents and then fetch source text for citation-grounded synthesis.

Search server. We build a FAISS (Douze et al., 2025) IVF-SQ8 index over the normalized embeddings and serve it through a lightweight API. This API defines the environment contract used throughout the paper: agents can search and read historical documents, but they cannot access articles published after the assigned cutoff date.

3.2 FINANCEGYM: Situation-Driven Benchmark Construction

Finance graph construction. We construct a finance entity graph from the corpus by first filtering to finance-relevant sources and then extracting entity-relation triples with Gemini-3.5-Flash (Team et al., 2023) under structured JSON output. Each edge stores a head entity, relation, tail entity, local context, source URL, and publication date. The extractor produces 5.74M raw edges; after filtering generic or malformed nodes, the working graph contains 4.37M edges. These edges span 1.11M unique entities from 1.20M source articles and serve as the basis for benchmark generation.

Situation mining. Rather than clustering the graph into broad themes, we mine situations that match how financial analysts discover research questions. The miner has three complementary modes: cross-category linkages between entities and event types, temporal narrative arcs around high-degree entities, and polar divergences such as upgrades versus downgrades or earnings beats versus misses. These modes instantiate the fine-grained situation types reported in the harness ablation (Appendix C): a multihop-path type from linkages, a temporal-narrative type, and several divergence (*tension*) types. An entity budget limits repeated sampling of the same primary name, preventing mega-cap firms from dominating the benchmark.

Publication-date cutoff selection and unconstrained generation. Each candidate situation is paired with a cutoff date. For each candidate day d , let $v(d)$ be its event volume, $e(d)$ entity diversity, and $r(d)$ relation entropy, and let $z(\cdot)$ denote a z -score across candidate days. Cutoff dates are then selected by a volume, entity-diversity, and relation-entropy objective:

$$\text{score}(d) = z(v(d)) \left(1 + e(d)\right) \left(1 + \frac{r(d)}{10}\right). \quad (1)$$

The volume term favors eventful days; the diversity and entropy terms penalize batch artifacts dominated by a single entity or one relation type.

Given a (situation, cutoff-date) pair, an LLM generates the benchmark record *without category priming*: the prompt does not provide a topic, sector, or reasoning taxonomy. The output contains an analyst-style question, a reference investment thesis, and a two-tier rubric. Pre-cutoff criteria test facts findable before the cutoff date; post-cutoff

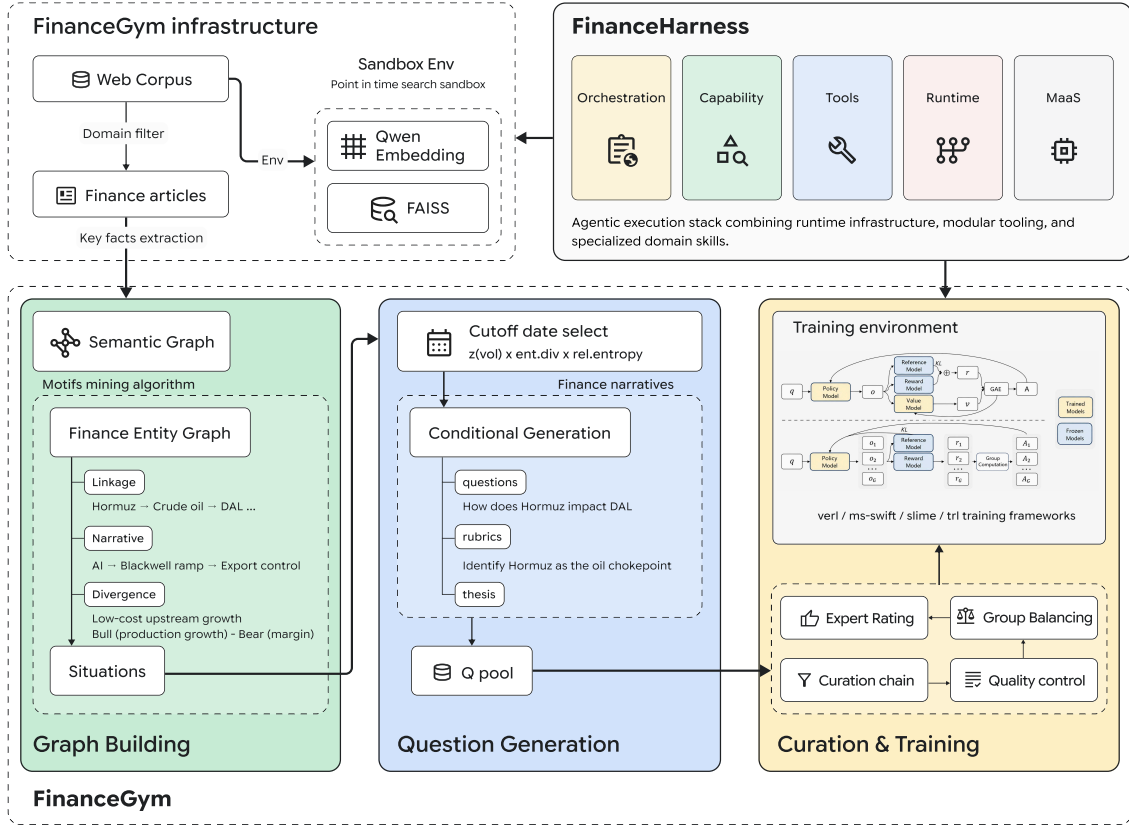


Figure 1: End-to-end benchmark construction. Web articles flow through a domain filter and FAISS + PIT embedding store into a semantic graph, where motif mining yields the finance entity graph and downstream situations; publication-date cutoff selection then drives unconstrained question generation, followed by curation, balancing, expert annotation, and final evaluation.

criteria test outcomes verifiable only after the cut-off date. This split is central to FINANCEGYM: it separates evidence retrieval from forward-looking synthesis while keeping both sides auditable.

Bottom-up taxonomy. After questions are generated, we assign taxonomy labels. LLMs classify batches of generated questions into natural categories, and the raw labels are consolidated into three axes: topic, sector, and reasoning type. This bottom-up taxonomy avoids imposing a prompt-time template on question generation while still allowing the final benchmark to be balanced and reported by interpretable groups.

Quality filtering and balancing. Generated questions pass through a sequence of LLM quality gates before selection for expert review. The filters test feasibility, institutional relevance, coherence, groundedness against evidence, and final balance across topic, sector, reasoning type, and cutoff date. This yields a data-determined 2,078-question publication pool from 29,669 unconstrained genera-

tions. For benchmarking, an integer linear program (ILP) selects a 500-question subset that maximizes quality subject to balance constraints across the taxonomy axes and monthly cutoff-date buckets.

Expert annotation. The balanced 500-question subset receives a final expert-annotation pass through an external professional data vendor, followed by LLM-assisted curation using the annotation labels. Annotators score question feasibility and clarity and mark each rubric item as feasible or infeasible with a short justification. Curation then keeps records with sufficiently clear questions and enough feasible rubrics to grade an agent report: 411 of the 500 annotated questions meet this bar, an 82% professional-annotation pass rate. Each sample requires approximately 1.2 hours of expert work, reflecting the difficulty of validating analyst-style questions and rubric items under a publication-date cutoff. We further balance this pool to the released FINANCEGYM benchmark of 400 expert-annotated questions with 2,464 annotated rubric items, balanced across 9 topics, 11

sectors (merged into 9 leaves), 6 reasoning types, and 12 monthly cutoff-date buckets (Figures 3, 4, and 5).

Release policy. As a final release gate, human experts review every released question along two axes. For quality, they confirm the question is well posed and answerable from public information, with a clear analytical target. For privacy and provenance, they confirm the question only *queries* information. In other words, it refers to publicly observable market events and asks the agent to research them, rather than embedding, quoting, or leaking content from the underlying corpus, which ensures that each question contains neither personal data nor copyrighted text. Questions that fail any check are removed.

After annotations, we do not publicly release the underlying corpus. We release the research questions and the report submission code. The grading runs through our leaderboard with private rubrics to preserve the integrity of our scoring criteria.

3.3 Evaluation Contract

Rubric-based scoring. An LLM judge evaluates each agent report against the question-specific rubric. For each criterion, the judge assigns a score on a 5-tier (0–4) scale with explicit anchors: 0 (not addressed), 1 (mentioned), 2 (partial), 3 (substantive), and 4 (fully grounded). The judge sees the question, thesis, cutoff date, target rubrics, and the agent’s report with cited URLs. Because the rubrics are professionally validated and evidence-linked, the LLM judge is used as a consistent rubric applier rather than as a source of new evaluation criteria.

The primary metric is the *outcome score*: the average over questions of the fraction of rubric points each report earns,

$$\text{Outcome} = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{c \in \mathcal{R}(q)} \text{score}(q, c)}{4 |\mathcal{R}(q)|}, \quad (2)$$

so each question contributes equally regardless of its number of rubric items. Scores can be disaggregated by criterion type, topic, sector, reasoning type, or cutoff-date bucket using the same formula.

Point-in-time protocol. Each question has an assigned cutoff date. During evaluation, the FAISS server enforces PIT filtering: every search result must have publication date $\leq$ the cutoff date, and agents do not access the live web. Pre-cutoff synthesis rubrics test retrieval and synthesis from pre-cutoff evidence. Post-cutoff reasoning rubrics test

whether the agent’s analysis anticipates developments that are only verifiable after the cutoff date. Any detected PIT leakage invalidates the run rather than being treated as an ordinary scoring error.

4 FINANCEHARNESS

In this section, we explain FINANCEHARNESS design and its connection to FINANCEGYM. The point-in-time (PIT) search sandbox defines what information an agent may retrieve before each question’s cutoff date, and FINANCEGYM defines how the resulting report is graded. FINANCEHARNESS is the executable harness built around that contract, which enables users to produce highly specialized reports for their financial research queries. It provides the same retrieval tools, orchestration loop, and rubric signal for both evaluation and optimization, so a model trained inside the harness sees the same tool-use distribution at test time.

4.1 Expert-guided Harness Design

Financial deep research is not only a retrieval problem: an analyst must connect relationships among entities, sector context, numerical evidence, and forward-looking risks into an auditable view. FINANCEHARNESS therefore exposes a layered finance tool interface rather than a single generic web-search action. The core loop covers PIT search, source reading, citation composition, and report finalization. Auxiliary finance tools and workflow skills add structured support for exact data retrieval, calculation, valuation, comparables, risk analysis, and scenario reasoning without expanding the base prompt for every query.

Model and serving layer. The serving layer hosts the orchestrator backbone and a lightweight reader used for long-document evidence extraction. It is deliberately opaque to the rest of the harness: the runtime sees only a request/response contract, so the backbone can be swapped across local serving stacks or closed-model SDKs without changing the orchestration logic.

Runtime layer. The runtime is the control plane. It owns the bounded agent loop, parsing and dispatch, schema validation, tier loading, prompt-mode selection, recovery, and the registries for tools and workflow skills. It contains no model intelligence; instead, it enforces cross-layer invariants such as schema conformance, tool-result chaining, citation finalization, and run-level budget limits.

Tool surface. The tool surface is tiered to keep the initial prompt small while preserving a broad set of financial capabilities. The always-loaded tier contains the core-loop actions. Deferred tools are exposed through a compact catalog, with full schemas loaded only after the model commits to a tool family. The extension tier supports MCP or CLI-backed connectors, including paid data vendors and internal systems.

Modes and skills. FINANCEHARNESS uses prompt-variant modes over a consistent tool surface. The research mode emphasizes web-first evidence gathering; the analytical mode emphasizes data and computation tools; the automatic mode lets the agent choose between them. Keeping the registry consistent across modes prevents a trajectory from being stranded by tools that disappear after a mode switch. Skills are reusable workflow specifications that the model can activate on demand. Each skill declares the tools and specialist context it expects, and routing is based on the skill description rather than a hard-coded name.

Grounding and robustness. The orchestrator can respond to a step with direct tool calls or by loading a workflow skill that brings in the tools needed for a structured analysis. For research-style reports, the harness composes numbered citations from visited sources, can retrofit citations when a draft omits the citation tool, and runs a light grounding-review pass that asks the backbone to soften or attribute claims it cannot tie to read sources or tool outputs. Recovery policies handle transient model failures, malformed tool calls, context-budget pressure, and truncation without changing the evaluation contract.

Operational guardrails. URL pre-fetch validates document IDs before expensive reads. In ablations, disabling this guardrail raises the visit error rate from 2.1% to 39.4%; the end score barely moves because the backbone often self-corrects, but trajectories become substantially more expensive. Search-budget caps show a different pattern: excessive searching correlates with question difficulty rather than causing low scores, so budget caps mainly control cost rather than quality.

4.2 Evaluation and Optimization

The harness is tightly coupled to the environment: tool calls hit the PIT search sandbox’s FAISS server directly, and the rubric judge used for benchmark

scoring is also available as a reward component. This matters because specialized research models are often trained against one tool stack and evaluated against another. In FINANCEHARNESS, swapping the backbone model changes only the learnable model layer. The search API, document-fetch contract, citation requirements, and scoring rubric stay fixed. Appendices H gives worked examples: complete cited reports produced by the harness, and the round-by-round tool trajectories behind them.

5 Experimental Setup

5.1 Benchmark Dataset

The publication dataset comprises 2,078 questions, the data-determined output of the end-to-end LLM quality filter chain (§3.2). For benchmarking, the upstream pipeline selects a 500-question ILP-balanced subset, which then passes through a two-batch external expert annotation process and three-stage LLM curation of §3.2 to produce FINANCEGYM: 400 expert-annotated questions with 2,464 annotated rubric items (mean 6.16 items per question), jointly balanced across 9 topics, 11 sectors, 6 reasoning types, and 12 monthly cutoff-date buckets. Each rubric item points to source evidence packaged with the record, including the cited article text, so the benchmark and any agent built against it can be audited or replayed *fully offline*, with no live-web fetches at evaluation or RL-rollout time. Sector coverage spans healthcare/biotech, technology/semiconductors, consumer discretionary, financial services, real estate, energy/natural resources, industrials/transportation, macro/policy, commodities, crypto/digital assets, and fixed income. The expert-annotation curation pass (§3.2) preserves the ILP balance to within ± 1 percentage point per sector.

5.2 Agent Baselines

Following Tongyi-DR’s framing (Team et al., 2025b), we group baselines by scaffolding complexity rather than by base-model identity. The three groups share our PIT-filtered FAISS retriever and FINANCEGYM as the headline evaluation set: **Trained OSS**: open-source models that internalize the research loop without an external scaffold, trained end-to-end with RL on agentic-search trajectories (Tongyi-DR (Team et al., 2025b), OpenResearcher (Li et al., 2026b), MiroThinker-1.7-mini (Team et al., 2025a, 2026)).

Foundation model with search harness: a fixed

30-step ReAct wrapper with one search tool, one final-answer tool, and 8 interchangeable backbones, which isolates model capability under a constant scaffold. We also report FINANCEHARNESS, our open-weight layered harness on a Qwen3.5-27B backbone, in this group for the open-weight comparison.

Frameworks: engineered scaffolds that make orchestration decisions in code while using an LLM for planning, retrieval, synthesis, or critique. This group includes TTD-DR (Han et al., 2025), GPT-Researcher,⁴ STORM (Shao et al., 2024), OpenClaw,⁵ and deepagents.⁶ TTD-DR, GPT-Researcher, STORM, OpenClaw, and deepagents all use Gemini-3-Flash as the LLM backbone.

Comparing the three groups on identical questions and the same corpus retriever isolates (a) how well training-time tool distributions transfer to a new search environment (Trained OSS vs. Frameworks); (b) how much hand-engineered scaffolding adds over a minimal ReAct loop on the same backbone (Foundation model with search harness vs. Frameworks); and (c) how performance varies with model identity at fixed scaffolding (within the foundation-model rows).

Two implementation details matter for interpreting later findings.

Tongyi-DR. Alibaba’s open-source deep research MoE (Team et al., 2025b) (30.5B total / 3.3B active) is trained with RL on agentic-search trajectories using Serper and Jina as the canonical tool stack. For our benchmark, we replace those tools with corpus-backed equivalents (Search → FAISS+PIT, Visit → SQLite) and serve the model at FP8 through vLLM. The tool substitution drives the distribution-shift analysis in App. B.1.

Deep Researcher with Test-Time Diffusion (TTD-DR). TTD-DR (Han et al., 2025) is an iterative deep-research agent over Gemini-3-Flash: plan, research, draft, critique, revise, and finalize. It runs up to three revision cycles, with a maximum of 50 LLM calls per question and top-5 retrieval per search. This is the main engineered-scaffold comparison point for § 6.1.

5.3 Evaluation Protocol

Agents are run once on the full 500-question ILP-balanced subset; we then aggregate scores against

FINANCEGYM’s 400 expert-annotated questions and their 2,464 annotated rubric items (§3.2) by subsetting the per-criterion judge outputs. All baselines run under identical search infrastructure (PIT-filtered FAISS over the same Qwen3-Embedding-4B index). The foundation-model matrix holds the wrapper fixed across 8 frontier and open-weight backbones. Each agent receives only the question text and cutoff date; the rubric, thesis, and supporting evidence are withheld for blind evaluation. The judge (Gemini Flash) scores each rubric criterion on the 5-tier (0–4) scale described in §3.3, with access to the question, thesis, full rubric, pre-cutoff edge evidence, post-cutoff edge evidence (foresight items only), and the agent’s full report with cited URLs. Headline metrics use the 400 expert-annotated questions and 2,464 annotated rubric items; the broader 500-question run is retained for auxiliary diagnostics. Headline scores are reported with bootstrap 95% confidence intervals ($n=1000$); paired bootstrap CIs test the significance of each pairwise agent delta.

6 Results and Analysis

We evaluate 23 baselines on FINANCEGYM under the protocol above. Scores are averaged with the 5-tier rubric (Eq. 2); the $\pm$ SE column reports bootstrap standard errors over questions. The main text reports the decision-relevant comparisons, while per-axis and harness-ablation breakdowns appear in the appendix.

Cost-quality Frontier. Fig. 2 shows the performance v.s. cost trade-offs among all benchmarked systems. We observe our FINANCEHARNESS stays on the frontier with Opus-5 achieves the best overall performance of 44.9%.

Table 1 reports all evaluated systems grouped by system category; the fixed-backbone FINANCEHARNESS ablation is deferred to §6.2.

As Table 1 shows, on a 27B open-weight backbone, FINANCEHARNESS ranks among the strongest systems on FINANCEGYM: it leads all open-weight models, slightly outperforms the best agentic search system, and trails only two proprietary backbones.

6.1 What Drives Performance

The main comparison is between model choice and research procedure under the same point-in-time corpus. Changing the backbone within the same search-tool wrapper produces a wider spread

⁴<https://github.com/assafelovic/gpt-researcher>

⁵<https://github.com/openclaw/openclaw>

⁶<https://github.com/hwchase17/deepagents>



Figure 2: Pareto-frontier for FINANCEGYM. FINANCEHARNESS reaches an overall frontier, indicating a favorable performance-to-cost trade-off.

Table 1: Headline results on FINANCEGYM by system category. Normalized rubric means (%); within each group, **bold** marks the best and underline the second best per column (Overall/Pre-cutoff/Post-cutoff). $\pm$ SE is the bootstrap SE of the Overall score.

Fine-tuned open-weight

System	Overall $\uparrow$	Pre-cutoff $\uparrow$	Post-cutoff $\uparrow$	$\pm$ SE
Tongyi-DR	28.2	<u>39.7</u>	10.7	0.8
OpenResearcher	<u>27.2</u>	39.9	<u>8.1</u>	0.7
MiroThinker	21.2	29.6	7.8	0.6

Open-weight with search

System	Overall $\uparrow$	Pre-cutoff $\uparrow$	Post-cutoff $\uparrow$	$\pm$ SE
FINANCEHARNESS	32.4	45.7	11.8	0.8
GLM-5	<u>30.4</u>	<u>44.7</u>	<u>8.5</u>	0.9
DeepSeek-v3.2	28.9	42.4	8.4	0.8
Qwen3-235B-A22B	26.8	39.7	7.3	0.7
Gemma-4-26B	25.7	37.6	7.7	0.6
gpt-oss-120b	18.7	27.8	5.2	0.8

Agentic search systems

System	Overall $\uparrow$	Pre-cutoff $\uparrow$	Post-cutoff $\uparrow$	$\pm$ SE
TTD-DR	31.5	45.5	<u>9.8</u>	0.7
GPT-Researcher	<u>30.4</u>	<u>42.2</u>	12.0	0.8
deepagents	28.1	40.7	8.6	1.1
OpenClaw	27.7	40.2	8.3	0.7
STORM	27.4	39.3	9.4	0.6

Proprietary with search

System	Overall $\uparrow$	Pre-cutoff $\uparrow$	Post-cutoff $\uparrow$	$\pm$ SE
Claude-Opus-4.7	34.1	50.1	<u>9.9</u>	0.8
Gemini-3.1-Pro	<u>33.2</u>	46.8	12.8	0.7
GPT-5.5	31.8	<u>47.5</u>	8.0	0.7
Gemini-3-Flash	30.2	43.7	9.8	0.7

than changing the scaffold around Gemini-3-Flash, where the engineered agentic systems cluster near a minimal ReAct loop and several fall below it. This suggests that current financial deep-research performance is still strongly constrained by the underlying model, while task-matched scaffolding can recover additional evidence and organize it more reliably.

The fine-tuned open-weight models show a related transfer issue: trained around live-web tool stacks, they are evaluated here through our corpus-

backed FAISS/PIT interface, where several general-purpose search backbones outperform them, consistent with the tool-distribution-shift analysis in §B.1.

The per-topic breakdown (Table 10) and the per-sector and per-reasoning-type breakdowns (Appendix Tables 4 and 5) follow the aggregate ordering within each system group, with causal and comparative analysis among the hardest reasoning types across systems.

Overall score also tracks backbone scale, but

only loosely. Plotting each system’s score against its backbone size (Fig. 7), FINANCEHARNESS sits on the upper-left efficiency frontier, matching much larger open-weight backbones and approaching the proprietary frontier despite its 27B backbone.

Finally, the pre-cutoff/post-cutoff split is stable across systems: across all baselines, pre-cutoff scores are several times higher than post-cutoff scores. Because the same PIT-filtered corpus is used for every agent, this gap reflects the benchmark’s core difficulty rather than the retrieval setup. Even leading agents leave most forward-looking rubric items only partially addressed or missed.

6.2 In-Environment Training

The same rubric judge can also serve as a training signal. Using the same retrieval tools and report format at rollout and evaluation, we further train FINANCEHARNESS with Group Relative Policy Optimization (GRPO) on a separate set of 172 machine-curated training instances that we build with the same environment and graph-construction harness used to construct the benchmark, with no human involvement and independent of FINANCEGYM, so that training and evaluation remain fully separate. The reward combines rubric-style evaluation against generated criteria (0.6 weight) with an LLM judge over report coherence, grounding, and trajectory quality (0.4 weight). We then evaluate the resulting policy on FINANCEGYM.

Table 2: Fixed-backbone (Qwen3.6-27B) ablation on FINANCEGYM: the model with search only (Vanilla), a naive harness (Naive), the full untrained FINANCEHARNESS (**Harness**, our main configuration), and FINANCEHARNESS after GRPO training on a separate, machine-curated training set independent of FINANCEGYM, with a mixed rubric/report-and-trace reward (RFT). Normalized rubric means (%).

Configuration	Total	Pre-cutoff	Post-cutoff
Vanilla (model with search)	25.3	36.1	8.7
Naive harness	29.6	41.8	10.7
Harness	32.4	45.7	11.8
+ RFT (GRPO training)	32.8	46.2	12.1

Table 2 reports four fixed-backbone configurations: the search-only model (Vanilla), a naive harness (Naive), the full untrained harness (Harness, our main configuration), and the harness after reinforcement fine-tuning with GRPO (RFT). GRPO training adds only 0.4 point over the untrained harness, so we treat it as a refinement rather than a headline result and leave a broader trained-policy

sweep to future work.

7 Conclusion

We presented FINANCEHARNESS, an expert harness for financial deep research, and FINANCEGYM, a point-in-time benchmark built from 400 expert-annotated questions and 2,464 rubric items. The benchmark pairs a reproducible PIT search sandbox with pre-/post-cutoff rubrics, allowing evaluation to separate pre-cutoff-date evidence retrieval from post-cutoff-date outcome anticipation. Across various baselines and FINANCEHARNESS, every system stays below 45% overall, confirming that financial DR is not solved by stronger retrieval alone. The fixed-backbone harness ablation further shows progressive improvement from a Qwen3.5-27B search-tool model (25.3%) to FINANCEHARNESS (29.6%) and then to FINANCEHARNESS with training (32.8%), supporting our core claim: evaluation, tool use, and optimization should share the same temporally controlled research environment.

Limitations

Corpus scope. The point-in-time search sandbox is built from a 2025 English-language web corpus that we collect ourselves, so coverage of non-English sources, paywalled professional data, and structured filings is limited.

Tool-distribution shift. Specialized deep-research models are evaluated out-of-distribution on our corpus-backed retriever rather than the live-web stacks they were trained against; their scores should therefore be read as transfer results under a controlled PIT interface.

Ethical Considerations

Annotator Quality. We contracted professional data vendor at a premium rate of \$152/hour. Annotators with specialized finance backgrounds spent an average of 77.6 minutes per sample auditing the candidate questions for feasibility and clarity. This intensive quality gate rejected 89 lower quality questions.

Copyright. To mitigate copyright and legal risks, the data-sharing protocol will never redistribute any publisher articles. The official release will consist only of research questions. We will share a reconstruction script that allows users to securely re-fetch and extract text from public records. Data is released under CC BY-NC 4.0 License.

Acknowledgments

We acknowledge the use of Gemini and Claude models for assisting with language polishing of paper writing. All generated content was critically reviewed, verified, and edited by the authors, who take full responsibility for the entire manuscript.

References

- Adrien Barbaresi. 2020. `htmldate`: A python package to extract publication dates from web pages. *Journal of Open Source Software*, 5(51):2439.
- Adrien Barbaresi. 2021. `Trafilatura`: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations*, pages 122–131.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, and 1 others. 2025. `Browsecomp-plus`: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*.
- João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and 1 others. 2025. `Deepresearchgym`: A free, transparent, and reproducible evaluation sandbox for deep research. *arXiv preprint arXiv:2505.19253*.
- Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Ijcai*, volume 15, pages 2327–2333.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.
- Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. `Deepresearch bench`: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.
- Google. 2026. Deep research max: a step change for autonomous research agents. <https://blog.google/innovation-and-ai/models-and-research/gemini-models/next-generation-gemini-deep-research/>.
- Rujun Han, Yanfei Chen, Zoey CuiZhu, Lesly Miculicich, Guan Sun, Yuanjun Bi, Weiming Wen, Hui Wan, Chunfeng Wen, Solène Maître, George Lee, Vishy Tirumalashetty, Emily Xue, Zizhao Zhang, Salem Haykal, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. 2025. `Deep researcher with test-time diffusion`. *Preprint*, arXiv:2507.16075.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. `Financebench`: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananeey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759.
- Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. 2026. `Meta-harness`: End-to-end optimization of model harnesses. *In Preprint*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2026a. `Deepresearch bench ii`: Diagnosing deep research agents via rubrics from expert report. *arXiv preprint arXiv:2601.08536*.
- Zhuofeng Li, Dongfu Jiang, Xueguang Ma, Haoxiang Zhang, Ping Nie, Yuyu Zhang, Kai Zou, Jianwen Xie, Yu Zhang, and Wenhui Chen. 2026b. `Open-researcher`: A fully open pipeline for long-horizon deep research trajectory synthesis. *arXiv preprint arXiv:2603.20278*.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. `Gaia`: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pages 9025–9049.
- Xuying Ning, Katherine Tieu, Dongqi Fu, Tianxin Wei, Zihao Li, Yuanchen Bei, and 1 others. 2026. `Code as agent harness`: Toward executable, verifiable, and stateful agent systems. *arXiv preprint arXiv:2605.18747*.
- OpenAI. 2025. Introducing deep research. <https://openai.com/index/introducing-deep-research>.
- Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. 2025. `Deepscholar-bench`: A live benchmark and automated evaluation for generative research synthesis. *arXiv preprint arXiv:2508.20033*.
- Perplexity. 2025. Introducing perplexity deep research. <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>.

- Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.
- Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.
- Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, and 1 others. 2025. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- MiroMind Team, S Bai, L Bing, L Lei, R Li, X Li, X Lin, E Min, L Su, B Wang, and 1 others. 2026. Mirothinker-1.7 & h1: Towards heavy-duty research agents via verification. *arXiv preprint arXiv:2603.15726*.
- MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, and 1 others. 2025a. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.
- Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025b. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.
- Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Al-barghouthi, Caiming Xiong, and Shafiq Joty. 2025. Liveresearchbench: A live benchmark for user-centric deep research in the wild. *arXiv preprint arXiv:2510.14240*.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kam-badur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.
- Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Fangda Ye, Yuxin Hu, Pengxiang Zhu, Yibo Li, Ziqi Jin, Yao Xiao, Yibo Wang, Lei Wang, Zhen Zhang, Lu Wang, and 1 others. 2026. Miroeval: Benchmarking multimodal deep research agents in process and outcome. *arXiv preprint arXiv:2603.28407*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A FINANCEGYM Details

This section summarizes the released FINANCEGYM set and positions it against prior benchmarks.

A.1 Dataset Portrait

Figures 3–5 summarize the released FINANCEGYM set along three axes: the expert-annotation quality that gates curation, the temporal spread of question cutoff dates, and the balance across topics, sectors, and reasoning types.

A.2 Benchmark Positioning

Table 3 highlights the gap FINANCEGYM targets. FinanceBench (Islam et al., 2023) is the closest finance-domain peer, but evaluates short-form QA over filings rather than analyst-style research. BrowseComp-Plus (Chen et al., 2025) and DeepResearchGym (Coelho et al., 2025) improve reproducibility through fixed corpora, but do not apply per-question publication-date cutoffs. DeepScholar-Bench (Patel et al., 2025) is closest on the PIT axis, but its live arXiv dependency leaves reproducibility tied to external state. DeepResearch Bench II (Li et al., 2026a) and ResearchRubrics (Sharma et al., 2025) provide rich rubrics, but do not isolate pre-cutoff-date evidence retrieval from post-cutoff-date outcome anticipation, which is the central difficulty in financial deep research.

B Additional Analysis

B.1 Tool-Distribution Shift

The trained OSS deep-research models underperform the strongest frameworks on FINANCEGYM despite being trained for agentic search. We interpret this as a tool-distribution-shift result rather than a general model-capability result. Tongyi-DR and MiroThinker were optimized around live-web search and page extraction conventions; in our evaluation, they must operate through a PIT-filtered corpus retriever with fixed document records and no live-web access. That substitution changes both the interaction pattern and the final-report format expected by the rubric.

The most visible consequence is attribution. FINANCEGYM’s score-4 anchor rewards claims that are correct, specific, and grounded in plausible source evidence. Frameworks such as TTD-DR and GPT-Researcher are explicitly prompted to produce source-attributed reports, while trained OSS systems often emit cleaner answer-style prose with

weaker inline attribution. This does not mean the trained models lack the underlying knowledge or search ability; it means their learned reporting interface is mismatched with a long-form, citation-sensitive financial rubric.

B.2 Hindsight and Foresight Remain Different Failure Modes

The stable gap between hindsight and foresight scores is the most important behavioral pattern in the benchmark. Systems improve much more readily on hindsight criteria, where better search, reading, and citation discipline help recover pre-cutoff-date evidence. Foresight criteria remain compressed even for stronger systems, suggesting that financial deep research is not solved by retrieval quality alone: agents must turn historical evidence into forward-looking judgments that are only verifiable after the cutoff date. This is why the fixed-backbone FINANCEHARNESS ablation is useful: the harness raises hindsight substantially, while training inside the harness is the first step that also moves foresight meaningfully.

B.3 Framework Trade-offs

Framework choice changes more than the final score. Iterative research frameworks such as TTD-DR achieve the strongest overall quality, but they are more expensive because they repeatedly plan, retrieve, draft, critique, and revise. GPT-Researcher is slightly weaker on aggregate quality but offers a better quality–cost trade-off. Other frameworks such as STORM and deepagents are useful counterpoints: they bring different composition strategies, but their default output styles are not always aligned with a rubric that rewards specific, source-grounded financial claims. We therefore treat framework comparisons as evidence about orchestration fit-to-task, not as a universal ranking of agent frameworks.

Tables 4 and 5 give the per-sector and per-reasoning-type scores for representative systems from each group (the top two Trained OSS, top two proprietary and top two open-weight foundation-model search systems, and top three Frameworks).

C FINANCEHARNESS Ablation Breakdowns

We expand Table 2 by topic, reasoning type, sector, and situation type. All rows use the same Qwen3.5-27B backbone and the same FINANCEGYM evaluation protocol.

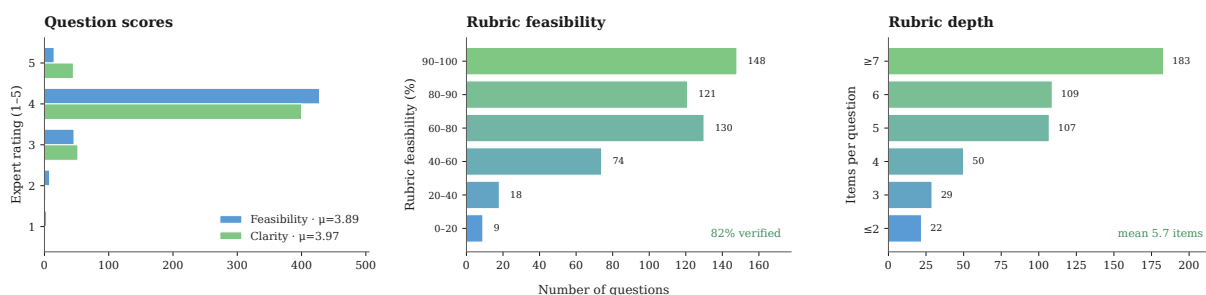


Figure 3: Expert-annotation quality over the annotated question pool: distribution of annotator feasibility and clarity ratings used as the quality-control gate that curates FINANCEGYM.

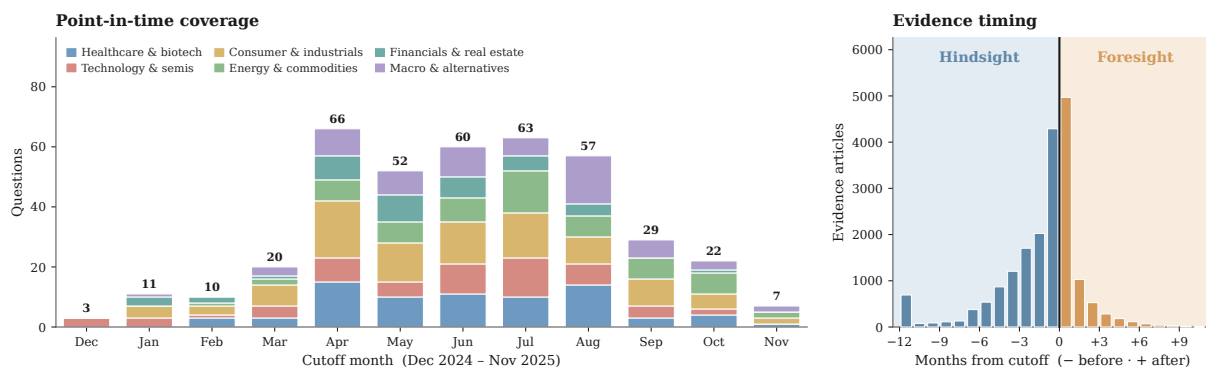


Figure 4: Temporal coverage of FINANCEGYM: distribution of question cutoff dates across the 12 monthly cutoff-date buckets.

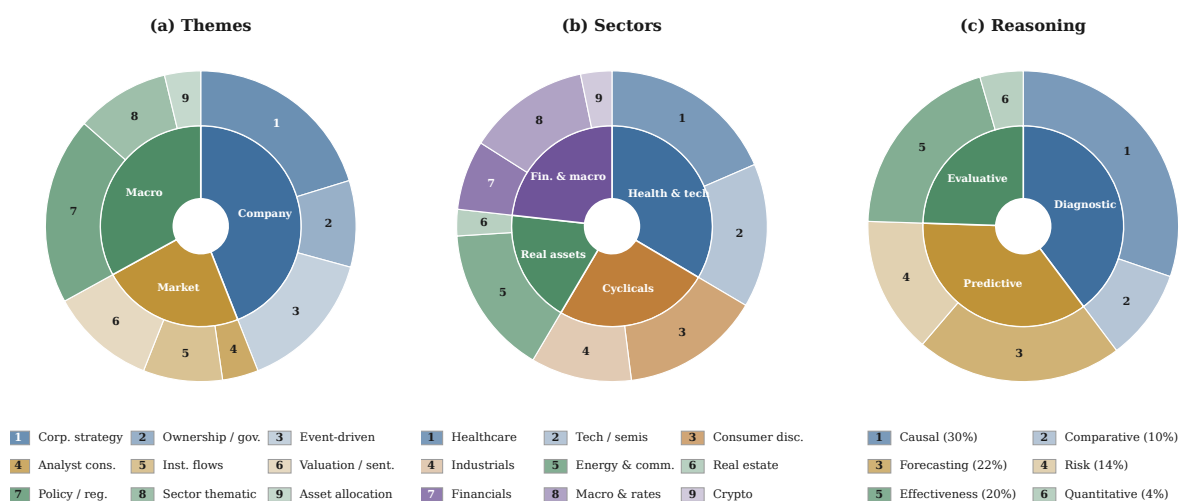


Figure 5: Distribution of the 400-question FINANCEGYM subset across the three bottom-up taxonomy axes. Topics and sectors are shown as two-ring sunbursts with super-groups; reasoning types appear as a donut.

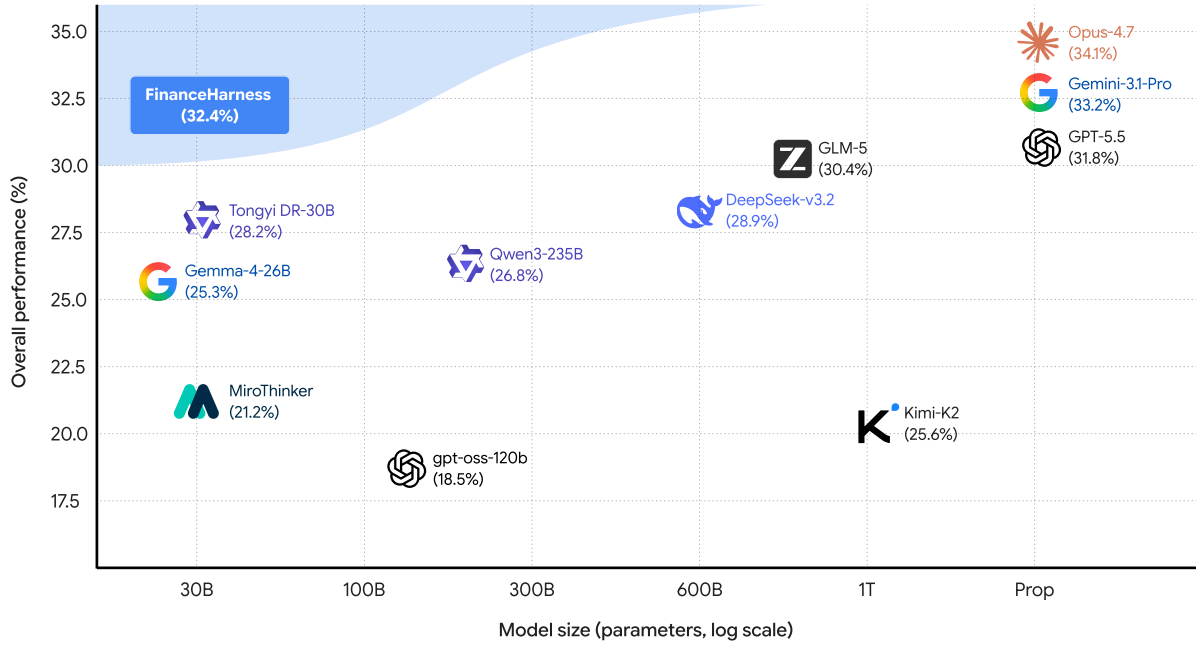


Figure 6: Overall outcome score versus backbone scale on FINANCEGYM. Each marker is one evaluated system, placed by backbone parameter count (x-axis, log scale; proprietary models grouped at right) and overall rubric score (y-axis). On a 27B open-weight backbone, FINANCEHARNESS reaches an overall score competitive with much larger open-weight models and approaching the proprietary frontier (shaded), indicating a favorable performance-to-scale trade-off.

Table 3: Comparison with representative deep-research and search-agent benchmarks. ✓/◇/✗ denote full, partial, and no support. *Reproducible* means evaluation uses a fixed corpus rather than live-web access. *PIT* means retrieval is constrained by a per-question publication-date cutoff. *Verifiable* means evaluation rests on externally checkable signals such as gold answers, binary rubric items, citation matches, or post-cutoff-date outcomes.

Benchmark	Domain	Long-form	Reproducible	PIT	Rubric	Verifiable
FinanceBench (Islam et al., 2023)	Finance	✗	✓	✗	Gold answer	✓
GAIA (Mialon et al., 2024)	General	✗	◇	✗	Gold answer	✓
BrowseComp (Wei et al., 2025)	General	✗	✗	✗	Gold answer	✓
BrowseComp-Plus (Chen et al., 2025)	General	✗	✓	✗	Gold answer + citation	✓
FRAMES (Krishna et al., 2025)	General	✗	✓	✗	Gold answer	✓
DeepResearchGym (Coelho et al., 2025)	General	✓	✓	✗	Citation + LLM-judge	◇
DeepResearch Bench (Du et al., 2025)	Multi-domain	✓	✗	✗	RACE + FACT	◇
DeepResearch Bench II (Li et al., 2026a)	Multi-domain	✓	✗	✗	9.4k binary rubrics	✓
ResearchRubrics (Sharma et al., 2025)	Multi-domain	✓	✗	✗	2.6k expert rubrics	✓
LiveResearchBench (Wang et al., 2025)	Multi-domain	✓	✗	✗	DeepEval checklist	✓
MiroEval (Ye et al., 2026)	Multi-domain	✓	✗	✗	Adaptive + process eval	✓
DeepScholar-Bench (Patel et al., 2025)	Academic	✓	✗	✓	Automated, 3 dimensions	✓
FINANCEGYM (Ours)	Finance	✓	✓	✓	Hindsight/foresight rubric	✓

Table 4: Per-sector averaged scores (%) on FINANCEGYM for the top two Trained OSS, top two proprietary and top two open-weight foundation-model search systems, and top three Frameworks. Bold = column-set winner per row.

Sector	Trained		F+S prop.		F+S open		Framework		
	Tongyi-DR	Miro	Cl-opus	Gem-Pro	GLM-5	DSeek	TTD-DR	GPTR	dpa
healthcare_biotech	33.5	21.9	38.7	36.9	36.1	35.2	43.7	43.1	28.6
technology_semiconductors	29.5	22.1	33.5	35.5	29.6	27.2	36.3	35.3	32.3
consumer_discretionary	28.0	25.0	35.5	33.3	33.3	30.7	37.2	35.8	28.8
energy_natural_resources	26.5	18.9	34.5	31.4	32.0	28.6	36.7	35.4	30.4
macro_policy	26.4	18.6	34.2	31.3	24.7	28.0	28.7	28.1	24.9
industrials_transportation	29.4	20.2	33.9	32.6	29.1	27.5	35.3	32.9	26.0
financial_services	25.6	24.1	32.7	34.0	30.0	29.4	39.6	33.9	31.3
commodities	21.0	17.5	24.7	31.4	25.5	22.8	26.1	24.5	22.3
crypto_digital_assets	24.7	18.1	22.6	23.8	18.2	19.2	23.2	20.3	19.2
real_estate	25.0	19.9	32.4	27.7	26.2	21.9	34.7	32.3	21.6
fixed_income	16.6	15.3	25.7	28.6	22.6	18.7	38.6	35.1	25.3

Table 5: Per-reasoning-type averaged scores (%) on FINANCEGYM. Same system columns as Table 4. Bold = column-set winner per row.

Reasoning type	Trained		F+S prop.		F+S open		Framework		
	Tongyi-DR	Miro	Cl-opus	Gem-Pro	GLM-5	DSeek	TTD-DR	GPTR	dpa
risk_assessment	34.9	21.4	38.1	38.0	35.4	31.9	40.5	40.7	32.0
effectiveness_evaluation	29.9	23.3	36.2	35.2	32.8	29.9	39.1	36.3	29.5
quantitative_analysis	27.2	16.7	31.3	29.8	30.5	30.6	38.1	37.2	26.8
predictive_forecasting	28.0	20.5	36.1	35.2	32.5	30.4	37.2	34.6	31.1
comparative_analysis	27.2	21.1	33.0	30.2	24.0	26.5	35.0	32.1	23.8
causal_analysis	24.6	21.0	30.2	29.7	26.8	26.3	32.1	31.4	24.9

Two patterns hold across every slice. Scores rise monotonically from Search to FINANCEHARNESS to FINANCEHARNESS+train in all four tables, so the harness improves every topic, reasoning type, sector, and situation type, and training improves them again with no regressions. The two steps contribute comparably, each adding three to four points on average, and the gains are broad rather than concentrated on a few easy slices. The hardest cells stay hard: commodities, crypto, and fixed income remain the weakest sectors, causal analysis the weakest reasoning type, and the sparse workforce-restructuring situation rises from 0% only to 20%.

D FINANCEHARNESS Architecture Details

Model and serving layer. The serving layer hosts the orchestrator backbone and a lightweight reader used for long-document evidence extraction. It is deliberately opaque to the rest of the harness: the runtime sees only a request/response contract, so the backbone can be swapped across local serving stacks or closed-model SDKs without changing the orchestration logic.

Runtime layer. The runtime is the control plane. It owns the agent loop, parsing and dispatch, schema validation, tier loading, and the registries for tools, skills, and subagents. It contains no model intelligence; instead, it enforces cross-layer invariants such as schema conformance and the depth limit on subagent delegation.

Tool surface. The tool surface is tiered to keep the initial prompt small while preserving a large finance capability surface. The always-loaded tier contains the core actions needed for PIT search, source reading, citation composition, loading, and delegation. Deferred tools are exposed through a compact catalog, with full schemas loaded only after the model commits to a tool family. The extension tier supports MCP or CLI-backed connectors, including paid data vendors and internal systems.

Skills and subagents. Skills are reusable workflow specifications that the model can activate on demand. Each skill declares the tools and specialist context it expects, and routing is based on the skill description rather than a hard-coded name. A subagent is a fresh agent loop with a curated tool subset, a filtered skill list, and its own context window. The runtime enforces depth-1 delegation, so a subagent cannot recursively spawn another

Table 6: FINANCEHARNESS ablation by topic. Scores are normalized rubric means (%).

Topic	Search	FINANCEHARNESS	FINANCEHARNESS + train
Corporate strategy competitive positioning	25.9	29.5	32.0
Macro regulatory policy	24.9	30.5	33.5
Event driven restructuring	33.0	36.8	39.3
Valuation sentiment divergence	20.9	25.9	28.4
Sector thematic analysis	23.3	27.3	29.7
Ownership governance stakeholder intelligence	27.1	32.4	34.6
Institutional flows market microstructure	18.4	21.6	25.2
Analyst research consensus dynamics	26.3	29.3	32.7
Asset allocation portfolio strategy	22.2	25.3	29.8

Table 7: FINANCEHARNESS ablation by reasoning type. Scores are normalized rubric means (%).

Reasoning type	Search	FINANCEHARNESS	FINANCEHARNESS + train
Causal analysis	22.4	27.2	29.6
Predictive forecasting	24.5	29.0	32.4
Effectiveness evaluation	27.0	31.0	33.4
Risk assessment	32.6	36.5	39.3
Comparative analysis	22.3	26.2	29.4
Quantitative analysis	24.5	27.6	31.6

Table 8: FINANCEHARNESS ablation by sector. Scores are normalized rubric means (%).

Sector	Search	FINANCEHARNESS	FINANCEHARNESS + train
Healthcare biotech	33.7	37.3	39.5
Technology semiconductors	25.9	29.6	31.9
Consumer discretionary	23.8	28.2	31.2
Energy natural resources	26.1	30.4	33.3
Macro policy	16.6	23.1	27.1
Industrials transportation	24.9	29.7	32.3
Financial services	26.6	31.4	33.7
Commodities	16.1	18.7	21.8
Crypto digital assets	17.9	20.7	24.7
Real estate	26.4	29.1	32.9
Fixed income	17.9	23.6	25.7

Table 9: FINANCEHARNESS ablation by situation type. Scores are normalized rubric means (%).

Situation type	Search	FINANCEHARNESS	FINANCEHARNESS + train
Multihop path	26.4	30.9	33.8
Temporal narrative	24.3	28.6	30.9
Tension analyst disagreement	24.2	28.3	30.6
Tension price target divergence	20.5	24.4	28.0
Tension performance divergence	19.7	22.8	26.4
Tension strategic portfolio shift	32.1	36.1	38.5
Tension ownership churn	31.1	35.2	38.3
Tension workforce restructuring	N/A	15.0	20.0

subagent.

Escalation modes. The orchestrator can respond to a step in three ways, ordered by context cost: an atomic tool call for one-shot lookup, a workflow that activates a reusable skill, or delegation to a context-isolated subagent. The baseline evaluation in §6 uses the common PIT search and fetch contract across systems; the deeper skill and subagent modes are used by FINANCEHARNESS when training or running the full harness.

Operational guardrails. URL pre-fetch validates document IDs before expensive reads. In ablations, disabling this guardrail raises the visit error rate from 2.1% to 39.4%; the end score barely moves because the backbone often self-corrects, but trajectories become substantially more expensive. Search-budget caps show a different pattern: search-storming correlates with question difficulty rather than causing low scores, so budget caps mainly control cost rather than quality.

E Overall Outcome Scores.

F Performance by Topics

G Annotation Instruction

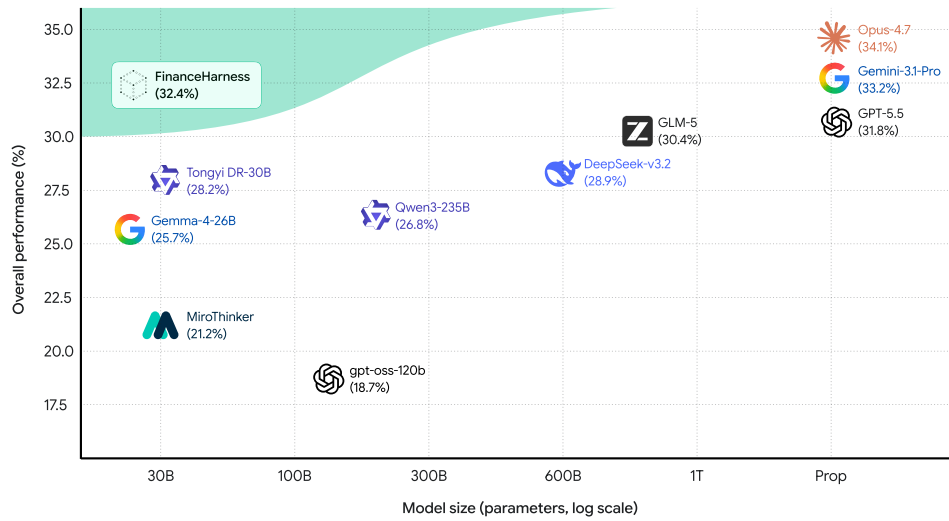


Figure 7: Overall outcome score versus backbone scale on FINANCEGYM. Each marker is one evaluated system, placed by backbone parameter count (x-axis, log scale; proprietary models grouped at right) and overall rubric score (y-axis). On a 27B open-weight backbone, FINANCEHARNESS reaches an overall score competitive with much larger open-weight models and approaching the proprietary frontier (shaded), indicating a favorable performance-to-scale trade-off.

← prev

next →

Q 8 / 10

NOT STARTED

Show list

How is PayPal's strategy to reach its \$2 billion revenue target for Venmo being influenced by merchant adoption, such as the JetBlue partnership, and the expansion into high-margin services like crypto transfers and expense tracking features?

Thesis. To evaluate Venmo's trajectory from a P2P tool to a high-margin merchant and digital asset platform as a driver for PayPal's valuation.

CUTOFF 2025-04-07	SECTOR financial_services	TOPIC corporate_strategy_competitive_positioning
REASONING causal_analysis	SITUATION temporal_narrative	
ENTITIES Venmo, PayPal, JetBlue, MoonPay		

QUESTION-LEVEL

QUESTION FEASIBILITY Answerable from public info before the cutoff?
1 2 3 4 5

QUESTION CLARITY Is the question clearly written and well-formed?
1 2 3 4 5

Keep in benchmark?
Yes Unsure No

Notes (optional)
Anything the scales above don't capture – e.g., "outcome item 7 looks fabricated;" "fact item 3 references info from after the cutoff;" or "question is more about politics than finance."

Missing required fields: question_feasibility, question_clarity, keep, rubric_feasibility[0], rubric_feasibility[1], rubric_feasibility[2], rubric_feasibility[3], rubric_feasibility[4], rubric_feasibility[5], rubric_feasibility[6], rubric_feasibility[7]

Skip Save & next →

RUBRIC (8 ITEMS)

FACT 0 before the cutoff date

PayPal set a 2027 revenue target of \$2 billion for Venmo in February 2025.

FEASIBILITY Relevant to the question and supported by the sources?
Feasible Not feasible
• supporting evidence (1 of 32 candidate sources, pre-cutoff)

FACT 1 before the cutoff date

JetBlue became the first airline to accept Venmo for online flight bookings in January 2025.

FEASIBILITY Relevant to the question and supported by the sources?
Feasible Not feasible
• supporting evidence (4 of 32 candidate sources, pre-cutoff)

FACT 2 before the cutoff date

Figure 8: Annotation Instruction and Interface.

Table 10: Per-topic averaged scores (%) on FINANCEGYM, by system group. Columns follow the Company/-Market/Macro topic levels of Figure 5 (Corp. strat., Own./gov., Event | Analyst, Inst. flows, Valu./sent. | Policy, Thematic, Alloc.). Each panel is one system group; shading (blue = higher, amber = lower), **bold** (best) and underline (second) are normalized per column *within the panel*.

(a) Fine-tuned open-weight									
System	Corp. strat.	Own./gov.	Event	Analyst	Inst. flows	Valu./sent.	Policy	Thematic	Alloc.
Tongyi-DR	27.6	32.5	<u>33.6</u>	27.5	20.9	<u>23.8</u>	29.2	26.0	31.1
OpenResearcher	25.9	27.8	35.1	27.3	19.0	24.0	28.8	25.6	25.5
MiroThinker	22.1	21.3	24.1	17.4	17.4	21.6	19.8	20.2	<u>25.9</u>
(b) Agentic search systems									
System	Corp. strat.	Own./gov.	Event	Analyst	Inst. flows	Valu./sent.	Policy	Thematic	Alloc.
TTD-DR	31.8	<u>34.5</u>	37.2	28.3	33.2	28.4	29.6	<u>28.1</u>	26.8
GPT-Researcher	<u>29.4</u>	35.2	36.3	30.8	28.6	<u>26.9</u>	28.9	28.0	<u>28.0</u>
deepagents	25.9	27.4	36.4	23.3	29.7	20.4	30.9	26.9	22.8
OpenClaw	26.8	29.0	33.8	25.2	22.4	24.8	28.5	25.8	28.5
STORM	25.7	26.2	34.3	<u>30.5</u>	22.6	23.8	27.6	28.8	26.4
(c) Proprietary with search									
System	Corp. strat.	Own./gov.	Event	Analyst	Inst. flows	Valu./sent.	Policy	Thematic	Alloc.
Claude-Opus-4.7	33.9	34.6	40.2	30.4	<u>22.7</u>	<u>28.4</u>	38.7	32.9	35.3
Gemini-3.1-Pro	<u>33.4</u>	<u>34.5</u>	38.6	30.8	25.3	28.5	35.4	<u>32.3</u>	32.2
GPT-5.5	32.3	33.3	36.5	30.0	22.5	25.5	<u>35.9</u>	30.5	29.7
Gemini-3-Flash	31.8	30.8	36.0	28.3	22.5	26.1	30.5	28.0	32.0
(d) Open-weight with search									
System	Corp. strat.	Own./gov.	Event	Analyst	Inst. flows	Valu./sent.	Policy	Thematic	Alloc.
FinanceHarness	31.7	34.7	39.1	33.4	25.9	28.6	33.6	29.6	29.6
GLM-5	<u>30.2</u>	<u>33.5</u>	<u>38.6</u>	28.7	19.9	23.9	<u>32.7</u>	28.8	26.7
DeepSeek-v3.2	27.2	30.9	35.6	<u>29.7</u>	19.5	<u>26.3</u>	32.5	26.0	24.4
Qwen3-235B-A22B	25.6	27.2	34.4	27.3	18.1	23.7	28.3	25.3	25.5
Gemma-4-26B	23.8	26.6	31.2	24.4	<u>20.2</u>	22.9	27.2	24.4	<u>29.0</u>
gpt-oss-120b	18.7	19.6	21.0	24.9	15.2	15.6	19.2	18.0	18.6

H FINANCEHARNESS Case-Study Reports

The boxed reports below are produced by FINANCEHARNESS, drawn from the cross-backbone coverage evaluation. Each example box gives the research question, the resulting cited report, and its sources.

H.1 Industry Deep Research

AI Infrastructure Value Chain: Beneficiaries Beyond the GPU

I. RESEARCH QUESTION

AI compute demand is driving a historic buildout of GPUs and data-center infrastructure. As a research analyst, work outward from the AI accelerators to map the broader value chain: identify the successive layers of industries and suppliers whose demand is pulled by this buildout. For each layer you surface, name the leading public companies, explain how their business connects to AI-compute growth, and assess their financial profile and positioning (growth, margins, valuation). Conclude with a synthesized view of where along the chain the exposure looks most attractive and durable, and the key risks.

II. CITED REPORT

AI Infrastructure Value Chain: Scarcity Rents Extend Well Beyond the GPU

Synthesis

The AI buildout is pulling demand through two interconnected chains:

- 1. The silicon chain: accelerators → HBM → advanced packaging/foundries → semiconductor equipment.
- 2. The physical-infrastructure chain: servers → networking/optics → electrical and cooling equipment → construction/grid interconnection → power generation → data-center real estate.

The most durable economics appear where technology, qualification requirements or physical constraints limit substitution: NVIDIA and Broadcom in compute silicon, TSMC in advanced manufacturing and packaging, ASML in lithography, Arista in AI networking, and Eaton/Vertiv in power delivery and cooling. The better valuation asymmetry currently appears in TSMC and selected power producers, while memory, optical components and server assemblers offer stronger near-term growth but more cyclical or commoditized economics. At the other extreme, data-center REITs benefit directly from capacity scarcity but combine high capital intensity, approximately 5–6x net leverage and elevated valuation multiples. They are comparatively weak expressions of the theme unless power-secured sites can earn exceptional returns.

Demand anchor

Based on reported company plans cited by CNBC, Alphabet, Amazon, Microsoft and Meta together indicate roughly \$734 billion of 2026 capital expenditure and finance leases. The definitions differ and not all of the spending is AI-related, but the scale establishes the demand envelope. The main question has shifted from whether infrastructure spending is rising to whether the associated revenue can generate adequate returns. Alphabet’s cash flow turned negative in Q2, while debt increased materially at both Alphabet and Amazon [9].

Alphabet illustrates the composition. Its Q2 capital expenditure was \$44.9 billion, approximately 60% for servers and 40% for data centers and networking. Google Cloud revenue grew 82% and backlog reached \$514 billion, suggesting that unusually high spending is still accompanied by capacity-constrained demand [8].

The physical consequence is equally significant. The IEA expects global data-center electricity consumption to more than double from 415 TWh in 2024 to approximately 945 TWh in 2030. Around 20% of planned projects could experience delays because transmission, transformers, cables and generating capacity are not being added fast enough [6].

Financial snapshot

The table uses the latest available market-data snapshot as of July 28, 2026. Growth is recent year-over-year revenue growth, margins are operating margins, and valuation is forward P/E unless otherwise indicated. These are screening metrics rather than fully normalized estimates; memory, utilities and hardware assemblers are especially cycle-sensitive.

Layer and selected companies	Revenue growth	Operating margin	Valuation
Accelerators: NVIDIA / AMD / Broadcom	85% / 38% / 48%	66% / 14% / 49%	15x / 33x / 19x
Memory: Micron	346%	80%	5x
Foundry: TSMC	36%	60%	18x
Equipment: ASML / Lam / KLA	21% / 24% / 12%	37% / 35% / 41%	27x / 32x / 37x
Servers: Dell / Super Micro / HPE	19% FY / 123% / 40%	12% EBITDA / 6% / 9%	17x / 8x / 11x
Networking: Arista / Marvell / Cisco	35% / 28% / 12%	43% / 15% / 25%	37x / 27x / 24x
Optics: Coherent / Lumentum	21% / 90%	14% / 22%	28x / 34x
Power and cooling: Vertiv / Eaton	30% / 17%	16% / 16%	29x / 24x
Construction/grid: Quanta Services / EMCOR	26% / 20%	4% / 9%	35x / 21x
Generation: Constellation / Vistra / GE Vernova	64% / 43% / 22%	22% / 27% / 8%	19x / 14x / 37x
Data-center REITs: Equinix / Digital Realty	12% / 30%	24% / 26%	29x / 29x EY/EBITDA
Cloud buyers: Microsoft / Alphabet / Meta	18% / 24% / 33%	46% / 34% / 41%	20x / 22x / 16x

The successive layers

1. Accelerators and custom AI silicon

Selected public companies: NVIDIA, AMD, Broadcom, Marvell.

This is the starting point and the most profitable layer in the current financial snapshot. NVIDIA supplies accelerators, networking products, systems and an extensive software platform. AMD is the principal listed merchant alternative in the screen, while Broadcom and Marvell provide exposure to custom silicon and networking.

NVIDIA’s Q1 fiscal 2027 Data Center revenue reached \$75.2 billion—approximately 92% of total revenue—and grew 92% year over year. Its 74.9% GAAP gross margin and roughly 66% operating margin demonstrate the current scarcity rents associated with its data-center platform [1].

Broadcom provides a counterweight to NVIDIA’s merchant model. Its AI revenue reached \$10.8 billion in fiscal Q2, more than doubling year over year, driven by custom accelerators, including Google’s TPU, and networking components [11]. Broadcom is therefore exposed both to the expansion of AI

clusters and to hyperscalers' efforts to deploy internally designed accelerators.

Positioning: This layer has the strongest current combination of growth, margins and technical barriers. Its weakness is customer concentration: a small number of hyperscalers control much of incremental spending and are simultaneously developing competing silicon.

2. High-bandwidth memory and data-center memory

Selected public companies: SK Hynix, Micron, Samsung Electronics.

An accelerator requires HBM, server DRAM and increasingly enterprise SSD capacity. HBM is especially demanding because multiple memory dies must be stacked, tested and integrated alongside the processor through advanced packaging.

SK Hynix's 2026 market-outlook page describes direct exposure through HBM3E, HBM4, server DRAM and enterprise SSDs. Citing Bank of America estimates, it places the 2026 HBM market at \$54.6 billion, up 58%, while also warning that capacity additions and competition could create price pressure after 2026 [18].

Micron's fiscal Q3 revenue rose from \$9.3 billion to \$41.5 billion year over year. Gross margin reached 84.6% and operating margin 80.4%, with management introducing multiyear customer agreements intended to make results more predictable [10]. Those margins should not be treated as normal through-cycle economics: Micron itself is investing at record levels in technology, products and supply, while the broader memory outlook acknowledges the possibility of future price correction [10,18].

Positioning: HBM is currently a bottleneck and may remain more differentiated than conventional memory. Nevertheless, the very low forward multiples reflect valid concerns about peak pricing, heavy capital expenditure and eventual supply response.

3. Foundry manufacturing and advanced packaging

Leading public exposure: TSMC.

AI accelerators require leading-edge logic and advanced packaging to connect compute dies with HBM. TSMC's role in both advanced process nodes and packaging makes it more than a conventional wafer foundry.

TSMC reported Q2 revenue of \$40.2 billion, a 67.7% gross margin and a 60.3% operating margin [2]. Management characterized AI demand as "extremely robust," raised expected 2026 revenue growth to slightly above 40%, and increased planned capital expenditure to \$60–64 billion. Approximately 70–80% is allocated to advanced processes and 10–20% to packaging, testing and related capacity [15].

The company screens at approximately 18x forward earnings despite growth and margins comparable with the highest-quality semiconductor companies in the market-data set. The valuation also reflects material risks: overseas expansion is expected to dilute margins, the N2 ramp will carry startup costs, and a large share of current production remains geographically concentrated.

Positioning: TSMC is arguably the best risk-adjusted tollbooth in the chain. Its foundry

model provides exposure to demand from cloud providers, accelerator customers and CPU suppliers without relying on a single chip architecture. Geographic concentration is the dominant risk.

4. Wafer-fabrication and process-control equipment

Selected public companies: ASML, Applied Materials, Lam Research, KLA.

More AI silicon eventually requires additional lithography, deposition, etch, inspection and metrology capacity. Equipment suppliers participate with a lag: end demand first raises utilization at foundries and memory manufacturers, which then place additional tool orders.

ASML reported €32.7 billion of 2025 sales, a 52.8% gross margin and a €38.8 billion year-end backlog. Q4 bookings included €7.4 billion of EUV orders. Management said stronger customer confidence in sustained AI demand had materially increased medium-term capacity plans and order intake [3].

The market-data screen shows operating margins of 35% at Lam and 41% at KLA, versus 37% at ASML. Their valuations of roughly 27–37x forward earnings reflect strong current returns and expectations that AI-related semiconductor investment will remain elevated.

Positioning: This is one of the more durable layers because advanced tools are technically complex and serve long-lived customer manufacturing programs. The trade-off is semiconductor-capital-equipment cyclicality and premium valuations.

5. Boards, servers, racks and systems integration

Selected public companies: Dell, Super Micro Computer, HPE.

These companies turn accelerators, CPUs, memory, network cards, power supplies and cooling components into deployable systems. AI servers generate exceptional revenue growth because accelerator content per server is extremely high, but much of that revenue is pass-through component value.

Dell shipped more than \$25 billion of AI-optimized servers in FY26, booked more than \$64 billion of orders and entered FY27 with a \$43 billion backlog. Its Infrastructure Solutions Group revenue grew 40%, but operating margin declined from 12.8% to 11.7%, demonstrating that volume does not automatically produce scarcity-level profitability [12].

Super Micro's market-data profile shows an 8% gross margin, a 6% operating margin and approximately 5x net debt/EBITDA. Dell's broader reported portfolio, larger cash generation and lower net leverage in the comparative screen provide a more balanced profile, although its margins remain far below those of semiconductor and networking suppliers.

Positioning: Strong demand exposure but weak structural economics. Buyers have alternatives, component suppliers capture much of the value, and working-capital requirements rise sharply during rapid growth. This is primarily a volume trade rather than a durable profit pool.

6. Ethernet switching, interconnect and optical components

Selected public companies: Arista Networks, Broadcom, Marvell, Cisco, Coherent, Lumentum.

Large clusters require both:

- **Scale-up networking** between accelerators within a rack or tightly coupled system.
- **Scale-out networking** connecting many racks into a larger training or inference fabric.

Arista's announced 1.6-terabit portfolio is designed for rack-scale AI infrastructure and clusters expanding from thousands to hundreds of thousands of accelerators, including air-, liquid- and hybrid-cooled systems. The products use Broadcom Tomahawk 6 switching silicon, illustrating how several companies can participate in the same network deployment [17]. Arista's approximately 64% gross margin, 43% operating margin and net-cash balance sheet distinguish it from lower-margin component suppliers.

NVIDIA's own networking revenue grew 199% to \$14.8 billion, substantially faster than its 77% growth in Data Center compute revenue during the same quarter [1]. Coherent supplies optical products used in data-center communications; its fiscal Q3 revenue grew 21%, gross margin reached 37.7%, and management said it was expanding capacity in response to AI-data-center demand [13].

Positioning: Arista combines attractive economics with a differentiated network operating stack, but trades near 37x forward earnings. Optical suppliers offer higher operating leverage but face rapid product transitions, customer concentration and changing interconnect architectures.

7. Electrical equipment and thermal management

Leading public companies: Vertiv, Eaton.

GPU density turns power conversion and heat removal into binding constraints. The equipment set includes switchgear, uninterruptible power supplies, power distribution and cooling systems.

Vertiv is the higher-purity listed exposure in this comparison. Q1 sales grew 30%, with Americas organic growth of 44% on data-center demand. Adjusted operating margin rose 430 basis points to 20.8%, while full-year organic growth guidance increased to 29–31%. Net leverage was only about 0.2x [4].

Eaton supplies power and energy management, backup power, software, safety, design and project-management services to data-center operators. The company explicitly identifies dense GPU clusters as creating new electrical, thermal and operational stress [16]. Its 24x forward multiple is less demanding than Vertiv's 29x, though its overall business is less concentrated on data centers.

Positioning: This is one of the most attractive second-order layers. Reliability requirements, manufacturing lead times and the need to redesign facilities for higher rack density support demand. Unlike a bet on one accelerator architecture, much of this spending remains necessary across competing compute platforms. The principal risk is that high valuations already discount several years of exceptional growth.

8. Data-center construction and grid interconnection

Selected public companies: Quanta Services, EMCOR, Comfort Systems USA.

These businesses provide exposure to transmission, substations, electrical work, mechanical systems and other large-load infrastructure. They benefit from the overlap among data-center construction, utility load growth and generation investment.

Quanta reported Q1 revenue of \$7.87 billion, up 26%, and record total backlog of \$48.5 billion. Management cited converging utility, generation and large-load markets as the demand driver and identified data centers among potential future project areas [5].

The market-data snapshot shows higher operating margins at EMCOR and Comfort Systems than at Quanta: approximately 9% and 8%, respectively, versus 4%. Quanta's roughly 35x forward multiple is high for its current operating margin, while EMCOR screens at approximately 21x.

Positioning: Backlogs provide visibility, but this remains a labor- and execution-intensive layer. Fixed-price contracts, skilled-labor shortages, permitting and project delays can erode the apparent backlog advantage.

9. Electricity generation and generation equipment

Selected public companies: Constellation Energy, Vistra, NextEra Energy, GE Vernova, Caterpillar, Cummins.

Data centers need continuous, high-quality power. The IEA expects electricity generation serving data centers to exceed 1,000 TWh by 2030. Natural gas and coal are expected to meet more than 40% of incremental demand through 2030, despite substantial renewable additions; nuclear becomes more important after 2030 [14].

Public-market exposure can therefore be sought through three broad groups:

- **Electricity producers:** Constellation, Vistra and NextEra.
- **Generation and grid equipment:** GE Vernova.
- **Backup and distributed-generation equipment:** Caterpillar and Cummins.

Constellation and Vistra screen at approximately 19x and 14x forward earnings, respectively—below many equipment beneficiaries. Their financial exposure, however, depends on regional power markets, contract structures, operating performance and regulation. GE Vernova offers more direct equipment exposure but trades around 37x forward earnings despite an 8% operating margin.

Positioning: This may offer some of the best valuation asymmetry in the physical chain, but it is not a clean or uniform exposure. Local transmission constraints mean national electricity growth will not translate evenly into company profits, and regulators can materially influence who pays for interconnection and generation investment.

10. Data-center landlords and interconnection platforms

Selected public companies: Equinix, Digital Realty, Iron Mountain.

These businesses monetize power-secured land, buildings and network interconnection. Equinix

reported 10% Q1 revenue growth, a 51% adjusted EBITDA margin and 12% AFFO growth. Approximately 60% of its largest deals were AI-related [7]. However, Equinix and Digital Realty both trade near 29x EV/EBITDA and carry net debt around 5–6x EBITDA. They must continue investing heavily to deliver capacity: Equinix expects roughly \$4.1 billion of 2026 capital expenditure [7].

Positioning: Demand is real, but economics depend on the spread between development returns and the cost of capital. Rising rates, power procurement, construction delays and customer concentration make this a less attractive expression than supplying scarce equipment to multiple developers and operators.

Where the exposure is most attractive

1. Best combination of durability and valuation: TSMC

TSMC combines 60% operating margins, above-40% expected 2026 revenue growth and an approximately 18x forward P/E. Its broad foundry and advanced-packaging role reduces dependence on which accelerator architecture ultimately gains share. Overseas expansion costs and geographic concentration prevent this from being a low-risk investment, but the operating franchise is unusually durable.

2. Highest-quality technology and ecosystem assets: ASML, NVIDIA and Broadcom

- **ASML's** EUV role, €38.8 billion backlog and installed-base business provide exposure to multiple chipmakers rather than a single end product [3].
- **NVIDIA** has the best current economics and the most integrated reported data-center platform in the group [1].
- **Broadcom** provides exposure to custom accelerators and the networking required around them [11].

Their central risk is that a handful of large customers possess substantial buyer power and are actively funding alternatives.

3. Best physical-infrastructure beneficiaries: Eaton and Vertiv

Power and cooling requirements increase with rack density across competing accelerator architectures. Eaton offers broader diversification and a lower multiple; Vertiv offers higher purity, faster growth and greater operational leverage. Neither is now undiscovered, making entry valuation important.

4. Best networking asset: Arista

Arista's operating margin, balance sheet and AI-focused product roadmap are exceptional. Its announced 1.6T systems address both scale-up and scale-out AI fabrics [17]. The limitation is a roughly 37x forward valuation and dependence on a concentrated set of large cloud and technology customers.

5. Attractive but higher-risk value: Vistra and Constellation

These companies provide electricity-generation exposure at less demanding multiples than many AI-

equipment beneficiaries. The exposure is nevertheless indirect and heavily influenced by regulation, local power prices, operating performance and transmission availability.

6. Tactical rather than durable: Micron, Coherent, Lumentum, Dell and Super Micro

These companies can produce the strongest earnings surprises during scarcity. They also face the greatest risk of margin normalization:

- Memory supply eventually responds.
- Optical standards and architectures change quickly.
- Server assemblers retain relatively little of the accelerator's economics.
- Working capital and customer concentration magnify downside during pauses.

7. Least compelling at current structure: data-center REITs

Equinix and Digital Realty have valuable power-secured portfolios, but high leverage, elevated EV/EBITDA multiples and continuing capital requirements limit upside relative to suppliers that can sell equipment into multiple new projects.

Key risks

1. **Hyperscaler return-on-capital failure.** A spending pause would propagate through servers, optics, semiconductors, equipment and construction, although timing would differ by backlog and contract structure. Rising debt and negative free cash flow show that investor tolerance is not unlimited [9].
2. **Efficiency and substitution.** Better models, inference optimization, lower-precision computation and custom ASICs could reduce GPU requirements per unit of AI output. Broadcom may benefit from custom silicon, but merchant accelerator vendors would face pressure.
3. **Customer concentration.** A small number of cloud companies have bargaining power over nearly every layer, particularly custom chips, optics, servers and real estate. Broadcom identified six core custom-chip customers as drivers of AI growth [11].
4. **Capacity overshoot.** HBM, optical components and server assembly are most vulnerable if suppliers extrapolate current shortages into permanent demand. SK Hynix's market outlook specifically identifies possible HBM price correction after 2026 as competition and capacity increase [18].
5. **Geopolitics and export controls.** The chain remains heavily dependent on TSMC's manufacturing footprint. China restrictions can also remove accelerator revenue directly: NVIDIA's Q2 outlook assumed no Data Center compute revenue from China [1].

6. **Power bottlenecks and regulation.** Transmission projects can take four to eight years, and grid constraints may delay data centers even when server demand remains strong [6]. Regulators can also determine how interconnection and generation costs are distributed.
7. **Technology transitions.** HBM4, N2, 1.6T Ethernet and liquid cooling require coordinated execution across multiple suppliers. The relevant company sources describe continuing ramps or future availability rather than fully completed transitions [15,17,18].
8. **Valuation compression.** Vertiv, Arista, Quanta and GE Vernova already trade at growth-stock multiples. Even successful earnings delivery may not protect them if the market lowers terminal-growth assumptions.

Bottom line

The AI infrastructure boom is not one trade but a hierarchy of profit pools. Economic value is currently highest where there is **proprietary technology plus scarce capacity**—accelerators, advanced foundry and packaging, lithography and high-performance networking. The next most durable opportunity is in **power distribution and cooling**, where demand is less dependent on which compute architecture wins.

The preferred risk-adjusted exposures are therefore **TSMC first; ASML, NVIDIA and Broadcom for technology scarcity; Eaton and Vertiv for the physical bottleneck; and Arista for network scaling.** Vistra and Constellation offer cheaper but more policy-sensitive exposure. Memory, optics and server assemblers are better treated as cyclical beneficiaries, while data-center REITs currently offer the least favorable combination of valuation, leverage and capital intensity.

III. SOURCES

- [1] NVIDIA Announces Financial Results for First Quarter Fiscal 2027
- [2] TSMC 2026 Q2 Quarterly Results - Taiwan Semiconductor ...
- [3] ASML reports €32.7 billion total net sales and €9.6 billion ...
- [4] Vertiv Holdings Co. - Vertiv Reports Strong First Quarter ...
- [5] QUANTA SERVICES REPORTS FIRST QUARTER 2026 RESULTS
- [6] Executive summary – Energy and AI – Analysis - IEA
- [7] Equinix Reports First-Quarter Results and... :: Equinix, Inc. (EQIX)
- [8] 2026 Q2 Earnings Call - Alphabet Investor Relations
- [9] Hyperscalers face higher capex scrutiny after Alphabet report panned
- [10] Micron Technology, Inc. Reports Record Results for the Third ...
- [11] Broadcom (AVGO) earnings report Q2 2026 - CNBC
- [12] Dell Technologies Delivers Fourth Quarter and Full-Year ...
- [13] Coherent Corp. Reports Third Quarter Fiscal 2026 Results
- [14] Energy supply for AI – Energy and AI – Analysis - IEA
- [15] Q2 2026 Taiwan Semiconductor Manufacturing Co Ltd Earnings ...
- [16] Data centers | Digitalization | Efficiency | Eaton
- [17] Arista Networks - Arista Introduces Next-Generation 1.6Terabit ...
- [18] 2026 Market Outlook – “Focus on the HBM-Led Memory Supercycle”

H.2 Valuation Compute Seam

Microsoft (MSFT): DCF, WACC & Sensitivity

I. RESEARCH QUESTION

Build an auditable five-year unlevered DCF for Microsoft (MSFT) from its latest fiscal-year results. Forecast revenue, EBIT margin, taxes, D&A, capex, and changes in working capital using clearly stated assumptions; derive WACC from current market inputs, a peer-based beta, prevailing Treasury yields, and an explicit equity-risk premium; use a 2.5% perpetual-growth rate; bridge enterprise value to equity value and intrinsic value per diluted share. Include a sensitivity table spanning WACC ± 2 percentage points and terminal growth from 1.5% to 3.5%. Compare the result with the current market price without making an investment recommendation.

II. CITED REPORT

Microsoft five-year unlevered DCF: \$233 per diluted share

Using Microsoft's FY2025 results as the operating base, a five-year unlevered DCF produces:

- **Enterprise value:** \$1.674 trillion
- **Equity value:** \$1.741 trillion
- **Intrinsic value:** \$233.19 per diluted share
- **Current market price:** \$381.70
- **DCF-to-market difference:** 38.9% below the current price

This is a valuation result, not an investment recommendation. The model's largest judgment calls are the **11.60% peer-beta-based WACC** and the pace at which elevated AI infrastructure spending normalizes.

1. FY2025 operating base

Microsoft's fiscal year ended June 30, 2025. Revenue was \$281.724 billion and operating income—used here as EBIT—was \$128.528 billion, producing a 45.62% EBIT margin. The reported income-tax provision was \$21.795 billion on \$123.627 billion of pretax income, an effective rate of 17.63%.^[2]

FY2025 anchor	Amount
Revenue	\$281.724B
EBIT	\$128.528B
EBIT margin	45.62%
Reported income-tax provision	\$21.795B
Effective tax rate	17.63%
D&A and other	\$34.153B
Capital expenditures	\$64.551B
Operating cash flow	\$136.162B
Reported OCF less capex	\$71.611B

D&A increased from \$22.287 billion in FY2024 to \$34.153 billion in FY2025, while capital expenditures increased from \$44.477 billion to \$64.551 billion.^[3] Those increases—53.2% and 45.1%,

respectively—support modeling a period in which depreciation catches up with the recent infrastructure build and capex gradually normalizes.

Working-capital baseline

The FY2025 cash-flow statement reported the following operating-asset and liability changes:^[3]

- Accounts receivable: \$(10.581)B
- Inventories: +\$0.309B
- Other current assets: \$(3.044)B
- Accounts payable: +\$0.569B
- Unearned revenue: +\$5.438B
- Other current liabilities: +\$5.922B

The net cash effect was a **\$1.387 billion outflow**, equivalent to 3.79% of FY2025's \$36.602 billion revenue increase. The forecast rounds this to a **4.0% investment for every dollar of incremental revenue**.

2. Forecast assumptions

These are explicit analyst assumptions rather than company guidance or consensus estimates.

Assumption	FY26E	FY27E	FY28E	FY29E	FY30E
Revenue growth	14.0%	13.0%	12.0%	10.0%	8.0%
EBIT margin	45.7%	45.9%	46.1%	46.3%	46.5%
Tax rate on EBIT	18.0%	18.0%	18.0%	18.0%	18.0%
D&A as % of revenue	13.5%	14.5%	15.0%	15.2%	15.0%
Capex as % of revenue	22.0%	20.0%	18.0%	16.5%	15.0%
Change in NWC	4% of incremental revenue in every year				

The revenue path represents an 11.38% five-year CAGR, decelerating from FY2025's 15% reported growth. Microsoft reported Azure annual revenue above \$75 billion, up 34%, providing context for maintaining double-digit consolidated growth through FY2028E.^[2]

The EBIT-margin assumption provides only modest expansion—45.62% to 46.5%—because scale and product mix benefits are assumed to be partly offset by rising depreciation from the infrastructure build. Capex remains above D&A through FY2029E and converges with D&A in FY2030E.

3. Five-year unlevered free-cash-flow forecast

Formula:

$$\text{UFCF} = \text{EBIT} \times (1 - \text{tax rate}) + \text{D\&A} - \text{capex} - \Delta \text{NWC}$$

Amounts are in billions of dollars.

	FY26E	FY27E	FY28E	FY29E	FY30E
Revenue	\$321.165	\$362.917	\$406.467	\$447.114	\$482.883
EBIT	146.773	166.579	187.381	207.014	224.540
Taxes on EBIT	(26.419)	(29.984)	(33.729)	(37.263)	(40.417)
NOPAT	120.354	136.595	153.653	169.751	184.123
D&A	43.357	52.623	60.970	67.961	72.432
Capex	(70.656)	(72.583)	(73.164)	(73.774)	(72.432)
Change in NWC	(1.578)	(1.670)	(1.742)	(1.626)	(1.431)
Unlevered FCF	\$91.477	\$114.964	\$139.717	\$162.313	\$182.692

The DCF calculation uses the displayed unrounded FCF schedule.

4. WACC derivation

Peer-based beta

The model uses the equal-weighted mean of current observed equity betas for five large cloud, enterprise-software and platform peers:

Peer	Equity beta
Alphabet	1.25
Amazon	1.46
Oracle	1.71
Salesforce	1.18
Adobe	1.43
Equal-weighted mean	1.406

This is an observed peer-equity-beta approach; the individual betas are not unlevered and relevered.

Cost of capital

Input	Value	Basis
Risk-free rate	4.68%	Current 10-year Treasury yield
Equity-risk premium	5.00%	Explicit model assumption
Peer beta	1.406	Equal-weighted peer mean
Cost of equity	11.71%	4.68% + 1.406 × 5.00%
Pretax cost of debt	5.18%	Treasury yield plus assumed 0.50% spread
Tax rate	17.63%	FY2025 effective tax rate
After-tax debt cost	4.27%	$5.18\% \times (1 - 17.63\%)$
Debt/equity	1.519%	$\$43.151\text{B debt} \div \$2.840\text{T market capitalization}$
Equity weight	98.50%	Market-value weight
Debt weight	1.50%	Market-value weight
WACC	11.5986%	Rounded to 11.60%

Microsoft's FY2025 debt consisted of \$2.999 billion of current long-term debt and \$40.152 billion of long-term debt; short-term debt was zero.[4]

5. Terminal value and enterprise-to-equity bridge

The terminal value uses a **2.5% perpetual-growth rate**:

$$\text{Terminal value}_{2030} = \frac{\$182.692\text{B} \times 1.025}{11.5986\% - 2.5\%} = \$2.058\text{T}$$

Valuation component	Amount
PV of five explicit FCFs	\$484.990B
PV of terminal value	1,188.980B
Enterprise value	\$1,673.970B
Add: cash and short-term investments	94.565B
Add: equity and other investments	15.405B
Less: debt	(43.151B)
Equity value	\$1,740.789B
Diluted shares	7.465B
Intrinsic value per diluted share	\$233.19

Microsoft reported \$30.242 billion of cash, \$64.323 billion of short-term investments and \$43.151 billion of total debt at June 30, 2025.[4] Equity and other investments were \$15.405 billion.[2] The valuation uses FY2025 weighted-average diluted shares of 7.465 billion rather than the 7.434 billion period-end basic share count.[5]

The terminal value represents **71.03% of enterprise value**, making the result materially sensitive to WACC and perpetual growth.

6. Sensitivity analysis

Intrinsic value per diluted share:

WACC \ Terminal growth	1.5%	2.0%	2.5%	3.0%	3.5%
9.6%	\$272	\$285	\$301	\$319	\$340
10.6%	\$241	\$251	\$263	\$276	\$291
11.6%	\$216	\$224	\$233	\$243	\$255
12.6%	\$196	\$202	\$209	\$217	\$226
13.6%	\$179	\$184	\$190	\$196	\$203

The requested sensitivity range is **\$179–\$340 per share**. Even the highest value in the grid—9.6% WACC and 3.5% perpetual growth—is 10.9% below the current \$381.70 market price.

Bottom line

The base model values Microsoft at **\$233.19 per diluted share**, versus a current price of **\$381.70**. The difference indicates that the market price reflects some combination of faster or more durable cash-flow growth, a lower required return, greater margin expansion, or less reinvestment than assumed here. No investment recommendation is made.

III. SOURCES

- [1] Microsoft 2025 Annual Report
- [2] FY25 Q4 - Press Releases - Investor Relations - Microsoft
- [3] FY25 Q4 - Cash Flows - Investor Relations - Microsoft
- [4] FY25 Q4 - Balance Sheets - Investor Relations - Microsoft
- [5] FY25 Q4 - Income Statements - Investor Relations - Microsoft