
SAB: Benchmarking LLM Agents on Startup Research and Ranking

Maxim Kalpin
Mento vc
max@mento.vc

Alexander Zhuravlev
Mento vc
alex@mento.vc

Dmitry Golomidov
Mento vc
dima@mento.vc

Sofya Bilich
Mento vc
sonya@mento.vc

Violetta Li
Mento vc
vi@mento.vc

Abstract

Venture funds increasingly use AI tools, including LLM agents, to research companies. Before relying on such an agent, a fund needs to know whether the agent finds the relevant evidence when most documents are irrelevant, stale or about a same-named company; whether it cites the sources its answers rest on; whether it answers “unknown” when evidence is missing; and whether it ranks companies better than a simple rule. We built SAB (Startup Agent Benchmark) at Mento, a venture fund, to answer these questions whenever our agents change. Its research part has 960 tasks across six analyst workflows, half on synthetic companies and half on a frozen snapshot of the fund’s data; agents search each task’s documents through two tools, and a deterministic grader checks answers and cited source IDs against private keys. Its ranking part gives a model, with no tools, the masked profiles of 100 companies and asks it to rank them by “operating momentum”, a term the prompt leaves undefined; across 101 such cohorts from late 2025, a ranking is scored by its Spearman correlation ρ with the companies’ subsequent growth in web traffic and headcount (1 for the realized order, 0 in expectation for a random one). Across 14 models, the three best average 96.2–97.2 of 100 research points and pass 94–96% of tasks under a strict rule (at least 80 points including 10 for evidence, every critical field correct, no nonexistent citation). These scores use answer keys corrected after an audit of the released keys, which score every model 4.8–7.5 points lower. Models rarely cite nonexistent sources but give a value for 9–33% of the facts we removed from the corpus. On ranking, all 13 models that completed the cohorts score below a heuristic that orders companies by past headcount growth (ρ from -0.16 to -0.003 , against $+0.06$), and 12 fall below the 95% range of random orderings. Their rankings favor larger companies, whose web traffic grew more slowly in percentage terms; a post hoc rule that ranks the smallest companies first scores $+0.24$. Scored on headcount growth alone, 10 of the 13 have point estimates above the heuristic’s, a difference we did not test, so the comparison with the heuristic depends on how the informally stated objective is measured. Mento now uses agents to screen companies against explicit criteria and to gather cited evidence; investors check the evidence and do the ranking.

1 Introduction

A venture fund reviews about a hundred companies for each investment it makes, and venture investors rate deal selection as the most important source of value creation, ahead of access to deals and support

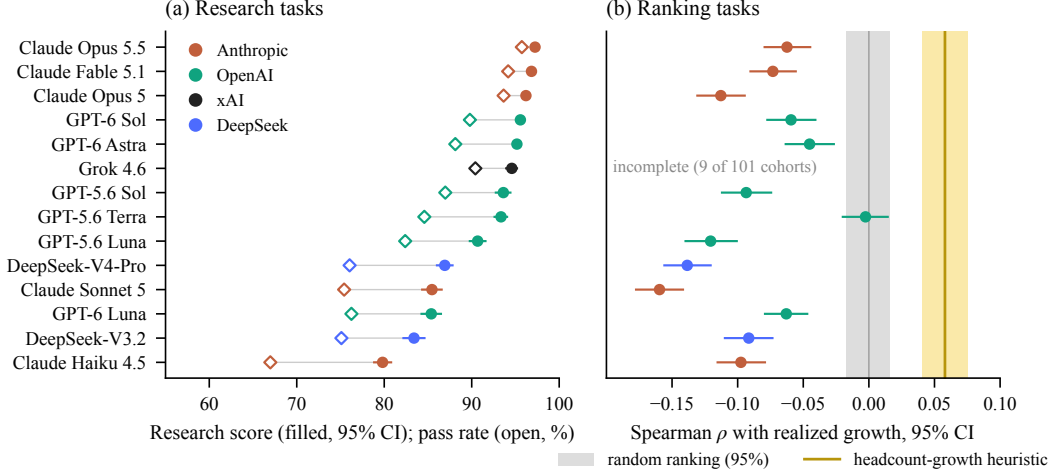


Figure 1: **Results on the two parts of SAB.** (a) Filled circles: research score (points per task, 80 for answer content and 20 for evidence, averaged over 960 tasks, corrected keys) with 95% stratified-bootstrap intervals. Open diamonds: strict pass rate (%; at least 80 points including 10 for evidence, every critical field correct, no nonexistent citation). Answering unknown everywhere scores 7.5. (b) Spearman ρ between each model’s ranking of 100 companies and their subsequent growth in web traffic and headcount (expected 0 for a random order), averaged over the 90- and 180-day horizons and 101 cohorts, with 95% bootstrap intervals over cohorts. Gray: 95% range of the same average for random rankings. Gold: headcount-growth heuristic (ranks companies by past headcount growth) with its interval. Every model ranks larger companies earlier on average, and larger companies had lower growth in web traffic, the more heavily weighted part of the target (§6).

after investing [13]. Much of the work behind selection is research on each candidate: what the company sells and to whom, how its market is moving, who it competes with and how its team and traction compare with peers. Funds increasingly hand this research to AI tools; in a vendor-run survey, 64% of investors reported using AI to speed up company research [1]. An agent’s errors can reach an investment memo as an unsupported figure, a stale metric presented as current or a confident answer where no data exist, and a screen built on its output can favor the wrong companies. Each has a precedent: a commissioned government report contained fabricated references [10], models answer confidently where they should abstain [37], and LLM-based financial backtests suffer from look-ahead bias [26, 27].

Before relying on a research agent, a fund therefore needs answers to four questions. Does the agent find the relevant evidence when most documents are irrelevant, stale or about a same-named company in another vertical? Does it cite the sources its answers rest on, and only sources that exist? Does it answer “unknown” when the evidence is missing? When it ranks companies, does it do better than a simple rule? Existing benchmarks answer parts of these questions. Startup-success studies score predictions made from given company profiles [3, 8, 16, 36], and a review of 57 such studies describes the field as methodologically fragmented [18]. Deep-research benchmarks evaluate agentic search on the open web or on general-purpose corpora [9, 14, 15, 29, 38]; only BrowseComp-Plus partly scores abstention, and none scores a ranking of companies against realized outcomes. Commercial deal-sourcing platforms offer research agents but, to our knowledge, publish no held-out evaluation of them.

We run SAB whenever a model, prompt or data source behind Mento’s agents changes; here we report 14 models in its frozen harness (§4). It has two separate parts: research tasks, whose answers and cited source IDs are graded against private keys, and ranking tasks, scored against the companies’ subsequent growth; the ranking tasks do not use the research answers. Our contributions are a benchmark with 960 research tasks, in six analyst workflows on synthetic and real company data, and 101 ranking cohorts scored against realized growth (§3); an evaluation protocol with task-scoped tools under fixed budgets, deterministic grading of answers and cited source IDs, a utility that credits unknown on withheld facts and an audit of the answer keys against model outputs (§4); and an evaluation of 14 models with paired statistics, costs, failure analysis and reference baselines (§5–6).

Table 1: Research task families. Each family has 80 synthetic and 80 real-world tasks.

Family	What the agent must do	Main difficulty
Product & customer	Identify the target among look-alikes; return its category, customer type, verticals and technologies	Same-named company in another vertical (synthetic)
Market & why-now	Select an exact peer set and compute peer medians, or compute a bottom-up market size and name the reasons the market is opening now	Near-miss peers; arithmetic over many documents
Competition	Map direct and indirect alternatives (synthetic); rank candidates with a disclosed similarity formula (real-world)	Partial matches on single factors
Team	Assess founder backgrounds and gaps (synthetic); deduplicate roles and compute headcount growth (real-world)	Namesakes; duplicate roles; missing biographies
Traction & risks	Reconcile dated metrics across providers and apply a stated risk rule	Stale figures from authoritative-looking sources
Sourcing screen	Collect every company that matches criteria and remove those already in the fund’s database	Near-miss companies; finding every match

Figure 1 summarizes the results. On research, the three best models average 96.2–97.2 points, yet every model gives a value for some withheld facts; on ranking, all 13 models that completed the cohorts score below the headcount-growth heuristic, although 10 have point estimates above it when scored on headcount growth alone.

2 Related work

Machine learning for startup selection. Startup screening with machine learning predates LLMs: a gradient-boosted classifier trained on 77,279 European early-stage companies outperformed the median of 111 venture investors in a screening survey [33], and EQT has built an in-house system [6, 7]. Investors who adopt such tools select more companies that resemble past successes and fewer novel ones [5]. LLM-based work predicts founder or startup success from profiles [8, 16, 36, 39] or simulates investment committees [3]. These studies evaluate predictions from given inputs; SAB evaluates the research that produces such inputs and, separately, rankings scored against outcomes.

Deep-research and agentic-search benchmarks. GAIA [29], BrowseComp [38], Assistant-Bench [41], Mind2Web 2 [14] and DeepResearch Bench [11] evaluate agents on the live web; BrowseComp-Plus [9] freezes the corpus to make retrieval comparable, and LiveBrowseComp [12] shows that closed-book knowledge inflates many search scores. FinSearchComp [15] and FinanceBench [17] target financial questions, and τ -bench [40] evaluates tool-using agents in simulated domains. Like BrowseComp-Plus, SAB uses a frozen corpus and checks cited source IDs; unlike it, SAB grades structured answers, credits abstention in the score and adds rankings scored against outcomes (Table 3 in Appendix A).

Forecasting and evaluation practice. ForecastBench [21] collects forecasts before the events resolve, so models cannot have seen the answers; backtests of LLM-based financial strategies are prone to look-ahead and survivorship bias [23–27], and both concerns apply to our ranking cohorts. We follow prior work on statistics [28, 30, 31], cost [19, 20], benchmark construction [4, 34, 42] and abstention scoring [22, 32, 37].

3 The SAB benchmark

Research tasks. The 960 research tasks cover six workflows of a venture analyst, one task family per workflow (Table 1), with 80 tasks per family in each of two tracks. Each task names a target company, asks one question, often with an explicit formula or decision rule, and fixes a JSON response schema in which every answer field also admits unknown. Answers are graded on 4,560 criteria, each scored by normalized exact match (strings), an explicit tolerance (numbers) or set F1 (lists). The keys mark 1,735 criteria as critical, typically those on the fields that the task’s question turns on. Appendix B gives example tasks.

Synthetic and real-world tracks. The *synthetic* track (480 tasks) uses fictitious companies produced by a generator, so that no model can know them in advance and every trap is placed deliberately. A synthetic environment holds a median of 49 documents; on average 5.5% of them support a graded answer, and the rest are distractors (SEO pages, directory listings, dead links, paywalled stubs and a same-named company in another vertical). Documents carry publication dates but no staleness labels; in traction tasks, an authoritative-looking figure is five months older than the verified one. The *real-world* track (480 tasks) is built from a frozen snapshot of 8,000 company records in Mento’s data warehouse, retrieved on 21 August 2026, with monthly company metrics through June 2026 and web-traffic panel observations through August 2026. The warehouse combines commercial company-data feeds (profiles and monthly metrics, web-traffic and SEO panels, funding rounds) with public professional profiles. A real-world environment holds a median of 14 documents, about 10 of them distractors derived from other records; in sourcing environments the distractors are instead companies that narrowly miss the screening criteria. The two tracks contain 30,529 documents in 960 environments.

Withheld facts. For every fifth task in each family, the generator tries to remove one document and labels a criterion unanswerable if the removal breaks its source-support rule: no supporting document is left or, for a criterion that needs several documents, one of them is gone. It rejects a removal if an affected value still appears verbatim in another document, or if no criterion or every criterion would be affected. Removal succeeded in 111 of the 192 attempts, mostly in synthetic tasks; under the corrected keys, 95 tasks contain at least one of the 221 unanswerable criteria (Appendix C). Withheld criteria are critical, so giving any value for one fails the task.

Ranking cohorts. Each of the 101 ranking tasks presents 100 privately held companies (10,100 in total, none repeated) with masked identities, founded in 2018–2025 with 2–200 employees and a baseline monthly web-traffic estimate between 10 and 5,000,000, and asks the model to rank them from most to least likely to show the strongest operating momentum over the next 90 and 180 days; the task contains every input and allows no search or fetch calls, so no search can surface later information. Each company comes with a description observed before the baseline date, with names and domains masked, and a signal panel: web traffic, headcount and total funding at the baseline date and, where available, three and six months earlier; founding year, employee range and industry; and two flags for whether the description mentions founders and prior founding experience. Five monthly waves (baselines August–December 2025) contain 20 cohorts each; one further cohort, the *recent cohort* (baseline December 2025), contains only companies whose descriptions first appeared after 1 June 2025, so that their text is less likely to be in the models’ training data; it is included in all averages and also shown separately (Appendix G). The target at each horizon is $0.625 F(\Delta T) + 0.375 F(\Delta H)$, where ΔT and ΔH are a company’s percentage growth in web traffic and headcount from the baseline and F is the percentile within the cohort. Investment returns take years to observe, so the target uses the two operating signals that the warehouse records monthly for every cohort company; its weights were set before any model was run on this release (Appendix G), and the prompt asks for operating momentum without stating the formula. A ranking is scored by the Spearman ρ between its order and the target, averaged over the two horizons (outcome snapshots three and six calendar months after the baseline) and then over cohorts: $\rho = 1$ means the ranking matches the realized order, and a random order has expected $\rho = 0$. We compare models with random rankings and with the release’s predefined *headcount-growth heuristic*, which ranks companies by headcount growth from their earliest stored month before the baseline (normally six months earlier) and places companies without that signal last. Companies that lacked positive traffic or headcount at the baseline or at either horizon were excluded; they are 28–37% of the eligible companies in each wave (Appendix G).

Environment and harness. Tasks are served through an MCP [2] server whose two tools are scoped to the current task’s environment (Figure 6 in Appendix B): `benchmark_search`, which returns up to 10 results, and `benchmark_fetch`, which returns one document. A synthetic task allows 8 searches and 12 fetches, a real-world task 12 and 20; the server rejects calls beyond these budgets. A frozen harness fixes the system prompt and task rendering, limits each conversation to 32 model turns and 32 tool calls (rejected calls included), validates the final answer against the JSON schema (an invalid answer is returned to the model with the error) and ignores repeated submissions. We evaluate release `sab-v0.7.0-baa0ceb6c489`, whose manifest records a hash commitment to the generation seed and the hash of every file, and score it with the corrected keys of §4.

4 Evaluation protocol

Scoring. Let criterion j of a task have weight w_j (the weights sum to 80) and accuracy $a_j \in [0, 1]$: 1 for an exact match or a number within tolerance, the F1 score for a list-valued answer, and, on a withheld criterion, 1 for unknown and 0 for any value. Let $e_j = 1$ if the answer cites a set of sources that the key marks as sufficient for criterion j (a single supporting document, or every document that a multi-source criterion needs) or, on a withheld criterion, answers unknown with no citation; otherwise $e_j = 0$. The task score is $\sum_j w_j a_j + \frac{1}{4} \sum_j w_j a_j e_j - \min(20, 10f)$, floored at zero, where f is the number of cited IDs that do not exist in the task’s corpus; the first sum gives up to 80 points for content and the second up to 20 for evidence. The grader checks that cited IDs exist and include the supporting ones; it does not check that the agent fetched them, and extra valid citations cost nothing. A task *passes* if its answer matches the schema, scores at least 80 points including at least 10 for evidence, answers every critical criterion correctly and cites no nonexistent ID. The **research score** is the mean task score over the 960 tasks; because each of the 12 track–family cells contains 80 tasks, it equals the macro average over cells. The release designates the pass rate as its primary leaderboard metric; we order models by research score because it gives partial credit, and report both. The **utility** of a task is $\sum_j w_j u_j / 80$, between -1 and 1 , with $u_j = +1$ for unknown on a withheld criterion, -1 for any value on a withheld criterion, $+0.2$ for unknown on an answerable criterion, 0 for a missing field and $2a_j - 1$ otherwise. On an answerable criterion, answering therefore pays more than abstaining only if the expected accuracy exceeds 0.6 . Where a corrected key also accepts unknown (Appendix C), that answer receives $a_j = 1$, $u_j = +1$ and, if uncited, $e_j = 1$, and it still counts as an abstention on an answerable criterion. As a reference, an agent that answers unknown to every field and cites nothing scores 7.5 points, with a utility of 0.26 ; it passes 2.6% of tasks, those in which every critical criterion is withheld and withheld criteria carry at least 64 of the 80 content points.

Key audit. After the runs we checked, against its documents and prompt, every criterion that few or no models answered correctly. We found keys that contradict the documents, unstated output conventions (label names, field formats, ratio precision, copied spans), mislabeled withheld facts, a cutoff date with two readings, citation rules that demand redundant documents or accept insufficient ones, and keys expecting 0 or *none* where the instructions ask for *unknown*. All research results use the corrected keys (v0.7.1), applied to the saved answers; Appendix C gives each correction and its effect.

Models and setup. We evaluate 14 models from four vendors: Claude Fable 5.1, Opus 5.5, Opus 5, Sonnet 5 and Haiku 4.5 (Anthropic); GPT-6 Astra, GPT-6 Sol, GPT-6 Luna, GPT-5.6 Sol, GPT-5.6 Terra and GPT-5.6 Luna (OpenAI); Grok 4.6 (xAI); and DeepSeek-V4-Pro and DeepSeek-V3.2 (DeepSeek). All models were accessed through APIs with two revisions of the frozen harness (Appendix D) that share the task prompt and tool contract; no chat interfaces, IDE agents or vendor agent products were involved, since each adds its own prompts and tools, and API scores do not reliably transfer even to chat interfaces [35]. Every model ran zero-shot at high reasoning effort through its provider’s native setting; for Claude Haiku 4.5, to which our gateway passes no effort setting, we used extended thinking with a $4,096$ -token budget (Appendix D). We ran each model once and scored the first accepted answer to each task. A task interrupted by a provider failure or by the turn or tool-call limit is restarted in a new conversation with a fresh turn limit, but the searches and fetches it has already spent still count; Appendix D lists these events. Grok 4.6’s ranking requests, each with 100 company profiles, repeatedly timed out at the provider; the model completed 9 of 101 cohorts, and we exclude it from the ranking analysis. Three models (DeepSeek-V4-Pro, DeepSeek-V3.2 and Grok 4.6) were also run *closed-book*: search returns nothing and fetch is refused.

Statistics and cost. All 14 models answer the same research tasks, so we use paired statistics. Intervals are pointwise 95% percentile intervals from 10,000 bootstrap replicates that resample tasks within each of the 12 cells, with the same draws for every model; the same replicates give rank intervals. Pairs of models are compared with a two-sided sign-flip permutation test on per-task score differences, with Holm correction over the 13 pairs adjacent in Table 2 and, separately, over all 91 pairs. Ranking intervals resample cohorts, keeping both horizons of a cohort together and using the same draws for every model and baseline; they describe variation across cohorts within the five observed months and do not capture variation across time periods. Costs are computed from logged token counts at each provider’s public list price (Appendix F); we did not use invoiced amounts.

Table 2: **Main results.** *Mean*: research score over 960 tasks (corrected keys; released scores in Table 6) with a 95% stratified-bootstrap interval. *Rank*: 95% bootstrap interval of the rank. *Pass*: strict pass rate (%). *Synth.*, *Real*: score per track. *Utility*: mean per-task utility (§4), which rewards unknown on withheld facts and penalizes values given for them, from -1 to 1 . *Fab.*: tasks citing a nonexistent source ID. *\$/task*: list-price cost per research task. *Ranking ρ* : mean Spearman correlation of the ranking with realized growth over 101 cohorts; $-$: did not complete the cohorts.

#	Model	Research score		Pass (%)	Track			Fab.	\$/task	Ranking ρ
		Mean [95% CI]	Rank		Synth.	Real	Utility			
1	Claude Opus 5.5	97.2 [96.9, 97.5]	1–2	96	98.7	95.8	0.97	0	0.113	−0.062
2	Claude Fable 5.1	96.8 [96.3, 97.3]	1–2	94	97.8	95.9	0.96	0	0.359	−0.073
3	Claude Opus 5	96.2 [95.7, 96.6]	3	94	97.2	95.2	0.94	0	0.180	−0.113
4	GPT-6 Sol	95.6 [95.0, 96.1]	4–5	90	95.3	95.9	0.95	0	0.051	−0.059
5	GPT-6 Astra	95.2 [94.7, 95.6]	4–6	88	93.8	96.6	0.95	0	0.166	−0.045
6	Grok 4.6	94.6 [93.8, 95.3]	5–7	90	96.0	93.2	0.93	0	0.112	–
7	GPT-5.6 Sol	93.6 [92.6, 94.5]	6–8	87	94.5	92.7	0.93	0	0.131	−0.093
8	GPT-5.6 Terra	93.4 [92.5, 94.2]	7–8	85	94.7	92.0	0.93	0	0.055	−0.003
9	GPT-5.6 Luna	90.7 [89.7, 91.7]	9	82	89.7	91.7	0.89	2	0.008	−0.121
10	DeepSeek-V4-Pro	86.9 [85.9, 87.9]	10–11	76	82.4	91.5	0.87	0	0.079	−0.138
11	Claude Sonnet 5	85.5 [84.2, 86.7]	11–12	75	85.9	85.0	0.80	0	0.083	−0.160
12	GPT-6 Luna	85.4 [84.1, 86.6]	11–12	76	89.7	81.0	0.83	1	0.003	−0.063
13	DeepSeek-V3.2	83.4 [82.1, 84.7]	13	75	86.5	80.3	0.81	0	0.107	−0.092
14	Claude Haiku 4.5	79.8 [78.7, 80.9]	14	67	81.7	77.9	0.73	0	0.065	−0.098
Answer unknown everywhere			7.5	2.6	8.1	7.0	0.26	0		
Headcount-growth heuristic										+0.058
Smallest-first rule (§6)										+0.242

5 Research results

Leaderboard. Claude Opus 5.5 averages 97.2 points, 0.41 above Claude Fable 5.1 (95% CI $[-0.02, 0.87]$, unadjusted $p = 0.08$; Appendix E). Three of the 13 differences between adjacent models are significant after Holm correction and remain so with correction over all 91 pairs ($p_{\text{Holm}} \leq 0.02$, the smallest adjusted value that 5,000 draws per pair can resolve): GPT-5.6 Terra over GPT-5.6 Luna, GPT-5.6 Luna over DeepSeek-V4-Pro and DeepSeek-V3.2 over Claude Haiku 4.5; the other adjacent orderings are not resolved at this sample size. Strict pass rates order the models differently from mean scores: Grok 4.6, sixth by mean score, is fourth by pass rate (90%). GPT-6 Sol scores 1.9 points above its predecessor GPT-5.6 Sol, but GPT-6 Luna scores 5.3 points below GPT-5.6 Luna, entirely on the real-world track (81.0 vs. 91.7; both score 89.7 on synthetic tasks).

Closed-book control. Run closed-book, DeepSeek-V4-Pro and DeepSeek-V3.2 (960 answers each) and Grok 4.6 (958 of 960) answer unknown almost everywhere, as instructed, and score 7.56, 7.49 and 7.53, against 86.9, 83.4 and 94.6 with the corpus and 7.53 for the all-unknown reference. Their scores therefore depend on the corpus; because the instruction forbids prior knowledge, this control does not measure what the models know.

Difficulty by family and track. Sourcing screens (90.5–100 in both tracks) and real-world product tasks (99–100) are near the ceiling for every model (Figure 2). Averaged over the two tracks, models differ most in market & why-now (53–97) and in competition (60–96). The lowest single scores are in synthetic market & why-now (32–100), real-world competition (28–94) and real-world traction (68–96); in synthetic market tasks, Claude Haiku 4.5 and DeepSeek-V4-Pro answer unknown on most answerable fields. The two tracks order the models similarly but not identically (Spearman $\rho = 0.78$, Figure 3b): DeepSeek-V4-Pro scores 82.4 on synthetic tasks and 91.5 on real-world ones, GPT-6 Luna 89.7 and 81.0.

Cost. At list prices a research task costs from \$0.003 (GPT-6 Luna) to \$0.359 (Claude Fable 5.1), a factor of more than 100 (Figure 3a). The Pareto frontier runs through GPT-6 Luna, GPT-5.6 Luna, GPT-6 Sol and Claude Opus 5.5, which scores highest at less than a third of Fable 5.1’s cost.

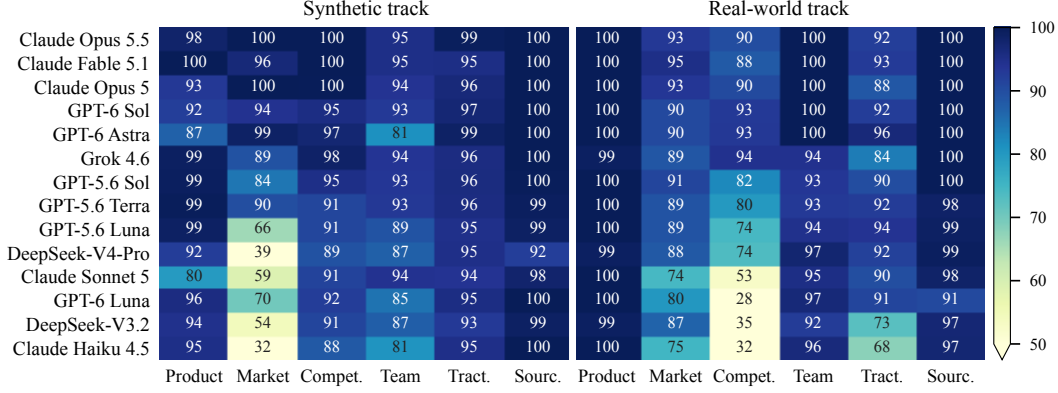


Figure 2: **Mean research score by family and track** (80 tasks per cell, corrected keys). Columns are the six families of Table 1. Sourcing screens and real-world product tasks are near the ceiling; synthetic market, real-world competition and real-world traction separate the models most. Cells below 50 share the lightest color.

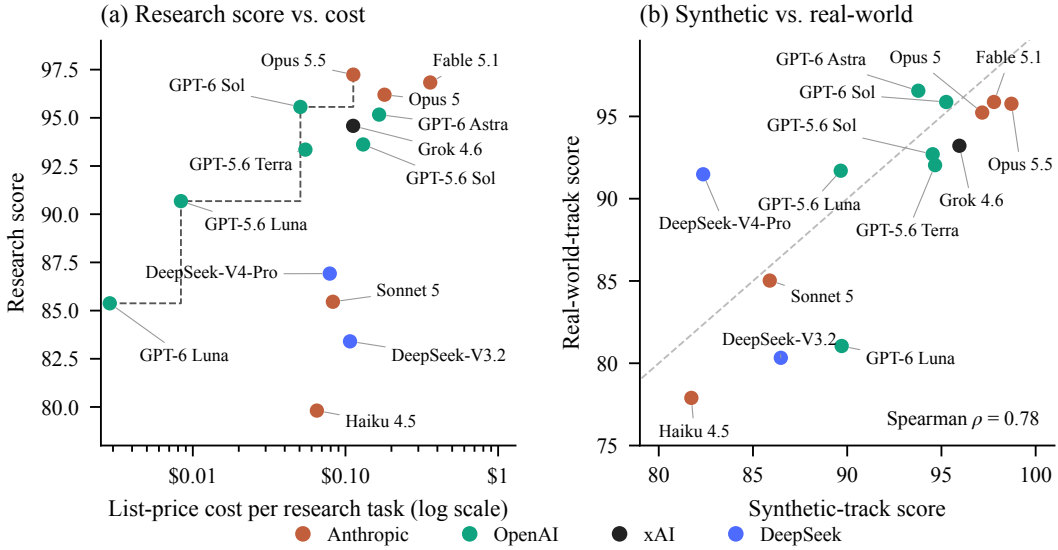


Figure 3: **Cost and agreement between tracks.** (a) Research score against list-price cost per research task (log scale); the dashed line is the Pareto frontier of the observed means. (b) Synthetic-track score against real-world-track score; the dashed line is $y = x$.

Citations and abstention. Citations of nonexistent source IDs are rare: no model makes one in more than two of its 960 tasks. Summed over models, the most common primary causes of failure are a wrong value in a critical field (528 failed tasks; 0.5–14% of tasks per model), unknown on an answerable critical field (507) and a value given for a withheld fact (464; Figure 4a). Every model gives a value for some withheld facts, 9–33% of the 221 withheld criteria: Claude Opus 5.5 (9%) and GPT-6 Astra (10%) do so least often, and the three GPT-5.6 models, Grok 4.6 and Claude Sonnet 5 (30–33%) most often (Figure 4b). Weaker models also err in the other direction: DeepSeek-V4-Pro, GPT-6 Luna, Claude Haiku 4.5 and DeepSeek-V3.2 answer unknown where the key expects a value on 7–10% of answerable criteria, against 0–2.2% for the five best models.

6 Ranking results

Raw results. The headcount-growth heuristic scores $\rho = +0.058$ [0.041, 0.075]. All 13 models that completed the cohorts score between -0.160 (Claude Sonnet 5) and -0.003 (GPT-5.6 Terra),

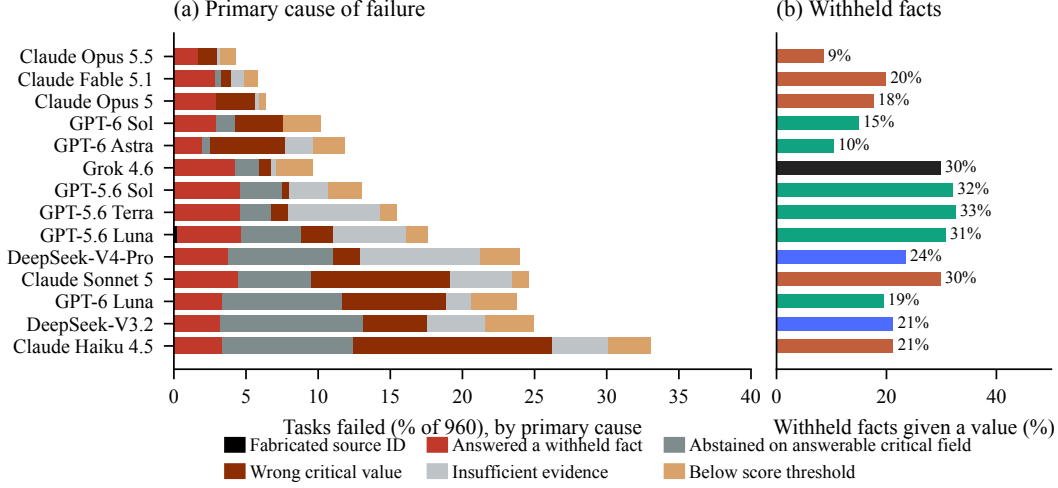


Figure 4: **Failure modes.** (a) Share of the 960 research tasks that fail, by primary cause; each failed task receives the first applicable label in legend order (read by rows), and bars are stacked in that order from the left. *Insufficient evidence*: fewer than 10 evidence points; *below score threshold*: fewer than 80 points. (b) Share of the 221 withheld criteria for which the model gave a value instead of unknown.

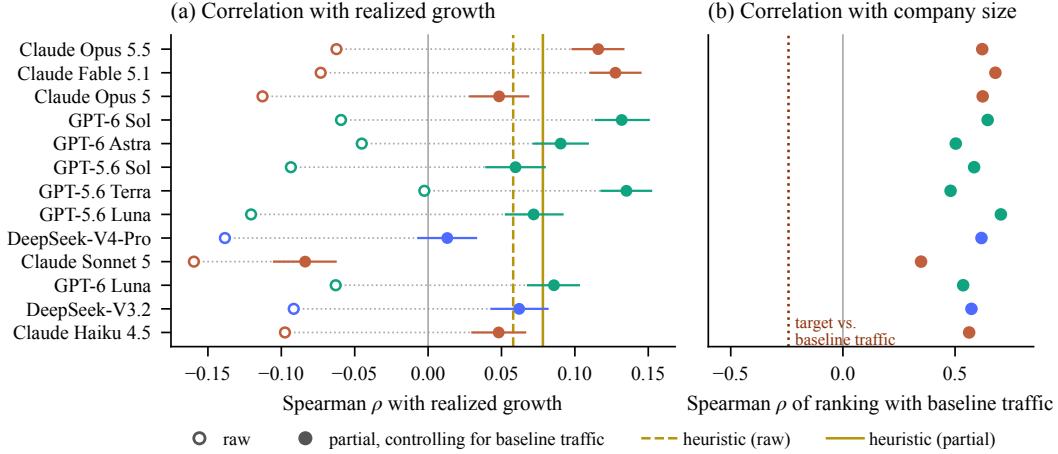


Figure 5: **Company size in the rankings.** (a) Open circles: mean Spearman ρ with realized growth, as in Figure 1b. Filled circles: partial correlation controlling for baseline web traffic, that is, the rank correlation left after the linear effect of baseline-traffic ranks is removed from both the ranking and the target, with 95% cohort-bootstrap intervals; gold dashed and solid lines show the heuristic, raw and partial. (b) Spearman ρ between each model’s ranking and baseline web traffic (positive: larger companies first); the dotted line is the correlation between the target and baseline traffic (-0.24).

and for each of them the paired 95% interval of the difference from the heuristic excludes zero (the smallest gap is GPT-5.6 Terra’s, -0.061 [$-0.080, -0.041$]). Of the 13 means, 12 fall below the 95% range of the random-ranking reference, $[-0.017, +0.016]$; GPT-5.6 Terra’s lies inside it. In each of the five monthly waves, every model’s mean is below the heuristic’s (Appendix G). Because a fund acts on the top of a list, we also report precision at 10 ($P@10$), the share of a ranking’s first ten companies that are among the cohort’s ten highest-target companies: 0.07 – 0.11 for the models, against 0.10 for a random order, 0.14 for the heuristic and 0.31 for the smallest-first rule defined below (Appendix G).

Company size. Every model places larger companies earlier: the correlation between each model’s ranking and baseline web traffic ranges from $+0.35$ to $+0.70$ (Figure 5b). In these cohorts smaller

companies grew faster in web traffic ($\rho = -0.32$ between traffic growth and baseline traffic), while headcount growth was slightly higher at larger companies ($\rho = +0.06$ between headcount growth and baseline traffic); the target correlates with baseline traffic at $\rho = -0.24$ (-0.21 at 90 days and -0.27 at 180 days). A post hoc *smallest-first rule* that ranks companies from smallest to largest baseline traffic scores $+0.24$ $[0.22, 0.26]$. Baseline traffic is the denominator of traffic growth; measurement noise, regression to the mean, real differences in growth and outcome-based selection could all contribute to the association, and our analysis does not separate them. We therefore report the smallest-first rule only as a diagnostic of how the target depends on size. The heuristic’s procedure applied to past web-traffic growth (the *traffic-growth rule*, also post hoc) scores -0.17 , below every model; like the heuristic, it places companies without a computable signal last.

Sensitivity to the target. Scored on traffic growth alone (Appendix G), every model’s ρ is negative (-0.25 to -0.08) and the heuristic’s is near zero ($+0.01$); with equal weights, every model stays below the heuristic ($+0.08$). Scored on headcount growth alone, on which larger companies did slightly better, every model’s ρ is positive ($+0.075$ to $+0.153$) and 10 of 13 have point estimates above the heuristic’s $+0.105$ (we did not compute paired intervals for this target). The rankings are thus positively related to later headcount growth; because the signal panel contains the past headcount series that the heuristic uses, this result may reflect that series alone.

Size-controlled correlation. The partial Spearman correlation with the target, controlling for baseline traffic, is positive for 12 of 13 models (Figure 5a). Six exceed the heuristic’s $+0.078$, and for four of them the paired 95% interval of the difference excludes zero: GPT-5.6 Terra ($+0.135$), GPT-6 Sol ($+0.132$), Claude Fable 5.1 ($+0.128$) and Claude Opus 5.5 ($+0.116$). We defined this analysis after seeing the raw results and treat it as exploratory.

7 Discussion: what Mento delegates to agents

SAB informs which work Mento’s agents do and which outputs investors review; we did not measure the accuracy or cost of that review. Even the highest-scoring model, Claude Opus 5.5, fails 41 of 960 tasks under the strict rule, among them 16 of the 80 real-world traction tasks, whose reconciled metrics go into an investment memo; in 16 of the 41 it gives a value for a withheld fact. Claude Opus 5.5 and GPT-6 Astra, which answer withheld facts least often, still give a value for 9–10% of them, so every answered field has to be traced to a supporting source. The models’ rankings correlate positively with company size and negatively with our growth target. Because the prompt did not state the target, we cannot separate forecasting errors from a different reading of the request. Mento’s agents therefore gather evidence with citations and screen companies against explicit criteria (the sourcing-screen family, on which every model scores at least 90.5); investors, who carry the fiduciary responsibility, check the sources and make the ranking and investment decisions.

8 Limitations and future work

Each model was run once, zero-shot, at one reasoning effort, so our intervals capture variation across tasks but not across runs. Provider endpoints do not guarantee identical checkpoints or defaults; the Azure endpoints also filter output content, although they blocked no output in the main runs. We have no human-analyst or non-agentic retrieval baseline, and no ranking model fitted on the signal panel. Closed-book runs cover three models from two vendors. The real-world track and the cohorts come from one fund’s data, which we cannot release in full. The key corrections were made after we saw the answers, follow our reading of the prompts and map free-text labels by keywords. An audit that starts from criteria few models answer correctly finds keys that are too strict and misses keys that are too lenient; only the citation rules were checked exhaustively, and that check lowered scores. We do not evaluate investment returns; the ranking target, short-term operating growth, is inversely associated with baseline traffic in these cohorts and is not stated to the model, and cohorts include only companies measurable at every horizon. The models’ training data may overlap the outcome period (November 2025 to June 2026); masking only partly mitigates this, and 2.3% of descriptions still name the company (Appendix G). The synthetic track may carry generator artifacts, and the research tasks are close to the ceiling for the best models. Next we add a forward ranking track [21], a human-analyst baseline, repeated runs and a prompt that states the target.

References

- [1] Affinity. The venture capital benchmark report: 2025 edition. <https://www.affinity.co/report/the-2025-investment-benchmark-report>, 2025.
- [2] Anthropic. Model Context Protocol. <https://modelcontextprotocol.io>, 2024.
- [3] Jae Yoon Bae, Simon Malberg, Joyce Galang, Andre Retterath, and Georg Groh. DIALECTIC: A multi-agent system for startup evaluation. arXiv preprint arXiv:2603.12274, 2026.
- [4] Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.
- [5] Maxime Bonelli. Data-driven investors. *The Review of Financial Studies*, 2025. doi: 10.1093/rfs/hhaf078.
- [6] Lele Cao, Vilhelm von Ehrenheim, Sebastian Krakowski, Xiaoxue Li, and Alexandra Lutz. Using deep learning to find the next unicorn: A practical synthesis. arXiv preprint arXiv:2210.14195, 2022.
- [7] Lele Cao, Gustaf Halvardsson, Andrew McCornack, Vilhelm von Ehrenheim, and Pawel Herman. Beyond gut feel: Using time series transformers to find investment gems. arXiv preprint arXiv:2309.16888, 2023.
- [8] Rick Chen, Joseph Ternasky, Afriyie Samuel Kwesi, Ben Griffin, Aaron Ontoyin Yin, Zakari Salifu, Kelvin Amoaba, Xianling Mu, et al. VCBench: Benchmarking LLMs in venture capital. arXiv preprint arXiv:2509.14448, 2025.
- [9] Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, et al. BrowseComp-Plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.
- [10] Krishani Dhanji. Deloitte to pay money back to Albanese government after using AI in \$440,000 report. The Guardian, 2025. URL <https://www.theguardian.com/australia-news/2025/oct/06/deloitte-to-pay-money-back-to-albanese-government-after-using-ai-in-440000-report>.
- [11] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations (ICLR)*, 2026.
- [12] HuiMing Fan, Xiao Wang, Zheng Chu, Qianyu Wang, Zhuoyao Wang, Ming Liu, Bing Qin, et al. LiveBrowseComp: Are search agents searching, or just verifying what they already know? arXiv preprint arXiv:2605.28721, 2026.
- [13] Paul A. Gompers, Will Gornall, Steven N. Kaplan, and Ilya A. Strebulaev. How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190, 2020. doi: 10.1016/j.jfineco.2019.06.011.
- [14] Boyu Gou, Zanning Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanov, Botao Yu, et al. Mind2Web 2: Evaluating agentic search with agent-as-a-judge. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025.
- [15] Liang Hu, Jianpeng Jiao, Jiashuo Liu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, et al. FinSearchComp: Towards a realistic, expert-level evaluation of financial search and reasoning. arXiv preprint arXiv:2509.13160, 2025.
- [16] Yagiz Ihlamur, Ben Griffin, and Rick Chen. PHBench: A benchmark for predicting startup Series A funding from Product Hunt launch signals. arXiv preprint arXiv:2605.02974, 2026.
- [17] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. arXiv preprint arXiv:2311.11944, 2023.
- [18] Seyed Mohammad Ali Jafari, Ali Mobini Dehkordi, Ehsan Chitsaz, and Yadollah Yaghoobzadeh. Deconstructing the crystal ball: From ad-hoc prediction to principled startup evaluation with the SAISE framework. arXiv preprint arXiv:2508.05491, 2025.
- [19] Sayash Kapoor, Benedikt Stroebel, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2025.

- [20] Sayash Kapoor, Benedikt Stroebel, Peter Kirgis, Nitya Nadgir, Zachary S. Siegel, Boyi Wei, Tianci Xue, Ziru Chen, et al. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *International Conference on Learning Representations (ICLR)*, 2026.
- [21] Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.
- [22] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025.
- [23] Yaxuan Kong, Hoyoung Lee, Yoontae Hwang, Alejandro Lopez-Lira, Bradford Levy, Dhagash Mehta, Qingsong Wen, Chanyeol Choi, et al. Evaluating LLMs in finance requires explicit bias consideration. arXiv preprint arXiv:2602.14233, 2026.
- [24] Sebastian Lehner and Alejandro Lopez-Lira. ChatGPT as a time capsule: The limits of price discovery. arXiv preprint arXiv:2604.21433, 2026.
- [25] Weixian Waylon Li, Hyeonjun Kim, Mihai Cucuringu, and Tiejun Ma. Can LLM-based financial investing strategies outperform the market in long run? arXiv preprint arXiv:2505.07078, 2025.
- [26] Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Profit Mirage: Revisiting information leakage in LLM-based financial agents. arXiv preprint arXiv:2510.07920, 2025.
- [27] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust LLMs’ economic forecasts? arXiv preprint arXiv:2504.14765, 2025.
- [28] Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.
- [29] Grégoire Mialon, Clémentine Fourier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.
- [30] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. arXiv preprint arXiv:2411.00640, 2024.
- [31] Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. arXiv preprint arXiv:2606.08679, 2026.
- [32] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, et al. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4.
- [33] Andre Retterath. Human versus computer: Benchmarking venture capitalists and machine learning algorithms for investment screening. SSRN Working Paper No. 3706119, 2020.
- [34] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024.
- [35] Jennifer Wang, Joachim Baumann, Daniel E. Ho, and Sanmi Koyejo. API benchmark scores do not reliably transfer to chatbot interfaces. arXiv preprint arXiv:2609.08861, 2026.
- [36] Xisen Wang, Yigit Ihlamur, and Fuat Alican. SSFF: Investigating LLM predictive capabilities for startup success through a multi-agent framework with enhanced explainability and performance. arXiv preprint arXiv:2405.19456, 2024.
- [37] Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. arXiv preprint arXiv:2411.04368, 2024.
- [38] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, et al. BrowseComp: A simple yet challenging benchmark for browsing agents. arXiv preprint arXiv:2504.12516, 2025.

- [39] Sichao Xiong, Yigit Ihlamur, Fuat Alican, and Aaron Ontoyin Yin. GPTree: Towards explainable decision-making via LLM-powered decision trees. arXiv preprint arXiv:2411.08257, 2024.
- [40] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ -bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025.
- [41] Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. AssistantBench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.
- [42] Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, et al. Establishing best practices for building rigorous agentic benchmarks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025.

A Comparison with related benchmarks

Table 3 compares SAB with eight of the benchmarks discussed in §2 on the five properties that motivate its design.

Table 3: SAB compared with related benchmarks. ✓ yes, ~ partially, ✗ no, – not applicable (no retrieval). Abstention scored means that declining to answer is scored explicitly; confidence calibration alone does not count.

Benchmark	Domain	Frozen corpus	Tool budget	Citations checked	Abstention scored	Outcome-scored predictions
GAIA [29]	general	✗	✗	✗	✗	✗
BrowseComp [38]	general	✗	✗	✗	✗	✗
BrowseComp-Plus [9]	general	✓	✗	✓	~	✗
Mind2Web 2 [14]	general	✗	✗	✓	✗	✗
FinSearchComp [15]	finance	✗	✗	✗	✗	✗
ForecastBench [21]	forecasting	–	–	–	✗	✓
VCBench [8]	venture	–	–	–	✗	✓
PHBench [16]	venture	–	–	–	✗	✓
SAB (ours)	venture	✓	✓	✓	✓	✓

B Harness, prompts and example tasks

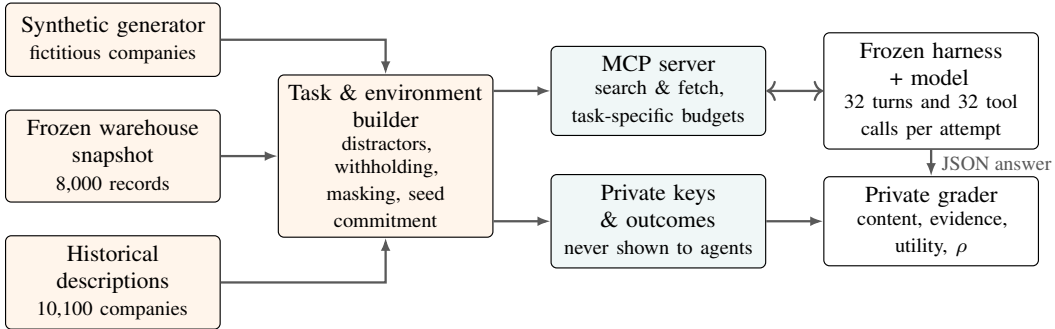


Figure 6: **SAB pipeline.** Tasks, corpora and keys are generated from private sources. In research tasks the agent reaches only its own task’s environment through two tools; ranking tasks deliver the company profiles in the task itself. Grading happens server-side against keys the agent never sees.

System prompt. Every model receives the same system prompt:

You are running a closed startup-analysis benchmark. Use only `benchmark_search` and `benchmark_fetch` for evidence. Never use prior knowledge or invent source ids. Search and fetch are scoped to the current task. When you have enough evidence, return only one JSON object matching the supplied `response_schema`. Cite `source_ids` in the evidence object for every answered field. If evidence does not establish a field, use exactly "unknown" and cite no source for that field. Do not call a submit tool; the frozen harness submits validated JSON.

The closed-book runs use the same prompt, so they measure behavior under an instruction that forbids prior knowledge.

Tools. `benchmark_search(query: string, limit: integer 1-10)` searches the corpus of the current task; `benchmark_fetch(source_id: string)` returns one document. The user turn is the line “Analyze this benchmark task. Return only the final JSON object when done.” followed by the task object serialized as JSON (prompt, instructions, response schema, tool budget, target and identifiers).

Synthetic-task instructions.

Use only sources returned by the benchmark MCP server. Return one JSON object matching response_schema. Copy extracted string values using the wording in the supporting source. Cite source_ids in evidence for every answer field. Do not infer a fact when the sources do not support it. If the sources do not establish a field, answer it with exactly "unknown" and cite no sources for it.

Real-world-task instructions.

Use only sources returned by the benchmark MCP server. Treat the frozen snapshot and metric cutoff as authoritative for this task. Do not convert missing or zero provider coverage into a negative fact. Return one JSON object matching response_schema. Cite source_ids in evidence for every answer field. If the sources do not establish a field, answer it with exactly "unknown" and cite no sources for it.

Example: market & why-now (synthetic).

Estimate the bottom-up serviceable annual market for Juniper Labs 1. Use only the current target segment and calculate $\text{bottom_up_sam_usd} = \text{target_accounts} \times \text{annual_contract_value_usd}$. Also return the two evidence-backed why-now drivers.

The prompt's "serviceable annual market" is quoted verbatim; the field is a serviceable addressable market (SAM). Response fields: market_scope, target_accounts, annual_contract_value_usd, bottom_up_sam_usd, why_now_drivers, evidence.

Example: traction & risks (synthetic). ARR is annual recurring revenue and NRR net revenue retention.

Assess current traction and risk for Beacon Pilot 1. Calculate year-over-year ARR growth from the verified ARR values, use independently verified customers rather than the company's claimed count, and apply these deterministic rules. Strongest signal: revenue_growth when ARR growth is at least 100%; otherwise customer_retention when NRR is at least 120%; otherwise enterprise_adoption when at least 20 customers are independently verified; otherwise product_usage. Main risk: customer_concentration when the largest customer is at least 35% of ARR; otherwise decelerating_growth when ARR growth is at most 25%; otherwise weak_retention when NRR is below 100%; otherwise unverified_claims when the claimed customer count is more than twice the verified count.

Response fields: current_arr_usd, yoy_arr_growth_pct, net_revenue_retention_pct, verified_customer_count, strongest_signal, main_risk, evidence. The environment contains an industry-newsletter ARR figure five months older than the one in the verified metrics registry; both documents carry publication dates, but neither is labeled stale.

Ranking prompt.

Using only the historical descriptions and the signal panel available before the stated as_of date, rank ALL candidate_ids from most to least likely to show the strongest operating momentum over the following 90 and 180 days. Names, domains, and outcomes are intentionally hidden.

The prompt does not state that the target is a weighted sum of within-cohort percentiles of percentage growth in web traffic and headcount.

C Withheld facts, scoring details and key audit

The generator attempts a removal in every fifth task of each family (16 tasks per track–family cell, 192 in total), under the acceptance rule given in §3. Table 4 lists where removals succeeded. In the other cells, no single document could be removed under these constraints, usually because each fact was supported by more than one document.

Table 4: Tasks in which a document was removed, by track and family (out of 80 tasks per cell). Under the v0.7.1 keys the withheld fields of the 16 synthetic team tasks are derivable from the surviving documents (correction 3 below), so 95 tasks retain an unanswerable criterion.

Track	Product	Market	Competition	Team	Traction	Sourcing	Total
Synthetic	16	16	16	16	15	0	79
Real-world	0	0	0	16	16	0	32

Empty lists. The scorer reads a non-list answer, including `unknown`, as an empty list on list-valued criteria. For the four answerable criteria whose expected answer is an empty list (52 of the 76,800 content points, all in real-world team tasks), `unknown` therefore earns full content credit. This adds at most 0.07 points to any model’s research score and to the all-unknown reference.

Key audit. For every track, family and criterion we computed the share of the 14 models’ answers that the released keys score as correct, separately for withheld criteria and for criteria whose key is 0 or none. We read the prompt, schema, documents and answers behind every group in which few or no answers were correct, and behind the families in which the best models failed most often. We counted a group as a defect when its key contradicts the documents or the stated rule, or when the key requires a convention that the prompt does not state; groups in which models fail for reasons the task intends (stale figures, namesakes, near-miss peers, the trap in a decision rule) were left unchanged.

Corrections. Table 5 lists the eight corrections; together they form the v0.7.1 keys used in the main text. (1) The real-world competition prompt states a similarity formula: 40 points for a shared market sub-vertical, 20 for the same customer type, 20 times the Jaccard overlap of the tag values, and 10 each for the same stage and company type. The keys compute the overlap over every tag type in the company record, while the profile documents show only market verticals, sub-verticals, technology types and product types, so in all 80 tasks at least one released similarity score cannot be reproduced from the documents. The keys also expect scorecard entries in the form “Company I 82.500” and break ties alphabetically, and the prompt states neither convention; we accept any tied candidate and a `closest_score` within 0.01 of the recomputed one. (2) In 13 of the 15 synthetic traction tasks with withheld facts, the strongest-signal or main-risk label stays answerable although its stated rule needs the removed metrics or customer document (24 fields, now unanswerable). (3) The withholding rule works on lists of supporting documents and misses some facts that the surviving documents still settle. In 7 real-world traction tasks the surviving domain-health or traffic documents determine the principal risk under the stated priority rule, and in 11 the observation months of the two surviving providers remain known. In synthetic team tasks the identity-risk rule asks whether the search results contain a same-named person, so once that page is removed the answer is `none` (9 fields), and the founders’ backgrounds and the missing capability remain stated in the surviving team page or profiles (14 fields). Under the v0.7.1 keys the synthetic team cell therefore has no withheld criterion, and 95 tasks retain one. (4) The real-world traction prompt states the risk priority but names only `none_observed` among the labels, and models answer “spread ≥ 3 ” or “inaccessible or parked domain”. We map each answer to the label whose keyword appears earliest in the answer text (domain, parked or inaccessible; spread, disagree, divergen- or conflict; both traffic, traffic growth or demand; headcount or team; none), and map domain or industry, technical or engineering, and go-to-market, GTM or sales to the synthetic team’s three capability labels. Provider months are matched by provider name and month, founder backgrounds by their background phrase, and the spread ratio, keyed to three decimals with a tolerance of 0.001, is accepted to two decimals. Every numeric tolerance is applied with a margin of 10^{-9} , so that an answer of 2.14 against a key of 2.135 is not rejected by binary rounding. (5) The real-world traction question names a date (June 2026) and the instructions call the metric cutoff authoritative, but the keys use the latest observations of the two web-traffic panels, and in 77 of 80 tasks one of them postdates that date. We also accept answers that apply the cutoff to every provider: `unknown` for the growth of a provider observed after it, and the spread, provider months and risk label recomputed without that observation (or `unknown` for the risk label when the recomputation leaves it open); each recomputed value earns evidence credit when the answer cites the documents that value rests on, even where it equals the released key’s value. (6) We checked every multi-document citation rule against the facts each criterion needs. Some demanded redundant documents: the team page alone or both founder profiles establish the founders, their backgrounds, the lead, the missing capability and the years of domain experience;

the domain-health page alone settles a domain risk; and peer membership and count follow from the profiles without the metric pages. Others accepted a document that does not establish the answer, so the corrected rules are stricter: peer medians need the peers’ metric pages, a traction decision label needs every document its rule reads, and the closest competitor needs every candidate profile, as the scorecard already did. (7) The real-world instructions forbid turning missing or zero coverage into a negative fact, yet 76 keys expect funding per employee of 0 when total funding is 0, and one expects counts of 0 from an empty team record; `unknown` is also accepted there. In the 27 synthetic team tasks without a domain expert, the key’s lead is `none`, a label the prompt never names and no model gave; `unknown` is also accepted for it. The sources state no duration of domain experience for these founders, so `unknown` is also accepted for the key’s 0 years. (8) The prompts ask for the source’s wording but do not fix the span: “manage site activation and protocol deviation review”, copied from a supporting document, is credited for the key “site activation and protocol deviation review”. An answer that gives an accepted alternative value is graded as correct on that criterion. An accepted `unknown` is scored like `unknown` on a withheld fact (full content credit, evidence credit without a citation, utility +1) but still counts as an abstention in the abstention rates.

Table 5: Corrections in the v0.7.1 keys. Each correction’s effect is the difference between the full v0.7.1 keys and the same keys without it: the number of the 13,440 model–task scores that differ, and the research score and pass rate (percentage points) it adds, averaged over the 14 models. The effects overlap; together the corrections add 6.6 points and 13.7 pass-rate points on average, and the stricter citation rules lower scores.

Correction	Change	Scores changed	Δ score	Δ pass
Competition keys	Similarity recomputed from the tags the profile documents show; any tied candidate accepted; closest score within 0.01; scorecard entries matched by company name (full or without legal suffix) and score (± 0.01)	895	+1.87	+4.1
Removed inputs	Synthetic traction labels whose stated rule needs a withheld document relabeled unanswerable (24 criteria)	159	+0.13	−0.1
Derivable withheld facts	Withheld fields that the surviving documents settle relabeled answerable: real-world risk label and provider months, synthetic identity risk, backgrounds and missing capability (41 criteria)	364	+0.54	+1.4
Unstated formats	Risk and capability labels mapped from free text by keywords; provider-month and background formats; spread ratio to two decimals; tolerances applied without binary rounding error	1,699	+2.83	+7.0
Cutoff reading	Real-world traction answers that exclude provider observations after the stated cutoff month also accepted	175	+0.35	+0.7
Citation rules	Each multi-document rule checked against the facts it needs: more lenient for founder facts (team page or both profiles), a domain risk (domain page) and peer membership (profiles only); stricter for peer medians (metric pages), traction decision labels (every document the rule reads) and the closest competitor (every candidate profile)	2,306	−0.50	−1.0
Absent facts	<code>unknown</code> accepted where the key expects 0 from zero funding or an empty team record, and for the unnamed <code>none</code> lead and 0 years of domain experience	434	+0.56	+1.5
Copied spans	A string answer credited when it contains the expected value and is copied verbatim from a supporting document	597	+1.12	+1.5

Effect. Table 6 compares the released and corrected scores. The released keys reproduce the grading service’s aggregates for every model to within 0.001 points; the corrections raise every model’s research score by 4.8–7.5 points and its pass rate by 5.7–17.6 points, and the two orders of models correlate at Spearman $\rho = 0.97$. Under the corrected keys, the scorer’s lenient string match (token-subset containment), which no reported number uses, would add 0–0.6 points and leave the order unchanged. The correction code will be released with v0.7.1; the corrected keys, like the released ones, stay private.

Table 6: Research score, strict pass rate and two of the most affected cells, under the released (v0.7.0) and corrected (v0.7.1) keys. Averaged over models, the corrections raise real-world competition by 19.9 points, real-world traction by 25.4 and synthetic team (not shown) by 21.3.

Model	Research score		Pass (%)		RW competition		RW traction	
	v0.7.0	v0.7.1	v0.7.0	v0.7.1	v0.7.0	v0.7.1	v0.7.0	v0.7.1
Claude Opus 5.5	90.1	97.2	78	96	63.7	89.9	54.8	92.2
Claude Fable 5.1	90.1	96.8	79	94	62.4	87.9	69.1	92.7
Claude Opus 5	89.2	96.2	76	94	63.4	90.3	65.8	88.4
GPT-6 Sol	88.2	95.6	74	90	63.4	92.9	65.3	92.3
GPT-6 Astra	89.4	95.2	71	88	63.1	93.0	73.2	96.3
Grok 4.6	88.4	94.6	75	90	63.2	94.2	70.2	83.8
GPT-5.6 Sol	86.4	93.6	73	87	56.4	82.1	64.1	90.0
GPT-5.6 Terra	86.2	93.4	73	85	55.5	79.7	69.6	92.4
GPT-5.6 Luna	83.4	90.7	66	82	56.8	74.3	72.3	93.8
DeepSeek-V4-Pro	79.8	86.9	63	76	58.6	73.6	62.1	91.7
Claude Sonnet 5	80.1	85.5	68	75	33.5	52.7	65.5	89.8
GPT-6 Luna	77.8	85.4	63	76	22.9	28.0	51.6	91.0
DeepSeek-V3.2	77.0	83.4	63	75	31.7	34.5	47.5	72.5
Claude Haiku 4.5	75.0	79.8	61	67	32.2	31.8	48.3	67.7

D Model settings and run log

Settings. Anthropic models run through the Messages API with adaptive thinking, effort high, temperature 1 and a cap of 16,384 output tokens per turn; our gateway does not pass the effort setting to Claude Haiku 4.5, which instead uses extended thinking with a 4,096-token budget. OpenAI models run through the Codex Responses API with `reasoning_effort=high`; this endpoint accepts neither a temperature nor an output cap. Grok and DeepSeek models run through Azure AI Foundry’s chat-completions API with `reasoning_effort=high`, temperature 0 and a 16,384-token output cap.

Interruptions and restarts. An interruption is a request that still fails after the harness’s retries, or a conversation that reaches the turn or tool-call limit; the interrupted task is restarted in a new conversation when the run resumes. Restarts count tasks and interruptions count events: one task can be interrupted several times (one Claude Haiku 4.5 task was interrupted 24 times by rate limits). A task in flight when a run was stopped and resumed is restarted without a recorded interruption. The GPT-5.6 runs used the previous harness revision (f2), which differs from f3 in retry policy, telemetry, client headers and task concurrency; the task prompt and tool contract are the same. Restarts matter mainly for DeepSeek-V3.2, most of whose 61 restarted research tasks were interrupted by its own repeated tool calls after its budget was exhausted: on the other 899 tasks it scores 87.3 instead of 83.4. Because the excluded tasks are exactly those it failed to finish, this is a mean over a selected subset and should not be read as a corrected score. For Grok 4.6, whose 19 restarts were caused by provider failures, excluding the restarted tasks raises its score by 0.1 points and leaves its rank unchanged.

Run log. Table 7 lists, per model, the endpoint, the harness revision and the number of tasks run concurrently (each task has its own conversation). For research tasks it gives the mean number of model turns, the total number of invalid final answers returned to the model, the number of restarted tasks and the interruptions by type; the last column counts interruptions on ranking tasks. Wall-clock time per run ranged from 2 to 177 hours, including pauses for rate limits and the time between a stop and a resume. DeepSeek-V3.2 kept calling the fetch tool after its budget was exhausted and recorded 57 turn- or tool-limit interruptions on research tasks; Grok 4.6 recorded 19 provider interruptions on research tasks (empty responses and rejections) and 28 on ranking tasks (27 of them timeouts); seven models were never interrupted on research tasks. Invalid final answers ranged from none for four OpenAI models to 955 across the 960 tasks of Claude Haiku 4.5. A fourth closed-book run (GPT-5.6 Luna) repeatedly hit the tool-call limit on empty searches; we stopped it after 920 of its 1,061 research and ranking tasks and do not report it.

Table 7: Endpoints and run log. f2 and f3 are revisions of the frozen harness; Workers is the number of tasks run concurrently.

Model	Endpoint	Harness	Workers	Research tasks				Ranking interr.
				Turns/ task	Invalid answers	Restarted	Interruptions	
Claude Opus 5.5	Anthropic Messages	f3	1	5.0	77	0	–	–
Claude Fable 5.1	Anthropic Messages	f3	1	4.5	593	3	timeout 1	1
Claude Opus 5	Anthropic Messages	f3	2	5.0	200	0	–	–
GPT-6 Sol	OpenAI Codex Responses	f3	4	5.9	0	0	–	–
GPT-6 Astra	OpenAI Codex Responses	f3	4	4.5	0	3	empty response 3	–
Grok 4.6	Azure AI Foundry	f3	4	5.7	147	19	empty response 11, provider rejection 8	28
GPT-5.6 Sol	OpenAI Codex Responses	f2	1	5.2	0	5	timeout 4	2
GPT-5.6 Terra	OpenAI Codex Responses	f2	1	4.9	3	0	–	–
GPT-5.6 Luna	OpenAI Codex Responses	f2	1	5.3	0	1	–	–
DeepSeek-V4-Pro	Azure AI Foundry	f3	4	7.9	795	6	provider rejection 12, turn/tool limit 2	–
Claude Sonnet 5	Anthropic Messages	f3	2	5.9	614	0	–	–
GPT-6 Luna	OpenAI Codex Responses	f3	4	5.2	1	0	–	–
DeepSeek-V3.2	Azure AI Foundry	f3	4	17.6	3	61	turn/tool limit 57, provider rejection 4	8
Claude Haiku 4.5	Anthropic Messages	f3	2	6.6	955	2	rate limit 24	–

E Pairwise comparisons

Table 8 compares each model with the next one in Table 2. Δ is the mean per-task score difference, the interval is the unadjusted paired stratified-bootstrap interval, and p comes from a two-sided sign-flip permutation test with 20,000 draws; p_{Holm} is adjusted over the 13 comparisons. With Holm correction over all 91 pairs (5,000 draws per pair), the same three adjacent differences remain significant ($p_{\text{Holm}} \leq 0.02$; with 5,000 draws the smallest resolvable raw p is $1/5,001$), and 76 of the 91 pairs differ significantly. As a rough guide, the per-task paired standard deviation is 17.5 points on average, so an unadjusted test at $\alpha = 0.05$ has 80% power for a difference of about 1.6 points.

Table 8: Comparisons between adjacent models.

Higher	Lower	Δ	95% CI	p	p_{Holm}
Claude Opus 5.5	Claude Fable 5.1	0.41	[−0.02, 0.87]	0.0767	0.4605
Claude Fable 5.1	Claude Opus 5	0.63	[0.13, 1.12]	0.0159	0.1590
Claude Opus 5	GPT-6 Sol	0.63	[0.06, 1.25]	0.0484	0.3876
GPT-6 Sol	GPT-6 Astra	0.40	[−0.18, 0.94]	0.2054	0.8216
GPT-6 Astra	Grok 4.6	0.58	[−0.19, 1.39]	0.2087	0.8216
Grok 4.6	GPT-5.6 Sol	0.97	[−0.14, 2.08]	0.0956	0.4780
GPT-5.6 Sol	GPT-5.6 Terra	0.27	[−0.87, 1.41]	0.6490	1.0000
GPT-5.6 Terra	GPT-5.6 Luna	2.67	[1.63, 3.68]	<0.0001	0.0006
GPT-5.6 Luna	DeepSeek-V4-Pro	3.76	[2.51, 4.99]	<0.0001	0.0006
DeepSeek-V4-Pro	Claude Sonnet 5	1.46	[0.11, 2.77]	0.0507	0.3876
Claude Sonnet 5	GPT-6 Luna	0.08	[−1.47, 1.68]	0.9245	1.0000
GPT-6 Luna	DeepSeek-V3.2	1.97	[0.28, 3.66]	0.0272	0.2452
DeepSeek-V3.2	Claude Haiku 4.5	3.59	[2.08, 5.11]	<0.0001	0.0006

F Cost accounting

Table 9 lists each provider’s public standard short-context list price as of September 2026. Output includes reasoning tokens, which every provider bills at the output rate. The GPT-6 endpoints report reasoning tokens separately but already include them in the output count; the Grok endpoint reports them separately and excludes them from output, so for Grok 4.6 we add them. Anthropic cache writes use the five-minute rate; the OpenAI and Azure endpoints do not report cache writes, so any such fees are omitted. The GPT-5.6 runs logged no cached-input counts, so their costs may be overstated if inputs were cached; for DeepSeek-V3.2 the input and cached-input prices are equal, so its missing count does not matter. Run totals cover research and ranking tasks (Grok’s ranking run is incomplete) and all restarted attempts; the research cost per task counts only tokens logged on research tasks, divided by 960.

Table 9: List prices in USD per million tokens. For OpenAI and Azure, the cache-read column is the cached-input price.

Provider	Model	Input	Cache read	Cache write	Output
Anthropic	Claude Fable 5.1	10	0.25	12.50	50
	Claude Opus 5.5	4	0.20	5	20
	Claude Opus 5	5	0.50	6.25	25
	Claude Sonnet 5	2	0.20	2.50	10
	Claude Haiku 4.5	1	0.10	1.25	5
OpenAI	GPT-6 Astra	10	1	–	50
	GPT-6 Sol	2	0.20	–	10
	GPT-6 Luna	0.10	0.01	–	0.50
	GPT-5.6 Sol	4	0.40	–	20
	GPT-5.6 Terra	2	0.20	–	12
	GPT-5.6 Luna	0.20	0.02	–	1.20
Azure AI Foundry	Grok 4.6	2	0.50	–	6
	DeepSeek-V4-Pro	1.74	0.145	–	3.48
	DeepSeek-V3.2	0.58	0.58	–	1.68

Table 10: Logged token usage and list-price cost per main run. For OpenAI and Azure models, Cache read is part of Input; for Anthropic models the counters are separate. —: counter not reported by the endpoint, or not logged by harness revision f2 (GPT-5.6). [†]Reported separately by the Grok and GPT-6 endpoints; billed in addition to output for Grok 4.6 and already included in output for GPT-6. The total sums unrounded run costs.

Model	Input	Cache read	Cache write	Output	Reasoning [†]	Run (\$)	Research \$/task
		million tokens, logged over the main run					
Claude Opus 5.5	15.0	17.8	12.8	1.36	—	155	0.113
Claude Fable 5.1	16.6	19.0	16.2	2.22	—	484	0.359
Claude Opus 5	16.7	20.2	15.2	2.31	—	246	0.180
GPT-6 Sol	28.5	3.0	—	1.08	0.52	62	0.051
GPT-6 Astra	20.2	0.3	—	0.62	0.11	230	0.166
Grok 4.6	55.5	13.8	—	0.64	3.66	116	0.112
GPT-5.6 Sol	30.2	—	—	2.06	—	162	0.131
GPT-5.6 Terra	26.0	—	—	1.38	—	69	0.055
GPT-5.6 Luna	38.4	—	—	2.87	—	11	0.008
DeepSeek-V4-Pro	77.3	31.3	—	4.31	—	100	0.079
Claude Sonnet 5	17.8	30.6	16.6	4.64	—	130	0.083
GPT-6 Luna	37.7	4.0	—	1.23	0.55	4	0.003
DeepSeek-V3.2	216.6	—	—	6.53	—	137	0.107
Claude Haiku 4.5	18.3	30.6	29.3	5.53	—	86	0.065
Total						1,990	

G Ranking details

Cohort selection. In each monthly wave, 15,577–17,475 companies met the baseline criteria (privately held, founded 2018–2025, 2–200 employees, baseline monthly web-traffic estimate between 10 and 5,000,000). Of these, 9,776–12,152 (63–72%) had positive traffic and headcount at the baseline and at both horizons, and 2,000 were sampled into 20 cohorts of 100. The recent cohort (baseline December 2025) was sampled from companies whose descriptions first appeared after 1 June 2025. Because inclusion depends on later measurements, the cohorts exclude companies that stopped reporting or shut down.

Target and baselines. For each cohort and horizon, every company receives its percentile within the cohort for percentage growth in web traffic and in headcount; the target is 0.625 times the first plus 0.375 times the second. The previous release weighted traffic, headcount and funding growth at 0.5, 0.3 and 0.2. An audit of the 459 positive funding labels in that release confirmed 47 (10%) from independent sources and contradicted none, so funding growth was dropped and the other two weights renormalized. The heuristic computes headcount growth from a company’s earliest stored month before the baseline to the baseline: six months when that month is stored, otherwise three. A company whose earliest month has no positive headcount, or that has no earlier month, is placed after every company with a signal, and ties are broken by candidate ID. The traffic-growth rule applies the same procedure to web traffic, and the smallest-first rule orders companies by increasing baseline web traffic; both were defined after the raw results were seen. P@10 is the share of a ranking’s first ten companies that are among the cohort’s ten highest-target companies, averaged over horizons; a random order has expected P@10 of 0.1. The 95% range of random rankings in Figure 1b comes from 2,000 Monte Carlo draws, each permuting the companies of every cohort before averaging ρ .

Sensitivity to the target weights. Table 11 re-scores the submitted rankings under other weights. Averaged over horizons and cohorts, traffic growth and headcount growth correlate with baseline traffic at $\rho = -0.32$ and $+0.06$, respectively; the release target correlates with baseline traffic at -0.21 at 90 days and -0.27 at 180 days, and with baseline headcount at $+0.01$.

Table 11: Mean ranking ρ under different weights for traffic and headcount growth, and precision among the first ten companies (P@10) for the release target. The first column is the target used throughout the paper.

	Mean ρ , traffic / headcount weight				P@10
	0.625 / 0.375	0.5 / 0.5	1 / 0	0 / 1	
Claude Opus 5.5	−0.062	−0.009	−0.153	+0.140	0.113
Claude Fable 5.1	−0.073	−0.020	−0.163	+0.134	0.100
Claude Opus 5	−0.113	−0.050	−0.214	+0.143	0.097
GPT-6 Sol	−0.059	−0.010	−0.143	+0.129	0.101
GPT-6 Astra	−0.045	+0.008	−0.139	+0.152	0.110
GPT-5.6 Sol	−0.093	−0.032	−0.196	+0.153	0.105
GPT-5.6 Terra	−0.003	+0.039	−0.080	+0.133	0.114
GPT-5.6 Luna	−0.121	−0.064	−0.209	+0.120	0.086
DeepSeek-V4-Pro	−0.138	−0.079	−0.227	+0.115	0.081
Claude Sonnet 5	−0.160	−0.099	−0.252	+0.114	0.082
GPT-6 Luna	−0.063	−0.026	−0.122	+0.086	0.096
DeepSeek-V3.2	−0.092	−0.052	−0.149	+0.075	0.070
Claude Haiku 4.5	−0.098	−0.049	−0.173	+0.104	0.066
Headcount-growth heuristic	+0.058	+0.079	+0.008	+0.105	0.139
Smallest-first rule	+0.242	+0.179	+0.320	−0.064	0.311
Traffic-growth rule	−0.173	−0.134	−0.221	+0.027	0.064
Random order (expected)	0	0	0	0	0.100

Size-controlled correlation. For each cohort and horizon we compute the partial Spearman correlation between a ranking and the target, controlling for baseline web traffic: $\rho_{xy \cdot z} = (\rho_{xy} - \rho_{xz}\rho_{yz}) / \sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}$, where x is the model’s preference (higher for companies ranked earlier), y the target and z baseline traffic. We average the two horizons within each cohort and bootstrap cohorts with the same draws for every model and the heuristic. As §6 notes, this analysis was defined after the raw results were seen.

By wave. Table 12 reports the raw ρ for each wave and for the recent cohort, and the size-controlled ρ over all cohorts.

Table 12: Ranking ρ by baseline wave (20 cohorts each), for the single recent cohort, and size-controlled ρ over all 101 cohorts.

Model Cohorts	2025-08 20	2025-09 20	2025-10 20	2025-11 20	2025-12 20	Recent 1	Size-controlled 101
Claude Opus 5.5	-0.015	-0.048	-0.082	-0.084	-0.087	+0.021	+0.116
Claude Fable 5.1	-0.033	-0.048	-0.100	-0.102	-0.086	-0.024	+0.128
Claude Opus 5	-0.066	-0.098	-0.156	-0.132	-0.117	-0.025	+0.048
GPT-6 Sol	+0.010	-0.028	-0.087	-0.106	-0.087	-0.017	+0.132
GPT-6 Astra	+0.017	-0.038	-0.097	-0.073	-0.038	+0.045	+0.090
GPT-5.6 Sol	-0.021	-0.087	-0.135	-0.120	-0.111	+0.040	+0.060
GPT-5.6 Terra	+0.046	-0.007	-0.018	-0.027	-0.014	+0.131	+0.135
GPT-5.6 Luna	-0.076	-0.110	-0.122	-0.172	-0.129	+0.007	+0.072
DeepSeek-V4-Pro	-0.073	-0.132	-0.170	-0.167	-0.154	-0.056	+0.013
Claude Sonnet 5	-0.129	-0.148	-0.211	-0.163	-0.145	-0.191	-0.084
GPT-6 Luna	-0.027	-0.060	-0.063	-0.077	-0.087	-0.074	+0.086
DeepSeek-V3.2	-0.076	-0.062	-0.086	-0.147	-0.094	+0.034	+0.062
Claude Haiku 4.5	-0.054	-0.116	-0.116	-0.119	-0.089	+0.012	+0.048
Heuristic	+0.084	+0.046	+0.006	+0.068	+0.081	+0.160	+0.078

Leakage audit. We flag a company as leaked if its legal name (without corporate suffixes, at least four characters) or the root of its domain (at least five characters) appears as a whole word in its masked description. This flags 230 of the 10,100 companies. Their mean composite target score is 0.51, against 0.50 for the rest. Every model places them later in its ranking, by 0.06–0.14 of the list length on average, and re-scoring every ranking without them, while keeping the original composite target scores, changes each model’s ρ by less than 0.006.

H Pilot runs

Before the main study, three Claude models were run with the previous harness revision, a 4,096-token output cap and provider-default reasoning settings. Under the corrected keys they scored 94.5 (Opus 5), 82.5 (Sonnet 5) and 78.9 (Haiku 4.5), against 96.2, 85.5 and 79.8 in the main study. The pilots are not used elsewhere in the paper; the differences combine the change in reasoning settings with the harness revision and the output cap.

I Data and ethics

Personal data. Real-world team tasks count people and roles from company records and expose no biographies; ranking descriptions are masked, with the leakage noted above. No human subjects were involved.

Use of language models. LLM coding and writing assistants were used to implement the evaluation harness and analysis scripts and to draft parts of this paper. All numbers were produced by the analysis code and checked by the authors. Our per-task recomputation under the released keys matches the grading service’s aggregates, which checks scoring consistency but not the source data; the corrected keys (Appendix C) are our own and have not been audited independently.

Broader impact. Better evaluation of sourcing agents can widen the set of companies a fund reviews. The ranking results show a risk as well: agents that favor larger companies could, without oversight, steer venture funding toward the larger startups in a cohort.