Beating the Median, Not the Best: Machine Learning versus Venture Capitalists in Startup Selection

Anatole Chapelier

November 2026

Abstract

Can a statistical screen trained only on public, point-in-time information match reconstructed venture-capital first-round books? We assemble a dataset of 305,010 companies from Y Combinator, historical Crunchbase snapshots, SEC Form D, Fundraising Fox, Field–Ritter IPO dates, and Wikidata, apply 448 documented outcome-date corrections whose effect on the primary tests is under 0.6 percentage points, and run a pre-declared walk-forward backtest. After the corrections, a histogram gradient-boosting model with enriched point-in-time features beats the median VC book on three-year follow-on financing and on five-year IPO-or-acquisition exits, and beats the 75th-percentile book on follow-on, but not on exits (26.96% versus 32.14%). The hit-rate edge does not become a robust multiple on estimated entry cost. Two further exercises qualify the claim. Parsed schedules of investments from 1,259 SEC filings of 20 public funds yield 591 fund–company positions; among 48 early-stage VC positions that subsequently leave the schedule, the median last-mark multiple is 0.67×, with the important caveat that disappearance from a filing is not realized proceeds. Six large language models then score anonymized first-round cards on 7,200 walk-forward deals. Claude Sonnet 5.5 and GPT-6.1 Sol post high top-k hit rates, but a recognition test shows that much of that edge comes from firms the model can name, no LLM significantly beats the gradient-boosting model after Holm or Benjamini–Hochberg correction, and post-2015 and post-cutoff AUCs are near chance. On a strongly de-identified rerun, where no model named any of 500 test firms, no LLM beats either VC book or the gradient-boosting model after Holm or Benjamini–Hochberg correction. The supported conclusion is narrow: public-data models can beat the *median* reconstructed investor as a screen. They do not beat the best books on five-year exits, and they are not a net-of-fee TVPI result.

Keywords: venture capital, startup screening, walk-forward backtest, multiple testing, large language models, SEC schedules of investments.

1 Introduction

A recurring claim in practitioner and academic work is that machine-learning models, trained on information available at the date of a financing, can screen startups at least as well as a typical venture capitalist. The claim is easy to overstate. Most of the literature measures success as a follow-on round, an IPO, or an acquisition, not as a fund's net IRR or TVPI after fees, carry, entry price, and dilution [3, 7]. Models are usually allowed to pick any company in a file, whereas real managers face a two-sided matching market in which access to late rounds and favoured syndicates is itself a source of persistence [11]. Look-ahead and survivorship can inflate

accuracy statistics [9, 22]. And when funds themselves adopt data tools, they may become better at selecting historically familiar firms while becoming *less* likely to finance rare, breakthrough exits [5].

This paper reports a public-data test designed around those caveats. The unit of observation is a company's first observed financing ("first event"). Features are restricted to records dated on or before that event. Training uses only earlier vintages whose outcome window has already closed (walk-forward). The primary outcomes are a follow-on financing within three years and an IPO or acquisition within five years; right-censored or undated outcomes are left as missing, not as zeros. The comparators are the vintage base rate, a permuted-label placebo, and the median and 75th-percentile reconstructed first-round books of investors who were active in the same year. A pre-declared primary specification—histogram gradient boosting (HGB) with the `v3_full` point-in-time feature set, all stages, top-50 equal-weighted portfolios, six one-sided tests against the two VC books—was written down before any v3 result was viewed (v1 and v2 results had already been seen). After an audit of 448 field-level corrections on 134 companies, the six primary conclusions are unchanged: every shift is under 0.6 percentage points.

The contributions are fourfold.

1. **A documented public dataset and a corrections audit.** We combine free and legally reusable sources into canonical company, round, investment, and people tables (305,010 companies; 346,409 rounds; 255,610 investments) and publish a 448-row change log. The primary-spec impact of the corrections is small and is reported in full.

2. **A pre-declared walk-forward comparison with reconstructed VC books.** On corrected data, the primary model averages a 68.7% three-year follow-on hit rate in 2003–2009 vintages (95% CI 66.6–70.9) against a 49.0% median VC book and a 64.2% 75th-percentile book, and a 26.96% five-year exit hit rate in 2005–2009 against 19.3% and 32.1%. Five of the six primary tests survive Holm correction under a joint bootstrap that also resamples VC books; the exception is the exit comparison with the 75th-percentile book.

3. **Direct evidence from SEC schedules of investments,** not from marketing materials. From 1,259 filings of 20 public business-development companies and registered funds we parse a position-level panel (31,962 security-level rows from 18 funds; 591 fund–company summary rows after reconciliation). Early-stage VC last-mark multiples on positions that then leave the schedule have a median of 0.67×. These are marks, not cash proceeds.

4. **A six-model "LLM as investor" benchmark** on anonymized, point-in-time cards, with a recognition test, a post-cutoff short-horizon cohort, and LLM-as-feature stacking. Official OpenRouter usage was $37.394 under a $45 cap. No same-sample LLM-versus-GBM hit-rate test survives Holm or Benjamini–Hochberg correction at 5%. Recognition of withheld names accounts for a large part of the apparent top-k advantage of the strongest frontier models.

We do not claim that an algorithm beats top-quartile funds on net TVPI, that the model could have obtained allocation in the same rounds, or that large language models are reliable autonomous investors. Section 2 reviews the literature we rely on. Section 3 describes the data and the corrections audit. Section 4 sets out the design. Sections 5–7 report the backtest, the fund-schedule and memo evidence, and the LLM benchmark. Sections 8–10 discuss uses, limits, and conclusions.

2 Related work

Direct comparisons of models with investors are still few. Retterath [15] gave 111 European VC professionals and an XGBoost screen the same one-page company descriptions and found that the algorithm outperformed the median (average) participant by 25% (29%) at a five-out-of-ten screening task. That is a screening result on a common, thin information set, not a fund-return result. Blohm et al. [4] compared an XGBoost survival model with 255 angels on a European platform (623 pre-screened opportunities, 51 realised investments) and reported a mean algorithm IRR of 7.26% against 2.56% for the angels; experienced angels who suppressed loss aversion and overconfidence still beat the model. The IRR used platform NAVs and the deals had already been staff-screened. Lyu et al. [9] trained a graph neural network on PitchBook data and reported AP@10 of 20.8% versus 10.7% for the investments of real investors, with success defined as an IPO or acquisition within five years of the first round. A later journal version appears in *ACM Transactions on Intelligent Systems and Technology* in 2025 (DOI 10.1145/3763001). Bissoto [3], in a walk-forward study of 55,858 U.S. Crunchbase rounds, finds that ML portfolios outperform most U.S. VCs on IPO hit rates—gains of about 14 percentage points for logit and up to almost 40 points for tree models—but that the within-investor benefit mean-reverts and fades after ML libraries became cheap in 2015–2017. The paper is still a working paper, and the counterfactual assumes access to any deal in the vintage.

A complementary, and cautionary, result is Bonelli [5]. When VC firms become "data-driven" (proxied by hiring data staff), they improve at selecting historically familiar startups—follow-on rates rise by 2.5 percentage points overall and by 4.2 points among backward-similar firms—but they become 2.5 points *less* likely to finance a startup that later IPOs or completes a profitable acquisition, relative to a 15% base. The object is the effect of tools adopted by funds, not a pure model backtest, and it is the right warning for our own claim: a screen that beats the median book on follow-on can still miss the rare exits that dominate fund returns.

Methodological papers that do not always compare models with named investors still set the standard we try to meet. Żbikowski and Antosiuk [22] and Arroyo et al. [1] emphasise time-aware splits and the removal of post-decision features on Crunchbase. Sharchilev et al. [18] add web mentions for seed follow-on prediction. High headline accuracies, such as the out-of-sample figures reported by Ross et al. [16], are known to be sensitive to look-ahead and class balance; Krishna et al. [8] is a canonical example of treating later financing stages as features. Xiang et al. [21] is an early demonstration that news text can help predict acquisitions, with the later caveat that full-sample cross-validation leaks the future [9]. Potanin et al. [14] report a walk-forward Series B/C simulation; entry is already late, and the advertised capital growth is not a net fund TVPI. Setty et al. [17] stress cost-sensitive errors rather than accuracy, on an Israeli sample. Bae et al. [2] evaluate a multi-agent LLM system on opportunities from five funds and report precision comparable to the humans; sample size is small, and memorisation of named firms is a recognised risk—a risk we test directly in Section 7.

Industry systems (EQT Ventures' Motherbrain; Correlation Ventures' quantitative co-investment model, described in a 2012 profile; firm-reported figures around Vela's VCBench) show that production screens exist, but they do not provide an independently audited comparison with vintage net TVPI [10, 20]. Official fund benchmarks remain the reference for that question: Cambridge Associates' U.S. VC index and PitchBook's global vintage tables report net-of-fee IRR and TVPI by vintage [6, 12, 13]. Harris et al. [7] document how strongly those outcomes vary by vintage, which is why we never pool years into a single random split. We cannot compute

Table 1: Analysis tables after the outcome-date corrections of 14 November 2026.

Table	Rows	Notes
companies	305,010	Canonical entities; 150,357 dated first events
rounds	346,409	146,482 rounds in 2016–2026
investments	255,610	Investor names mainly through December 2015, plus YC
investors	53,933	Includes the Fundraising Fox firm directory
company_people	605,159	Crunchbase 2013 relationships plus Form D related persons
entity_matches (parquet)	169,418	Full cross-source match table
formd_offerings	104,717	Startup-like Form D offerings, latest amendment
IPOs (dated)	6,813 (4,961)	Field–Ritter supplies most new offer dates
Acquisitions (dated)	15,082 (14,348)	Post-2015 M&A remains thin
Shutdowns (dated)	8,289 (3,060)	Dates still sparse
Follow-on / exits	58,041 / 21,150	Exit = IPO or acquisition
Unicorn flag	542	Wikipedia list; not used as a feature

Counts are from the corrected `companies.parquet` and sibling tables. The CSV `entity_matches.csv` (81,941 rows) is a stale subset and should not be used; the parquet file is complete. First events in 2016–2026: 68,211.

a comparable net TVPI from the public files used here, and we do not pretend to.

3 Data

3.1 Sources and coverage

The company file was refreshed on 14 November 2026 from free and legally reusable sources only. Nothing in it was fabricated. The core combines the Y Combinator public directory (via the community yc-oss API), the Crunchbase 2013 Snapshot, the 4 December 2015 Crunchbase export (CC BY-NC), and SEC Form D offerings from 2008 Q1 through 2026 Q3, including Form D/A amendments and `RELATEDPERSONS` [19]. Post-2015 funding and outcome sources are the Fundraising Fox Open Dataset v2026-11 (CC BY 4.0; SEC- and press-verified rounds, without co-investor edges), Field–Ritter IPO offer dates through 23 December 2025, Wikidata listings and dated acquisitions or dissolutions, and the Wikipedia unicorn list (CC BY-SA). OpenVC was not used. A planned EDGAR S-1/F-1/424B4 scrape was initially blocked and is not invented here; prospectus-level valuations, where present, live in a separate valuations layer that is not required for the primary hit-rate tests.

Table 1 summarises the corrected tables. Source flags on companies include YC (6,397), Crunchbase 2015 (66,236), Crunchbase 2013 (195,333), Form D (52,596), Fundraising Fox (71,949), Field–Ritter (4,227), Wikidata (1,487), and Wikipedia unicorns (542). After December 2015, Crunchbase investor-level edges end: a multi-VC book can be rebuilt consistently only through the Crunchbase window (and for YC thereafter). Post-2015 follow-on is observable for Form D, Fundraising Fox, and YC companies; post-2015 M&A is not a census (dated acquisitions in 2016–2026 are few) and missing acquisitions are coded as unknown, not as failures.

Entity matching uses unique keys only—domains, permalinks, slugs, exact unique names, and name-plus-state—with no fuzzy matching. Names that collide across several company identifiers are left unmatched. A short data-dictionary summary is in Appendix A; the full column-level dictionary is the project's `DATA_DICTIONARY.md`.

3.2 Point-in-time construction

The decision date is the first funding event in `rounds`. Horizon labels (`follow_on_3y`, `exit_5y`, and the unused-in-primary `ipo_7y` and `acquired_5y`) are 0/1 or null. A label is null when the horizon would run past that company's source-specific observer wall, or when an exit is known but undated, so that a mechanical zero would be misleading. Features used in the backtest are records dated on or before the first event: sector and geography, first-round amount and type, syndicate size and investors' *past* deal and exit counts as of that date, founder counts and dated prior roles, Form D related-person counts from filings on or shortly after the event, trailing market counts, and—in secondary specifications only—walk-forward text scores. Current team size, current directory status, and unicorn flags are never used as features. Crunchbase categories and some undated founder titles can still be edited after the fact; that residual look-ahead is listed as a limitation, not assumed away.

3.3 Corrections audit

An audit of outcome dates produced 448 field-level changes on 134 companies, each with a rule, a reason, and a source URL (`processed/corrections/corrections.csv`; script `04_apply_corrections.py`). The pre-correction companies table is backed up. Before writing, the label logic was re-checked on all 305,010 rows (zero discrepancies) and then recomputed for touched rows. The rules were: three manual fixes (Rocket Lab's 2020 "acquisition" by Rocket Lab USA was an internal reorganisation and was removed, and its IPO date was set to the 2021-08-25 Nasdaq listing; Instacart's 2023-01-01 placeholder was replaced by the 2023-09-19 Field–Ritter offer date; Atlassian's 1980-10-07 date, a reused-ticker match, was replaced by 2015-12-09); 100 Ritter reused-ticker conflicts, of which 57 dates were replaced by a consistent similar-named row and 43 listings were left as undated positives; 15 January-1 placeholders replaced by a same-ticker Ritter offer date within one year; 16 self-acquisitions (acquirer name equals the company's own name) removed; and 217 derived-field updates on 131 companies.

A further 2,390 rows were flagged and *not* changed, for lack of an alternative source (exits dated before the first event, January-1 acquisition dates, pre-1990 IPOs with much later first events, and related inconsistencies). A companion file lists 57 acquirer-name variants for manual review; many appear to be genuine deals in which the acquirer later took the target's name. Unresolved flags are not silently "fixed."

Of the 134 corrected companies, 98 sit in the v3 universe (131,204 companies). Inside the test windows, `follow_on_3y` did not change; `exit_5y` changed for five companies in 2005–2009 and eight in 2010–2015. Some features also moved (trailing acquisition counts, and investor past-exit features on about 4,000 rows). Table 2 reports the primary-spec impact. VC benchmark values are unchanged to six decimals. Every change in the model mean is under 0.6 percentage points, and all six primary conclusions—including which tests survive Holm—are unchanged. Unless a sentence says otherwise, headline backtest numbers below are the *post-correction* primary run. The three-model v3 grid, ablations, leakage checks, and the permuted-label placebo were subsequently re-run on the corrected files (Section 5); a standalone v2 backtest was not repeated.

Table 2: Impact of the 448 corrections on the pre-declared primary specification (HGB, `v3_full`, all stages, top-50, seeds 0–4).

Test	Model hit		Joint Δ		Holm p	Survives
	pre	post	pre	post	post	
FO 2003–09 vs median VC	68.11	68.69	+19.15	+19.72	0.0006	yes
FO 2003–09 vs p75 VC	68.11	68.69	+3.92	+4.49	0.0014	yes
FO 2010–15 vs median VC	87.07	86.80	+21.77	+21.50	0.0006	yes
FO 2010–15 vs p75 VC	87.07	86.80	+8.11	+7.84	0.0006	yes
EX 2005–09 vs median VC	26.72	26.96	+7.39	+7.63	0.0006	yes
EX 2005–09 vs p75 VC	26.72	26.96	−5.42	−5.18	0.9988	no

Hit rates in percent. Joint Δ is the mean difference versus the VC book, in percentage points, with VC books also resampled (1,000 replicates). Holm p is within the six-test primary family. Source: `processed/corrections/impact_primary*.csv` and `backtest_v3_postcorr/results/`.

4 Methodology

4.1 Walk-forward design and outcomes

For each vintage year T, the model is trained only on first events from earlier years whose outcome horizon has already elapsed by T, and is tested on first events in year T. Features are point-in-time as of the test company's first event. The two primary outcomes are:

- `follow_on_3y`: a later funding event more than 45 days after the first event, and within three years;
- `exit_5y`: a dated IPO or acquisition within five years. The label is 1 if either event succeeds inside the window, 0 only if both can be observed as 0, and null otherwise.

The primary windows with reconstructable multi-VC books are follow-on vintages 2003–2009 and 2010–2015, and exit vintages 2005–2009. A 2016–2023 follow-on track on Form D, Fundraising Fox, and YC companies is reported as a secondary, company-feature track: there is no reconstructed multi-VC book in the open post-2015 files, so the comparators are the vintage base rate and the YC book.

4.2 Models and portfolios

The pre-declared primary model is scikit-learn's `HistGradientBoostingClassifier`, averaged over seeds 0–4. The primary feature set is `v3_full`: the v2 company and syndicate features plus point-in-time team features from Crunchbase 2013 and Form D, round-context features, investor track records as of the decision date, and geography/macro controls. Text scores are *excluded* from the primary set because descriptions are scraped after the decision date. Secondary cells, all pre-declared as exploratory relative to v3, include logistic regression and random forests, company-only and text-augmented sets, stage slices, and top-20 portfolios.

Headline k is a fixed 20 or 50 companies per vintage (primary: 50), equal-weighted. That is more stable than matching the median active VC's book size of about six or seven first-round deals. For the LLM benchmark, k is scaled so that the sample portfolio is the same quantile as the v3 top-50 in the full universe.

4.3 VC books, base rate, and placebo

In each vintage, an active VC is an investor with at least five first-round deals that year in the file. The VC book hit rate is the share of that investor's first-round companies that realise the outcome; we then take the median and the 75th percentile across active VCs. The vintage base rate is the share of labelled test companies that realise the outcome. A permuted-label placebo, fitted on the post-correction data, is the relevant null for "beats the base rate": noise-trained models still prefer feature-rich companies, and richer profiles have better outcomes in these files, so a model can beat the base rate without ranking skill. The official primary model (HGB, `v3_full`, all stages, top-50) still clears that placebo comfortably: 68.69% versus a 42.46% placebo (base 40.27%) on 2003–2009 follow-on, 86.80% versus 54.00% (base 47.08%) on 2010–2015 follow-on, and 26.96% versus 16.72% (base 13.35%) on 2005–2009 exits. Source: `backtest_v3_postcorr/results/summary_placebo.csv` and `summary_main.csv`.

4.4 Inference

Pick-level bootstrap confidence intervals resample companies within vintage and seed (1,000 replicates). For the primary family, a joint bootstrap also resamples VC books, so uncertainty in the benchmark is not ignored. P-values are one-sided (H_0: model mean $\leq$ benchmark). Holm–Bonferroni is applied to the six primary tests at $\alpha = 0.05$. Across the broader post-correction v3 grid, both Holm and Benjamini–Hochberg (5% FDR) are reported; that family has 8,464 correlated tests, so Holm is conservative and BH is the more informative family-wide count. The LLM family is described in Section 7.

Options A–C, carried over from v2 and treated as secondary, are not fund returns. Option A is an estimated-cost multiple (lower and upper bounds, with a 100× cap) on realised exits within five years, with zeros for non-exits, on an estimated entry valuation that can be wrong by a factor of about two, and with about two-thirds of five-year exits lacking a disclosed price. Option B is undisclosed-as-zero exit value per pick. Option C is the rank correlation between the model's exit score and SEC pre-IPO holders' multiples on IPO companies only. None of these is a net TVPI.

5 Results

5.1 Main comparison with VC books

Figure 1 and Table 3 report the post-correction primary specification. On three-year follow-on, the model is well above both VC books in both periods with reconstructable books: 68.7% (66.6–70.9) versus a 49.0% median and a 64.2% 75th percentile in 2003–2009, and 86.8% (85.2–88.5) versus 65.3% and 79.0% in 2010–2015. Joint-bootstrap differences are +19.7 points (15.8–22.8) and +4.5 points (2.0–8.0) in the earlier window, and +21.5 (18.2–23.8) and +7.8 (3.6–10.4) in the later window; all four Holm p-values are ≤ 0.0014. On five-year IPO-or-acquisition exits in 2005–2009, the model averages 26.96% (24.6–29.4) against a 19.3% median book and a 32.1% 75th-percentile book. The +7.6 point gap versus the median (4.8–10.3) survives Holm; the −5.2 point gap versus the 75th percentile (−8.5 to −1.9) does not (Holm $p = 0.9988$). In short: the screen beats the median investor on both outcomes, and beats the top-quartile book on follow-on, but not on exits.

The 2016–2023 follow-on track, which uses company traits only at the primary-model setting

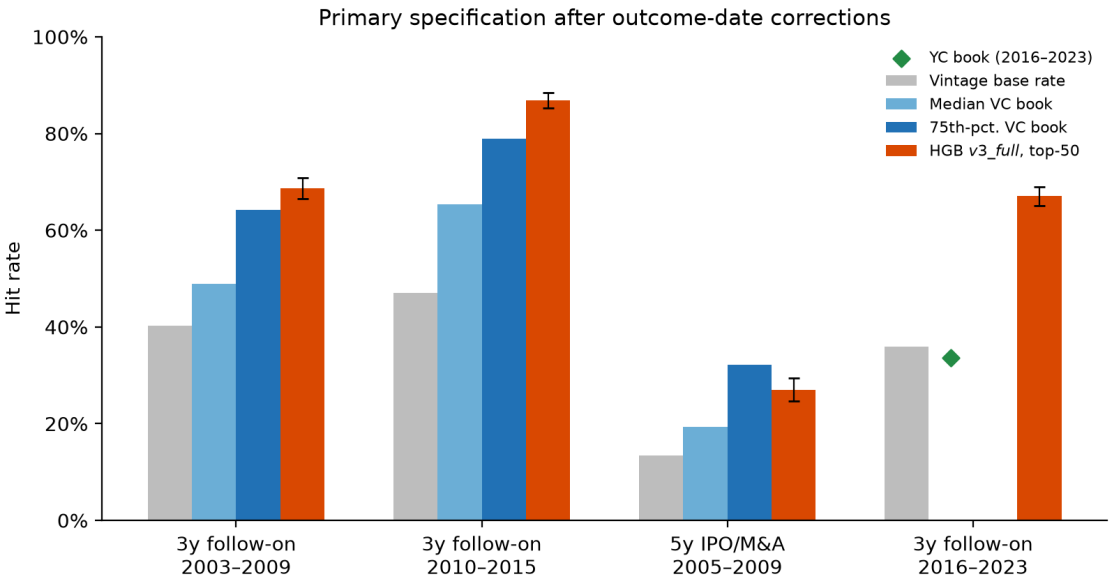


Figure 1: Post-correction primary specification (HGB, `v3_full`, all stages, top-50). Bars for the model include pick-level 95% confidence intervals from `summary_main.csv`. There is no reconstructed multi-VC book after 2015; the diamond is the YC book on the 2016–2023 follow-on track. Source: `backtest_v3_postcorr`.

Table 3: Primary-spec hit rates after corrections (HGB, `v3_full`, all stages).

Outcome	k	Vintages	Model (95% CI)	Base	Median VC	p75 VC	YC
FO 3y	50	2003–09	68.7 (66.6–70.9)	40.3	49.0	64.2	62.2
FO 3y	20	2003–09	69.6 (66.4–72.9)	40.3	49.0	64.2	62.2
FO 3y	50	2010–15	86.8 (85.2–88.5)	47.1	65.3	79.0	65.9
FO 3y	20	2010–15	86.2 (83.5–88.7)	47.1	65.3	79.0	65.9
EX 5y	50	2005–09	27.0 (24.6–29.4)	13.3	19.3	32.1	24.6
EX 5y	20	2005–09	26.2 (22.6–29.8)	13.3	19.3	32.1	24.6
FO 3y	50	2016–23	67.1 (65.0–69.0)	35.9	—	—	33.6
FO 3y	20	2016–23	70.1 (66.8–73.3)	35.9	—	—	33.6

Percentages. Model intervals are pick-level bootstrap CIs, unweighted means across vintages, five seeds. Base rates and VC/YC books are from the same post-correction summary. The 2016–2023 track has no multi-VC book. Source: `backtest_v3_postcorr/results/summary_main.csv`.

reported here, still lifts the hit rate to 67.1% (top-50) or 70.1% (top-20) against a 35.9% base and a 33.6% YC book. That is not a multi-VC test, and the universe is U.S.-filing- and press-heavy.

Top-20 intervals are wider than top-50, as expected. A single-vintage 2010 exit cell (28.0% top-20, 25.2% top-50, against a 20.0% median and 40.0% p75 book) is reported in the same summary file but is not part of the six-test primary family.

5.2 Robustness, from the v3 grid

The richer v3 features were re-tested after the corrections in a grid of 3,150 main fits (three models, six feature sets, five seeds; written as tag `grid` so the official primary `main` files were not overwritten), plus 2,625 ablation and leakage fits and 1,050 placebo fits. Figure 2 is built from those post-correction runs. The qualitative conclusions are unchanged.

Ablations of each new group, relative to `v2_full` and paired by seed, are small and inconsistent

(documented in the project's `RESULTS_V3.md`). The only clear AUC gain for HGB is follow-on in 2003–2009 (+0.022 for `v3_full` versus `v2_full`, paired by seed); even there, the top-50 hit rate barely moves (+0.63 points). Random-forest exit AUC in 2005–2009 rises slightly (0.646 to 0.658). Logistic regression deteriorates, and collapses on 2013–2014 follow-on, because Form D features are rare in Crunchbase-era training data and common later; dropping that group restores performance. Tree models are unaffected. Permutation importance remains dominated by the v2 base features.

Leakage checks can catch a leak when one is planted: adding Crunchbase 2013 snapshot size and round counts raises 2003–2009 follow-on AUC by 0.107 and the top-20 (top-50) hit rate by 19.3 (20.3) points, with univariate AUC of 0.820 for the snapshot round count. No point-in-time feature comes close (the next highest univariate AUCs among non-`lk_` features are about 0.62). Scrubbed descriptions do not help, even though 51.3% of exited companies' raw descriptions mention outcome terms (43.8% for non-exits). About half of the small team-feature gain survives a stricter rule that keeps only Crunchbase records *created* on or before the decision date.

On multiples, there is still no robust edge. In the post-correction primary run, the Option A lower bound capped at 100× averages 0.17× in 2005–2009 (CI 0.13–0.22) against a 0.28× base and a 0.30× mean VC book. Across the post-correction full grid, five-year exit versus the VC 75th percentile still has one BH survivor out of 252 tests and zero Holm survivors; that single BH cell is exploratory and should not be quoted as a confirmed result. Option C's v2 logistic correlation between model score and the pre-IPO multiple (+0.122, 724 IPO companies, vintages 2005–2018) falls to +0.015 with `v3_full` features.

Across the 8,464 post-correction model-versus-benchmark tests, 5,856 are raw wins, 3,922 survive Benjamini–Hochberg at 5% FDR, and 1,762 survive Holm. Survivors concentrate in follow-on (232 of 252 tests against the median VC survive BH; 76 of 252 against the 75th percentile) and in exits against the median VC (107 BH). Exit against the 75th-percentile VC has one BH survivor out of 252 and zero Holm survivors. "Beats the median, not the best, on exits" is therefore not an artefact of the primary cell. Source: `backtest_v3_postcorr/results/multiple_testing_all_tests_grid.csv`.

A company-only model (no syndicate reputation) already produced a large follow-on lift in earlier runs; investor features are not the whole story, and using them assumes the screen can see—and join—the same syndicate.

5.3 Post-correction

Table 2 already states the result: moving 448 fields on 134 companies changes the primary model means by 0.24 to 0.57 percentage points and does not change any Holm decision. Follow-on labels in the test windows did not change; a handful of exit labels did, and a few market-wide features moved with them. The official headline remains the post-correction primary specification. The robustness, ablation, leakage, and placebo layer reported above is the same grid re-run on the corrected files; official primary `summary_main.csv` files were left untouched.

6 Evidence from VC decisions

6.1 SEC schedules of investments

Public business-development companies and registered funds file schedules of investments in 10-K, 10-Q, N-CSR and related forms. We parse those schedules from 1,259 filings of 20 CIKs

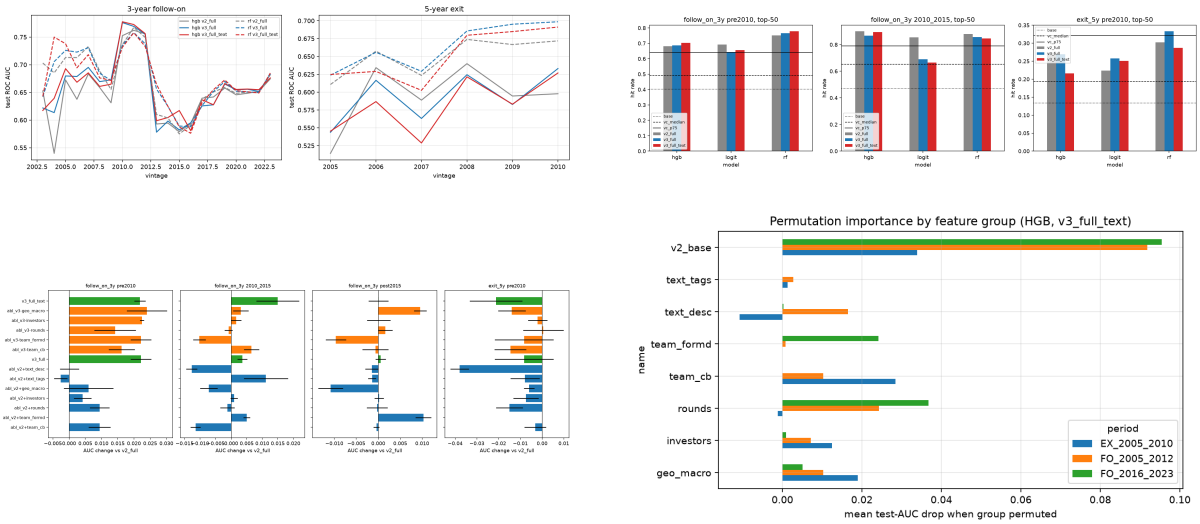


Figure 2: Post-correction v3 grid (3,150 main fits, plus ablation, leakage, and placebo). Top: AUC by vintage and hit rates versus VC books. Bottom: group ablations and permutation importance. Source: `backtest_v3_postcorr/figures/`.

already on disk (no re-download). N-PORT reports, which typically lack a cost basis, are not parsed. A filing contributes position rows only when a table under a "schedule of investments" title has cost and value columns. For each fund-date we keep one source filing (the report whose period equals that date, then the latest amendment). A snapshot is *reconciled* when the sum of parsed fair values, including cash-like rows, is within 3% of the filing's own "Total investments" line and, when that line also reports cost, parsed cost is within 10%.

The resulting panel has 31,962 security-level rows from 18 funds, of which 9,795 lie in reconciled snapshots and 9,635 remain after dropping 160 description-like names. Rolling up to fund × company produces 591 summary rows. Reconciled snapshots are 156 of 626 fund-dates (401 of the 626 had a total-investments line). Coverage is uneven: SuRo/GSV Capital (now Neostellar) reconciles 62 of 62 dated snapshots; MVC 45 of 84; Harris & Harris / 180 Degree Capital 18 of 36 (2011–2016); Excelsior Venture Partners III 11 of 31; Private Shares Fund 7 of 20; Capital Southwest 10 of 84; Rand 3 of 46. Several funds (Firsthand, Crossroads, Destiny Tech100, ARK Venture, Fundrise Innovation, Hatteras, Equus, MACC, Westford, and the Technology Funding partnerships) do not reconcile at all and are kept in the panel but excluded from the statistics. Harris & Harris before 2012 often reports value without a per-position cost, so early multiples are missing.

A position is coded as `exited_or_written_off` if it is absent from the next reconciled snapshot no more than 400 days later. That is a disappearance, not a realised sale. Schedules do not report proceeds; a last mark just before an IPO or sale can differ from cash received, and a mark of zero before disappearance is treated as a write-off. Among 48 early-stage VC positions with that exit status (20 at Excelsior, 28 at 180 Degree Capital), the median last-mark multiple (fair value / cost at the last reconciled mark) is 0.67×. Pooled last-mark multiples on those exited positions are 0.18× at Excelsior and 1.09× at 180 Degree. Later-stage and buyout styles look different: Neostellar's median exited mark is 0.86× with a 1.22× pooled exited multiple; Private Shares Fund's median exited mark is 1.27×. Capital Southwest, classified as PE/buyout from its own filings and excluded from the VC statistics, has a 1.00× median exited mark. These figures describe marks in public filings. They are not fund TVPI, not IRR, and not a claim about net proceeds to LPs.

Table 4: Published-memo and anti-portfolio catalogue, rebuilt 15 November 2026.

Table	Rows	Notes
memos	42	22 Bessemer, 7 Root Ventures, 4 1confirmation
anti-portfolio passes	36	24 Bessemer, 6 Version One, plus 6 others
deal terms	77	41 memo decisions, 36 passes; 61 project-matched
model scores (exploratory)	54 companies	50 scored for follow-on, 25 for exit

Facts only, with source URLs; full memo text is not reproduced. Model percentiles are from the project's first-event walk-forward scores, not from the memo-round decision. Source: `vc_decisions/out/` and `model_scores_summary.json`.

Names are matched to `companies.parquet` only when the normalised name is unique; SPV rows use the stated underlying company when one is given. Match rates vary widely (about 11% at Excelsior, 71% at 180 Degree, 83% at Private Shares Fund).

6.2 Memos and anti-portfolios

A companion catalogue of published investment memos and anti-portfolio passes was rebuilt from the project's hand-extracted records (`vc_decisions/scripts/02_build_tables.py` and `records.py`): 42 memos (22 from Bessemer Venture Partners, 7 from Root Ventures, 4 from 1confirmation, and 9 from other funds), 36 documented passes (24 from Bessemer), and 77 deal-term rows (41 memo decisions and 36 passes), with source URLs on every record and at most a short attributed quote. Sixty-one of the 77 deal-term rows match a unique company in the project file. The project's `valuations/out` tables were not on disk when the catalogue was rebuilt, so the cross-check against that valuation layer is empty; entry valuations and outcomes that come from the memos and anti-portfolio pages themselves are still present. An exploratory, read-only scoring of matched companies with the project's earlier walk-forward model (`backtest/results/universe_features.parquet`) covers 54 companies: 50 scored for three-year follow-on and 25 for five-year exit. Among scored companies, 76% realised follow-on and 19% realised a five-year exit. Median HGB percentiles are high for both invested companies (77.95 on follow-on, $n = 28$) and documented passes (82.6 on follow-on, $n = 17$), as expected in a catalogue of ex-post successes. The model still scores the first observed financing rather than the memo round, and n is too small for a test. Table 4 records the counts and Table 10 in the appendix lists group-level percentiles and every documented pass that could be scored. The memo, pass, and deal-term tables were restored by re-running the catalogue build from the hand-extracted, source-linked records, so the case study is reproducible from `vc_decisions/scripts/`. It remains a case-study pointer, not a second backtest.

7 Large language models as investors

Six models scored anonymized, point-in-time first-round cards through OpenRouter on 15 November 2026: Claude Sonnet 5.5, GPT-6.1 Sol, Llama 4 Maverick, Qwen3-235B-Instruct-2507, DeepSeek V3.2, and Grok 4.7. The first five each received 7,200 main walk-forward deals (stratified across follow-on vintages 2003–2023), 2,000 recent 2024–2026 deals, and a 2,500-deal recognition test. Grok was run only on a stratified subsample of 682 main and 189 recent deals (`effort=low`, no recognition), because a full-size run would not fit the $45 cap; every Grok-versus-GBM number is on that matched subsample. Official usage was $37.394. Prompts withhold names, URLs, YC batch identifiers, and any post-round information; the specification

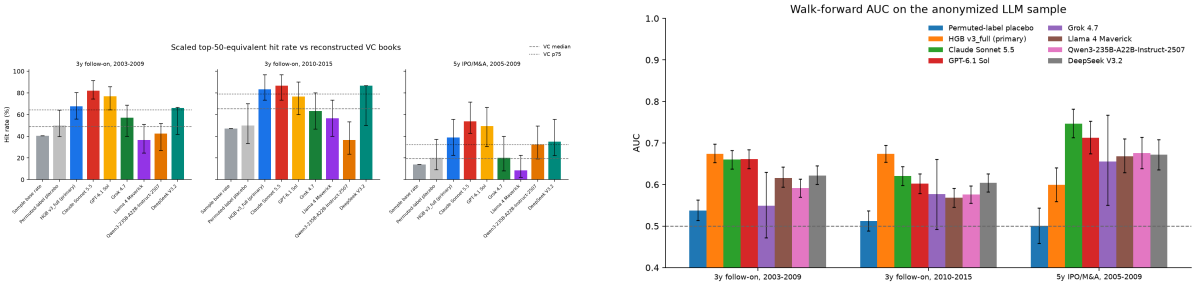


Figure 3: LLM benchmark, 15 November 2026. Left: scaled top-50-equivalent hit rates. Right: unweighted mean of per-vintage AUCs. Grok is a stratified 682-deal subsample. Source: `llm_benchmark/figures/`, reproduced from `RESULTS_LLM.md`.

is summarised in Appendix C. Temperature is zero. Top-k is scaled to the same quantile as the v3 top-50. The GBM row is the primary HGB `v3_full` model, scored on the same companies.

7.1 Headline ranking and hit rates

Figure 3 and Table 5 report scaled top-50-equivalent hit rates and AUCs, as unweighted means across vintages, with 95% percentile bootstrap CIs. On 2003–2009 follow-on, Claude reaches 82.0% (74.4–91.2) and GPT-6.1 Sol 76.8% (64.2–85.5), against 67.5% (55.7–80.1) for the GBM on the same 7,200-deal sample, 49.0% for the median VC book, and 64.2% for the 75th-percentile book. AUCs tell a more cautious story: Claude 0.660 (0.637–0.682) and GPT 0.661 (0.638–0.683) sit slightly *below* the GBM's 0.674 (0.652–0.697). The most striking descriptive gap is on 2005–2009 exits, where Claude's AUC is 0.747 (0.712–0.782) against 0.599 (0.558–0.640) for the GBM on 1,678 companies. That gap is not a pre-registered, multiplicity-adjusted hit-rate test.

After Holm correction at 5% in the 42-test primary LLM-versus-VC family (27 raw wins, 10 BH survivors, 7 Holm survivors), Claude beats both VC books on 2003–2009 follow-on and on 2005–2009 exits; GPT-6.1 Sol and the mean of the five full-size models beat the median VC on 2003–2009 follow-on; the ensemble also beats the median VC on exits. Several other raw wins survive only BH. **None of the 21 same-sample LLM-versus-GBM hit-rate tests survives Holm or BH.** The smallest raw p-values are Claude's 2003–2009 follow-on (0.029) and 2005–2009 exit (0.048); both have Holm $p > 0.6$ once the family is corrected.

Open-weight models and Grok do not beat the GBM. Llama and Qwen often have mid-range AUCs (0.57–0.68) but worse top-k hits than the median VC, consistent with poorly ordered tails or heavy ties. On Grok's 682-deal subsample, follow-on AUC is 0.55–0.58 against 0.66–0.74 for the GBM on the same deals, with wide intervals. A few DeepSeek intervals are pathological (point estimate outside the bootstrap CI) and should not be quoted on their own. Llama left 310 of 7,200 main scores unparseable; GPT omitted 420 EXIT lines. Those deals were not re-queried.

7.2 Recognition, post-cutoff, and stacking

Cards were anonymized, but a model can still recall a firm from a distinctive description. In the recognition test the same card is shown again and the model is asked to name the company or answer `UNKNOWN`. A guess counts if it is not `UNKNOWN` and fuzzy-matches the withheld name or domain. GPT-6.1 Sol named 11.2% of 2,500 asked companies (280), Claude 5.7% (143), DeepSeek 1.5% (38), Llama 1.3% (33), and Qwen 1.0% (25). Grok was not tested.

Once recognised companies are dropped, top-k advantages shrink sharply while AUCs fall only modestly (Table 7 in the appendix). On the asked 2003–2009 follow-on slice, Claude's scaled

Table 5: LLM scaled top-50 hit rates and AUCs (unweighted vintage means, 95% CIs).

Model	FO 2003–09	FO 2010–15	EX 2005–09	FO 2016–23
Hit rate, percent				
Base / med. VC / p75 VC	40.3 / 49.0 / 64.2	47.1 / 65.3 / 79.0	13.9 / 19.3 / 32.1	35.9 / — / —
HGB `v3_full`	67.5 (55.7–80.1)	83.3 (73.3–96.7)	39.0 (22.5–55.6)	57.5 (42.5–75.0)
Placebo	50.0 (39.7–64.2)	50.0 (33.3–70.0)	20.1 (9.2–37.0)	35.0 (25.0–57.5)
Claude Sonnet 5.5	82.0 (74.4–91.2)	86.7 (73.3–96.7)	53.8 (42.8–71.6)	55.0 (40.0–67.5)
GPT-6.1 Sol	76.8 (64.2–85.5)	76.7 (60.0–90.0)	49.5 (30.5–66.5)	65.0 (47.5–80.0)
Ensemble (mean 5)	73.8 (63.6–83.8)	80.0 (63.3–93.3)	47.6 (30.8–66.1)	60.0 (42.5–75.0)
DeepSeek V3.2	66.0 (41.7–66.5)	86.7 (50.0–83.3)	35.1 (22.3–55.6)	67.5 (40.0–70.0)
Grok 4.7 (subs.)	57.1 (39.9–68.6)	63.3 (46.7–80.0)	20.0 (8.0–40.0)	57.5 (40.0–70.0)
Qwen3-235B-2507	42.6 (26.9–51.7)	36.7 (23.3–53.3)	32.4 (19.1–49.3)	42.5 (30.0–60.0)
Llama 4 Maverick	36.4 (24.5–50.8)	56.7 (40.0–73.3)	8.5 (2.0–22.5)	50.0 (30.0–60.0)
AUC				
HGB `v3_full`	0.674 (0.652–0.697)	0.673 (0.653–0.695)	0.599 (0.558–0.640)	0.629 (0.606–0.654)
Claude Sonnet 5.5	0.660 (0.637–0.682)	0.620 (0.597–0.643)	0.747 (0.712–0.782)	0.536 (0.512–0.559)
GPT-6.1 Sol	0.661 (0.638–0.683)	0.602 (0.578–0.625)	0.713 (0.673–0.752)	0.532 (0.509–0.556)
Ensemble (mean 5)	0.654 (0.631–0.676)	0.610 (0.588–0.632)	0.718 (0.683–0.752)	0.528 (0.503–0.552)
DeepSeek V3.2	0.621 (0.600–0.645)	0.604 (0.582–0.625)	0.671 (0.635–0.708)	0.537 (0.514–0.561)
Llama 4 Maverick	0.616 (0.594–0.641)	0.568 (0.545–0.591)	0.668 (0.628–0.710)	0.495 (0.473–0.520)
Qwen3-235B-2507	0.592 (0.569–0.613)	0.576 (0.555–0.597)	0.676 (0.638–0.713)	0.526 (0.505–0.547)
Grok 4.7 (subs.)	0.549 (0.472–0.629)	0.577 (0.491–0.660)	0.655 (0.550–0.767)	0.567 (0.495–0.644)
Placebo	0.537 (0.513–0.562)	0.512 (0.488–0.536)	0.501 (0.458–0.543)	0.466 (0.442–0.489)

VC books are the official v3 reconstructed books on the full universe. DeepSeek's 2010–2015 hit-rate interval is pathological (the point estimate lies outside the CI) and should be ignored as a standalone figure. Source: `llm_benchmark/RESULTS_LLM.md`.

top-50 hit falls from 74.4% to 62.4% (AUC 0.665 to 0.640) and GPT's from 69.8% to 47.1% (AUC 0.649 to 0.619), which is at the median VC book. On 2005–2009 exits, Claude's hit on the asked slice falls from 40.0% to 20.0% (AUC 0.682 to 0.673)—from well above the 75th-percentile book down to the median—while ranking quality is almost unchanged. Holm-surviving full-sample hit-rate wins should not be read as skill on unrecognised firms.

The cleaner leakage control is the post-cutoff cohort: first rounds after each model's knowledge cutoff, labelled with 6- or 12-month follow-on only when the full window has ended (as-of 22 September 2026). Frontier models with mid-2026 cutoffs have *no* fully observed window. The open models we can test are near chance: Llama (cutoff August 2024) has AUC 0.513 on 1,515 deals at six months and 0.491 on 870 at twelve months; DeepSeek (March 2025) 0.470 / 0.454 on 1,077 / 432; Qwen (June 2025) 0.569 / 0.562 on 809 / 164. This is a small, right-censored, Form D / Fundraising Fox / YC-heavy universe, not a VC-book comparison, and it is a null leakage-control result for the models we can actually test. On the main 2016–2023 follow-on window, LLM AUCs are 0.50–0.54, against 0.629 for the GBM.

Adding the LLM score as a walk-forward extra GBM feature (a two-feature logistic stack, or a rank average) raises 2003–2009 follow-on AUC a little—Claude stack 0.720 versus GBM 0.674; GPT stack 0.711; five-model ensemble stack 0.706—and does not help in 2016–2023 (Claude stack 0.606 versus GBM 0.629). Local API costs were $13.56 (Claude), $13.08 (GPT), $5.62 (Grok), $1.72 (DeepSeek), $1.14 (Llama), and $0.60 (Qwen); the log total of $35.72 can lag the official $37.394 usage figure.

The honest reading is that anonymized LLM screens can look strong on older, famous vintages; that much of the top-k edge is recognition; that they do not significantly beat the gradient-boosting model once comparisons are corrected; and that after 2015, and after each model's cutoff, they look like chance.

7.3 Blind cards

The cards in the main LLM run withheld names and URLs, but not cities, exact dates, investor names, or free-text descriptions, and recognition was not rare (GPT 11.2%, Claude 5.7% of 2,500 asked deals). A second prompt, `blind_v1`, further coarsens the card to vintage buckets, a broad region, amount bins, coarse team and investor-quality tiers, and sector group only, with no product text. The specification is in the project's `PROMPT_SPEC_BLIND.md`. Grok was not rerun. Llama, Qwen, and DeepSeek scored the same 7,200 main-sample deals; Claude and GPT were scored on a 1,545-deal stratified subsample sized to a remaining budget. Official OpenRouter usage, cumulative with the first run, ended at $45.799 of a $48.50 cap.

On a fresh 500-deal recognition pretest, no model named the firm (0 of 500 for all five models; DeepSeek's two non-`UNKNOWN` replies were refusals to parse the card, not company names). Because the coarsening is heavy, the 7,200 deals collapse to 4,241 distinct cards. The 5% frontier gate was passed, and scoring proceeded.

Figure 4 and Table 6 compare original and blind scores on the *same* companies (paired within-vintage bootstrap, 1,000 replicates). On 2003–2009 follow-on, most AUCs fall: GPT from 0.643 to 0.588 ($p = 0.002$, $n = 514$), DeepSeek from 0.621 to 0.589 ($p = 0.008$, $n = 2{,}399$), and a three-model open ensemble (Llama, Qwen, DeepSeek) from 0.632 to 0.603 ($p = 0.004$, $n = 2{,}203$). Claude's matched hit rate goes from 71.4% to 65.7% ($n = 517$), still above the 49.0% median VC book at the point estimate, but the drop is not significant ($p = 0.49$), and Claude does not beat the GBM on those 517 deals (77.1%, $p = 0.938$). Llama's 2003–2009 follow-on AUC also falls (0.616 to 0.576, $p = 0.002$); Qwen's does not move. In 2010–2015, 2016–2023, and on exits, blind-minus-original AUC changes are small and not significant. Because blind cards also remove information (exact year, city, exact amounts, description), the fall is an upper bound on the recognition effect rather than a clean estimate of it. After 2015, blind AUCs remain near chance (Claude 0.555, GPT 0.571, DeepSeek 0.528) against 0.629 for the GBM on the full sample. Some hit rates, especially Llama's, are unstable because of score ties and a small scaled k; AUC is the more reliable comparison.

In the 36-test primary family of blind-card LLM hit rates against the median and 75th-percentile VC books, there is one raw win at 5% (Claude on 2005–2009 exits versus the median book, raw $p = 0.042$) and no Holm or Benjamini–Hochberg survivor; in the 18 same-sample LLM-versus-GBM hit-rate tests there is no raw win. On follow-on AUC every blind LLM is below the GBM in every window (0.51–0.59 versus 0.63–0.67). The one pattern that survives blinding is exit ranking: on 2005–2009 five-year exits, blind AUCs of 0.65–0.68 remain above the GBM's 0.60 on the same deals (raw one-sided paired-bootstrap p between 0.001 and 0.043 for the open models and Claude, 0.15 for GPT), as on the original cards. That is a descriptive, unadjusted AUC comparison, not a pre-declared test, and the hit-rate version does not survive. Adding a blind LLM score as a walk-forward extra GBM feature still yields only a small AUC bump on older vintages (Qwen stack 0.689 and the open-model ensemble stack 0.684, versus GBM 0.674 on 2003–2009 follow-on) and none after 2015 (ensemble stack 0.619 versus GBM 0.629). The original-card Holm wins for Claude should not be read as skill that remains once the card is stripped toward the GBM's information set.

8 Practical applications

Three uses follow from the results, and several do not.

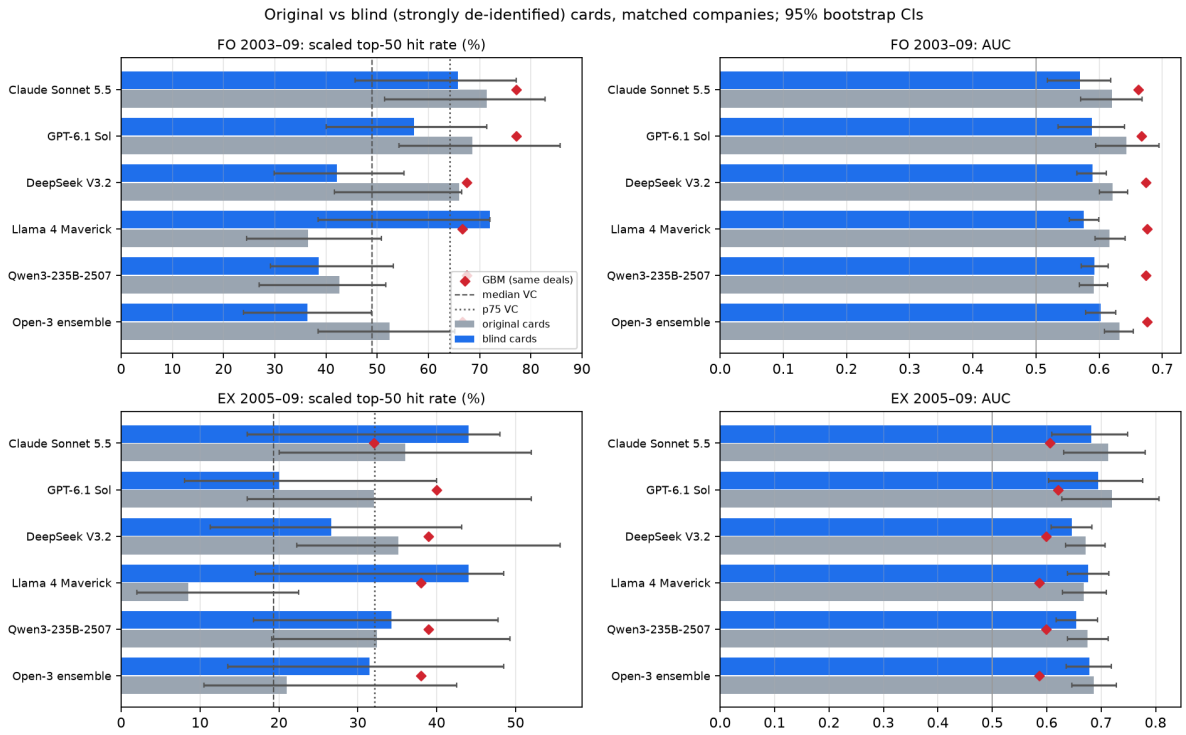


Figure 4: Original versus blind (strongly de-identified) cards on matched companies (2003–2009 follow-on and 2005–2009 exits; hit rate and AUC), with 95% bootstrap CIs. Diamonds are the GBM on the same deals; dashed and dotted lines are the median and 75th-percentile VC books. Source: `llm_benchmark/results_blind/blind_vs_original.csv`.

First, a walk-forward public-data model is a reasonable *screen* for first-round follow-on risk, including without syndicate features, and it beats the median reconstructed book on five-year exits as well. It is a triage tool: it can move a partner's time toward a short-list. It is not a substitute for allocation, terms, or board work, and it does not justify a claim to top-quartile exit picking.

Second, last-mark multiples from public fund schedules are a rare, unselected window onto cost and fair value, and they should be used as marks. They are useful for documenting write-offs and stale winners. They should not be dropped into a TVPI calculation as if they were distributed proceeds.

Third, frontier LLMs are not, on this evidence, ready to replace a gradient-boosting screen. They may be worth stacking as an extra feature on older vintages, where we observe a modest AUC gain, and they are a poor stand-alone ranker after 2015. Any production use should run a recognition test, drop or flag named firms, and treat post-cutoff performance—not pre-2010 hit rates—as the relevant skill check. At $45.799 of cumulative official API spend we could not give every model a full, post-cutoff, long-horizon test; a budget that buys memorisation of famous outcomes is not a budget that buys an investor.

What the results do not support is allocating a fund as if the model were a top-quartile manager, or reporting Option A multiples as if they were net returns. Official vintage benchmarks [6, 13] remain the right reference for that question, and we cannot compute them here.

Table 6: Original versus blind cards, matched companies (unweighted vintage means).

Model	n	AUC orig	AUC blind	Hit orig	Hit blind
3y follow-on, 2003–2009					
Claude Sonnet 5.5	517	0.621	0.569	71.4	65.7
GPT-6.1 Sol	514	0.643	0.588	68.6	57.1
DeepSeek V3.2	2,399	0.621	0.589	66.0	42.1
Open-3 ensemble	2,203	0.632	0.603	52.5	36.4
Llama 4 Maverick	2,204	0.616	0.576	36.4	72.0
Qwen3-235B-2507	2,400	0.592	0.593	42.6	38.6
3y follow-on, 2016–2023					
Claude Sonnet 5.5	512	0.563	0.555	72.5	67.5
GPT-6.1 Sol	511	0.562	0.571	60.0	57.5
DeepSeek V3.2	2,399	0.537	0.528	67.5	40.0
Open-3 ensemble	2,382	0.523	0.533	60.0	45.0
5y IPO/M&A, 2005–2009					
Claude Sonnet 5.5	359	0.713	0.682	36.0	44.0
GPT-6.1 Sol	320	0.720	0.694	32.0	20.0
DeepSeek V3.2	1,678	0.671	0.646	35.1	26.6
Open-3 ensemble	1,550	0.686	0.678	21.0	31.5

Hits are scaled top-50-equivalent percentages. The open-3 ensemble is the mean of Llama, Qwen, and DeepSeek. GPT's 2003–2009 follow-on AUC drop (-0.055) has two-sided $p = 0.002$; DeepSeek's (-0.032) has $p = 0.008$; Llama's (-0.040) $p = 0.002$. Llama's blind hit rates are tie-driven (its modal score covers 49% of deals) and should not be read on their own. GBM AUC on the same deals: 0.662–0.676 (2003–2009 follow-on), 0.586–0.620 (exits), 0.629–0.644 (2016–2023). Source: `results_blind/blind_vs_original.csv`.

9 Limitations

The limits are first-order, not footnotes.

- **Survivorship and coverage.** Public Crunchbase over-represents visible, funded companies. Fundraising Fox and Form D are U.S.-filing- and press-heavy. Wikidata and Wikipedia M&A coverage after 2015 is thin; missing acquisitions are null, not zeros. Feature-rich firms have better labelled outcomes, which is why the placebo—not the base rate—is the right null.
- **Deal access.** The model may pick any company that appears in the file. Nanda et al. [11] show that preferential access, not only selection skill, sustains VC performance. We cannot test whether the model would have been invited into the same syndicate.
- **Hit rate is not a fund return.** We have no reliable pre-money valuations, ownership, fees, or carry, and we do not compare Option A with Cambridge Associates or PitchBook net TVPI [7].
- **Last-mark multiples** from SEC schedules are not realised proceeds, cover a non-random set of public funds, and reconcile for only 156 of 626 fund-dates.
- **LLM memorisation.** Anonymization does not block parametric knowledge of famous firms. On the original cards, recognition explains a large part of the top-k edge. A stricter blind-card recognition test on 500 deals was 0% for all five models; ranking edges on older vintages still shrink, and no blind-card LLM hit-rate test survives Holm. Grok was not recognition-tested and was not included in the blind rerun.
- **Cost and sample.** Official spend on the original LLM run was \$37.394 of a \$45 cap. The subsequent blind-card rerun brought cumulative official usage to \$45.799 of a \$48.50 cap.

Grok was restricted to a 682-deal subsample on the original cards and was skipped on the blind run; Claude and GPT were then limited to a 1,545-deal stratified subsample. A minority of Llama and GPT replies on the original run were unparseable and were not re-queried. GPT-6 Astra, the more expensive flagship, was not run.

- **Short post-cutoff horizon.** As-of 22 September 2026, Claude, GPT-6.1 Sol, and Grok have no fully observed 6- or 12-month post-cutoff window. Open-model AUCs of 0.45–0.57 are a null on a short, low-base-rate label, not a replication of the main headline.
- **Residual look-ahead in public features.** Categories and some undated titles can be edited after the fact. Team coverage collapses after the 2013 Crunchbase snapshot. The v3 robustness, ablation, leakage, and placebo grid was re-run after the corrections; the official primary files were left in place.
- **Knowledge-cutoff sources** for some LLM rows (notably Grok) rest on a compilation site rather than a confirmed provider model card, as flagged in `RESULTS_LLM.md`.

10 Conclusion

A walk-forward model trained on public, point-in-time features beats the median reconstructed VC first-round book on three-year follow-on and on five-year exits, and beats the 75th-percentile book on follow-on, after a documented corrections audit whose effect is under 0.6 percentage points. It does not beat the 75th-percentile book on exits, and the hit-rate edge does not become a robust entry-cost multiple. Parsed SEC schedules add a sober case study: early-stage VC last-mark multiples on positions that then leave the filing have a median of 0.67×, and those marks are not cash. Large language models, scored on anonymized cards, can look like strong top-k screens on older vintages, but recognition of withheld names accounts for much of that appearance; a stricter de-identified prompt, on which no model named the firm in a 500-deal test, shrinks most of those advantages, and no blind-card LLM hit-rate test survives Holm correction. They do not significantly outperform gradient boosting after multiple-testing correction, and after 2015 they are close to chance. The title is the result: the median, not the best.

Table 7: Recognition test: scaled top-50 hit rate and AUC on asked companies, with and without firms the model named correctly.

Model	Period	Hit all	Hit ex. rec.	AUC all	AUC ex. rec.
Claude (5.7% named)	FO 2003–09	74.4	62.4	0.665	0.640
Claude	EX 2005–09	40.0	20.0	0.682	0.673
GPT-6.1 Sol (11.2%)	FO 2003–09	69.8	47.1	0.649	0.619
GPT-6.1 Sol	EX 2005–09	36.0	28.0	0.671	0.634
Llama (1.3%)	FO 2003–09	25.2	25.2	0.572	0.567
Llama	EX 2005–09	20.0	16.0	0.648	0.641
Qwen (1.0%)	FO 2003–09	61.4	65.3	0.577	0.576
Qwen	EX 2005–09	36.0	32.0	0.683	0.678
DeepSeek (1.5%)	EX 2005–09	32.0	28.0	0.658	0.660

DeepSeek's 2003–2009 follow-on subset hit rate is a pathological 100% point estimate (bootstrap interval 39–69%) and is omitted here. N asked = 2,500 per model (Grok not run). Source: `RESULTS_LLM.md`.

Table 8: Walk-forward LLM score as an extra GBM feature (two-feature logistic stack).

Model	Period	GBM AUC	LLM AUC	Stack AUC
Claude Sonnet 5.5	FO 2003–09	0.674	0.660	0.720
Claude Sonnet 5.5	FO 2010–15	0.673	0.620	0.699
Claude Sonnet 5.5	FO 2016–23	0.629	0.536	0.606
GPT-6.1 Sol	FO 2003–09	0.676	0.661	0.711
GPT-6.1 Sol	FO 2016–23	0.630	0.532	0.612
Ensemble (mean of 5)	FO 2003–09	0.674	0.654	0.706
Ensemble (mean of 5)	FO 2016–23	0.629	0.528	0.608

Intervals and hit rates, and the rank-average fusion, are in `RESULTS_LLM.md`. The stack is trained on earlier vintages and applied to vintage T; it does not use future labels.

A Data-dictionary summary

Dates are ISO `YYYY-MM-DD` (or null). Amounts are reported U.S. dollars, not inflation-adjusted. Boolean outcomes are 0/1. Horizon labels are 0/1 or null when the window cannot be certified (right-censoring, or a known but undated exit). `company_id` is a canonical key (`cb_`, `yc:`, `formd_`, or `ff:`). `name_norm` is the matching key. `first_event_date` is the decision date. `follow_on_3y` and `exit_5y` are the primary labels. `yc_status_at_pull`, `team_size_at_pull`, and unicorn fields are pull-time attributes and are not used as features. Full column definitions, observer walls, and source caveats are in `DATA_DICTIONARY.md` and `DATA_QUALITY.md`.

B Additional tables

Table 7 reports recognition-conditional hit rates and AUCs. Table 8 reports the walk-forward GBM-plus-LLM stack for the two strongest full-size models and the five-model mean, which is the comparison emphasised in the text; the remaining models are in the LLM results report. Table 9 summarises reconciled SEC-fund positions. Table 10 gives the memo and anti-portfolio case study.

Table 9: Reconciled SEC-fund position summaries (fund × company).

Fund (style)	Dates	Positions	Exited	Med. exit mark	Match
MVC Capital (VC then mezz.)	45	100	74	1.00×	0.15
Excelsior VP III (early VC)	11	38	20	0.67×	0.11
Neostellar / SuRo (late, secondaries)	62	152	119	0.86×	0.57
Private Shares Fund (late, secondaries)	7	82	31	1.27×	0.83
Capital Southwest (PE/buyout)[a]	10	134	27	1.00×	0.22
Rand Capital (VC then mezz.)	3	27	0	—	0.19
180 Degree Capital (early VC)	18	58	28	0.60×	0.71

[a] Classified as PE/buyout from the fund's own description and excluded from the early-VC statistics.

Marks are last reported fair value / cost, not realised proceeds. Match is the share of positions matched to a unique project company. Source: `vc_decisions/sec_funds/out/fund_summary.csv`.

Table 10: Memo and anti-portfolio case study: exploratory walk-forward HGB percentiles of catalogued companies.

Label	Group	n	Realised (%)	Median HGB pct.	In top decile (%)
FO 3y	Invested	28	64	78.0	43
FO 3y	Passed	17	88	82.6	29
FO 3y	Invested by one, passed by another	3	100	88.7	33
FO 3y	Invested after an earlier pass	2	100	60.1	50
EX 5y	Invested	13	15	54.5	15
EX 5y	Passed	10	14	79.8	20
EX 5y	Invested by one, passed by another	2	100	92.2	50

Documented passes scored on follow-on (fund, pass year, first-event year, HGB percentile, realised)
Ikanos Communications (Bessemer, 2000, 2003, 99.4, yes); Airbnb (Bessemer, USV, 2010/2011, 2009, 99.3, yes); Uber (USV, —, 2009, 90.4, yes); Okta (Bessemer, 2009, 2009, 90.3, yes); Coinbase (Bessemer, Foundry, 2012, 2012, 90.3, yes); Instacart (Bessemer, Version One, 2012/2013, 2012, 89.6, yes); Facebook (Bessemer, 2004, 2004, 84.0, yes); Tesla (Bessemer, 2006, 2004, 83.2, yes); Kayak (Bessemer, —, 2004, 82.6, yes); A123 Systems (Bessemer, 2004, 2005, 81.6, yes); Honey (Version One, 2015, 2012, 68.4, yes); Carta/eShares (Version One, 2013, 2013, 56.8, yes); Zoom (Bessemer, 2014, 2011, 54.9, yes); Wealthsimple (Version One, 2014, 2014, 54.3, yes); Peloton (Foundry, —, 2012, 53.0, yes); Atlassian (Bessemer, 2010, 2010, 6.5, no); Scale AI (Version One, 2016, 2016, 5.3, no).

Percentile of the company's walk-forward HGB score within its first-event vintage cohort (project's earlier model, `backtest/`); it scores the first observed financing, which can differ from the memo or pass round. Realised is the share with the label equal to 1 among scored companies. Catalogue of ex-post famous firms: selection on outcomes is severe, so no test is run. Source: `vc_decisions/out/model_scores.csv` and `model_scores_summary.json`.

C Prompt specification

Prompt version `v1`, built by `llm_benchmark/prompts.py`. Each card is usable at the first observed financing and withholds identifiers. Fields included: year of first financing; country, state, city, and a metro bucket (no street addresses); sector group; first-round type and stage; first-round amount, rounded, or "not disclosed"; point-in-time team facts with no names (founder count, dated prior founding, technical versus business mix, dated flags for a prior big-tech role, a QS 2012/13 top-25 degree, a PhD or MBA, and Form D officer/director counts); investor quality with no names (named-investor count, the most experienced investor's prior deal and exit counts as of that date, a top-5% prior-exit flag, and a "well-known seed accelerator, name withheld" flag—Y Combinator is never named). A short description is included only when a scrubbed text field exists: URLs and emails are removed, names and domain stems are replaced with "the company", and later-outcome sentences are dropped, capped at 800 characters.

Fields excluded: company, founder, or investor names, websites, domains, or YC batch identifiers; any field dated after the first financing; snapshot-size and pull-time status fields. The model is asked for `FOLLOW_ON: 0.xxx` and `EXIT: 0.xxx` on main deals, for a 12-month follow-on probability on 2024–2026 deals, and, in the recognition test, to name the company or answer `UNKNOWN`. The system message tells the model to ignore later outcomes even if it recognises the firm. The authoritative text is `llm_benchmark/PROMPT_SPEC.md`.

D Reproducibility

Code and data live in the project workspace. The dataset build comprises the numbered scripts under `scripts/` (download through corrections). The pre-declared primary specification is `PRIMARY_SPEC.md` in the `backtest_v3` folder; post-correction primary results are `backtest_v3_postcorr/results/`. The LLM benchmark is `llm_benchmark/`, with the report in `RESULTS_LLM.md`. SEC-fund scripts and the rebuilt `out/` panel are `vc_decisions/sec_funds/`. A compressed snapshot (processed tables, post-correction backtests, key v3/v2 result tables, LLM code, results, figures, and a compressed response cache, excluding `.venv`, raw SEC downloads, and secrets) is split under `dist/2026-11-15/` as `ai-vs-vc-dataset-2026-11-15.tar.xz.001`–`.014`, with SHA-256 checksums in `SHA256SUMS` and reassembly instructions in `REASSEMBLE.txt`. The rebuilt SEC-fund panel is the add-on `ai-vs-vc-sec-funds-panel-2026-11-15.tar.xz`. Official OpenRouter usage for the original LLM run was $37.394 of a $45 cap; after the blind-card rerun, cumulative official usage was $45.799 of a $48.50 cap. Blind scores and tables are in `llm_benchmark/results_blind/`. No API keys are stored in the archive or in the cached response bodies.

References

[1] Javier Arroyo, Francesco Corea, Guillermo Jiménez-Díaz, and Juan A. Recio-García. Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, 7:124233–124243, 2019. doi: 10.1109/ACCESS.2019.2938659.

[2] Jae Yoon Bae, Simon Malberg, Joyce Ann Clarize Galang, Andre Retterath, and Georg Groh. DIALECTIC: A multi-agent system for startup evaluation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics: Industry Track*, 2026. URL `https://aclanthology.org/2026.eacl-industry.53/`.

[3] Luiz Bissoto. Predictability and persistence in deal selection, 2026. URL `https://www.paris-december.eu/sites/default/files/papers/2026/7268_luizbissoto_2026_complete.pdf`. Working paper, EPFL/SFI, 29 May 2026. Presented at Paris December 2026 Finance Meeting.

[4] Ivo Blohm, Torben Antretter, Charlotta Sirén, Dietmar Grichnik, and Joakim Wincent. It's a peoples game, isn't it?! a comparison between the investment returns of business angels and machine learning algorithms. *Entrepreneurship Theory and Practice*, 46(4):1054–1091, 2022. doi: 10.1177/1042258720945206.

[5] Maxime Bonelli. Data-driven investors. *The Review of Financial Studies*, 39(7):1909–1969, 2026. doi: 10.1093/rfs/hhaf078.

[6] Cambridge Associates. US venture capital benchmark book, 31 march 2026. Technical report, Cambridge Associates, 2026. URL https://www.cambridgeassociates.com/wp-content/uploads/2026/08/WEB-2026-Q1-USVC-Benchmark-Book.pdf.

[7] Robert S. Harris, Tim Jenkinson, and Steven N. Kaplan. Private equity performance: What do we know? *Journal of Finance*, 69(5):1851–1882, 2014. Also NBER Working Paper 17874.

[8] Amar Krishna, Ankit Agrawal, and Alok Choudhary. Predicting the outcome of startups: Less failure, more success. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2016.

[9] Shiwei Lyu, Shuai Ling, Kaihao Guo, Haipeng Zhang, Kunpeng Zhang, Suting Hong, Qing Ke, and Jinjie Gu. Graph neural network based VC investment success prediction, 2021. URL https://arxiv.org/abs/2105.11537. arXiv:2105.11537.

[10] MIT Technology Review. An algorithm to pick startup winners. *MIT Technology Review*, 2012. URL https://www.technologyreview.com/2012/07/09/185027/an-algorithm-to-pick-startup-winners/. Profile of Correlation Ventures.

[11] Ramana Nanda, Sampsa Samila, and Olav Sorenson. The persistent effect of initial success: Evidence from venture capital. *Journal of Financial Economics*, 137(1):231–248, 2020.

[12] National Venture Capital Association. NVCA yearbook 2025, 2025. URL https://nvca.org/wp-content/uploads/2025/03/2025-NVCA-Yearbook.pdf.

[13] PitchBook. PitchBook benchmarks: Venture capital, Q1 2025 (with preliminary Q2 2025 data). Technical report, PitchBook Data, 2025. URL https://files.pitchbook.com/website/files/pdf/Q1_2025_PitchBook_Benchmarks_with_preliminary_Q2_2025_data_Venture_Capital_20365.pdf. Global VC pooled net IRR and TVPI as of 31 March 2025.

[14] Mark Potanin, Andrey Chertok, Konstantin Zorin, and Cyril Shtabtsovsky. Startup success prediction and VC portfolio simulation using CrunchBase data, 2023. URL https://arxiv.org/abs/2309.15552. arXiv:2309.15552.

[15] Andre Retterath. Human versus computer: Benchmarking venture capitalists and machine learning algorithms for investment screening, 2020. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3706119. SSRN working paper 3706119.

[16] Greg Ross, Sanjiv Das, Daniel Sciro, and Hussain Raza. CapitalVX: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7:94–114, 2021.

[17] Ronald Setty, Yuval Elovici, and Dafna Schwartz. Cost-sensitive machine learning to support startup investment decisions. *Intelligent Systems in Accounting, Finance and Management*, 2024. doi: 10.1002/isaf.1548.

[18] Boris Sharchilev, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten de Rijke. Web-based startup success prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2283–2291, 2018.

[19] U.S. Securities and Exchange Commission. Form D data sets, 2026. URL https://www.sec.gov/data-research/sec-markets-data/form-d-data-sets. Quarterly tabular extracts of Regulation D filings.

[20] Vela Research. VCBench: A benchmark for founder-level startup success prediction, 2025. URL https://arxiv.org/abs/2509.14448. arXiv:2509.14448. Firm-associated preprint; not a peer-reviewed journal article.

[21] Guang Xiang, Zeyu Zheng, Miaomiao Wen, Jason Hong, Carolyn Rose, and Chao Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on TechCrunch. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2012. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14306.

[22] Kamil Żbikowski and Piotr Antosiuk. A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, 58(4): 102555, 2021. doi: 10.1016/j.ipm.2021.102555.