arXiv:2509.14448v1 [cs.AI] 17 Sep 2025

VCBench: Benchmarking LLMs in Venture Capital

Rick Chen[1], Joseph Ternasky[2], Afriyie Samuel Kwesi[2], Ben Griffin[1], Aaron Ontoyin Yin[2], Zakari Salifu[2], Kelvin Amoaba[2], Xianling Mu [1], Fuat Alican[2], Yigit Ihlamur[2]

[1] *University of Oxford* [2] *Vela Research*

Abstract. Benchmarks such as SWE-bench and ARC-AGI demonstrate how shared datasets accelerate progress toward artificial general intelligence (AGI). We introduce **VCBench**, the first benchmark for predicting founder success in venture capital (VC), a domain where signals are sparse, outcomes are uncertain, and even top investors perform modestly. At inception, the market index achieves a precision of 1.9%. Y Combinator outperforms the index by a factor of 1.7×, while tier-1 firms are 2.9× better. **VCBench** provides 9,000 anonymized founder profiles, standardized to preserve predictive features while resisting identity leakage, with adversarial tests showing more than 90% reduction in re-identification risk. We evaluate nine state-of-the-art large language models (LLMs). DeepSeek-V3 delivers over six times the baseline precision, GPT-4o achieves the highest $F_{0.5}$, and most models surpass human benchmarks. Designed as a public and evolving resource available at vcbench.com, **VCBench** establishes a community-driven standard for reproducible and privacy-preserving evaluation of AGI in early-stage venture forecasting.

1. Introduction

Benchmark datasets have played a defining role in the progress of machine learning (ML). By turning open-ended challenges into standardized and measurable tasks, they have enabled reproducible comparisons and driven entire fields forward. As models advance, there is growing demand for benchmarks that not only test raw accuracy but also allow systematic comparisons between machine and human performance.

Venture capital (VC) is a compelling testbed for evaluating expert forecasting, offering a real-world setting to measure whether models can match or surpass human reasoning under uncertainty. Decisions rely on sparse and uncertain signals from founder backgrounds and early company data, while the financial stakes are high. Even leading investors perform modestly. The market index achieves 1.9% precision at inception, while Y Combinator reaches 3.2% (1.7× the index) and tier-1 VC firms are at 5.6% (2.9×). Recent models (Mu et al., 2025; Griffin et al., 2025) show that founder profiles alone can yield strong predictive signals, but the field lacks a standardized benchmark. These limits highlight both the practical importance and the scientific value of VC. It provides a domain where forecasting skill can be measured against expert performance, and exceeding that benchmark would represent a step toward human-level decision-making.

Building a benchmark for VC faces unique challenges. Founder data is abundant but noisy. LinkedIn provides wide coverage, yet its entries are inconsistent and unregulated. Crunchbase offers structured business metrics, but with limited scope. Together, these sources create four key issues: (i) *format irregularity*, where the same degree or role appears in many variants; (ii) *entry irregularity*, such as inclusion of internships or short courses; (iii) *coverage imbalance*, with many missing or incomplete fields; and (iv) *data contamination*, where large language models (LLMs) can re-identify founders from profile text and bypass the intended prediction task. As summarized in Table 1, these challenges differ from those in previous benchmarks and require a different approach.

In this paper we introduce **VCBench**[1], a standardized and anonymized benchmark for founder-success prediction. The dataset contains 9,000 founder profiles, with 810 labeled as successful. A key challenge is ensuring that models cannot bypass the prediction task by re-identifying founders. To address this, we conduct adversarial re-identification experiments, which reduce identifiable founders by 92% while preserving predictive features. Building on this foundation, we design a pipeline of standardization, filtering, enrichment, and anonymization. We then evaluate nine state-of-the-art LLMs as shown in Figure 1. To encourage further study, we release a public leaderboard for continuous evaluation of both LLMs and purpose-built VC models.

---

[1] Correspondence to: Rick Chen (rick.chen.ac@gmail.com); Yigit Ihlamur (yigit@vela.partners).

Benchmark	Data Filtering	Data Standardization	Coverage Improvement	Data Anonymization
ARC-AGI-2	–	–	–	–
SWE-Bench	✓	–	–	–
Microsoft SDBench	–	–	✓	–
VCBench	✓	✓	✓	✓

Table 1. Comparisons with Existing LLM Benchmarks

Our contributions

1. We introduce **VCBench**, the first anonymized benchmark for founder-success prediction in VC.
2. We develop a generalizable pipeline for data cleaning and anonymization, and validate it with adversarial re-identification tests.
3. We benchmark nine state-of-the-art LLMs, several surpassing human baselines, and release a public leaderboard at vcbench.com.

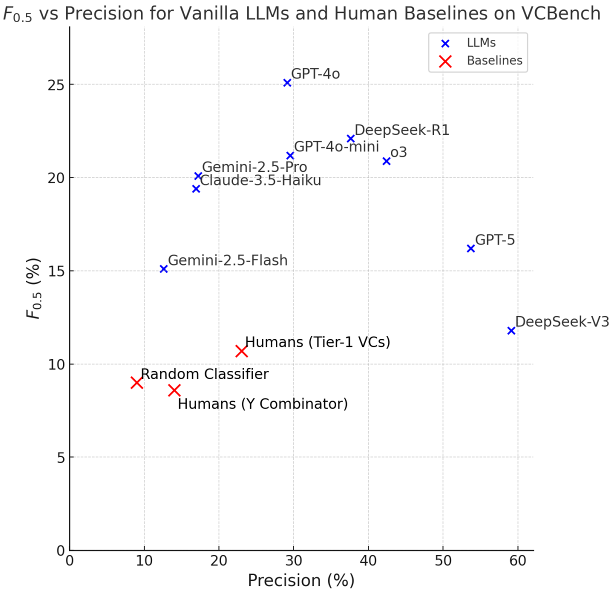

Figure 1. Predictive performances of nine vanilla LLMs on VCBench, with human-level baselines. The human-level baseline results are scaled linearly to reflect the inflation of success rate from the real-world (1.9%) to VCBench (9%).

2. Related Work

Benchmarks in machine learning. Benchmark datasets have long been central to machine learning progress. In vision and language, VQA (Agrawal et al., 2016) and ImageNet (Russakovsky et al., 2015) defined early standards, while in reasoning and problem solving, ARC-AGI (Chollet, 2019; Chollet et al., 2025) has become a touchstone. Applied domains have also introduced specialized benchmarks, such as SDBench for medical diagnosis (Nori et al., 2025) and SWE-bench for software engineering (Jimenez et al., 2024). These benchmarks share the common goal of translating an open-ended task into a reproducible and measurable challenge that accelerates research. However, they remain concentrated on perception and reasoning tasks. To our knowledge, no existing benchmark evaluates decision-making under uncertainty in socioeconomic contexts such as venture capital, where outcomes are rare, signals are noisy, and even expert performance is modest. **VCBench** fills this gap by establishing a standardized and anonymized testbed for early-stage startup success prediction.

Venture capital prediction. Several works have explored predictive models for startup success, utilizing a variety of machine learning techniques and data formats (Dellermann et al., 2021; Razaghzadeh Bidgoli et al., 2024; Potanin et al., 2023; Maarouf et al., 2025; Arroyo et al., 2019). Closely related works, Mu et al. (Mu et al., 2025) and Griffin et al. (Griffin et al., 2025), show that founder profiles alone can yield strong predictive signals. However, these studies rely on proprietary or nonstandard datasets, limiting reproducibility and comparability. Unlike previous work, VCBench provides a standardized and anonymized dataset that enables fair comparison across models and establishes human-level baselines for evaluation.

Data anonymization and leakage. A growing body of research highlights risks of data contamination in LLM evaluation. Models may memorize entities and bypass the intended task by re-identification rather than reasoning. In the VC setting, this problem is particularly acute, as founder profiles are often public. VCBench addresses this challenge through a multistage anonymization pipeline validated by adversarial re-identification experiments, achieving a 92% reduction in identifiable founders while preserving predictive features.

Summary. In contrast to previous benchmarks, predictive models and anonymization studies, VCBench integrates these threads into a single resource. It offers the first standardized benchmark for founder-success prediction, designed to evaluate models fairly against human expertise while preventing identity leakage.

3. Dataset

Scale and success definition. VCBench contains 9,000 anonymized founder profiles with a 9% baseline success rate (810 founders). Each founder is paired with their most recently founded company, which determines the success label. A founder is considered successful if the company was acquired or had an initial public offering (IPO) above \$500M valuation, or raised more than \$500M in funding. Companies that raised \$100K–\$4M at inception but did not achieve an exit, IPO, or substantial follow-on funding within eight years of founding are classified as unsuccessful. The eight-year horizon reflects expert practice: it typically takes 8–12 years for startups to reach an IPO, a large-scale funding, or an exit. Those that fail to do so are not considered outlier successes in VC.

Coverage. Most companies were founded in the U.S. between 2010 and 2018. In this period, roughly 5,000 U.S. startups raised more than \$100K annually, suggesting a base population of 45,000. **VCBench** therefore covers about a fifth of this population, making it statistically representative of the early-stage startup landscape in the U.S. Distributions of industries and founding years are presented in Figures 3 and 4 in Appendix B.

Fields. Each founder record contains structured fields capturing founder attributes and company outcomes:

- `success`: Binary label indicating founder success as defined above.
- `industry`: Sector of the startup.
- `ipos` and `acquisitions`: Details of the founder's prior experience in IPOs and exits as a co-founder.
- `educations_json`: List of education records with `degree`, `field`, and `QS ranking`.
- `jobs_json`: List of job records with `role`, `company size`, `industry`, and `duration`.

Sources and filtering. Founder and company data was collected from licensed and publicly available data sources. LinkedIn provides education and job histories. Crunchbase helps with enrichment and contributes IPO, acquisition, and funding records. For each founder, information is restricted to data available prior to the founding of the company, simulating real-world early-stage prediction.

Formats. To accommodate both LLMs and ML models, VCBench is distributed in two complementary formats. The first is `anonymized_prose`, a natural-language summary of each founder constructed from structured fields, designed for direct LLM input. The second format exposes structured JSON fields, enabling feature-level access for custom ML models.

4. Methodology

We construct VCBench in four stages: (i) data coverage improvement, (ii) format standardization and filtering, (iii) anonymization at both the entry and dataset levels, and (iv) iterative field selection guided by adversarial testing. Figure 2 visualizes the full pipeline on an example profile, while Table 2 summarizes vocabulary compression after standardization and filtering.

4.1. Data Coverage Improvement

LinkedIn offers wide coverage but irregular entries. Crunchbase is structured yet incomplete. Empty fields can obscure the signal. For example, an empty acquisition field may indicate the absence of exits or missing information. To mitigate this without speculative inference, we adopt two steps:

1. **Data source cross-checking.** When a field is empty in one source, we retrieve the corresponding value from the other source when a direct match is apparent.
2. **Cross-record consistency.** We enforce consistent values for the same entity across profiles (e.g., the industry label attached to the same organization) to reduce label drift.

These steps reduce irregularities and missing values while maintaining a conservative stance toward uncertain or conflicting entries.

4.2. Format Standardization and Data Filtering

LinkedIn entries contain spelling variants, inconsistent punctuation, and heterogeneous representations of degrees and roles (e.g., "PhD", "p.h.d.", "Doctor of Philosophy"). We adopt a two-step process:

1. **Deterministic canonicalization:** Trim whitespace, normalize conjunctions ("and", "&", "/"), punctuation, and common aliases for degrees and roles.
2. **LLM-assisted flagging:** LLM identifies strings requiring reformatting and tags nonformal entries using explicit exclusion categories (e.g., "Intern", "Course", "Visiting").

Through this process, non-formal education and job entries are flagged and removed to avoid systematic bias. After filtering, founders with no remaining job record are excluded from the dataset to ensure well-formed professional histories. Overall, this stage reduces label fragmentation and consolidates noisy vocabularies while preserving predictive structure (Table 2).

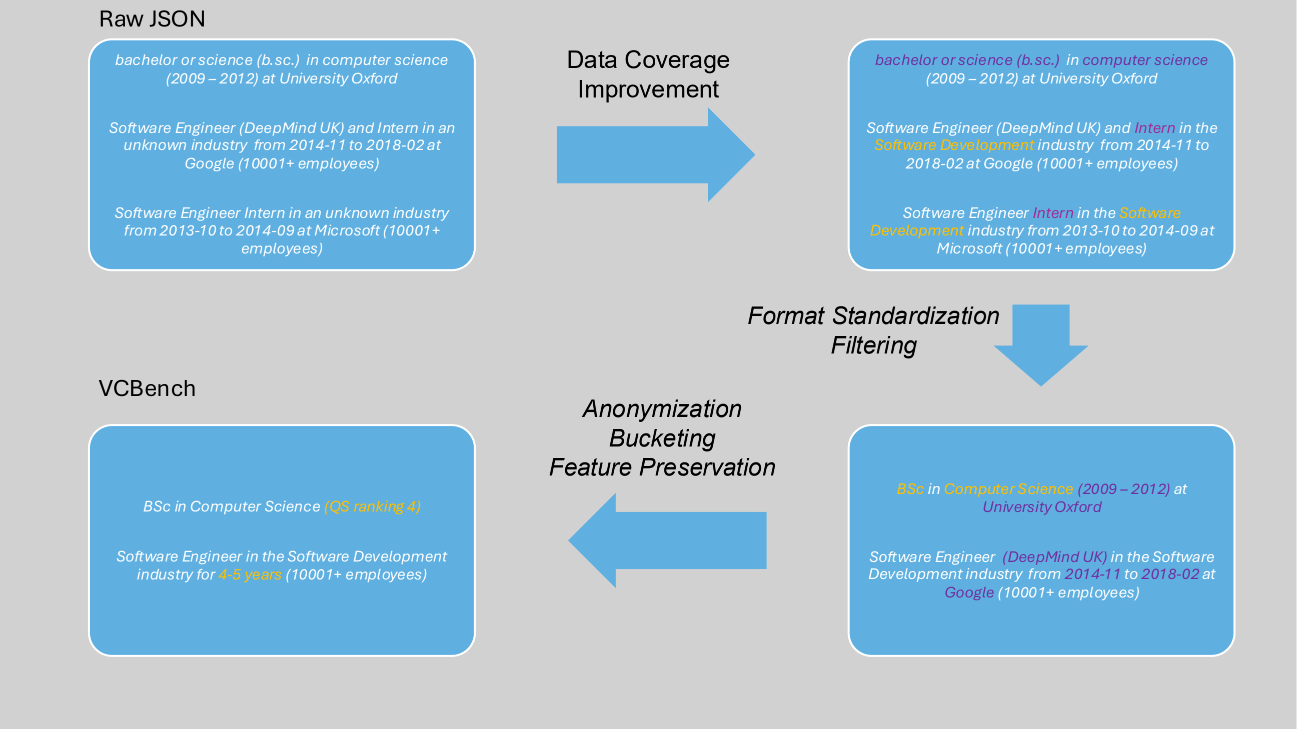
Raw JSON
bachelor or science (b.sc.) in computer science (2009 – 2012) at University Oxford
Software Engineer (DeepMind UK) and Intern in an unknown industry from 2014-11 to 2018-02 at Google (10001+ employees)
Software Engineer Intern in an unknown industry from 2013-10 to 2014-09 at Microsoft (10001+ employees)
Data Coverage Improvement
bachelor or science (b.sc.) in computer science (2009 – 2012) at University Oxford
Software Engineer (DeepMind UK) and Intern in the Software Development industry from 2014-11 to 2018-02 at Google (10001+ employees)
Software Engineer Intern in the Software Development industry from 2013-10 to 2014-09 at Microsoft (10001+ employees)
Format Standardization
Filtering
VCBench
Anonymization
Bucketing
Feature Preservation
BSc in Computer Science (QS ranking 4)
Software Engineer in the Software Development industry for 4-5 years (10001+ employees)
BSc in Computer Science (2009 – 2012) at University Oxford
Software Engineer (DeepMind UK) in the Software Development industry from 2014-11 to 2018-02 at Google (10001+ employees)

Figure 2. Data Cleaning Pipeline

Record Type	Original No. unique entries	Final No. unique entries	Percentage Reduction
industry	314	61	80.6%
education degree	2155	404	81.3%
education field of study	6360	3969	37.6%
job role	21259	16374	23.0%
education record	20573	15620	24.1%
job record	45975	41183	10.4%

Table 2. An entry-level summary of the format standardization and data filtering process.

4.3. Anonymization

We consider three adversary classes by increasing identification capability: general-purpose LLMs (e.g., GPT-4o), reasoning models (e.g., o3), and tool-assisted models with web search. Our approach combines entry-level and dataset-level anonymization.

Entry-level anonymization. We remove founder names, company names, locations, and dates from all fields, including nested text in `educations_json` and `jobs_json`. This prevents direct string lookup (e.g., job titles such as "Software Engineer (Microsoft India)").

Dataset-level anonymization. To reduce linkage via rare attribute combinations, we group simple numeric fields (`ipos` and `acquisitions`). For `industry`, we follow the following procedure:

1. Embed industry labels using an OpenAI embedding model.
2. Apply agglomerative hierarchical clustering with cosine similarity.
3. Refine clusters using o3 reasoning.
4. Perform a final human pass.

This yields 61 clusters, each containing at least 10 founders.

Education prestige is preserved using QS university rankings, with aliases handled by an LLM to unify naming conventions (e.g., "California Institute of Technology" vs. "Caltech"). Job start and end dates are converted into duration buckets, expressed in years, which preserve career trajectory information while concealing exact timelines.

4.4. Iterative Anonymization and Feature Selection Process

We refine the dataset with *anonymization unit tests* that explicitly target founder identification rather than success prediction. Initial experiments (Appendix D) using the o3 model due to its powerful reasoning power (with and without web search) identified leakage pathways such as specific institution names, company names, funding/IPO/acquisition values, and exact dates. We therefore removed or bucketed these fields and re-tested as described in the prior sections.

At each iteration, we adopted the following loop: propose a change, run identification tests, and retain the change only if it reduces re-identification while preserving predictive structure. Founders who were identified on at least two occasions are removed to reduce the risk of data contamination.

5. Experiments and Ablation Analysis

The purpose of the experiments is to validate whether VCBench achieves strong anonymization while preserving predictive signal. We designed *anonymization unit tests* in which models are explicitly instructed to re-identify founders rather than predict success. This setting provides a direct measure of the residual leakage risk.

Experimental design. We sampled 300 successful founders, as these represent the most high-profile cases and thus the hardest anonymization challenge. For each profile, we tested whether an adversary could correctly infer the founder's identity. We employed two adversaries representing distinct threat models: DeepSeek-R1, a reasoning model without web access (offline), and Gemini-2.5-Pro with grounding (web-search, online). The evaluation metric is the proportion of correctly identified founders.

Ablation formats. We compared five input formats, each progressively modifying the representation:

- **JSON:** structured education and job records before anonymization but excluding IPOs and acquisitions.
- **Base (anonymized JSON):** The anonymized format with the founder names and institution names removed, adding IPOs and acquisitions.
- **Base + QS rankings (bucketed):** The anonymized format enriched with QS ranks grouped into intervals (e.g. 1-20, 20-100).
- **Base + QS rankings (unbucketed):** The anonymized format enriched with exact QS ranks.
- **Final Format (VCBench):** The final anonymization format with unbucketed QS ranks and bucketed job durations.

Input	Online %	Offline %
Processed JSON	77.0	17.2
Base	18.3	1.2
QS (bucketed)	15.4	2.3
QS (unbucketed)	12.7	–
Final Format (VCBench)	15.1	1.3

Table 3. Anonymization unit testing results: average identification rates for different input formats. Bold formats were tested three times for reliability.

Results. Identification rates are reported in Table 3. Moving from the pre-anonymization JSON to the final format reduced online identification by approximately 80% and offline identification by 92%. An unexpected but beneficial finding was that explicit QS rankings reduced identification rates further. Reasoning logs showed that models often misused current QS lists, which in turn blurred re-identification. This allowed us to preserve educational prestige while strengthening anonymity.

Representativeness. Our sample of 300 covers more than a third of the successful founder cohort, enhancing the generalizability of our results to the broader successful population. Moreover, small-scale experiments that we ran suggest that the unsuccessful founders are about three times less likely to be identified, supporting our decision to monitor and control the level of anonymization of successful founders.

Interpretation. By removing identifiers, bucketing rare values, and clustering industries, multistage anonymization reduces leakage even under strong adversaries. The final VCBench format balances data contamination with feature diversity, supporting robust benchmarking and fair comparison.

6. The VCBench Leaderboard

After constructing VCBench, we evaluated the predictive performance of nine state-of-the-art LLMs. The dataset of 9,000 founders was split into six folds of 1,500 each, preserving the 9% success rate. Performance is measured using the $F_{0.5}$ score, which weights precision twice as heavily as recall:

$$F_{0.5} = (1 + 0.5^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(0.5^2 \cdot \text{Precision}) + \text{Recall}}.$$

This emphasis reflects the VC setting, where false positives are more costly than false negatives. The average results are reported in Table 4, with fold-specific outcomes in Appendix C.

Model	Precision (%)	Recall (%)	$F_{0.5}$ (%)
GPT-4o	29.1	16.2	**25.1**
DeepSeek-R1	37.6	8.4	22.1
GPT-4o-mini	29.5	10.1	21.2
o3	42.4	7.0	20.9
Gemini-2.5-Pro	17.2	59.0	20.1
Claude-3.5-Haiku	16.9	48.6	19.4
GPT-5	53.7	4.3	16.2
Gemini-2.5-Flash	12.6	**69.1**	15.1
DeepSeek-V3	**59.1**	3.0	11.8

Table 4. Predictive performance of nine LLMs on VCBench, averaged across six folds and ordered by average $F_{0.5}$. Results should be interpreted cautiously when extrapolating from the 9% dataset success rate to the real-world 1.9% success rate.

Results. GPT-4o achieved the highest $F_{0.5}$ (25.1). DeepSeek-V3 recorded the highest precision (59.1%) but with a low recall, while Gemini-2.5-Flash obtained the highest recall (69.1%) but modest precision.

Cost and efficiency. Inference cost and latency are summarized in Table 5. GPT-4o-mini offers strong performance at low cost, while DeepSeek-V3 offers cost-effective reasoning with better performance over other reasoning models such as Gemini-2.5-Pro and GPT-5.

Leakage check. Comparatively high precision obtained on fold 4 documented in Appendix C (Table 6) motivated closer inspection. We reviewed prediction reasoning logs and ran additional anonymization unit tests on 300 founders in that fold. Identification rates remained low (4.3% for o3; 7.3% for Gemini-2.5-Pro with web search), and no explicit identifications appeared in model outputs. The elevated precision was instead due to a concentration of founders with extreme outcomes (e.g., prior IPO/acquisition events above $500M), confirming that improvements reflected predictive signal rather than identity leakage.

Interpretation. These results demonstrate that anonymized founder profiles preserve enough predictive signal for LLMs to outperform human experts in startup investing. GPT-4o achieves the best $F_{0.5}$ score (25.1) with a precision of 29%, representing a 3.2× improvement over the precision baseline and exceeding the 2.9× performance of tier-1 VC firms (Figure 1). DeepSeek and Gemini models highlight different points on the precision–recall frontier: DeepSeek-V3 produces highly precise predictions with low coverage, whereas Gemini-2.5-Flash recalls nearly 70% of successful founders at the expense of precision. Crucially, leakage tests confirm that these gains are not explained by identity re-identification.

Mitigation of Future Leakage. An essential consideration for modern benchmarks is mitigating the risk of leakage into the pre-training corpus of future LLMs, which would introduce inherent biases into model comparisons. We address this by only releasing half of the dataset (4,500 founders) publicly. The remaining half is reserved for private evaluation. Leaderboard scores are computed on the private fold, preserving the integrity of model assessments.

Model	Cost In/Out	Latency
GPT-4o	$2.50 / $10.00	3.59 s
DeepSeek-R1	$0.55 / $2.19	37.83 s
GPT-4o-mini	$0.15 / $0.60	3.04 s
o3	$2.00 / $8.00	6.90 s
Gemini-2.5-Pro	$1.25 / $10.00	10.73 s
Claude-3.5-Haiku	$0.80 / $4.00	3.36 s
GPT-5	$1.25 / $10.00	1.54 s
Gemini-2.5-Flash	$0.30 / $2.50	8.32 s
DeepSeek-V3	$0.27 / $1.10	10.07 s

Table 5. Cost per 1M tokens and latency for models evaluated on VCBench, ordered by average $F_{0.5}$.

7. Discussion

We assess **VCBench** in terms of benchmark validity and design trade-offs. Its interpretation is bounded by several factors: the inflated success rate (9% vs. 1.9% in the real-world), potential sampling bias from LinkedIn/Crunchbase, right-censoring from the eight-year horizon, and residual noise after standardization. To extend utility without weakening predictive value, we suggest anonymized proxies for company prestige, scalable clustering for high-cardinality fields, trajectory-level features that balance privacy and signal, broader model coverage, and a sequential simulation mode for decision-making under resource constraints. Together, these directions clarify how current scores should be read and outline a path toward a more realistic and robust benchmark.

7.1. Limitations

VCBench carries several known caveats:

- **Prevalence shift.** The benchmark sets the success rate at 9% rather than the real-world 1.9%. This inflated prevalence stabilizes testing and improves statistical significance, but also introduces caveats. Precision does not scale linearly with prevalence, so relative multipliers (e.g., 3.2× for GPT-4o versus 2.9× for tier-1 VCs) may not hold under the true 1.9% distribution. Reported gains should therefore be interpreted cautiously when extrapolated back to deployment settings.
- **Comparability of human baselines.** The normal-

ization procedure assumes that human investors and LLMs face the same distribution of opportunities. In reality, VCs self-select their deal flow, and access is constrained by competition, reputation, and human bandwidth. This structural difference means that even after normalization, comparisons between model and human performance may misrepresent the actual gap in practice.

- **Bias in collected data.** Although VCBench covers a statistically significant share of U.S. founders from 2010 to 2018, it still inherits biases from LinkedIn and Crunchbase. Coverage is stronger for technology startups and founders with public visibility, while non-tech sectors, self-funded companies, and underrepresented demographics are more likely to be absent or inconsistently represented. These biases may skew the benchmark toward patterns that generalize poorly beyond the VC mainstream.
- **Temporal bias.** The eight-year horizon used to define success introduces a right-censoring effect. Companies founded later in the window are more likely to be labeled unsuccessful simply because their outcomes have not yet had time to materialize. For example, a startup founded in 2017 is penalized relative to one founded in 2012, which has had additional years to mature and demonstrate success. While the eight-year cutoff reflects the median time to major outcomes such as IPO or acquisition, it inevitably places more recent cohorts at a disadvantage.
- **Residual irregularities.** Despite multistage cleaning, large-scale founder data remains noisy. Stochastic LLM standardization and irregular entries make it difficult to guarantee full consistency.

7.2. Future Work

Several directions can extend VCBench and address the limitations mentioned above. First, continued community-driven refinement will be important. Releasing updates as feedback arrives will improve data coverage and reduce residual irregularities, making VCBench a "living benchmark" rather than a static artifact. Second, anonymization can be enhanced by incorporating company prestige signals (e.g., Fortune-500 or NASDAQ-100) that preserve organizational context without revealing identities. Third, the current clustering approach works well for industries but not for high-cardinality fields like job roles or education. Developing scalable clustering methods would enable more complete dataset-level anonymization.

On the feature side, more advanced feature engineering could transform structural and temporal patterns (such as job sequences, career trajectories, or prestige proxies) into predictive variables that reduce re-identification risk while strengthening model accuracy. On the evaluation side, the leaderboard should expand to include both additional LLMs and domain-specific VC models, enabling richer comparisons across approaches. Finally, moving beyond static classification, we envision a VC-simulation mode in a tournament-style format where startups appear sequentially and investment decisions must be made under budget constraints. This would capture the real dynamics of venture investing and test whether models can make consistent, resource-limited decisions. Looking further, we are exploring a *gamified arena* where human investors, LLMs, and purpose-built VC models compete side by side, creating an open platform for benchmarking decision-making under uncertainty in real time.

8. Conclusion

We introduced **VCBench**, the first standardized and anonymized benchmark for founder-success prediction in venture capital. The dataset was constructed using a multistage anonymization pipeline, validated by adversarial tests, which reduced re-identification risk by over 90% while preserving predictive signal.

Using this benchmark, we evaluated nine state-of-the-art LLMs and found that several outperform not only the market index but also the leading VC firms, with GPT-4o achieving the highest $F_{0.5}$ score. These results show that anonymized founder profiles are sufficient to surpass human-expert baselines in early-stage venture forecasting.

By releasing both the dataset and a public leaderboard, we provide a foundation for reproducible research in this high-stakes domain. **VCBench** is designed as a community-driven benchmark that will evolve with feedback, richer features, and new evaluation modes, including simulation and human–AI competitions, offering a path toward more realistic tests of decision-making under uncertainty.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. Vqa: Visual question answering, 2016. URL https://arxiv.org/abs/1505.00468.

Arroyo, J., Corea, F., Jiménez-Díaz, G., and Recio-García, J. Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, PP:1–1, 08 2019. doi: 10.1109/ACCESS.2019.2938659.

Chollet, F. On the measure of intelligence, 2019. URL https://arxiv.org/abs/1911.01547.

Chollet, F., Knoop, M., Kamradt, G., Landers, B., and Pinkard, H. Arc-agi-2: A new challenge for frontier ai reasoning systems, 2025. URL https://arxiv.org/abs/2505.11831.

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., and Leimeister, J. M. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method, 2021. URL https://arxiv.org/abs/2105.03360.

Griffin, B., Ternasky, J., Alican, F., and Ihlamur, Y. Random rule forest (rrf): Interpretable ensembles of llm-generated questions for predicting startup success, 2025. URL https://arxiv.org/abs/2505.24622.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Maarouf, A., Feuerriegel, S., and Pröllochs, N. A fused large language model for predicting startup success. *European Journal of Operational Research*, 322(1): 198–214, April 2025. ISSN 0377-2217. doi: 10.1016/j.ejor.2024.09.011. URL http://dx.doi.org/10.1016/j.ejor.2024.09.011.

Mu, X., Ternasky, J., Alican, F., and Ihlamur, Y. Policy induction: Predicting startup success via explainable memory-augmented in-context learning, 2025. URL https://arxiv.org/abs/2505.21427.

Nori, H., Daswani, M., Kelly, C., Lundberg, S., Ribeiro, M. T., Wilson, M., Liu, X., Sounderajah, V., Carlson, J., Lungren, M. P., Gross, B., Hames, P., Suleyman, M., King, D., and Horvitz, E. Sequential diagnosis with language models, 2025. URL https://arxiv.org/abs/2506.22405.

Potanin, M., Chertok, A., Zorin, K., and Shtabtsovsky, C. Startup success prediction and vc portfolio simulation using crunchbase data, 2023. URL https://arxiv.org/abs/2309.15552.

Razaghzadeh Bidgoli, M., Raeesi Vanani, I., and Goodarzi, M. Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship*, 13:80, 2024. doi: 10.1186/s13731-024-00436-x. URL https://doi.org/10.1186/s13731-024-00436-x.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Appendix

A. Prompts and Input Formats

A.1. Example raw JSON founder profile

```
{
  "industry": "Research Services",
  "twitter_url": null,
  "jobs": [
    {
      "title": "Professor",
      "company": "Duke University",
      "company_industry": "Higher Education
    ",
      "started_at": "2013-07-13",
      "ended_at": "still working"
    },
    {
      "title": "Founder, President, CEO",
      "company": "Applied Quantum
    Technologies",
      "company_industry": null,
      "started_at": "2006-08-13",
      "ended_at": "2020-02-13"
    },
    {
      "title": "Associate Professor",
      "company": "Duke University",
      "company_industry": "Higher Education
    ",
      "started_at": "2010-07-13",
      "ended_at": "2013-06-13"
    },
    {
      "title": "Assistant Professor",
      "company": "Duke University",
      "company_industry": "Higher Education
    ",
      "started_at": "2004-06-13",
      "ended_at": "2010-06-13"
    },
    {
      "title": "MTS and Technical Manager",
      "company": "Lucent Technologies /
    Bell Labs",
      "company_industry": "
    Telecommunications",
      "started_at": "1999-03-13",
      "ended_at": "2004-03-13"
    }
  ],
  "educations": [
    {
      "university": "Young Dong High School
    ",
      "degree": null,
      "fields": null,
      "started_on": null,
      "ended_on": null
    }
  ]
}
```

A.2. Example anonymized prose format

```
"This founder leads a startup in the IT
    Services & Digital Solutions industry.
Education:
* BA in History (Institution QS rank 42)

Professional experience:
* Board Member for <2 years in the `
    Management, Strategy & Professional
    Services` industry (51-200 employees)
* Board Member for 2-3 years in the `
    Education & Training Services` industry
     (51-200 employees)
* Angel Investor, Advisor for 4-5 years
* VP for <2 years in the `Software
    Development` industry (1001-5000
    employees)
* Co-Founder, CEO for 6-9 years in the `
    Software Development` industry (51-200
    employees)
* Board Member for <2 years
* Director (Corporate Development) for <2
    years
* Co-Founder for <2 years

They have overseen one acquisition as a
    founder: one 15M - 50M in value."
```

A.3. Anonymization unit testing prompt

```
You are given the profile of an
    entrepreneur who founded at least one
    startup after 2010. This includes
    information like the founder's
    education and job histories.

### Your Task
1. Examine the profile in JSON format below
    .
2. Deduce the most likely identity of the
    founder (using web search as an aid).
3. Respond in the exact format specified
    under "Required Output".

### OUTPUT FORMAT
Put the founder's full name right at the
    beginning. Avoid abbreviations or
    nicknames for first and last names. Do
    not include middle names. Do not
    include any other text before it.
Do not return the text "Founder's name".
    Always give a guess, even if you are
    not sure.
In addition, include the following
    information:
Reasoning: <one brief paragraph explaining
    how the data led you to this name>
Confidence: <Low, Medium, High>

### Inputs
{profile}
```

B. VCBench Distributions

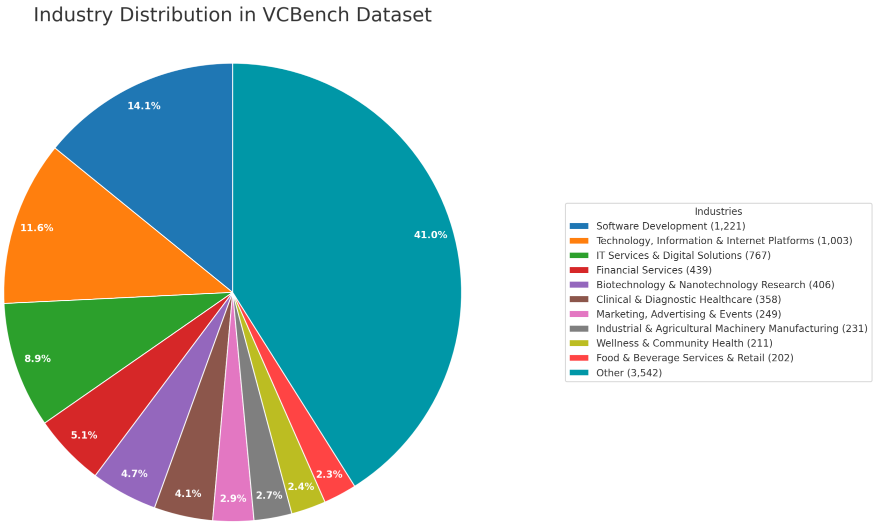

Figure 3. Distribution of industries in VCBench after bucketing.

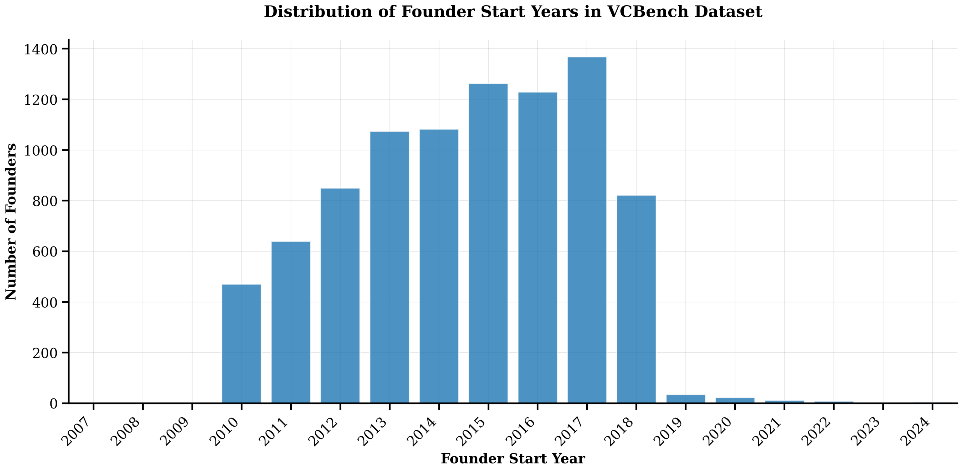

Figure 4. Distribution of startup founding years in VCBench.

C. Per-fold Results for Vanilla LLMs

Table 6 reports per-fold precision, recall, and $F_{0.5}$ for each model (six folds; 1,500 founders per fold; 9% prevalence). Averages match with Table 4 in the main text.

Model	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Fold 6 (%)	Avg (%)
Claude-3.5-Haiku							
Precision	17.0	17.1	17.6	17.5	12.9	**19.5**	16.9
Recall	50.4	50.4	51.1	48.1	37.8	**54.1**	48.6
$F_{0.5}$	19.6	19.7	20.2	20.0	14.8	**22.3**	19.4
DeepSeek-V3							
Precision	77.8	50.0	**–**	42.9	**83.3**	41.7	59.1
Recall	**5.2**	3.0	**0.0**	2.2	3.7	3.7	3.0
$F_{0.5}$	**20.5**	12.0	**0.0**	9.2	15.7	13.7	11.8
DeepSeek-R1							
Precision	27.0	30.0	22.2	**50.0**	46.2	**50.0**	37.6
Recall	7.4	6.7	4.4	11.1	8.9	**11.9**	8.4
$F_{0.5}$	17.7	17.6	12.4	29.4	25.1	**30.4**	22.1
Gemini-2.5-Flash							
Precision	**13.9**	12.7	13.1	12.6	10.5	12.9	12.6
Recall	**74.8**	72.6	72.6	66.7	57.8	70.4	69.1
$F_{0.5}$	**16.6**	15.3	15.7	15.0	12.5	15.5	15.1
Gemini-2.5-Pro							
Precision	**17.9**	17.5	**17.9**	17.3	15.5	17.5	17.2
Recall	60.7	**62.2**	**62.2**	57.0	51.1	60.7	59.0
$F_{0.5}$	**20.9**	20.4	20.8	20.1	18.0	20.4	20.1
GPT-4o-mini							
Precision	**35.2**	26.5	27.5	26.5	31.9	29.4	29.5
Recall	**14.1**	6.7	8.2	9.6	11.1	11.1	10.1
$F_{0.5}$	**27.1**	16.6	18.6	19.6	23.2	22.1	21.2
GPT-4o							
Precision	31.1	25.7	**31.5**	27.9	27.4	31.0	29.1
Recall	17.0	14.1	17.0	14.1	14.8	**20.0**	16.2
$F_{0.5}$	26.7	22.0	26.9	23.3	23.4	**28.0**	25.1
GPT-5							
Precision	**63.6**	33.3	50.0	58.3	**63.6**	53.3	53.7
Recall	5.2	2.2	2.2	5.2	5.2	**5.9**	4.3
$F_{0.5}$	19.6	8.8	9.4	19.1	19.6	**20.5**	16.2
o3							
Precision	35.5	26.1	35.7	47.1	**58.3**	52.0	42.4
Recall	8.2	4.4	3.7	5.9	**10.4**	9.6	7.0
$F_{0.5}$	21.2	13.2	13.1	19.7	**30.3**	27.7	20.9

Table 6. Per-fold and average results for all models. All results are recorded as percentages. Note that DeepSeek-V3 encountered an undefined precision on Fold 3 (no positives predicted), so the average precision was calculated across the remaining 5 folds; average recall and $F_{0.5}$ were still calculated across all folds.

D. Example Initial Experiment with o3

The following demonstrates one of the initial founder identification experiments. Using web searches, the OpenAI o3 model correctly identified the given founder in under two minutes, despite the absence of any explicit identifier. The exceptionally successful exit was the point of failure in this case.

```
User:
You are an expert in venture capital tasked
     with identifying successful founders
    from their unsuccessful counterparts.
    All founders under consideration are
    sourced from LinkedIn profiles of
    companies that have raised between
    $100K and $4M in funding. A successful
    founder is defined as one whose company
     has achieved either an exit or IPO
    valued at over $500M. Given the
    following founder {'Industry': '
    Biotechnology Research', 'Education':
    [], 'Work Experience': [{'title': 'VP
    Applications', 'company': 'consumer
    genomics unicorn (>$3 B exit)', '
    company_industry': 'Biotechnology
    Research', 'started_at': '2010', '
    ended_at': '2012'}, {'title': 'First
    Employee, Founding Architect, Director
    of R&D', 'company': 'digital PCR
    startup (>$150 M M&A exit)', '
    company_industry': 'Biotechnology
    Research', 'started_at': '2006', '
    ended_at': '2010'}]}

Who is this founder?"


OpenAI o3:
I'm considering several possibilities for
    the "consumer genomics unicorn" with a
    >$3B exit. Ancestry.com was sold in
    2012 for $1.6B, and later for $4.7B in
    2020, which doesn't fit. The Grail spin
    -off from Illumina could work, but its
    exit didn't happen until 2020 with a
    $7B deal. Sequenom, a PCR startup,
    raised $150M for an exit around 2010
    but was sold in 2017. Natera could also
     be a contender, though its valuation
    was smaller initially. Still, the Grail
     sale seems the closest.

(...)

The search results point to Serge Saxonov
    as matching both the QuantaLife and 23
    andMe roles, though the timeline is
    reversed in the data. This could be a
    misorder in the dataset, but the
    identity seems to be correct. I also
    considered Ben Hindson, though he didn'
    t work at 23andMe. It looks like Serge
Saxonov is the founder. Next step: let'
s confirm by checking for other people
who may have had roles at both
QuantaLife and 23andMe.
```