VCBench - Benchmarking LLMs in Venture Capital

Rick Chen [1] **Joseph Ternasky** [2] **Fuat Alican** [2] **Aaron Ontoyin Yin** [2] **Yigit Ihlamur** [2]

Abstract

Large-Language-Models (LLMs) have been extensively benchmarked against human performance in domains such as software engineering (SWE-bench), medical diagnosis (SDBench) and general reasoning (ARC-AGI). Venture capital (VC) represents a high-stake domain of critical decision-making, in which the use of LLMs remains comparatively under-exploited. Despite its importance, there is no publicly available, standardised, and anonymised dataset for systematically evaluating predictive models. In this paper, we introduce VCBench, a standardised benchmark dataset comprising 9 000 anonymised founder profiles. The dataset construction process integrates deterministic hard-coding (reliable) with LLM-reasoning (practical), enabling flexible implementations of data standardisation, enrichment, filtering, and clustering. To evaluate the robustness of anonymisation, we conducted iterative founder re-identification experiments as an ablation study, achieving an approximately **80%** reduction in the number of identifiable founders. We further established reference LLM prediction levels by running nine state-of-the-art LLMs through the benchmark; the best average precision with respect to the baseline was obtained by GPT-5 ($\mathbf{6.4}\times$), while o3 performed with the highest $F_{0.5}$ score. Finally, we release a public VCBench leaderboard to facilitate future experimentations with both LLMs and specialised VC-prediction systems.

1. Introduction

Benchmark studies are essential for evaluating and comparing different models on a shared task. The construction of standardised, reliable datasets is critical to ensuring the credibility of experimental conclusions. Traditional benchmark datasets have served as important guidance and driving forces in AI development (Agrawal et al., 2016; Russakovsky et al., 2015). In the new age of Large-Language-Models (LLMs), the demand for diverse, high-quality datasets has increased substantially. Numerous LLM-specific benchmarks have emerged in response, spanning domains including AGI (Chollet, 2019; Chollet et al., 2025), medical diagnosis (Nori et al., 2025) and software engineering (Jimenez et al., 2024).

Venture capital is another domain in which LLMs have high-potential to support critical decision making. Predicting the success of early-stage startups is valuable for optimising resource allocation, yet challenging due to limited available information. Recent models such as (Mu et al., 2025) and (Griffin et al., 2025) have demonstrated high-precision predictions based only on founder profiles. However, there is an ongoing need for a well-regulated testing ground for different founder prediction models, including vanilla LLMs. Developing such a benchmark for venture capital faces several intrinsic and distinctive challenges, as highlighted in Table 1. One major data source, LinkedIn, provides extensive coverage but suffers from irregularities stemming from its unregulated nature. Conversely, Crunchbase offers reliable and valuable business metrics but exhibits limited coverage. More specifically, there are four key contributing issues:

- **Data Format Irregularity**: The same entity or founder status could have different representations in the dataset, for example PhD vs p.h.d. vs Doctor of Philosophy.
- **Data Entry Irregularity**: Many data entries require further filtering, for example non-university level degrees (e.g. high school experiences, courses) or non-full-time jobs (e.g. internships).
- **Data Coverage Imbalance**: The data collection process yields many empty entries, leading to uneven coverage across entities.
- **Data Contamination in LLMs**: Both our internal experiments (to be published) and the experiments in this paper demonstrate that LLMs are able to identify founders from profile descriptions alone, even

[1]Department of Mathematics, University of Oxford, Oxford, United Kingdom [2]Vela Research, San Francisco, United States. Correspondence to: Rick Chen <rick.chen@seh.ox.ac.uk>, Yigit Ihlamur <yigit@vela.partners>.

with founder names removed. This would likely offset model evaluations, where an LLM leverages its pre-training corpus to deduce founder identities, and thereby bypasses the intended prediction challenge.

In this paper, we introduce VCBench – a regularised, anonymised dataset designed to address these issues. The dataset comprises 9 000 founder records sourced from LinkedIn and Crunchbase, 810 (9%) of which are labelled as successful. Its primary objective is to prevent LLM founder identifications while preserving sufficient features for effective founder predictions. In the Methodology section, we outline our data-cleaning pipeline, which includes format standardisation, data filtering, data enrichment, data anonymisation and feature preservation (Figure 4). This pipeline may be generalised to other domains facing similar data collection challenges.

To balance anonymisation with precision retention, we implemented an iterative unit-testing process to determine the final set of fields to include. This involves explicitly prompting LLMs to attempt founder identifications, thereby allowing us to investigate and refine the security level of the dataset. Our ablation analysis demonstrated an approximately **80%** reduction in the number of identified founders when web-search is enabled, and an over **95%** reduction when web-search is disallowed.

Finally, we evaluated the predictive performance of nine state-of-the-art LLMs. Figure 1 illustrates the results: o3 topped the cohort with the highest $F_{0.5}$ score and $\mathbf{5.9}\times$ the baseline precision. This marks the beginning of our publicly accessible leaderboard, enabling continuous evaluation of both state-of-the-art LLMs and human-designed systems developed by VC firms.

Our Contributions:

- **Anonymised benchmark for VC predictions**
- **Generalisable data-cleaning and anonymisation methods**
- **Publicly accessible leaderboard for VC models**

2. Related Work

Existing Benchmarks for LLMs Numerous benchmarks have been constructed to assess the capabilities of LLMs across diverse domains. The well-regarded ARC-AGI benchmarks (Chollet, 2019; Chollet et al., 2025) display the progress towards Artificial-General-Intelligence (AGI) by challenging LLMs with carefully designed puzzles. SWE-bench (Jimenez et al., 2024) focuses on the ability of LLMs to address real-world Github coding issues, while the recent SDBench (Nori et al., 2025) is designed for evaluating LLM medical diagnosis.

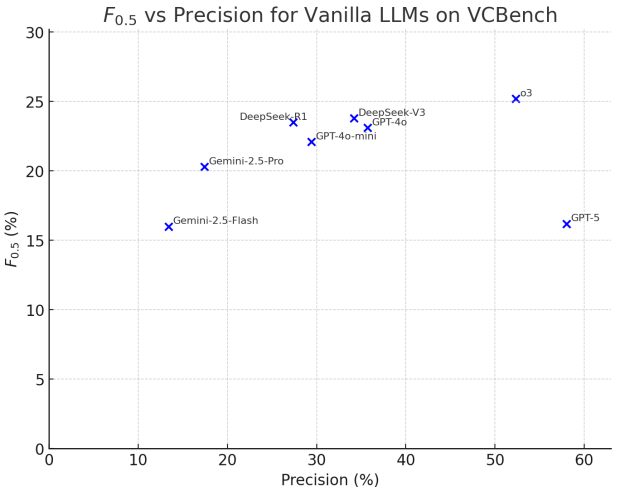

Figure 1. Predictive performances nine vanilla LLMs on VCBench

While these existing benchmarks offer valuable inspirations, Table 1 highlights the distinct combination of challenges faced by VCBench. ARC-AGI comprises human-designed puzzles (Chollet, 2019; Chollet et al., 2025), thereby bypassing all data-cleaning challenges. SWE-bench applied extensive data filtering (Jimenez et al., 2024) but benefitted from the inherently regulated structure of Github repositories and programming code. Medical diagnosis – a domain comparatively similar to VC – demands expert-level anonymisation to protect patient identities. While SDBench introduces synthetic data generation to counter data coverage irregularities, its source of data was already anonymised. Moreover, feature preservation is comparatively straightforward in medical contexts, as anonymisation tends to remove less information; in contrast, within VC, identifiers such as company and education institute names are usually important predictive features themselves.

Startup Investment with LLMs: LLMs have recently demonstrated their potential as VC analysts. Three different LLM-powered VC prediction models – (Xiong et al., 2024), (Griffin et al., 2025) and (Mu et al., 2025) – managed to surpass baseline and average VC success rates by substantial margins (up to $20\times$ baseline success rate). These models represent specialised, human-designed VC models that VCBench is intended to support. Using VCBench, future research could calibrate the performances of existing VC models through reduced data contamination and improved data quality and consistency. We also hope that VCBench will provide a fair environment for bespoke model comparisons, supporting and accelerating VC model developments.

Data Contamination The primary purpose of anonymisation in VCBench is to eliminate data contamination/leakage, a phenomenon where LLMs access known information from

Table 1. Comparisons with Existing LLM Benchmarks

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

their pre-training corpus, leading to inflated model performances. A detailed introduction and review of the concept of data contamination can be found in (Palavalli et al., 2024).

3. Dataset

The dataset consists of 9 000 founders with a 9% baseline success rate (810 successful founders). Each founder is paired with their most recently founded company, which is used to determine the success of the founder. All companies 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, or raised more than $500M in funding.**

The majority of records correspond to companies founded in the U.S. between 2010 and 2018 (Figure 3). Each year, roughly 5,000 U.S. startups raisee more than $100K in funding; this suggests a base population size of 45,000. Our dataset therefore covers roughly a fifth of this population and is statistically representative.

Each founder record has the following associated fields:

```
[anonymised_uuid,success,industry,
linkedin_num_of_connections,ipos,
acquisitions,educations_json,
jobs_json,anonymised_prose]
```

Founder and organisation data are initially collected from LinkedIn (public) and Crunchbase (licenced) via their respective APIs, with responses returned in JSON format. LinkedIn served as the primary source for education and job descriptions, while Crunchbase served as a backup source for these. In addition, Crunchbase provides valuable business metrics such as previous IPOs, acquisitions and funding records. For each founder, their profile is filtered to only include information before the founding date of the founder's company in inspection. This effectively simulates the lack of information in early-startup predictions.

To accommodate both LLMs and human-designed VC models, VCBench offers two available formats. The first, `anonymised_prose`, is an anonymised summary of the founder in natural language. The summary is composed using hard-code based on the other fields in the dataset. This format is designed for direct LLM usage. The second format comprises the remaining columns, with education and job records stored in JSON. This format offers precise feature-level access for primarily human-designed, bespoke VC prediction models. Each education record includes three fields: `degree`, `field`, `qs_ranking`. Each job record includes four fields: `role`, `company_size`, `industry`, `duration`. The distributions of the top industries and founding years in VCBench are presented in Figure 2 and 3 respectively.

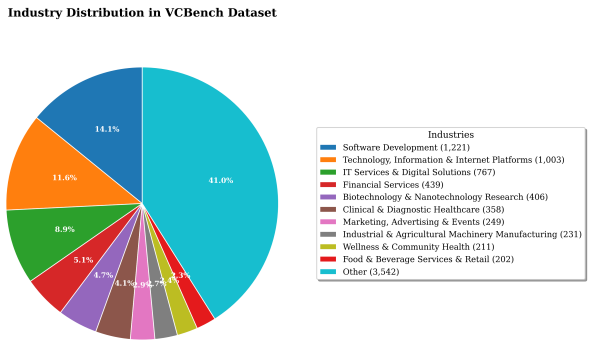

Figure 2. Industries in VCBench

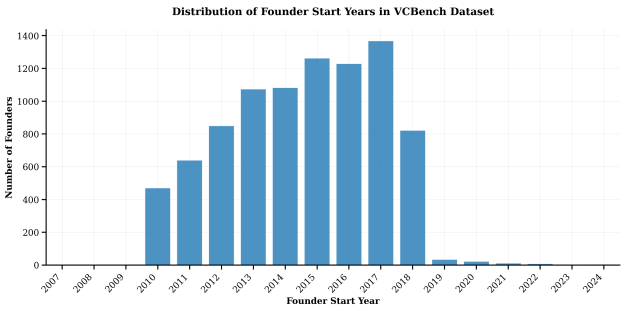

Figure 3. Founding years in VCBench

4. Methodology

In this section, we outline our methodology for constructing VCBench. Figure 4 visualises the data cleaning pipeline for an explicit example profile.

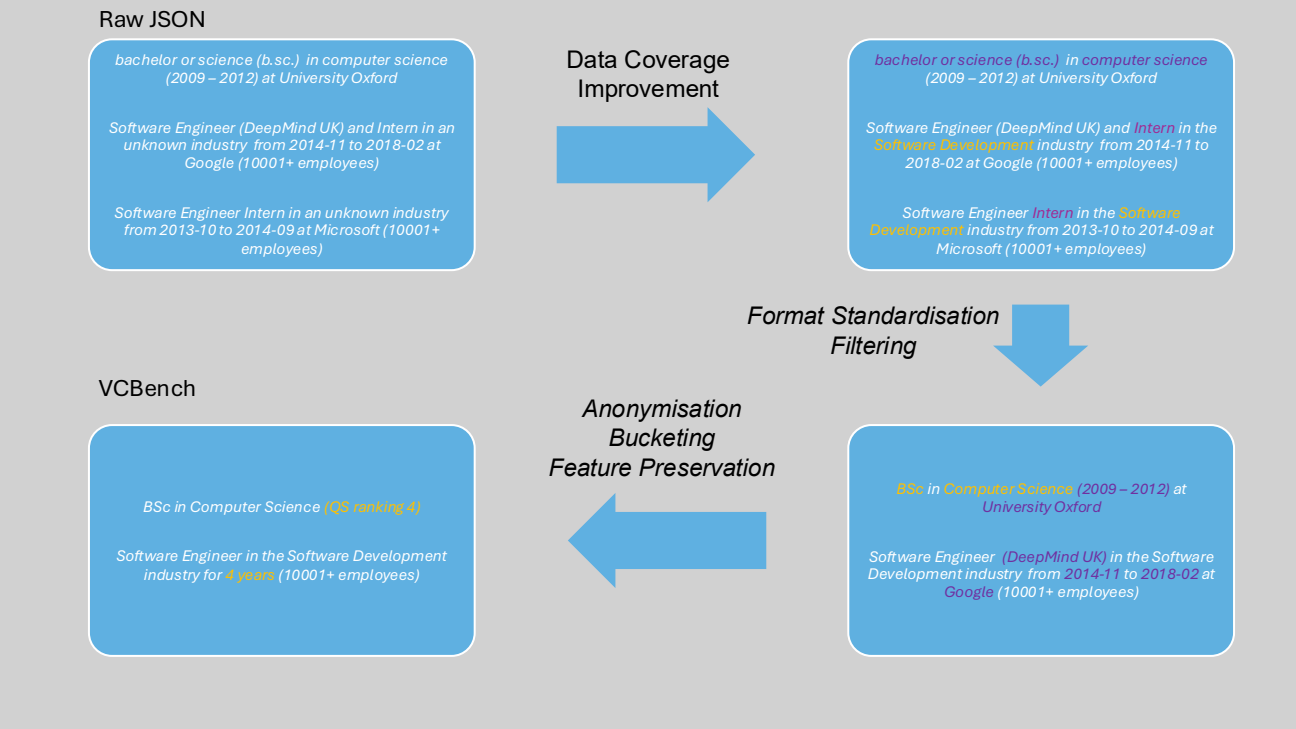

Figure 4. Data Cleaning Pipeline

4.1. Data Coverage Improvement

The unregulated nature of LinkedIn data and the limited coverage of Crunchbase data called for additional efforts to obtain fair and representative data. Preliminary analysis revealed a significant proportion of empty fields in the collected data. For example, an empty Crunchbase acquisition string could indicate either that a founder has no acquisition experience or that such information is missing from the Crunchbase record. The latter scenario could introduce significant irregularities to the field, deteriorating its predictive effectiveness.

Although full elimination of this issue is unlikely in the VC context, we implemented pragmatic field-filling methods to mitigate the impact of missing data. These included:

- **Data Source Cross-Checking:** For example, if founderA has a missing LinkedIn industry field, we would try to retrieve the industry from Crunchbase by matching records.
- **Cross-record Consistency:** For example, if Google is given the industry ”Software Development” under founderA’s job record and the empty industry under founderB’s record, we would ensure consistency by pulling existing industries across records.

4.2. Format Standardisation and Data Filtering

LinkedIn data is based on user input, which poses challenges for systematic data analysis. Entries may contain additional whitespaces, spelling errors, or non-English languages. Conjunctions such as ”and” appear in various forms (e.g. &, commas, /). The same academic degree may have more than 20 different representations in the dataset, for instance ”PhD” vs. ”p.h.d.” vs. ”Doctor of Philosophy”.

In addition, many founders include non-university level degrees, non-full-time jobs or internships in their record. To avoid introducing bias towards these founders, such records need to be identified and removed. This is especially challenging for education degrees, where ambiguous course names (often abbreviated) make filtering difficult.

We adopted the practical solution of prompting an LLM to address these issues; the likelihood of hallucination in reformatting is low, while the reasoning capabilities of LLMs are valuable for distinguishing formal degrees and job experiences from non-formal ones. Specifically, the LLM was instructed to identify fields that required reformatting or modifying and return the modified version. To eliminate invalid entries, we implemented a two-step procedure: **(i)** the LLM is prompted with specific exlusion categories (e.g. ”Intern”, ”Course”, ”Visiting”), and **(ii)** records flagged with these labels are deterministically removed from the dataset.

This approach has the advantage that certain features (e.g. experience in an accelerator programme) can be easily reintroduced if desired.

At the end of the filtering process, we removed founders with no remaining job record. Overall, the combined process of format standardisation and data filtering removed and compacted significant proportions of records (Table 2), improving data quality.

4.3. Anonymisation

We conceptualise anonymisation as a two-part process. In order of increasing identification capability, currently available LLMs can be classified into three categories: general-purpose LLMs (e.g. GPT-4o), reasoning models (e.g. o3) and tool-assisted models (e.g. Gemini Grounding with Google Search). Removing all explicit identifiers such as founder and company names is essential, a process which we label as "entry-level anonymisation". This mitigates founder re-identifications for weaker LLMs. However, stronger LLMs – particularly those augmented with web-search – could still infer founder identities through pattern matching or by exploiting rare combination of founder traits. To counter these models, dataset-level anonymisation is required in addition to reduce the presence of such rare patterns.

Dataset-Level Anonymisation Traditional datasets pushed forward the concept of k-anonymity (Domingo-Ferrer & Torra, 2008), roughly speaking a state where an individual record is indistinguishable from at least $k-1$ others. Although $k-anonymity$ has been criticised for its limitations and ambiguous definitions (Domingo-Ferrer & Torra, 2008) , it remains a reasonable strategy for mitigating LLM re-identifications. That being said, due to the individuality of founder profiles, achieving full k-anonymity for VCBench is highly improbable without extensive feature removal.

We bucketed simple numerical or count-based fields like `linkedin_num_of_connections`, `ipos` and `acquisitions`, since precise values in these fields could otherwise lead to immediate identification.

For `industry` and the industry field in `jobs_json`, we deployed the following procedure:

1. Conversion of industry labels into word vectors via the OpenAI embedding model;
2. Agglomerative hierarchical clustering based on cosine similarity;
3. Iterative cluster refinement using o3 reasoning;
4. Final adjustment through human inspection.

This resulted in 61 industry clusters, each associated with at least 10 different founders.

Unfortunately, this process cannot be generalised to more complex fields such as job roles or education fields. Even after standardising the format, word embedding often gives unsatisfactory clustering results; these results cannot be easily refined as the number of unique entries is too large for either human or o3 inspection. For these fields, we applied only data standardisation as a primitive way to reduce the number of unique entries.

Entry-Level Anonymisation and Feature Preservation Following all preceding preprocessing steps, we removed all founder names, company names, locations and dates from the dataset. A critical consideration is to remove such identifiers from `educations_json` and `jobs_json` as well. For instance, two explicit identifiers exist in the job role "Software Engineer (Microsoft India)". This task was delegated the same LLM employed during the format standardisation and filtering step. The more involved question was how to preserve valuable predictive features in this anonymisation process.

For education records, we incorporated institution prestige by retrieving QS university rankings, based in the 2000s to align with the graduation period of the founders in the dataset. To account for variations in naming conventions (e.g. California Institute of Technology vs. Caltech) and prominent sub-institutions (e.g. Harvard Business School), we prompted an LLM to provide aliases for each ranked institution

For job records, the start and end dates were converted into durations in years and then bucketed. This preserves important signals such as career growth trajectories. Due to time constraints, we did not incorporate company prestige following the removal of company names – we leave this for future iterations of the dataset.

4.4. Iterative Anonymisation and Feature Selection Process

We carefully selected the fields in VCBench through iterative refinement and adversarial testing. At the beginning of the project, we use the o3 chat interface to identify several selected founders based on their profiles; this included both web-search enabled and disabled tests. We then analysed the model reasoning logs to understand how identifications were made, and asked o3 to suggest input-formatting recommendations that would mitigate future identifications. This process quickly ruled out the inclusion of any specific institution name, company name, funding/ipo/acquisition value, and date. Furthermore, it revealed that even a coarse employment outline could enable identifications of the most high-profile founders. Consequently, VCBench adopted a bal-

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. A entry-level summary of the format standardisation and data filtering process.

anced design, tolerating low-level founder re-identification risks in exchange for more predictive feature and more structurally realistic founder profiles.

To evaluate the validity of anonymisation, we instructed LLMs to focus exclusively on founder identification rather than success prediction. In the Experimental section, we present the stage-by-stage testing results as an ablation analysis, which provide empirical support for to the final field selection.

5. Experiments and Ablation Analysis

We iteratively refined the trade-off between anonymisation and feature preservation through anonymisation unit tests on a sample of 300 successful founders from the collected data. For each founder profile, we prompted an LLM to infer the most likely identity of the founder, requesting for both a predicted name and an accompanying reasoning log. Our prompt is attached in Appendix B. Since the most high-profile founders tend to be successful ones, robust anonymisation for this statistically representative subset would likely extend to unsuccessful ones too. We experimented with five profile formats:

- **Processed JSON (before anonymisation)**: contained standardised and filtered education and job records in JSON format, together with `industry` and `linkedin_num_of_connections`, but excluding ipos and acquisitions.
- **Base (anonymised JSON)**: The base anonymised prose format, obtained by removing founder and company names, as well as adding `ipo` and `acquisition`.
- **Base + QS rankings with bucketing**: The base format enriched with QS rankings of education institutions, bucketed into ranges (e.g., 1-20, 20-100).
- **Base + QS rankings without bucketing**: The base anonymised prose format with precise QS rankings included.
- **Final Format (VCBench)**: the eventual format adopted for VCBench, stored under the column `anonymised_prose`. This includes the base format, unbucketed QS rankings, and bucketed job durations.

Note that founder names were never intentionally provided in any format, though they could appear unintentionally within job records for the Processed JSON format, for instance when the founder has a previous organisation named after them.

For the anonymisation tests, we employed two models: DeepSeek-R1, representing an offline reasoning model, and Gemini-2.5-pro with grounding (web-search), representing an online tool-assisted model. Together, these simulate the two most likely identity-attacks on the dataset. The identification rates across different input formats are presented in Table 3. Moving from the initial pre-anonymisation version to the final format yielded reducations of roughly 80% (resp. 92%) reduction in the number of online (resp. offline) identified founders.

An unexpected finding was the fact that online identification rates dropped significantly after including explicit QS rankings. Examination of reasoning logs revealed that LLMs occasionally misuse explicit QS rankings by consulting the most recent QS ranking list, and thus inferring incorrect institutions. This effect is beneficial for our purposes, as it allowed us to preserve institution prestige as a feature while simultaneously lowering re-identification rates.

Given the capacity to trade dataset size for stronger anonymisation, we removed a large portion of the re-identified founders. We prioritised removing founders who were identified on at least two occasions.

6. The VCBench Leaderboard

After constructing VCBench, we evaluated the predictive performance of nine state-of-the-art LLM models on the dataset. The 9 000 founders were divided into six folds, each containing 1 500 founders with a 9% baseline success rate (135 successful founders). Model performances were evaluated primarily using the $F_{0.5}$-score, a balanced

Input Format	Online Test Identification %	Offline Test Identification %
Processed JSON	77.0	17.2
Base	18.3	1.2
Base + QS rankings with bucketing	15.4	2.3
Base + QS rankings without bucketing	12.7	–
Final Format (VCBench)	15.1	1.3

Table 3. Anonymisation unit testing results – average identification rates for different input formats. The formats in bold were tested three times to ensure reliability.

domain-appropriate metric combining precision and recall as follows:

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

This gives precision – our secondary metric – twice the weight. Such emphasis is well alighed with the VC context, where false positives are more costly than false negatives. The average precisions, recalls, and $F_{0.5}$-scores are presented in table 4, while fold-specific results are provided in Appendix A.

After observing strong performances by the OpenAI models on fold 4, we specifically inspected the prediction reasoning logs for those founders. In addition, we performed anonymisation unit tests on a sample of 300 founders in fold 4, with o3 and Gemini-2.5-Pro with Grounding obtaining identification rates of 4.3% and 7.3% respectively. Importantly, no explicit signs of identification were found in the prediction reasoning logs, and a significant portion of those founders had prior IPO/acquisition experiences exceeding $500M. These evidence suggest that the elevated precision scores were driven primarily by the prevalence of extreme success cases rather than identification breaches.

7. Discussion

7.1. Limitations

- **Inflated Success Rate**: In reality, only about 1.9% of early-startups meet our success criterion. However, we adopted the higher success rate due to the current performance level of vanilla LLMs – it can otherwise lead to unstable testing results. This limits the extent to which the model performances generalise to real-world predictions.
- **Inherent bias in collected data**: Although VCBench covers a statistically significant subset of U.S. based founders winthin the target time frame, there could still be bias introduced by differences in the underlying distributions.
- **Inherent bias from founding year**: Successful companies typically take 8 years to reach the $500M success threshold. As a result, later founded companies may be penalised unfairly, as their success might not have materialised yet in the time frame.
- **Residue data irregularity** Despite our efforts, it is extremely difficult to eliminate (and verify) that all data irregularities have been removed, due to the large volume of data, the stochastic nature of LLMs, and significant levels of data irregularities from LinkedIn and Crunchbase.

7.2. Future Work

- **Preservation of company prestige**: Incorporate company prestige (e.g. NASDAQ-100, Fortune-500) by pulling relevant information before removing company names.
- **Better clustering method**: Develop a clustering technique generalisable to fields with large numbers of unique entries, which would enable reliable dataset-level anonymisation for thosee fields with minimal human-intervention.
- **Feature engineering**: Transform structural and temporal information (like a job record sequence) into comprehensive, descriptive features (like number of jobs, average duration, highest company prestige). Preliminary experiments suggested that this method may completely eliminate residue re-identificatio risks while retaining (or even improving) prediction precisions.
- **Expanding the leaderboard**: Add to the leaderboard by testing more LLMs and specialised VC-models.
- **Adding a VC-simulation mode**: Introduce a simulation mode that mimics the sequential nature of real-world VC decision making: startups appear over time and resources are limited. Such a mode would introduce the additional challenge of fast serial decision making, enabling more realistic model evaluations.

Model	Precision	Recall	$F_{0.5}$	Cost In / Out	Latency
o3	52.3	8.3	**25.2**	$2.00 / $8.00	6.90 s
DeepSeek-V3	34.2	10.9	23.8	$0.27 / $1.10	10.07 s
DeepSeek-R1	27.4	15.2	23.5	$0.55 / $2.19	37.83 s
GPT-4o	35.7	9.8	23.1	$2.50 / $10.00	3.59 s
GPT-4o-mini	29.4	11.2	22.1	$0.15 / $0.60	3.04 s
Gemini-2.5-Pro	17.4	62.2	20.3	$1.25 / $10.00	10.73 s
Claude-3.5-Haiku	15.9	55.2	18.5	$0.80 / $4.00	3.36 s
GPT-5	**58.0**	4.2	16.2	$1.25 / $10.00	1.54 s
Gemini-2.5-Flash	13.4	**72.8**	16.0	$0.30 / $2.50	8.32 s

Table 4. Vanilla LLMs performances on VCBench, averaged over 6 folds of 1 500 founders, sorted by $F_{0.5}$ (descending).

8. Conclusion

In this paper, we introduced VCBench, a standardised, anonymised VC dataset for early-startup success predictions. With the primary objective of anonymising while preserving sufficient features, we applied a generalisable, multi-stage data cleaning process. The process involved data coverage improvement, data standardisation, data filtering, data decontamination and clustering. We perform anonymisation unit tests to continuously and directly monitor the level of anonymisation throughout the construction process, allowing us to gradually introduce more features without reintroducing substantial identification risks. This method successfully mitigated LLM founder identifications for both offline and web-search-enabled LLMs, resulting in over 80% and 92% reductions respectively in the number of identifiable founders.

After the construction process, we benchmarked the predictive performance of nine state-of-the-art LLMs on VCBench. OpenAI o3 currently leads the cohort with the highest $F_{0.5}$ score and $5.9\times$ the baseline precision. With the leaderboard and part of the benchmark made public, we hope that VCBench will enable fair evaluations of both LLMs and bespoke VC prediction models.

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.

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.

Domingo-Ferrer, J. and Torra, V. A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pp. 990–993, 2008. doi: 10.1109/ARES.2008.97.

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.

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.

Palavalli, M., Bertsch, A., and Gormley, M. R. A taxonomy for data contamination in large language models, 2024. URL https://arxiv.org/abs/2407.08716.

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.

Xiong, S., Ihlamur, Y., Alican, F., and Yin, A. O. Gptree: Towards explainable decision-making via llm-powered decision trees, 2024. URL https://arxiv.org/abs/2411.08257.

A. Appendix: Vanilla LLMs Result Tables

The following tables summarise the detailed test results for the vanilla LLM experiments.

Test Set	Precision	Recall
1	30.0	13.33
2	20.45	6.67
3	28.57	10.37
4	25.53	8.89
5	**39.13**	13.33
6	32.79	**14.81**
Average	29.41	11.23

Table 5. GPT-4o-mini results

Test Set	Precision	Recall
1	29.41	11.11
2	27.27	6.67
3	27.27	6.67
4	**48.72**	**14.07**
5	41.67	11.11
6	40.0	8.89
Average	35.72	9.75

Table 6. GPT-4o results

Test Set	Precision	Recall
1	80.0	5.93
2	40.0	2.96
3	25.0	0.74
4	54.55	4.44
5	**81.82**	**6.67**
6	66.67	4.44
Average	58.01	4.20

Table 7. GPT-5 results

Test Set	Precision	Recall
1	52.0	9.63
2	41.67	7.41
3	47.37	6.67
4	45.45	7.41
5	**68.75**	8.15
6	58.33	**10.37**
Average	52.26	8.27

Table 8. o3 results

Test Set	Precision	Recall
1	12.96	70.37
2	13.21	72.59
3	**13.99**	**79.26**
4	13.97	74.07
5	12.43	68.15
6	13.61	72.59
Average	13.36	72.84

Table 9. gemini-2.5-flash results

B. Example Prompts and Input Formats

B.1. Example raw JSON founder profile string


```
{
  "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"
```

Test Set	Precision	Recall
1	**17.98**	65.93
2	17.43	62.22
3	17.70	**63.70**
4	17.25	58.52
5	16.98	60.00
6	17.14	62.96
Average	17.41	62.22

Table 10. gemini-2.5-pro results

Test Set	Precision	Recall
1	35.09	14.81
2	31.71	9.63
3	28.12	6.67
4	29.27	8.89
5	37.21	11.85
6	**43.90**	**13.33**
Average	34.22	10.86

Table 11. DeepSeek-V3 results

Test Set	Precision	Recall
1	27.27	13.33
2	22.58	15.56
3	25.30	15.56
4	27.03	14.81
5	25.00	11.85
6	**36.99**	**20.0**
Average	27.36	15.19

Table 12. DeepSeek-R1 results

Test Set	Precision	Recall	$F_{0.5}$
1	15.69	53.33	18.26
2	16.49	58.52	19.26
3	16.46	58.52	19.22
4	15.84	54.07	18.44
5	13.20	45.19	15.38
6	**17.66**	**61.48**	**20.60**
Average	15.89	55.19	18.53

Table 13. Claude-3.5-Haiku results

```
    }
  ],
  "educations": [
    {
      "university": "Young Dong High School",
      "degree": null,
      "fields": null,
      "started_on": null,
      "ended_on": null
    }
  ]
}
```

B.2. Example Anonymised Prose Format

```
"This founder leads a startup in the IT Services & Digital Solutions industry.
They maintain 2001-5000 LinkedIn connections.
Education:
* BA in History (Institution QS rank 42)

Professional experience:
* Board Member for <2 years in the 'Management, Strategy & Professional Services' industry (5
* 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: o
```

B.3. Anonymisation unit testing prompt

```
You are given the profile of an entrepreneur wh

### Your Task
1. Examine the profile in JSON format below.
2. Deduce the most likely identity of the found
3. Respond in the **exact** format specified un

### OUTPUT FORMAT
Put the founder's full name right at the beginn
Do not return the text "Founder's name". Always
In addition, include the following information:
Reasoning: <one brief paragraph explaining how
Confidence: <Low, Medium, High>

### Inputs
{profile}
```