JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Generative AI Agents for Startup Opportunity Identification and Early-Stage Decisions: A Literature-Grounded Framework, an Empirical Information-Ceiling Benchmark, and a Pre-Specified Agent Experiment

Perumala Vishal | Rishabh Misra | rishabheco@gmail.com

Abstract

Generative AI agents are increasingly proposed as decision support for screening startup ideas, yet it is unclear how much any screening method can be expected to add when only idea-stage information is available. This paper (i) synthesises literature on opportunity theory, lean-startup practice, large language model (LLM) agents and generative AI in knowledge work; (ii) proposes a framework in which a role-specialised agent pipeline supports a human decision-maker; (iii) reports an empirical benchmark on real data, using 922 US startups (596 acquired, 326 closed) from a public Crunchbase-derived dataset, that estimates how well structured information available at different stages discriminates outcomes; and (iv) specifies a within-subject experiment (planned N = 48) to test agent support directly. Under repeated stratified cross-validation, sector and state alone discriminated acquired from closed firms only weakly (ROC AUC = 0.59, 95% CI 0.55–0.63; permutation $p \approx .002$), and adding early funding signals gave a small, statistically fragile gain that did not replicate in a temporal hold-out. Adding later-stage signals raised AUC to about 0.82, but these signals are not observable at screening time and partly reflect hindsight. The empirical study does not test AI agents; it provides a reference point for judging claims about agent accuracy. The agent experiment has not been run, so conclusions about agents remain provisional. Practically, the findings suggest that value from agent support is more likely to come from generating and structuring evidence than from predicting outcomes from idea descriptions.

Keywords: *generative AI; AI agents; large language models; opportunity identification; startup evaluation; entrepreneurial decision-making; lean startup; predictive validity*

1. Introduction

Entrepreneurship research frames the field as the study of how opportunities for future goods and services are discovered, evaluated and exploited (Shane & Venkataraman, 2000). Evaluation is where scarce founder time and capital are committed, and it occurs with thin evidence and strong cognitive biases. Practitioner methods such as the lean startup (Ries, 2011; Blank, 2013) and business-model design (Osterwalder & Pigneur, 2010) replace prediction with fast, evidence-generating experiments, consistent with the view that entrepreneurship is a process of experimentation (Kerr et al., 2014). Even so, founders, mentors and early-stage investors must screen many ideas cheaply before experiments can be run, and surveyed venture capitalists report relying heavily on judgement in doing so (Gompers et al., 2020).

Since the transformer architecture (Vaswani et al., 2017) and few-shot language models (Brown et al., 2020), LLMs can draft market analyses, list competitors and critique plans in seconds. Agentic scaffolds interleave reasoning with actions such as search (Yao et al., 2023; Schick et al., 2023), ground outputs in retrieved documents (Lewis et al., 2020) and add memory and reflection (Park et al., 2023); a survey is given by Wang et al. (2024). Whether such agents improve early-stage decisions is an open empirical question. A prior question is how much any method can add, given what is knowable about an idea at the time of screening. This paper addresses both.

1.1 Research questions

- RQ1. What does the literature imply about the capabilities and limits of generative AI agents for evaluating startup opportunities?
- RQ2. How should an agent-assisted evaluation workflow be structured to support human judgement and remain auditable?
- RQ3 (empirical, Study 1). Using real data, how well does structured information available at different stages discriminate acquired from closed startups, and how robust is that discrimination?
- RQ4 (experimental, Study 2, protocol only). Does structured multi-agent support improve evaluation quality relative to human-only and single-LLM conditions?

1.2 Contributions

The contributions are: a structured synthesis; a conceptual framework with falsifiable propositions; a reproducible empirical benchmark on real data with explicit leakage controls; and a pre-specified agent experiment. The paper is explicit about what the evidence does and does not show: Study 1 does not involve LLMs or agents, and Study 2 has not been conducted.

2. Literature Review

2.1 Opportunity identification and evaluation

Discovery theories treat opportunities as features of markets that alert individuals can notice (Kirzner, 1973), whereas creation theories hold that opportunities are enacted through entrepreneurs' actions (Alvarez & Barney, 2007). The distinction matters for AI support: an agent that searches existing data aligns naturally with discovery logic, but is structurally weaker where an opportunity has no data trace because it is still being created. Sarasvathy (2001) contrasts causation, which selects means to reach a goal, with effectuation, which lets goals emerge from available means. Rubric-based agent scoring embodies causal logic; effectual founders may use agents differently, for example to widen the set of imaginable partners and uses.

2.2 Early-stage decision-making under uncertainty

Prospect theory and the heuristics-and-biases tradition document systematic departures from normative choice (Kahneman & Tversky, 1979; Kahneman, 2011), and Kahneman et al. (2021) document noise, meaning unwanted variability between judges examining the same case. Lean practice mitigates both by structuring learning into build–measure–learn cycles and testing the riskiest assumptions first (Ries, 2011; Blank, 2013). A decision-support agent could standardise screening and force explicit assumption lists, but may also inherit biases from training data (Bender et al., 2021).

2.3 Generative AI agents and their evaluation

Prompting that elicits intermediate steps improves LLM reasoning (Wei et al., 2022). ReAct interleaves reasoning traces with tool actions (Yao et al., 2023), Toolformer reports models learning when to call tools (Schick et al., 2023), and retrieval-augmented generation conditions outputs on retrieved passages (Lewis et al., 2020). Generative-agent architectures add memory, reflection and planning (Park et al., 2023). Foundation-model reports also catalogue systemic risks such as output homogenisation and error propagation (Bommasani et al., 2021). LLMs are sometimes used to judge other LLM outputs; Zheng et al. (2023) report substantial agreement with human preferences but also document position and length biases.

2.4 Evidence on generative AI in knowledge work

Noy and Zhang (2023) found that a chatbot reduced time on professional writing tasks and raised rated quality. Brynjolfsson et al. (2023) examined the effects of generative AI on customer-support productivity. The study provides evidence on AI-assisted knowledge work, although its findings should not automatically be generalized to startup opportunity evaluation. Dell'Acqua et al. (2023) describe a 'jagged technological frontier' in a study of management consultants: AI helped on tasks inside the frontier but made consultants less likely to be correct on a task outside it. Davenport and Ronanki (2018) classify business AI as process automation, cognitive insight or cognitive engagement; early-stage evaluation is closest to the second. Work directly on entrepreneurs' strategic decisions with AI is limited; Csaszar et al. (2024) examine AI-assisted strategic decision-making among entrepreneurs and investors, but I have not verified their detailed findings and the entry should be checked before it is relied upon.

2.5 Risks: hallucination, leakage and governance

Hallucination, meaning fluent but unsupported content, is a documented property of text generation systems (Ji et al., 2023); a fabricated market size can anchor a decision. A related methodological hazard in any predictive study is data leakage, where

information unavailable at prediction time inflates apparent accuracy (Kapoor & Narayanan, 2023). Study 1 is designed with this hazard in mind. Governance frameworks emphasise human oversight and documentation (NIST, 2023; Regulation (EU) 2024/1689).

Table 1. Synthesis of the literature and implications

Stream	Representative sources	Implication for this study
Opportunity theory	Kirzner (1973); Shane & Venkataraman (2000); Alvarez & Barney (2007)	Agents fit discovery logic best; creation-type opportunities need human enactment
Decision-making and bias	Kahneman & Tversky (1979); Kahneman et al. (2021); Gompers et al. (2020)	Standardised rubrics may reduce noise; agents may add own bias
Lean and effectual practice	Ries (2011); Blank (2013); Kerr et al. (2014); Sarasvathy (2001)	Evidence generation, not prediction, is the core mechanism
LLM and agent methods	Wei et al. (2022); Yao et al. (2023); Lewis et al. (2020); Park et al. (2023)	Role-specialised, retrieval-grounded pipelines are plausible designs
Generative AI at work	Noy & Zhang (2023); Brynjolfsson et al. (2023); Dell'Acqua et al. (2023)	Gains are task-dependent; a frontier exists beyond which AI misleads
Risk, leakage, governance	Ji et al. (2023); Kapoor & Narayanan (2023); NIST (2023)	Evidence trails, human verification and leakage controls are required

2.6 Research gap

Existing studies show that generative AI can raise speed and sometimes quality on bounded tasks and can mislead outside its competence. Missing are (a) a transparent estimate of how much discrimination between startup outcomes is available from structured information at idea stage, against which agent claims can be judged, and (b) controlled comparisons of human-only, single-LLM and multi-agent evaluation. This paper provides the former empirically and the latter as a protocol.

3. Conceptual Framework and Hypotheses

Figure 1 depicts the proposed workflow. An idea brief and curated sources enter an agent layer with four roles. The researcher gathers evidence and writes an evidence log with source identifiers; the analyst estimates market size with explicit assumptions and scores the idea against a rubric (Section 4.5); the critic argues against the idea and lists its riskiest assumptions; the synthesiser produces a recommendation with a stated confidence and open questions. Outputs go to a human decision-maker, whose validation feeds back to the inputs.

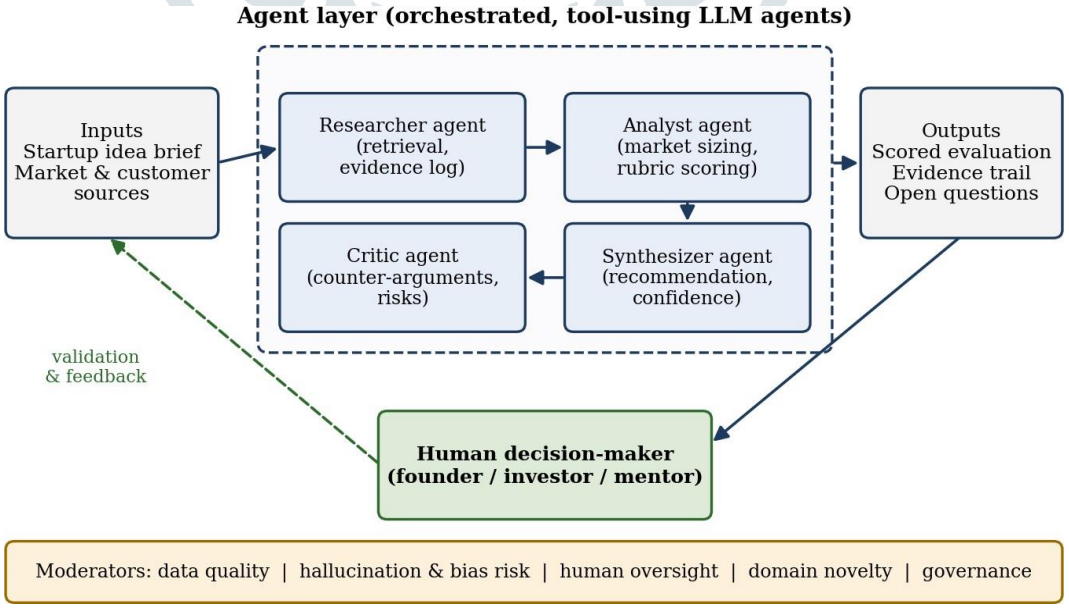

Figure 1. Conceptual framework: a role-specialised agent layer supporting a human decision-maker.

Hypotheses for Study 1 were specified before the models were run (by the author, not externally pre-registered). Propositions for Study 2 are theoretical expectations.

Table 2. Hypotheses and propositions

ID	Statement	Test
H1	Idea-context information (sector, state) discriminates acquired from closed startups better than chance, but only modestly (AUC < 0.70).	Cross-validated AUC; permutation test
H2	Adding early funding signals yields a material AUC gain (≥ 0.05) that survives a temporal hold-out.	Paired bootstrap ΔAUC; temporal hold-out
H3	Later-stage signals raise AUC substantially, but this reflects information unavailable at screening time (hindsight).	Paired bootstrap ΔAUC; leakage analysis
P1–P2	Study 2: multi-agent support (C3) yields lower error against expert consensus than human-only (C1) and single-LLM (C2), with fewer unsupported claims.	MAE; hallucination rate
P3–P4	Study 2: C2 and C3 reduce time vs C1; overconfidence is greatest under C2.	Minutes; calibration gap

4. Methods

4.1 Overview

Study 1 is an empirical benchmark on real secondary data. Study 2 is a within-subject experiment specified as a protocol. The comparison of configurations in Section 6 is a structured conceptual comparison (Eisenhardt, 1989; Yin, 2018) and is not a field case study.

4.2 Study 1: data

The dataset is a 923-record Crunchbase-derived table of US startups commonly distributed as 'startup data.csv' (a copy was retrieved from the GitHub repository listed in the references on 23 October 2026). Each record has an outcome label of acquired or closed and 47 other fields covering location, sector flags, founding and funding dates, funding rounds and amounts, network ties ('relationships'), milestones and a flag for a top-500 list. One duplicate company identifier was removed, leaving N = 922 (596 acquired, 326 closed; acquired share 64.6%). Founding years range from 1984 to 2013 (median 2006). The original data collection procedure and licence could not be verified, so the data should be treated as a convenience sample; see Section 8.

Fields excluded from all models to avoid leakage: the closure date (missing for 98.5% of acquired firms and never missing for closed firms, hence a proxy for the outcome), the pre-encoded label column (identical to the outcome), the top-500 flag (its construction is undocumented) and identifiers and coordinates. Feature sets were defined by when the information could plausibly be known: Set A, idea context (nine sector flags and four state flags); Set B, A plus early funding signals (age at first funding, angel funding, venture-capital funding); Set C, B plus later-stage signals (number of funding rounds, log total funding, round A–D indicators, network ties, milestones, average round participants and the age at last funding and at first and last milestone, with missing values median-imputed and flagged). Set C is deliberately hindsight-contaminated: these fields are recorded over the firm's whole history and cannot be known at screening time.

4.3 Study 1: models and statistical analysis

Models were L2-regularised logistic regression (standardised inputs) and random forests (Breiman, 2001; 300 trees, minimum leaf size 5), implemented in scikit-learn (Pedregosa et al., 2011). The primary metric was ROC AUC (Fawcett, 2006) from 5 × 5 repeated stratified cross-validation, with out-of-fold predictions averaged across repeats and 95% CIs from 1,000 bootstrap resamples of companies (Efron & Tibshirani, 1993). Because these CIs do not fully capture fold-to-fold variation, fold-level standard deviations are also reported. Significance versus chance used a label-permutation test with 500 permutations (Ojala & Garriga, 2010), whose smallest attainable p is 1/501 ≈ .002. Incremental value was assessed with paired bootstrap differences in AUC. Robustness was assessed with a temporal hold-out (train on firms founded in 2006 or earlier, n = 534; test on firms founded later, n = 388). Secondary descriptive analyses used chi-square tests for binary features and Mann–Whitney tests (Mann & Whitney, 1947) for continuous features, with Holm correction (Holm, 1979) across the 17 univariate tests; a multivariable logistic regression on Set B (statsmodels; Seabold & Perktold, 2010) reports odds ratios descriptively. Permutation importance (drop in AUC on held-out folds) summarises influential features. The random seed was fixed (42), and the analysis script and outputs accompany this manuscript.

4.4 Study 2: agent experiment protocol

Participants (planned N = 48 students or novice founders) evaluate six one-page idea briefs from different sectors, written from synthetic or anonymised ideas so outcomes cannot be looked up. Each participant completes two briefs under each of three conditions: C1 human-only with ordinary web search; C2 a general-purpose LLM chat assistant; C3 a structured

pipeline of researcher, analyst, critic and synthesiser agents producing an evidence log (Appendix A). Because the three conditions differ in workflow and evidence access as well as agent configuration, the experiment compares complete decision-support workflows; it is not designed to isolate the causal effect of using multiple agents alone. Task instructions, time limits and permitted external resources will be specified consistently wherever applicable. Order and brief assignment follow a six-order Latin square with eight participants per order (Figure 2). Reference scores come from three blind domain experts using the rubric in Section 4.5;

reliability is assessed with the intraclass correlation (Shrout & Fleiss, 1979) and, for categorical flags, Cohen's kappa (Cohen, 1960; Landis & Koch, 1977).

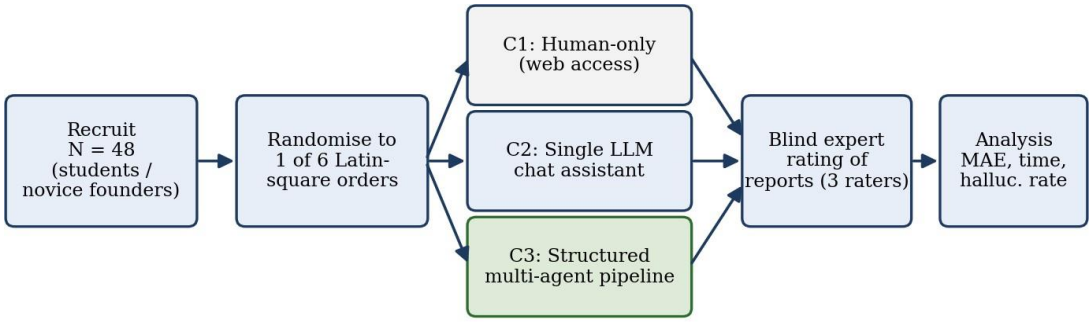

Figure 2. Study 2 design: Latin-square crossover with three conditions and blind expert rating.

Measures: mean absolute error (MAE) against the expert-panel mean score; minutes on task; unsupported-claim rate (the proportion of factual claims that two fact-checkers cannot verify from the available evidence, with a third resolving disagreements); and a calibration gap between stated confidence and a normalized accuracy score. For paired comparisons of a medium effect (d = 0.5; Cohen, 1988), a two-sided test at alpha = .05 and power = .80 needs about 34 participants; with a Bonferroni-adjusted alpha of .0167 for three contrasts it needs about 45 (computed from the noncentral t distribution), so N = 48 is planned to balance the Latin square. The effect size is an assumption, not an estimate from prior work. The primary analysis is a linear mixed-effects model with random intercepts for participant and brief, with Wilcoxon signed-rank tests (Wilcoxon, 1945) as non-parametric checks and thematic analysis (Braun & Clarke, 2006) for open comments. Ethics approval and informed consent are required before data collection.

Before analysis, the accuracy score will be mapped to the same 0–100 scale as stated confidence, using a pre-specified scoring rule; the calibration gap will be calculated as stated confidence minus this normalised accuracy score.

4.5 Rubric for expert scoring

Table 3. Evaluation rubric (proposed, unvalidated instrument)

Dimension	Weight	Scoring anchor (0 = very weak, 100 = very strong)
Problem severity	20	Evidence that the target customer has a frequent, costly problem
Market attractiveness	20	Defensible size and growth logic with stated assumptions
Solution–customer fit	15	Plausible match between proposed solution and customer needs
Competitive differentiation	15	Clear, hard-to-copy advantage versus identified alternatives
Feasibility	15	Technical, regulatory and resource feasibility for a small team
Evidence quality	15	Share of claims traceable to a verifiable source

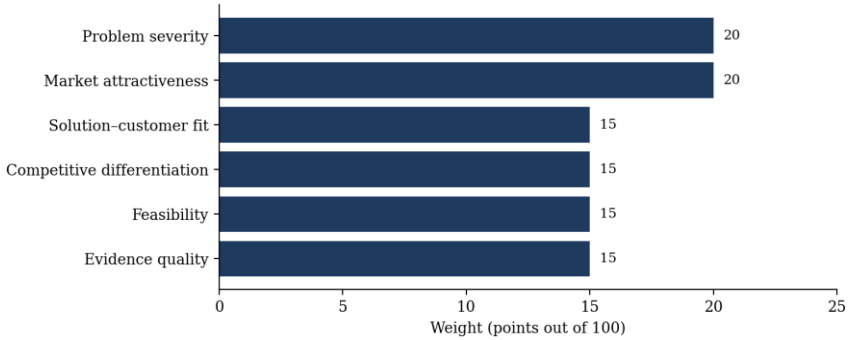

Figure 3. Proposed rubric weights (design parameters, not results).

5. Results of Study 1 (Real Data)

5.1 Sample and descriptive findings

Table 4 summarises the analysis sample. Several descriptive differences are large but concern signals that accumulate after founding. Firms with a recorded round A were acquired more often than those without (73.3% vs 55.7%), and the gap grows with later rounds (round D: 84.8% vs 62.4%; all Holm-adjusted $p \leq .0002$). Acquired firms also had more network ties (median 7 vs 3; rank-biserial r = 0.56), more milestones (median 2 vs 1) and higher total funding (median log10 USD 7.10 vs 6.70). By contrast, sector flags were unrelated to outcome (e.g., software 66.0% vs 64.4% acquired, unadjusted p = .70), and the unadjusted California difference (68.1% vs 60.8%; p = .020) did not survive Holm correction (p = .16).

Table 4. Sample and selected univariate comparisons

Feature	Acquired	Closed	Effect	Holm p
Sample size (N = 922)	596 (64.6%)	326 (35.4%)	—	—
Round A funding (% acquired, yes vs no)	73.3 vs 55.7	—	χ^2 = 31.1	< .001
Round D funding (% acquired, yes vs no)	84.8 vs 62.4	—	χ^2 = 18.1	< .001
Network ties (median)	7	3	r_rb = 0.56	< .001
Milestones (median)	2	1	r_rb = 0.40	< .001
log10 total funding USD (median)	7.10	6.70	r_rb = 0.31	< .001
Age at first funding, yrs (median)	1.45	1.44	r_rb = -0.00	= 1.000
California (% acquired, yes vs no)	68.1 vs 60.8	—	χ^2 = 5.4	= .161
Software sector (% acquired, yes vs no)	66.0 vs 64.4	—	χ^2 = 0.2	= 1.000

Note. r_rb = rank-biserial correlation. Holm correction applied across 17 univariate tests. Differences are descriptive and not causal.

5.2 Predictive discrimination (H1–H3)

Table 5. Cross-validated discrimination by feature set (5 × 5 repeated stratified CV)

Feature set	Model	AUC (95% CI)	Fold SD	Brier	Bal. acc.
A: Idea context (sector, state)	Logistic regression	0.591 (0.553–0.632)	0.039	0.223	0.580
A: Idea context (sector, state)	Random forest	0.584 (0.545–0.625)	0.036	0.223	0.564
B: + early funding signals	Logistic regression	0.612 (0.575–0.649)	0.041	0.220	0.587
B: + early funding signals	Random forest	0.614 (0.577–0.656)	0.033	0.215	0.604
C: + later-stage signals (hindsight)	Logistic regression	0.789 (0.759–0.820)	0.043	0.170	0.710
C: + later-stage signals (hindsight)	Random forest	0.817 (0.786–0.847)	0.042	0.154	0.745

Note. Brier score of a constant prediction equal to the base rate is 0.229. Bal. acc. = balanced accuracy at a threshold equal to the base rate.

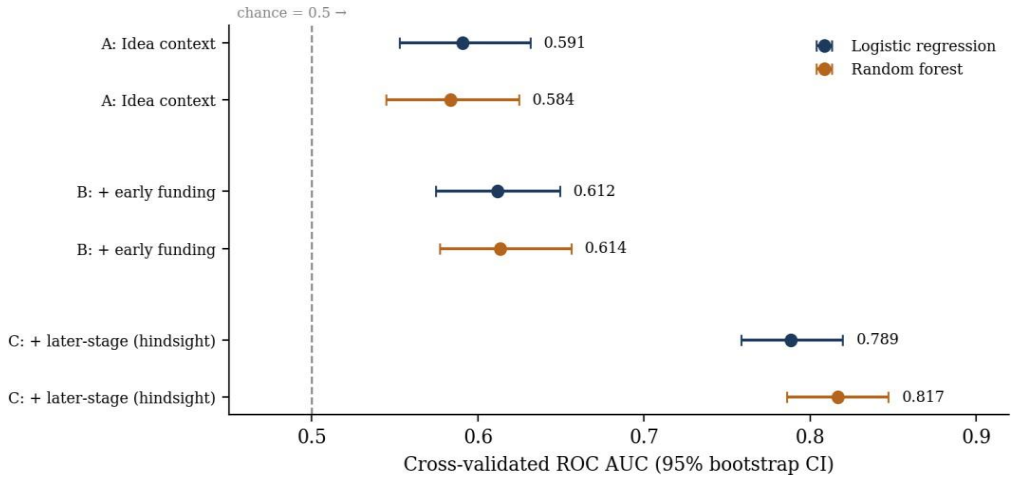

Figure 4. Cross-validated ROC AUC with 95% bootstrap CIs by feature set and model.

H1 was supported in direction and magnitude: Set A discriminated better than chance (logistic AUC = 0.591, 95% CI 0.553–0.632; permutation p ≈ .002; random forest 0.584) but the discrimination was weak, and Brier scores were barely below the constant-prediction benchmark. H2 was not clearly supported: adding early funding signals changed AUC by +0.021 for logistic regression (95% CI 0.001 to 0.043) and +0.030 for random forest (95% CI -0.005 to 0.065), below the pre-specified 0.05 threshold, with the random-forest interval including zero. H3 was supported: later-stage signals raised AUC by +0.177 (logistic) and +0.203 (random forest), with intervals well above zero.

Table 6. Incremental AUC (paired bootstrap, 1,000 resamples)

Comparison	Model	ΔAUC	95% CI
B vs A	Logistic regression	+0.021	0.001 to 0.043
B vs A	Random forest	+0.030	-0.005 to 0.065
C vs B	Logistic regression	+0.177	0.139 to 0.214
C vs B	Random forest	+0.203	0.158 to 0.243

5.3 Robustness: temporal hold-out

When models were trained on earlier-founded firms and tested on later ones (Table 7), Set A performance was stable (AUC ≈ 0.60), whereas Set B did not hold up (logistic 0.563; random forest 0.490, at chance). Set C remained high. The early funding signals therefore add little dependable information beyond sector and state in this sample.

Table 7. Temporal hold-out (train: founded ≤ 2006, n = 534; test: founded > 2006, n = 388)

Feature set	Model	AUC (95% CI)
A: Idea context (sector, state)	Logistic regression	0.602 (0.541–0.665)
A: Idea context (sector, state)	Random forest	0.600 (0.542–0.660)
B: + early funding signals	Logistic regression	0.563 (0.505–0.623)
B: + early funding signals	Random forest	0.490 (0.429–0.549)
C: + later-stage signals (hindsight)	Logistic regression	0.809 (0.765–0.849)
C: + later-stage signals (hindsight)	Random forest	0.829 (0.787–0.870)

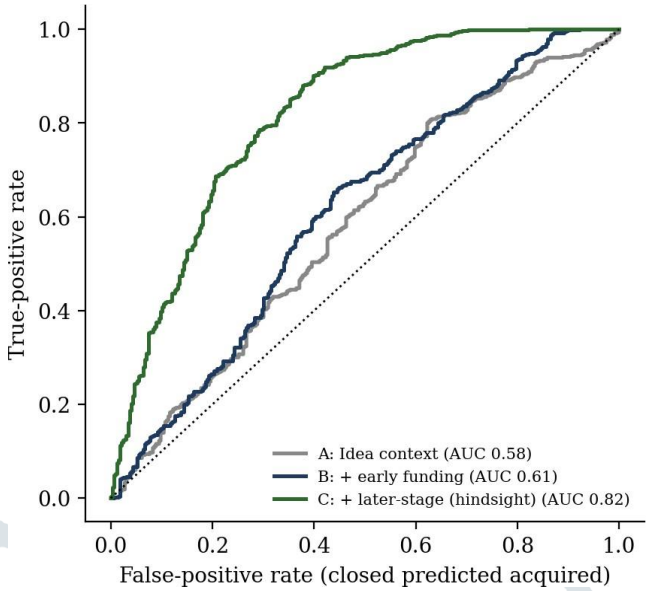

Figure 5. Out-of-fold ROC curves (random forest) by feature set.

5.4 Multivariable and importance analyses (exploratory)

In a multivariable logistic model on Set B (McFadden pseudo-R^2 = 0.053), several associations were estimated, with the reference category being startups outside the four named states and outside the named sectors. Odds of acquisition were higher in California (OR 2.11, 95% CI 1.50–2.99), New York (OR 2.95, 1.74–5.02) and Massachusetts (OR 3.21, 1.77–5.82), and for the enterprise sector (OR 2.35, 1.27–4.34); angel funding was associated with lower odds (OR 0.57, 0.40–0.80). These estimates are unadjusted for multiplicity and are not causal; the angel-funding result in particular may reflect how outcomes were selected into this sample. Permutation importance (Figure 6) shows that age at first funding and location dominated Set B, while network ties dominated Set C.

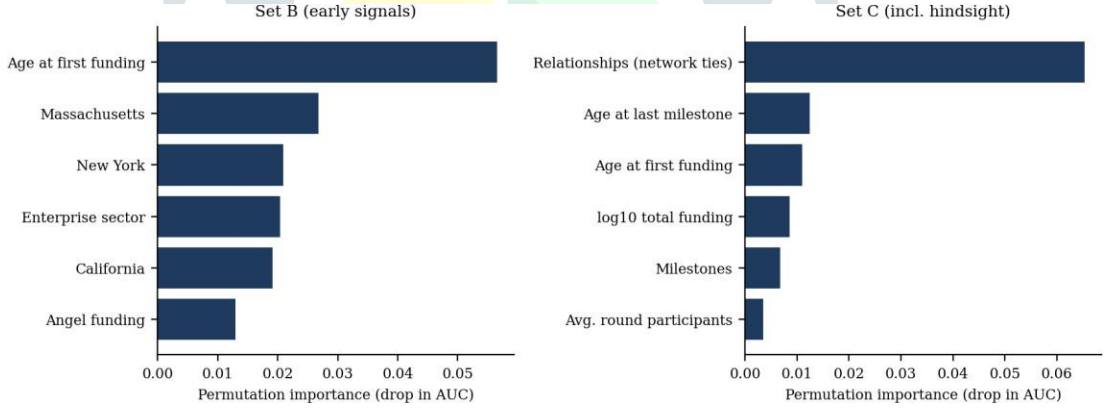

Figure 6. Permutation importance (random forest, held-out folds): top features in Sets B and C.

6. Comparative Analysis of Decision-Support Configurations

Table 8 compares the three Study 2 configurations. Ratings are the author's qualitative judgements derived from the cited literature; they are hypotheses to be tested, not measurements.

Table 8. Qualitative comparison (author's judgement; untested)

Dimension	C1 Human-only	C2 Single LLM chat	C3 Multi-agent pipeline
Breadth of search	Limited by time	High, ungrounded	High, retrieval-grounded
Structured critique	Depends on person	On request only	Built in (critic role)
Traceability of claims	Moderate	Low	High if evidence log enforced
Hallucination exposure	Low	Higher	Reduced, not eliminated
Speed	Slowest	Fast	Fast; setup cost
Novel-market inference	Strong with experience	Weak (Dell'Acqua et al., 2023)	Weak; needs human enactment
Setup and governance cost	Low	Low	Highest

7. Discussion

Study 1 provides a calibration point. In this sample, structured information knowable around founding (sector and state) carried only weak signal about whether a firm ended up acquired or closed, and early funding signals added little that survived a temporal hold-out. The large gains arose only when later-stage fields were included, and those fields are partly consequences of success (firms that are on a path to acquisition accumulate rounds, ties and milestones). This is the leakage pattern warned about by Kapoor and Narayanan (2023), and it implies that headline accuracy figures for startup-success prediction should be read with attention to what the model is allowed to see.

For AI agents, the implication is conditional and should not be overstated. The data are structured fields, not text, and no LLM was tested, so Study 1 cannot show what agents can or cannot do. What it does suggest is that an agent restricted to idea-stage information is unlikely to discriminate outcomes much better than the weak benchmark found here unless it brings new evidence, and that claims of high agent accuracy deserve scrutiny of what information was available. This fits the lean view that the useful move is to generate evidence cheaply (Ries, 2011; Kerr et al., 2014), which is the role the researcher and critic agents are designed to support. Whether agents actually do this well is the question for Study 2.

7.1 Implications

- Founders and investors: use agents to widen search and surface the riskiest assumptions, require sources for every market figure, and convert assumptions into tests.
- Researchers: report which fields were available at prediction time, and test with temporal hold-outs.
- Educators and managers: teach verification alongside agent use, and keep prompts, model versions and outputs for auditability (NIST, 2023).

8. Limitations

- Study 1 does not test AI agents or LLMs; it benchmarks structured-feature screening on historical data.
- The outcome contrast (acquired vs closed) excludes firms still operating and treats acquisition as success, which is imperfect; selection into the sample is undocumented, and the angel-funding association may reflect this.
- The dataset is a secondary copy of Crunchbase-derived data whose collection procedure, licence and exact time frame I could not verify; most firms were founded before 2014, so findings may not describe current markets.
- Set B variables may still partly postdate founding; the temporal hold-out is a partial, not complete, control.
- Bootstrap CIs on pooled out-of-fold predictions understate fold-to-fold variability (fold SD ≈ 0.03–0.04), and the Study 1 hypotheses were specified by the author rather than externally pre-registered.
- Study 2 has not been conducted; its propositions, the rubric weights and the assumed effect size are unvalidated.
- The literature review was not a systematic database search, and sources after mid-2026 are not covered.

9. Conclusion

Using 922 real startup records, this paper shows that information available around founding discriminated acquired from closed firms only weakly, that early funding signals added little robust information, and that strong apparent accuracy appears only when later-stage, hindsight-contaminated fields are included. Generative AI agents may help screening mainly by widening search and structuring critique and evidence, but whether they improve decisions remains untested here; a pre-specified experiment is provided to answer that question. Until it is run, agent outputs are best treated as drafts for human verification.

References

Verification notes: entries were compiled from the author's recall of the literature, not a live database. Check each against its publisher or DOI before submission and add DOIs. Entries marked † are those for which details are least certain.

Alvarez, S. A., & Barney, J. B. (2007). Discovery and creation: Alternative theories of entrepreneurial action. Strategic Entrepreneurship Journal, 1(1–2), 11–26.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). ACM.

Blank, S. (2013). Why the lean start-up changes everything. Harvard Business Review, 91(5), 63–72.

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv:2108.07258. Braun, V., & Clarke, V. (2006).

Using thematic analysis in psychology. Qualitative Research in Psychology,

3(2), 77–101.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Brown, T. B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work (NBER Working Paper No. 31161). National Bureau of Economic Research.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Csaszar, F. A., Ketkar, H., & Kim, H. (2024). Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. Strategy Science. †

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108–116.

Dell'Acqua, F., McFowland, E., Mollick, E. R., et al. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Harvard Business School Working Paper No. 24-013).

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall.

Eisenhardt, K. M. (1989). Building theories from case study research. Academy of Management Review, 14(4), 532–550.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? Journal of Financial Economics, 135(1), 169–190.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

Ji, Z., Lee, N., Frieske, R., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. †

Kerr, W. R., Nanda, R., & Rhodes-Kropf, M. (2014). Entrepreneurship as experimentation. Journal of Economic Perspectives, 28(3), 25–48.

Kirzner, I. M. (1973). Competition and entrepreneurship. University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18(1), 50–60.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1).

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. Journal of Machine Learning Research, 11, 1833–1863.

Osterwalder, A., & Pigneur, Y. (2010). Business model generation. John Wiley & Sons.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). ACM.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Ries, E. (2011). The lean startup. Crown Business.

Sarasvathy, S. D. (2001). Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. Academy of Management Review, 26(2), 243–263.

Schick, T., et al. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference. †

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. Academy of Management Review, 25(1), 217–226.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428.

Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. Wang, L., et al. (2024).

A survey on large language model based autonomous agents. Frontiers of Computer
Science, 18, 186345. †

Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83.

Yao, S., et al. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR 2023).

Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE.

Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 36 (Datasets and Benchmarks Track).

Startup dataset (923 records; Crunchbase-derived 'startup data.csv'). Copy retrieved from https://github.com/sumitjhaleriya/Startup-Success-Prediction-using-Machine-Learning (file: startup data.csv). Original collection procedure and licence not verified by the author.

Appendix A. Template prompts for the multi-agent pipeline (Study 2, C3)

Design template, not a tested prompt; refine in a pilot.

- Researcher: Given the idea brief, retrieve and list sources on customer problem, market size and competitors. For each claim, give a source identifier or write 'no source found'. Never invent a source.
- Analyst: Using only the evidence log, estimate market size with every assumption stated, then score the idea on the six rubric dimensions with a one-sentence justification each.
- Critic: Argue the strongest case against the idea. List the five riskiest assumptions and one low-cost test for each.
- Synthesiser: Produce a recommendation (pursue, test further, or drop), a stated confidence from 0 to 100, and a list of open questions that need human verification.

Appendix B. Study 2 results template (to be completed after data collection)

Outcome	C1 mean (SD)	C2 mean (SD)	C3 mean (SD)	Test statistic, p, effect size
MAE vs expert panel	[ ]	[ ]	[ ]	[ ]
Time on task (min)	[ ]	[ ]	[ ]	[ ]
Hallucination rate (%)	[ ]	[ ]	[ ]	[ ]
Calibration gap	[ ]	[ ]	[ ]	[ ]