VCAF: A Multi-Agent Framework for Venture Capital Decision-Making Using Synthetic Startup Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Venture capital (VC) investment decisions rely heavily on evaluating early-stage startup data, which is frequently sparse, incomplete, or proprietary. To address this challenge, we introduce **InvestAI Corpus**, a synthetic dataset comprising 158 startup profiles generated using a multi-step large language model (LLM) pipeline with human validation, alongside the **Venture Caption Agents Framework** (VCAF), a multi-agent decision-making system powered by Claude-3.7-Sonnet. When evaluated on InvestAI Corpus with complete information, VCAF achieves 74.05% accuracy and an 80.56% F1 score, surpassing baseline human VC performance. The framework provides a systematic backtesting approach for venture capital analysis while generating interpretable investment recommendations that capture the nuanced, qualitative factors critical to early-stage investment decisions.

1 Introduction

Venture capital (VC) investment seeks to identify transformative startups (e.g., Tesla) while avoiding high-profile failures and fraudulent ventures (e.g., Theranos). Investors typically rely on subjective assessments of early-stage factors such as market potential, founding team quality, and product innovation. However, incomplete information and cognitive biases constrain decision-making, resulting in modest success rates (∼60–65% for exits exceeding $50M) [Potanin et al., 2023].

Recent AI methods (e.g., FinQA [Chen et al., 2021], StockNet [Xu and Cohen, 2018]) focus largely on structured financial data and mature markets, overlooking the inherently qualitative nature of early-stage ventures. To bridge this gap, we present **InvestAI Corpus**, a comprehensive benchmark of startup information, and the **Venture Caption Agents Framework** (VCAF), which leverages large language models (LLMs) to analyze unstructured venture data and generate actionable investment insights.

2 Related Work

LLM-powered multi-agent systems have shown promise for complex financial tasks. For example, FinCon [Yu et al., 2024] and TradingAgents [Xiao et al., 2024] employ multi-agent architectures for investment analysis, focusing primarily on trading and public markets rather than early-stage VC. In the VC domain, prior work has applied machine learning (ML) to startup success prediction [Sarisa et al., 2024, Bai and Zhao, 2021] and portfolio simulation [Potanin et al., 2023] using structured datasets (e.g., Crunchbase). While effective for numeric predictions, these approaches struggle with incomplete, multi-dimensional startup information [Wang et al., 2024, Ozince and Ihlamur, 2024]. To our knowledge, no existing system uses agent-based LLMs specifically for detailed VC decision support, motivating our InvestAI Corpus set and VCAF framework.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

3 InvestAI Corpus: Dataset for AI-Driven VC Evaluation

InvestAI Corpus consists of 158 synthetic startup profiles labeled by outcome. Among these, 57.1% are labeled successful and 42.9%. A startup is considered successful if it meets at least one of the following criteria: (1) **Initial Public Offering (IPO)**: a valuation exceeding \$50 million or funds raised over \$100 million at the time of offering; (2) **Acquisition (ACQ)**: an acquisition price greater than the company's total funds raised or exceeding \$100 million in absolute value; (3) **Unicorn Status (UNIC)**: a valuation exceeding \$1 billion, verified using Crunchbase data.

The geographic distribution reveals a pronounced concentration in Silicon Valley, with 44 companies situated in the SF Bay Area and 13 in New York. Temporally, the sample is biased toward the 2020–2025 period, capturing the recent wave of startup formation and investment (detailed statistics are provided in Appendix B).

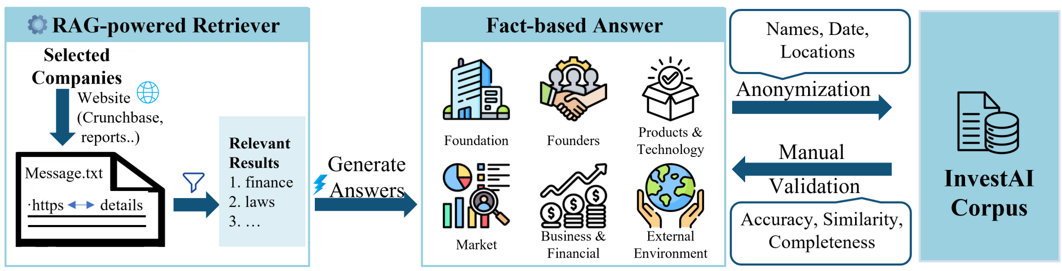

Figure 1: Synthetic startup data generation pipeline.

We adopt a multi-stage generation pipeline (see Figure 1) separating retrieval, generation, and validation, as mentioned in synthetic data generation [Liu et al., 2024]:

- **RAG-powered Retriever:** The pipeline begins by selecting target companies and collecting structured input data from multiple sources, such as Crunchbase and press releases. Information is stored in structured text files and processed with a top-K retrieval strategy by *Doubao-1.5-Thinking-Pro-0415* [Team, 2025], ensuring that LLM-generated profiles are grounded in factual external sources.
- **Fact-based Answer Generation:** Retrieved data are processed by a fact-based answer module powered by *DeepSeek-R1-0528* [Guo et al., 2025], an LLM optimized for structured output. Profiles cover six evaluation dimensions: *Foundation*, *Founders*, *Products & Technology*, *Market*, *Business & Financial*, and *External Environment* (Appendix C).
- **Expert Validation and Anonymization:** To emulate VC decision-making, we first anonymize the generated profiles by removing identifiable attributes (e.g., company and founder names). The anonymized profiles are then reviewed by domain experts, who assess their accuracy, completeness, and consistency against a curated reference set (Table 1). This process ensures that the dataset remains both reliable and privacy-preserving. For detailed definitions of the evaluation metrics, see Appendix A.

  This procedure highlights not only the effectiveness of the dataset for backtesting venture evaluation models, but also the broader applicability of our assessment methodology.

Table 1: Quality Assurance Performance of LLM-Generated Profiles

Outcome	Company	Accuracy	Integrity	Correlation	Score
Success	Tesla	86.67%	100%	83.33%	90%
	DJI	83.33%	100%	83.33%	88.89%
	Airbnb	83.33%	100%	91.67%	91.67%
	Snowflake	90%	100%	80%	90%
	ByteDance	90%	100%	83.33%	91.11%
Failure	Pets.com	80%	100%	86.67%	88.89%
	Vine	76.67%	100%	80%	85.56%
	Juicero	93.33%	100%	86.67%	93.33%
	WeWork	90%	100%	90%	93.33%
	Theranos	86.67%	100%	83.33%	90%

4 Venture Caption Agents Framework

Figure 2 illustrates the multi-agent decision framework of VCAF. At its core, we adopt Claude-3.7-Sonnet as the primary reasoning engine, which achieves the best accuracy and f1-score on InvestAI Corpus, outperforming other models such as GPT-4o and Llama-3.1 (Table 2). The framework is organized into four key modules: *Due Diligence Team*, *Investment Evaluation Manager*, *Risk Management Team*, and *Investment Decision Committee*. Each module contains specialized agents focused on distinct aspects of investment analysis, whose outputs are integrated to produce a final recommendation with prompt in Appendix H. This modular design mitigates potential biases, such as over-optimism in startup valuation.

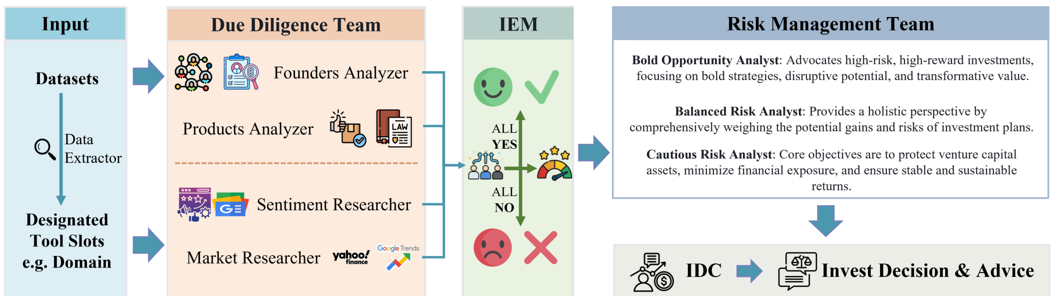

Figure 2: Architecture of the Venture Caption Agents Framework (VCAF).

4.1 Due Diligence Team

The Due Diligence Team consists of four specialized agents conducting comprehensive startup assessments. To minimize data leakage, information retrieval is restricted to designated tool slots, which act as controlled invocation points for external APIs. :

- **Sentiment Researcher:** Extracts insights from Google News and social media, performing sentiment analysis, trend detection, and event impact evaluation to gauge public perception and emerging risks.
- **Market Researcher:** Aggregates market analytics and financial data of Google Trends and Yahoo Finance to evaluate sector trends, consumer interest, and competitor activity. Generates quantitative metrics on market size and growth potential.
- **Founders Analyzer:** Evaluates the founding team's expertise, experience, and track record using InvestAI Corpus. Produces credibility scores and team profiles to assess operational capability.
- **Product Analyzer:** Assesses product or technology maturity, innovation potential, business-model viability, and regulatory considerations, identifying potential barriers and vulnerabilities.

4.2 Investment Evaluation Manager

The Investment Evaluation Manager (IEM) functions as a central analytical unit, evaluating the outputs from the Due Diligence Team. While it does not issue final decisions, it generates reasoned preliminary recommendations. In cases of unanimous consensus among analysts, the IEM endorses the collective view with a concise rationale. If disagreements exist, the IEM performs a balanced evaluation of supporting and opposing arguments, culminating in a recommendation that reflects the relative weight of evidence.

4.3 Risk Management Team

This module includes multiple analysts with distinct risk profiles, containing Risk-Seeking, Neutral and Conservative. Each analyst evaluates factors such as market volatility, credit risk, and operational vulnerabilities. Their complementary assessments are aggregated to inform risk-aware decision-making, ensuring that potential hazards are appropriately considered in the final recommendation.

4.4 Investment Decision Committee

The Investment Decision Committee (IDC) integrates the outputs of the Due Diligence Team, IEM, and Risk Management Team. It synthesizes these inputs to estimate a probability of success and a valuation range for the target company. The committee then issues the final investment decision, accompanied by an actionable recommendation that also considers current market conditions and analyst forecasts.

5 Evaluation and Results

LLM Performance. We evaluated 12 representative LLMs on the InvestAI Corpus startup classification task (Table 2). Claude-3.7-Sonnet achieves the highest F1-score (78.80%) with a balanced precision (66.14%) and high recall (97.67%), reaching 71.50% overall accuracy. Llama-3.3-70B and Claude-3.5-Haiku achieve perfect recall (100.00%) but lower precision, reflecting over-optimistic predictions of successful startups. GPT-4o demonstrates moderate performance (58.86% accuracy, 72.34% F1), tending to over-predict successes. Smaller LLMs such as Llama-3.1-8B achieve 55.70% accuracy (69.30% F1), indicating limitations in capturing risk factors. Error analysis shows that false positives in Claude are mainly due to overestimated market sizes, highlighting the challenge of balancing optimism and risk in automated investment predictions.

Agents Framework Evaluation. We evaluated three system variants: (1) the baseline Claude-3.7-Sonnet classifier, (2) VCAF without the Risk Management module, and (3) the full VCAF framework. The baseline achieves high recall (97.67–98.84%) but moderate accuracy (71.50%) and precision (66.14%), exhibiting an over-optimistic bias with many false positives. VCAF without the Risk module improves accuracy (74.68%) and precision (73.47%), reducing false positives at the expense of lower recall (83.72%). The full VCAF balances this trade-off, achieving 74.05% accuracy, 80.56% F1 (highest), and 98.80% recall (second highest), effectively capturing true positives while minimizing false positives. Both VCAF variants outperform typical human VC accuracy (60–65%) [Lahr and Trombley, 2020], demonstrating the practical value of our multi-agent framework for investment screening and due diligence.

Table 2: Model performance on InvestAI Corpus startup classification. Twelve representative LLMs and our framework (based on Claude-3.7-sonnet) are compared.

Model	TP ↑	FP ↓	TN ↑	FN ↓	Accuracy ↑	Precision ↑	Recall ↑	F1-score ↑
Llama-3.1-8B-Instruct	79	63	9	7	55.70	55.63	91.86	69.30
Llama-3.3-70B-Instruct	86	70	2	0	55.70	55.13	**100.00**	71.07
Mistral-7B-Instruct	85	71	1	1	54.43	54.49	98.84	70.25
Mistral-small-3.1-24B-Instruct	84	61	11	2	60.13	57.93	97.67	72.73
Qwen3-32B	83	56	16	3	62.66	59.71	96.51	73.78
Qwen3-235B-A22B	85	56	16	1	63.92	60.28	98.84	74.89
Claude-3.5-haiku	86	69	3	0	56.33	55.48	**100.00**	71.37
Claude-3.7-sonnet	84	43	29	2	71.50	66.14	97.67	<u>78.80</u>
Gemini-2.0-flash	80	43	29	6	68.99	65.04	93.02	76.56
Gemini-Pro-1.5	81	58	14	5	60.13	58.27	94.19	72.00
GPT-3.5-turbo	83	60	12	3	60.13	58.04	96.51	72.49
GPT-4o	85	64	8	1	58.86	57.05	98.84	72.34
VCAF w/o risk module	72	26	46	14	**74.68**	**73.47**	83.72	78.26
VCAF	85	40	30	1	<u>74.05</u>	<u>68.00</u>	98.80	**80.56**

6 Conclusion

We presented InvestAI Corpus and VCAF as tools for venture capital analysis. InvestAI Corpus addresses the lack of detailed public startup dataset, while VCAF leverages multi-agent LLMs to analyze qualitative information. On InvestAI Corpus, VCAF achieves up to 74.7% accuracy and 80.56% F1-score, surpassing typical human performance. By providing transparent rationales, the framework supports VC analysts in screening and due diligence. Limitations include reliance on synthetic data and computational cost; future work will involve real-world validation and efficiency improvements.

References

Sarah Bai and Yijun Zhao. Startup investment decision support: Application of venture capital scorecards using machine learning approaches. *Systems*, 9(3), 2021. ISSN 2079-8954. doi: 10.3390/systems9030055. URL https://www.mdpi.com/2079-8954/9/3/55.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. *Proceedings of EMNLP 2021*, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Henry Lahr and Timothy E. Trombley. Early indicators of fundraising success by venture capital firms. *Journal of Corporate Finance*, 65:101672, 2020. ISSN 0929-1199. doi: https://doi.org/10.1016/j.jcorpfin.2020.101672. URL https://www.sciencedirect.com/science/article/pii/S0929119920301164.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data, 2024. URL https://arxiv.org/abs/2404.07503.

Ekin Ozince and Yiğit Ihlamur. Automating venture capital: Founder assessment using llm-powered segmentation, feature engineering and automated labeling techniques, 2024. URL https://arxiv.org/abs/2407.04885.

Mark Potanin, Andrey Chertok, Konstantin Zorin, and Cyril Shtabtsovsky. Startup success prediction and vc portfolio simulation using crunchbase data, 2023. URL https://arxiv.org/abs/2309.15552.

Manikanth Sarisa, Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Shravan Kumar Rajaram, and Mohit Surender Reddy. Deep learning to predict the success of startup companies based on smote and xb-rfe model. In *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, pages 1–5, 2024. doi: 10.1109/ICMNWC63764.2024.10872045.

ByteDance Seed Team. Doubao-1.5-pro model overview. https://seed.bytedance.com/zh/special/doubao_1_5_pro, 2025. Accessed Month Day, Year.

Xisen Wang, Yigit Ihlamur, and Fuat Alican. Ssff: Investigating llm predictive capabilities for startup success through a multi-agent framework with enhanced explainability and performance. *arXiv preprint arXiv:2405.19456*, 2024.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.

Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1183. URL https://aclanthology.org/P18-1183/.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W. Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, Denghui Zhang, Koduvayur Subbalakshmi, Guojun Xiong, Yueru He, Jimin Huang, Dong Li, and Qianqian Xie. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making, 2024. URL https://arxiv.org/abs/2407.06567.

A Evaluation Metrics Definitions

- **Accuracy:** Field match rate, scored as 1 (full match), 0.5 (partial match), or 0 (no match).
- **Integrity:** The ratio of non-empty fields to the total number of defined fields.
- **Correlation:** Keyword overlap, calculated as the Jaccard similarity (intersection over union) of keywords between generated and manual profiles.
- **Score:** Average of the above three metrics, representing overall profile quality.

B Data Analysis of InvestAI Corpus

This figure provides a concise overview of the InvestAI Corpus dataset. Startup outcomes show a small fraction (5-10%) achieving successful exits (IPO/acquisition), with most remaining active or failed. Founding year trends indicate a rise in startups since the 2000s, peaking around 2015-2020. Regionally, the U.S., particularly Silicon Valley, dominates (>50%), followed by Europe and Asia. Industry composition highlights software/technology as the leading sector, followed by healthcare/biotech and fintech.

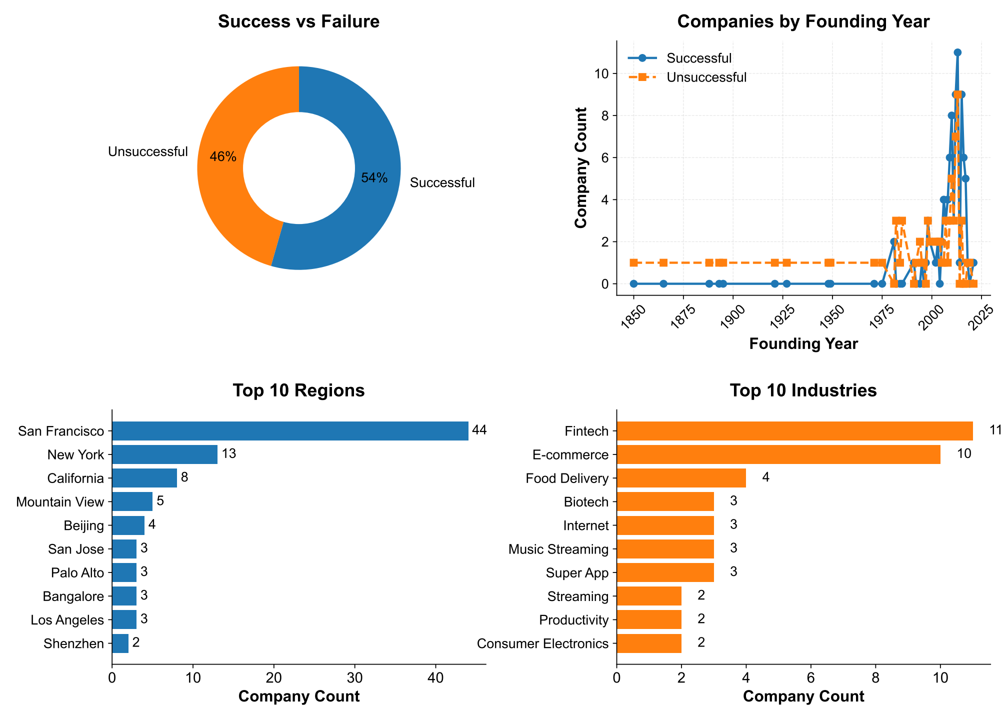

Figure 3: Overview of InvestAI Corpus, including startup outcomes, founding year trends, regional distribution, and industry composition.

C Detailed Data Schema and Dimensions

Each company profile in InvestAI Corpus is structured along six evaluation dimensions. For each dimension, we provide specific sub-items inspired by the structured prompt:

1. **Business Model and Strategy**:
    - Value proposition, revenue model, cost structure, and competitive advantage
    - Profit model (e.g., subscription, transaction commission)
    - Business scalability and unit economic efficiency
2. **Product and Innovation**:

- Product/service quality, market fit, and innovation pipeline
- Technological advantages, patents, and intellectual property
- Development progress (prototype, release timeline)
- Production or supply chain innovations

3. **Team and Execution**:
   - Founders' expertise, academic background, and team cohesion
   - Online footprint (GitHub, LinkedIn, etc.) and network of advisors/investors
   - Previous entrepreneurial experience, awards, and achievements
   - Operational capabilities and execution track record
4. **Market and Competition**:
   - Market situation (total addressable market, target users)
   - Competitor analysis (direct and indirect competitors)
   - Market barriers and entry risks
5. **Financial Performance**:
   - Revenue growth, profitability, and funding history
   - Burn rate and annual recurring revenue
   - Financial projections and key metrics (unit economics)
6. **Macro and Industry Context**:
   - Industry growth trends and regulatory environment
   - Macro-economic and socio-cultural factors affecting the industry
   - Policy support and government incentives
7. **Metadata and Verification**:
   - Data sources (primary or secondary)
   - Confidence score (1-5, with 1 = unverified)
   - Cross-source contradictions or inconsistencies

D Voting Analysis of Investment Evaluation Manager

This table analyzes the non-unanimous voting patterns of the Investment Evaluation Manager (IEM) when aggregating predictions from multiple domain-specific analysts. Vote distribution shows the number of analysts voting to Invest versus Reject for each case.

The table indicates that the IEM tends to adopt a conservative approach in non-unanimous cases. For example, in a 2:2 split among analysts, the IEM chooses to Invest in 14 out of 41 cases, while in a 3:1 split, it invests in only 7 out of 17 cases. Overall, the majority of non-unanimous decisions (86 out of 107) are rejections, demonstrating that the IEM mitigates over-optimistic bias by systematically weighting analyst votes and erring on the side of caution.

Table 3: Non-unanimous voting distributions by the Investment Evaluation Manager. Each row shows the distribution of votes for a given voting pattern.

Vote Distribution(Invest vs Reject)	Invest	reject	Total
1:3	49	0	49
2:2	27	14	41
3:1	10	7	17
Total	86	21	107

E Scatter Plot of LLMs and VCAF Performance Comparison

The scatter plot compares accuracy and F1-scores across various large language model families, including Claude, Gemini, GPT, LLaMA, Mistral, Qwen3, and VCAF, highlighting VCAF's superior performance on the InvestAI Corpus dataset.

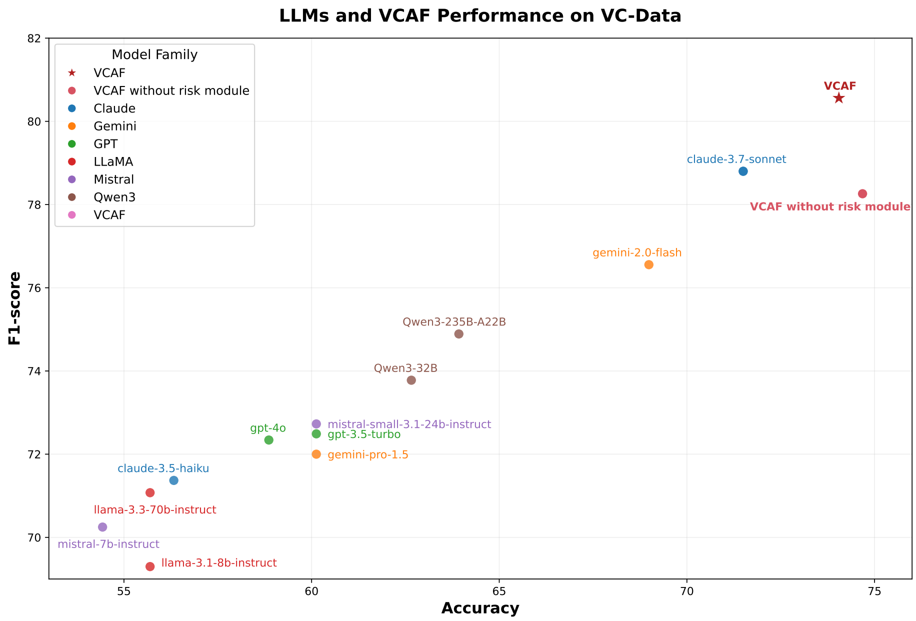

Figure 4: Evaluation results for large language models and VCAF on the InvestAI Corpus dataset.

F Analysis of Investment Decisions and Key Variables

Investment decisions are associated with higher success probabilities and larger valuation ranges compared to non-investments, confirming the model's decision logic.

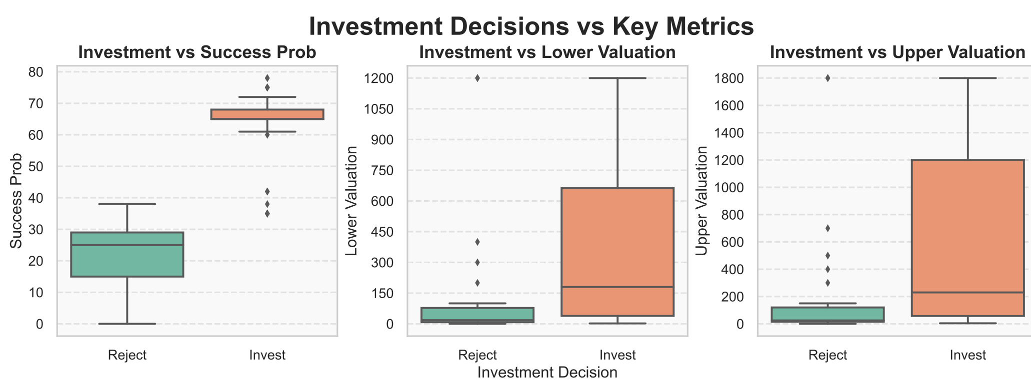

Figure 5: Boxplot analysis of investment decisions versus key variables: (a) success probability, (b) lower valuation, and (c) higher valuation.

G Additional Model Performance Evaluation

Table 4 reports performance metrics for selected models, including two VCAF variants and Qwen3-235B-A22B.

As shown, removing the risk-management module ('VCAF without risk module') yields higher accuracy and F1-score relative to the full VCAF. This occurs because the risk module introduces conservative adjustments that slightly reduce true positives. However, the full VCAF demonstrates a more balanced performance profile, achieving higher recall and a trade-off that aligns with practical investment risk considerations. This highlights the framework's ability to moderate aggressive predictions while maintaining robustness.

Importantly, when applied to a different base model (Qwen3-235B-A22B), VCAF still improves key metrics, demonstrating the framework's general effectiveness in enhancing model performance while preserving a controlled balance between precision and recall.

Table 4: Performance Metrics of Qwen Model and VCAF Variants on InvestAI Corpus.

Model	TP	FP	TN	FN	Accuracy	Precision	Recall	F1-Score
Base (Qwen3-235B-A22B)	85	56	16	1	63.92	60.28	**98.84**	74.89
VCAF without risk module	79	32	40	7	**70.25**	<u>66.39</u>	91.90	**77.09**
VCAF	82	46	26	4	<u>68.35</u>	64.06	<u>95.35</u>	<u>76.63</u>

H Agent Prompts

This section provides examples of prompts used for some representative agents in our multi-agent framework, demonstrating how agent behavior is guided for consistent task execution.

H.1 Data Extractor

The data extractor obtains structured information from startup descriptions.

Prompts for Data Extractor

You are a professional market data extractor skilled at obtaining key information from texts. Please extract the following from the startup description:

- **Time information (trade_date):** establishment date, key funding date, or product launch date in `YYYY-mm-dd` format. Leave blank if unavailable.
- **Regional information (startup_area):** headquarters location, main operating region, or market coverage.
- **Domain information (startup_domain):** main business field or industry (e.g., AI, biomedicine).
- **Keywords (startup_keywords):** key terms related to products, technologies, or business models.

Output the result in `JSON` format.

H.2 Product Analyzer

The Product Analyzer evaluates product fundamentals based solely on the provided description.

Prompts for Product Analyzer

Analyze the flagship product of the startup. Focus strictly on the provided description; do not speculate about company identity or make external comparisons. Document assumptions if details are missing.

H.3 Market Researcher

The Market Researcher evaluates market environment and trends.

Prompts for Market Researcher

Analyze up to eight market indicators derived from the description:

- **Market Size Estimate:** potential market size
- **Market Growth Potential:** qualitative/quantitative outlook
- **Competitive Intensity:** level of competition
- **Consumer Demand Signal:** inferred demand
- **Policy Support Impact:** regulatory incentives
- **Macro Trend Influence:** economic/environmental trends
- **Competitor Activity Level:** inferred competitor actions
- **Barriers to Entry:** structural challenges

Provide a detailed, actionable report with a Markdown table of findings. Do not speculate beyond the provided description.

H.4 Investment Evaluation Manager

This agent aggregates expert analyses to make the final investment decision.

Prompts for Investment Evaluation Manager

If all four specialists (Market, Founder, Tech, Risk) agree on *invest* or *no-invest*, return the consensus decision with a brief rationale.
If disagreement exists, evaluate arguments for and against the investment and make a balanced decision.

H.5 Investment Decision Committee

The Committee evaluates risk analysts' debate to provide a final recommendation.

Prompts for Investment Decision Committee

Provide a clear recommendation: **INVEST** or **NOINVEST**.
Include:

- Success probability (0–100%) for the potential outcome
- Valuation range (in monetary terms)
- Summary of key arguments from analysts
- Documented reasoning with references to description data and past lessons

Generate a structured, actionable investment report (investment plan) including the probability, valuation, rationale, and recommended next steps.