DIALECTIC: A Multi-Agent System for Startup Evaluation

Jae Yoon Bae[1,2,*], Simon Malberg[1,*], Joyce Galang[1,3,*], Andre Retterath[2], Georg Groh[1]

[1]Technical University of Munich, [2]Earlybird Venture Capital, [3]UVC Partners
[*]These authors contributed equally to this work.
Correspondence: jaeyoonbae99@gmail.com

Abstract

Venture capital (VC) investors face a large number of investment opportunities but only invest in few of these, with even fewer ending up successful. Early-stage screening of opportunities is often limited by investor bandwidth, demanding tradeoffs between evaluation diligence and number of opportunities assessed. To ease this tradeoff, we introduce DIALECTIC, an LLM-based multi-agent system for startup evaluation. DIALECTIC first gathers factual knowledge about a startup and organizes these facts into a hierarchical question tree. It then synthesizes the facts into natural-language arguments for and against an investment and iteratively critiques and refines these arguments through a simulated debate, which surfaces only the most convincing arguments. Our system also produces numeric decision scores that allow investors to rank and thus efficiently prioritize opportunities. We evaluate DIALECTIC through backtesting on real investment opportunities aggregated from five VC funds, showing that DIALECTIC matches precision of real VCs in predicting startup success.

1 Introduction

The global venture capital (VC) industry is expanding rapidly alongside intensified competition for attractive deals. The market is projected to grow from USD 337 billion in 2024 to USD 1.46 trillion by 2033, a compound annual growth rate of 17.6% (IMARC Group, 2024). Entrepreneurial activity has also surged, with annual U.S. business formations rising from 3.5 million in 2019 to 5.2 million in 2024, an increase of nearly 40% (U.S. Census Bureau, 2025). Traditional VC decision-making processes are challenged in this setting. Investors face high time pressure and information overload, both associated with suboptimal decision (Zacharakis and Shepherd, 2001). These conditions have increased interest in computational approaches for scalable investment evaluation.

Among these approaches, machine learning methods have emerged as a promising direction. Prior studies demonstrate strong predictive performance and, in some cases, even surpass human investors (Antretter et al., 2019; Retterath, 2020; Zacharakis and Shepherd, 2001; Arroyo et al., 2019; Dellermann et al., 2021; Sharchilev et al., 2018). Yet, these non-iterative models diverge from how investment decisions are formed by human VCs. In practice, conviction emerges through iterative hypothesis formation, challenge, and refinement as new information appears (Chong and Tuckett, 2014).

Recent advances in *large language model* (LLM) orchestration tools enable iterative and interpretable reasoning. Frameworks such as *LangChain* (Chase, 2022) support the decomposition of complex tasks, the generation of intermediate conclusions, and the iterative refinement of responses while making the underlying logic explicit. They allow multi-step reasoning and dialectical interaction, a setup in which LLMs can articulate arguments, generate counterpoints, and produce transparent reasoning traces.

This paper introduces ***D****ecision* ***I****teration with* ***A****rgument-****L****evel* ***E****vidence and* ***C****ounter-****T****hinking for* ***I****nvestment* ***C****onclusions* (**DIALECTIC**), an LLM-based system that models iterative and argumentative elements of venture evaluation. Our system draws on principles of dialectical reasoning, an approach shown to be effective for complex, unstructured problems that benefit from structured confrontation of differing perspectives (Jarupathirun and Zahedi, 2007). The contributions of this work are:

- A structured LLM reasoning system that models how investors build and refine investment theses through argumentation.
- An empirical evaluation demonstrating predictive performance in venture screening.

Overall, the proposed system brings data-driven VC methods closer to industry practice. Furthermore, it enables the process of iterative argumentation in early-stage screening, which has traditionally been restricted to later stages of the funnel due to limited investor bandwidth. This shift allows investors to apply iterative reasoning earlier in the process, improving both diligence quality and screening efficiency.

2 Related Work

Prior studies propose different machine learning approaches to predict startup outcomes, drawing on public data sources such as *Crunchbase* (Arroyo et al., 2019; Żbikowski and Antosiuk, 2021; Retterath, 2020), *Twitter* (Antretter et al., 2019), web data (Sharchilev et al., 2018), and *Google Search* (Gavrilenko et al., 2023), and often reporting promising prediction accuracy (see Table 4 in the Appendix for an overview). Most studies trained gradient tree boosting models (e.g., *XGBoost*) (Corea et al., 2021; Arroyo et al., 2019; Żbikowski and Antosiuk, 2021; Retterath, 2020) and interpreted predictions using feature-importance rankings with features such as geography, industry, or founder background (Żbikowski and Antosiuk, 2021; Sharchilev et al., 2018; Gavrilenko et al., 2023).

Some newer studies have used LLMs to extract structured features or embeddings from unstructured data, while still resorting to machine learning models such as *XGBoost* for prediction (Ozince and Ihlamur, 2024; Maarouf et al., 2025). Xiong and Ihlamur (2023) used LLMs to assess founder-idea fit, also providing pro and contra arguments for interpretability. In follow-up work (Xiong et al., 2024), they focus on extracting traits associated with successful entrepreneurs. Both studies look at individual founders rather than startup companies.

Beyond VC, LLM-based decision-making frameworks have been proposed for fields such as business or finance. *DeLLMa* combines LLMs with decision-theoretic reasoning (Liu et al., 2025), while *STRUX* extracts facts from companies' earnings calls and produces weighted pro and contra aspects for buy or sell decisions (Lu et al., 2025).

A promising approach to improving LLM reasoning is the introduction of multi-agent systems (Han et al., 2024). Instead of relying on a single model, several LLMs interact through collaboration, debate, or specialization. In adversarial or collaborative debating, agents defend opposing stances and a separate judge model or heuristic evaluates the quality of their arguments (Chan et al., 2023; Liang et al., 2024).

3 DIALECTIC

DIALECTIC is inspired by how real VCs make investment decisions. They collect information about a startup, form narrative investment hypotheses, and refine these hypotheses through debate with other VCs until making a decision. DIALECTIC proceeds in three phases: **fact collection**, **reasoning**, and **decision-making**. During fact collection, DIALECTIC gathers factual knowledge about a company and organizes these facts hierarchically in a question tree. In the reasoning phase, it synthesizes raw facts into arguments pro and contra an investment, which it iteratively self-critiques, evaluates, and refines, letting only the best arguments survive. Finally, it makes a decision based on a comparison of the best pro and contra arguments; see Figure 1 for an illustration.

In the following, we formally introduce DIALECTIC. Let $X = \{x_i\}_{i=1}^{N}$ be the set of investable companies, each described by multiple features $x_i^{(d)}, d = 1, ..., D$. The goal is to predict the ground truth label $y_i \in \{\text{successful}, \text{unsuccessful}\}$ signaling whether the company will be successful and should be invested in or not.

3.1 Fact Collection Phase

We denote the universe of natural-language questions as $\mathcal{Q}$, the universe of natural-language answers as $\mathcal{A}$, and the set of industries as I. For a given company x in industry $x^{(0)} \in I$, we start by providing DIALECTIC with a set of **seed questions** $Q_0 \subset \mathcal{Q}$. Specifically, we ask four questions about the general company, team, product, and market (see Appendix B.1 for details) to cover the main aspects typically considered by VC investors (Retterath, 2020). Inspired by *ProbTree* (Cao et al., 2023) and *Socratic Questioning* (Qi et al., 2023), we define two LLM-based agent operations:

- The **decomposer** $Q : Q_0 \times I \to \mathcal{Q}^*$ takes a seed question $q \in Q_0$ and hierarchically decomposes it into a finite set of M_q sub-questions relevant in the industry, thus creating an industry-specific **question tree**[1] $\{q_l\}_{l=1}^{M_q} = Q(q, x^{(0)})$ with the decision-relevant questions that should be answered.

[1]For simplicity, we do not explicitly model the hierarchical

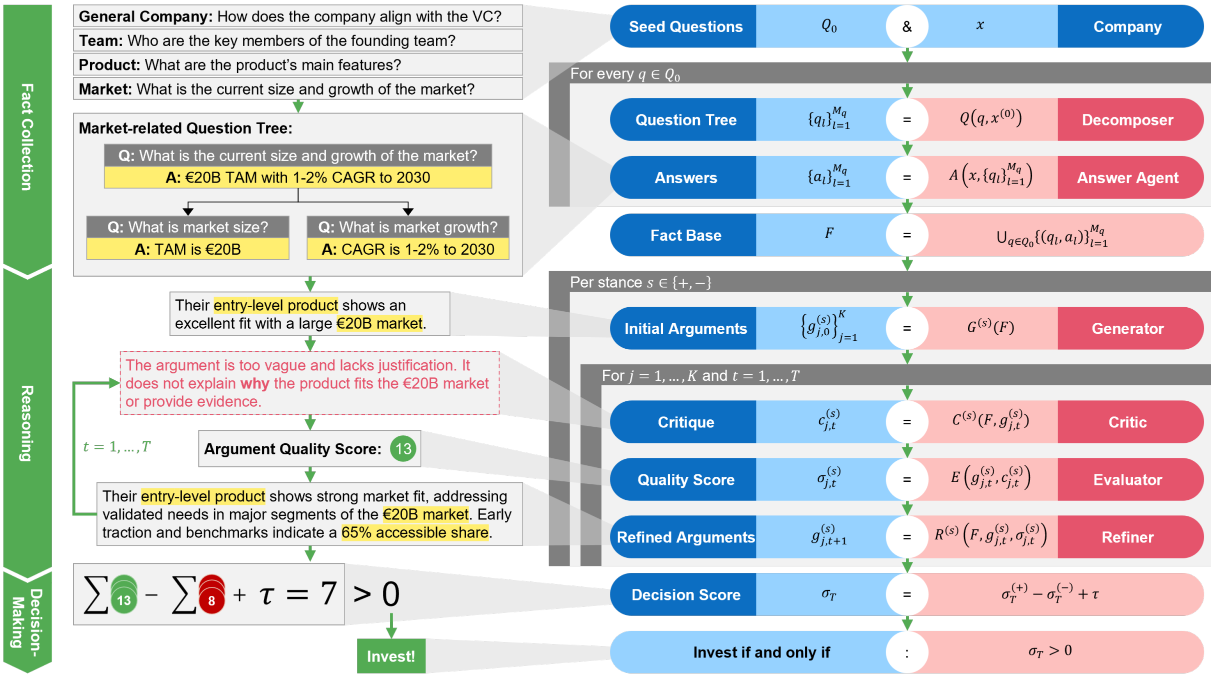

Figure 1: Overview of the DIALECTIC method. The right side shows the flow of operations. Agents are shown in red, agent inputs/outputs are shown in blue, and loops are shown in green. The left side illustrates the key outputs of the agents.

- The **answer agent** $A : X \times \mathcal{Q}^* \to \mathcal{A}^*$ looks at the company features $x^{(d)}$ and uses them to generate answers $\{a_l\}_{l=1}^{M_q}$ to all questions in the tree. It also has access to a web search tool that it can use agentically. Like *ProbTree* (Cao et al., 2023), it answers question trees in a post-order traversal, aggregating answers from child nodes when generating answers for parent nodes. We provide further details in Appendix B.2.

When executed for all seed questions, the agents produce a rich hierarchically structured **fact base** $F \subset \mathcal{F}$ ($\mathcal{F}$ is the universe of all possible question-answer pairs) about the company x:

$$F = \bigcup_{q \in Q_0} \{(q_l, a_l)\}_{l=1}^{M_q}$$

3.2 Reasoning Phase

In the reasoning phase, DIALECTIC combines facts (possibly from different question trees) into arguments taking a stance on whether the VC should invest in a company or not. Let $s \in \{+, -\}$ denote the pro or contra stance, $\mathcal{G}$ the universe of natural-language arguments, and $\mathcal{C}$ the universe of natural-language critiques of these arguments. We define four LLM-based agent operations:

- The **generator** $G^{(s)} : \mathcal{F} \to \mathcal{G}^K$ takes the fact base and generates K **arguments** $\{g_j^{(s)}\}_{j=1}^{K} = G^{(s)}(F)$ per stance s, citing various facts from the fact base. This is inspired by Park et al. (2023) who recursively synthesize observations into higher-level reflections.
- The **critic** $C^{(s)} : \mathcal{F} \times \mathcal{G} \to \mathcal{C}$ criticizes an argument, producing a **critique** $c_j^{(s)} = C^{(s)}(F, g_j^{(s)})$ against it, possibly also citing facts from the fact base. The critic thereby acts as a *devil's advocate* (Kim et al., 2024) sparking a debate about the argument.
- The **evaluator** $E : \mathcal{G} \times \mathcal{C} \to \mathbb{N}$ takes an argument and corresponding critique and judges the convincingness of the argument with a **quality score** $\sigma_j^{(s)} = E(g_j^{(s)}, c_j^{(s)}) \in \mathbb{N}$. Internally, it uses a 14-criteria evaluation scheme based on the argument quality taxonomy by Wachsmuth et al. (2017). See Appendix B.3 for further details.
- The **refiner** $R^{(s)} : \mathcal{F} \times \mathcal{G} \times \mathbb{N} \to \mathcal{G}$ refines a given argument trying to improve its quality. It produces a **refined argument** $\tilde{g}_j^{(s)} = R^{(s)}(F, g_j^{(s)}, \sigma_j^{(s)})$.

organization of questions in our notation but represent question trees as simple sets. However, our code implementation preserves the full hierarchy.

As the refinement can be repeated, we use the notation $g_{j,t+1}^{(s)} = R^{(s)}(F, g_{j,t}^{(s)}, \sigma_{j,t}^{(s)})$ instead, where the index $t = 1, ..., T$ denotes the iteration.

Starting with an initial set of K_0 arguments $\{g_{j,0}^{(s)}\}_{j=1}^{K_0} = G^{(s)}(F)$, DIALECTIC iteratively critiques, evaluates, and refines the arguments. It hereby follows a *survival-of-the-fittest* logic, keeping only the best K_t arguments (the **survivors** S_t) after each iteration t:

$$S_{t+1}^{(s)} = TopK(\{g_{j,t+1}^{(s)} : g_{j,t}^{(s)} \in S_t^{(s)}\}, K_{t+1}),$$

where $TopK(\{\cdot\}, K_{t+1})$ denotes the K_{t+1} arguments with the highest quality scores $\sigma_{j,t+1}^{(s)}$ in $\{\cdot\}$. With arguments iteratively improving and K_t decreasing over the iterations, DIALECTIC converges to a narrow selection $S_T = S_T^{(+)} \cup S_T^{(-)}$ of high-quality pro and contra arguments. This mimics a debate in a VC investment committee where different members have different stances on the investment and continue to bring forward arguments until the room converges to a dominant narrative.

3.3 Decision-Making Phase

After T iterations of debate, a few dominant arguments for either stance have emerged. To determine which stance has the better arguments, we look at the sum of the argument quality scores for all surviving arguments and compare the pro and contra stances, calculating the **decision score** σ_T:

$$\sigma_T = \sigma_T^{(+)} - \sigma_T^{(-)} + \tau,$$

where $\sigma_T^{(s)} = \sum \sigma_{j,T}^{(s)}$ is the sum of the quality scores of all surviving arguments $g_{j,T}^{(s)} \in S_T^{(s)}$ and τ is a **decision threshold** capturing VC's preference for a margin of safety. Finally, DIALECTIC will decide to invest if and only if $\sigma_T > 0$.

3.4 Hyperparameters & Implementation

The above definition of DIALECTIC presents three main hyperparameters: The number of arguments kept per iteration K_t, the number of iterations T, and the decision threshold τ. In our implementation we set $K_t = 5$ for $t \neq T$ and test different values of K_T, T, and τ. For the LLM, we use OpenAI's `gpt-5-mini-2025-08-07` (OpenAI, 2025). We set the `temperature` parameter to 0.0 for the answer agent and to 0.5 for all other agents. We report all used prompts in Appendix C.

4 Evaluation Setup

We evaluate our method in a backtesting experiment by predicting startup success from historic data and benchmarking against real VC investors. Our dataset includes 259 startups that were added to real VCs' watchlists[2] between January 1, 2021 and December 31, 2021. The VCs considered joining the initial funding rounds (*seed* or *pre-seed*) of these startups, which were raised some time between January 1, 2021 and February 28, 2023.

Dependent variable Similar to prior work (Sharchilev et al., 2018; Gavrilenko et al., 2023), we define a startup as *successful*, if it has subsequently raised a *series A* or later round by September 1, 2025, otherwise as *unsuccessful*. With startup success as the dependent variable, our setup is a binary classification. Among all 259 startups, 25% were successful.

Independent variables To predict startup success, the following features are known for each startup: company name, short and long description, industry domain, team description, website content, and web search results (Table 2 in the Appendix presents descriptions of all features). These features are extracted from the VCs' watchlists, *Crunchbase.com*, startup homepages, and the *Perplexity Sonar API*. To prevent *look-ahead bias* (Żbikowski and Antosiuk, 2021), we use historic data snapshots and time filters to ensure all features have been available to the VC at the time of the investment decision. See Appendix A for a detailed description of the dataset and its creation.

Baselines We compare the classification performance of our method against the performance of the real VCs. The VCs invested in 6 of the 259 startups, and 2 of these were successful. Also, we compare against simple input-output (IO) prompting (see Listing 10 in the Appendix for the prompt).

Dataset split We split the data into a validation set with 129 startups and a test set with 130 startups using random stratified sampling, so that both sets have an equal ratio of successful startups and each set includes 3 startups that the VCs invested in.

Metrics To measure the performance of DIALECTIC and its baselines, we primarily look at

[2]The watchlists comprises data of five different VC funds and the real VCs' performance reported in this paper represents a weighted average across these funds.

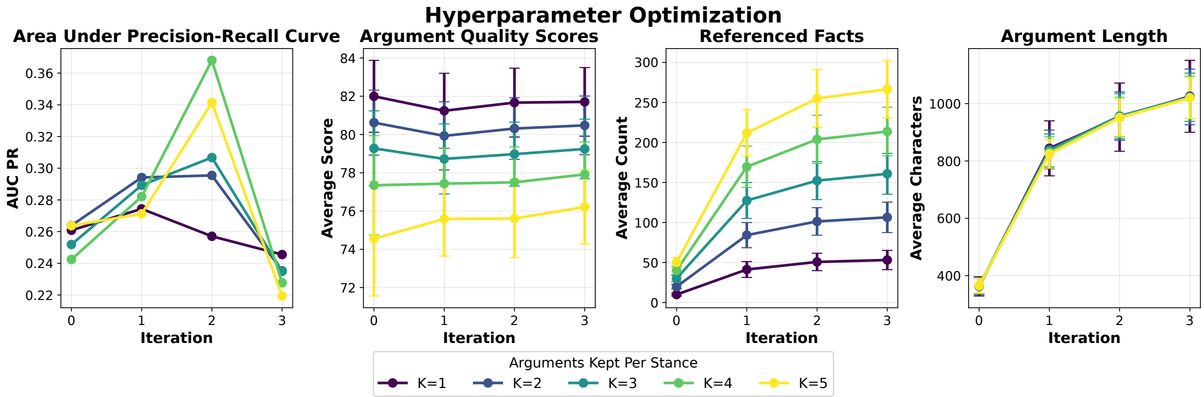

Figure 2: Results from the hyperparameter optimization, showing AUC-PR, raw argument scores, QA pair count, and argument length for different numbers of arguments K_t and iterations T.

precision and recall for different values of the decision threshold τ, as well as the area under curve (AUC) of the precision-recall (PR) line. We also assess the argument quality scores, number of cited facts, length of the arguments, as well as the distribution of decision scores. We first identify a well-performing combination of our hyperparameters T and K_T on the validation set. We then evaluate this configuration on the test set.

5 Results

Our results cover hyperparameter optimization, comparative predictive performance, and an analysis of which facts DIALECTIC uses.

5.1 Hyperparameter Optimization

We optimize the system by varying the number of surviving arguments per side (K_T) and the number of refinement iterations (T). These parameters control how broadly the system explores arguments and how deeply it refines them. Figure 2 summarizes the effect of varying these hyperparameters. Precision–recall performance shows a clear pattern: AUC-PR increases consistently from $T = 0$ to $T = 2$ and declines for $T \geq 3$. The best result occurs at $T = 2$ with $K_2 = 4$, which we use for subsequent experiments.

The diagnostic plots display coherent trends across other measures. Argument quality scores increase with more iterations, with only mild variation across K_t. Referenced facts rise with both parameters, suggesting stronger arguments rely on broader evidence. Argument length jumps from $T = 0$ to $T = 1$, driven by the introduction of structured justification, and grows more slowly thereafter as later iterations add elaboration rather than new insights. Because length and referenced facts increase smoothly while AUC-PR declines only at $T \geq 3$, the performance drop likely reflects over-refinement effects (e.g., redundancy or drift) rather than simple argument inflation.

5.2 Predictive Performance Against Baselines

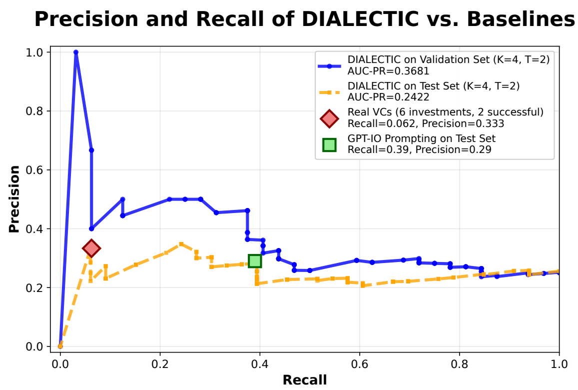

Figure 3: Precision and recall of DIALECTIC across all possible values of the decision threshold τ in comparison to the human VCs and GPT IO prompting baseline

Figure 3 reports the predictive performance of DIALECTIC on the validation set and the held-out test set. The system attains an AUC-PR of 0.2422 on the test set, with precision comparable to human investors and the GPT-IO prompting baseline. Performance is higher on the validation set, where it achieved higher AUC-PR and even outperformed the real VCs. The operating points in Figure 3 show that, in the high-precision, low-recall region, the system behaves similarly to the baselines. Unlike the baselines, it produces a full ranked frontier rather than a single operating point, which allows practitioners to choose a decision threshold tailored to screening capacity. Overall, DIALECTIC is comparable to baselines in predictive performance while offering a full decision frontier and ranking rather than a single operating point.

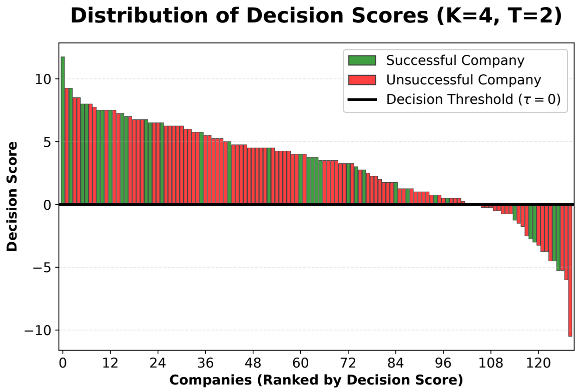

Figure 4: Distribution of decision scores.

Figure 4 shows the distribution of decision scores for the best-performing configuration. Successful companies cluster at the top of the ranking, (left end of the plot), while unsuccessful ones appear toward the bottom (right side of the plot). This separation shows that higher scores correspond to a higher likelihood of success. In practice, selecting a threshold simply involves choosing a cutoff along this ranking. A higher threshold prioritizes the strongest opportunities and filters out most low-scoring cases.

5.3 Evidence Utilization

Aspect	Usage	Availability	Ratio
General	34.40%	34.88%	0.986
Team	20.52%	20.17%	1.017
Product	29.77%	29.43%	1.011
Market	15.31%	15.51%	0.987

Table 1: Utilization of factual evidence in arguments. Aspect refers to the question trees built for the four seed questions. Usage measures the share of all cited facts, availability measures the relative size of question trees, and ratio is the ratio of usage and availability.

Table 1 summarizes how the model uses different evidence categories when generating factual references. General company and product information dominate, accounting for nearly 65% of all references, mirroring their share in the fact base. The "ratio" column reports the ratio between how often an aspect is referenced and its relative representation in the fact base. Values slightly above one (team: 1.017; product: 1.011) indicate that these aspects are referenced more frequently than their availability alone would predict. Market information is slightly under-used (efficiency < 1). Overall, ratios cluster near one, indicating proportional use of available evidence. Notably, team-related evidence receives the highest relative usage, which aligns with established findings that investors frequently prioritize founder and team attributes when forming investment judgments (Gompers et al., 2020).

6 Conclusion

This paper introduced DIALECTIC, an LLM-based multi-agent system for early-stage startup screening. The system integrates fact extraction, argument generation, iterative critique, and scoring into one pipeline. Evaluated on an industry-sourced dataset, the system achieves performance comparable to human investors while producing interpretable argument structures.

A central contribution of the approach is the introduction of iterative argumentation at the beginning of the investment funnel. Since investor bandwidth limits such deliberation during initial screening, it usually occurs later in the funnel. Enabling it earlier provides a structured foundation for preliminary assessments and supports reasoning under uncertainty.

Operationally, the system reduces time to initial assessment and produces a ranking when deal volume exceeds human screening capacity. It supports both fixed-threshold (returning only companies exceeding a certain decision score) and fixed-quantity (returning only the top N companies according to their decision scores) screening modes, reflecting constraints encountered in practice. Since missing strong opportunities is costlier than evaluating weak ones, the ranking mechanism aligns with recall-oriented objectives common in top-of-funnel screening. The generated arguments and evidence also support later stages such as due diligence or memo preparation.

Limitations

Venture capital is a domain of high uncertainty and low signal-to-noise ratios. This is also reflected by the relative instability of our results. Hyperparameter changes and sample composition can have a considerable impact on model performance, highlighted by the difference in AUC-PR seen between the validation and the test sets.

In investment decision tasks, avoiding look-ahead bias is a critical priority. We had to sacrifice more than 90% of the original companies in our dataset as a result of our efforts to minimize look-ahead bias and ensure feature completeness, leaving only 259 companies in the final dataset (see Appendix A). Similarly, this also restricted the corresponding baseline of real VCs' performance to data of only 6 invested companies.

While we took extensive measures to minimize look-ahead bias, a minor risk of bias remains. As we had access to a historic *Crunchbase* snapshot from January 24, 2022 only, about half of the companies in our dataset announced their seed or pre-seed funding round before that date making it theoretically possible that some of the *Crunchbase* information represents a lookahead. However, we consider it unlikely that this information provides substantial information about the future as the potentially affected data fields do not provide any information on future funding rounds. At the same time, our GPT-IO prompting baseline may be affected by look-ahead bias, if `GPT-5-mini` was trained on data about the companies in our dataset, possibly overstating the performance of this baseline. Long-term studies capturing reliable data and evaluating startup performance over a long timeframe would be needed to certainly rule out any risk of look-ahead bias.

As done by several previous studies (see Table 4 in the Appendix for an overview), we modeled startup evaluation as a success prediction task with a binary definition of success. Reality is more complex, as success may come in various levels and over longer timeframes. Also, the real impact a VC investor would have had on a startup's success after investing in it is in many cases an unobservable counterfactual that cannot be assessed in a backtesting study.

Lastly, the scope of our analysis is limited to early-stage VC investments (*pre-seed/seed* to *series A*) in European companies and may not generalize to other forms of VC or private equity.

References

Torben Antretter, Ivo Blohm, Dietmar Grichnik, and Joakim Wincent. 2019. Predicting new venture survival: A twitter-based machine learning approach to measuring online legitimacy. *Journal of Business Venturing Insights*, 11:e00109.

Javier Arroyo, Francesco Corea, Guillermo Jimenez-Diaz, and Juan A. Recio-Garcia. 2019. Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, 7:124233–124243.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. 2023. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12541–12560, Singapore. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Harrison Chase. 2022. Langchain. `https://github.com/langchain-ai/langchain`. Accessed: 2025-01-15.

Kimberly Chong and David Tuckett. 2014. Constructing conviction through action and narrative: how money managers manage uncertainty and the consequence for financial market functioning. *Socio-Economic Review*, 13(2):309–330.

Francesco Corea, Giorgio Bertinetti, and Enrico Maria Cervellati. 2021. Hacking the venture industry: An early-stage startups investment framework for data-driven investors. *Machine Learning with Applications*, 5:100062.

Dominik Dellermann, Nikolaus Lipusch, Philipp Ebel, Karl Michael Popp, and Jan Marco Leimeister. 2021. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. *Preprint*, arXiv:2105.03360.

Emily Gavrilenko, Foaad Khosmood, Mahdi Rastad, and Sadra Amiri Moghaddam. 2023. Improving startup success with text analysis. *Preprint*, arXiv:2312.06236.

Paul A. Gompers, Will Gornall, Steven N. Kaplan, and Ilya A. Strebulaev. 2020. How do venture capitalists make decisions? *Journal of Financial Economics*, 135(1):169–190.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. Llm multi-agent systems: Challenges and open problems. *CoRR*, abs/2402.03578.

IMARC Group. 2024. Venture capital investment market: Global industry trends, share, size, growth, opportunity and forecast 2025–2033. `https://www.imarcgroup.com/venture-capital-investment-market`. Market size projections 2024–2033.

Suprasith Jarupathirun and Fatemeh "Mariam" Zahedi. 2007. Dialectic decision support systems: System design and empirical evaluation. *Decision Support Systems*, 43(4):1553–1570. Special Issue Clusters.

Alex Kim, Keonwoo Kim, and Sangwon Yoon. 2024. DEBATE: Devil's advocate-based assessment and text evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1885–1897, Bangkok, Thailand. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. 2025. DeLLMa: Decision making under uncertainty with large language models. In *The Thirteenth International Conference on Learning Representations*.

Yiming Lu, Yebowen Hu, Hassan Foroosh, Wei Jin, and Fei Liu. 2025. STRUX: An LLM for decision-making with structured explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 131–141, Albuquerque, New Mexico. Association for Computational Linguistics.

Abdurahman Maarouf, Stefan Feuerriegel, and Nicolas Pröllochs. 2025. A fused large language model for predicting startup success. *European Journal of Operational Research*, 322(1):198–214.

OpenAI. 2025. Gpt-5-mini (2025-08-07). Large language model.

Ekin Ozince and Yiğit Ihlamur. 2024. Automating venture capital: Founder assessment using llm-powered segmentation, feature engineering and automated labeling techniques. *Preprint*, arXiv:2407.04885.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. The art of SOCRATIC QUESTIONING: Recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. Association for Computational Linguistics.

Andre Retterath. 2020. Human versus computer: Benchmarking venture capitalists and machine learning algorithms for investment screening. SSRN working paper.

Boris Sharchilev, Michael Roizner, Andrey Rumyantsev, Denis Ozornin, Pavel Serdyukov, and Maarten de Rijke. 2018. Web-based startup success prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2283–2291, New York, NY, USA. Association for Computing Machinery.

U.S. Census Bureau. 2025. Business formation statistics. `https://www.census.gov/econ/bfs/index.html`. Business applications and formations data, accessed 2025.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Sichao Xiong and Yigit Ihlamur. 2023. Founder-gpt: Self-play to evaluate the founder-idea fit. *Preprint*, arXiv:2312.12037.

Sichao Xiong, Yigit Ihlamur, Fuat Alican, and Aaron Ontoyin Yin. 2024. Gptree: Towards explainable decision-making via llm-powered decision trees. *Preprint*, arXiv:2411.08257.

Andrew L Zacharakis and Dean A Shepherd. 2001. The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, 16(4):311–332.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Kamil Żbikowski and Piotr Antosiuk. 2021. A machine learning, bias-free approach for predicting business success using crunchbase data. *Information Processing & Management*, 58(4):102555.

A Dataset Creation

For our backtesting experiment, we prepared a dataset of real startups from the watchlists of five VC funds (in the following we will refer to these watchlists in singular). The dataset was created by extracting and merging data from four sources: (1) the **watchlist** of a real VC, (2) **Crunchbase** (`crunchbase.com`), an online database aggregating information about businesses, (3) historical snapshots of the **startup websites** retrieved through the *Internet Archive*'s *Wayback Machine* (`web.archive.org`), and (4) **search results** obtained through the Perplexity Sonar API. Table 2 includes a summary of all extracted data fields and their origination dates.

A.1 Preventing Look-Ahead Bias

When working with historic data, it is important to consider which information was and was not available to the VC at the time the investment decision had to be made. Including information about the startups originating from a time after the VC's decision would constitute a *look-ahead bias* (Żbikowski and Antosiuk, 2021). Doing so could put the backtested method at an unfair advantage compared to the real VC, because it could leak information about the startup's future success. To prevent such a bias, we carefully filter the information available during backtesting by using historic data snapshots.

Cutoff date In order for the real VC to participate in the initial funding round, the investment decision had to be made at some point between the VC becoming aware of the startup (the date the startup was added to the VC's watchlist) and the announcement of the initial funding round, likely closer to the latter. Therefore, we consider the announcement date of the initial funding round as a cutoff and use it to restrict the information that we include in the dataset:

- **Watchlist:** We limit the data to startups that were added to the watchlist between January 1, 2021 and December 31, 2021 and had not yet raised a *series A* or later by the time they were added.

- **Crunchbase:** We had access to two *Crunchbase* snapshots taken on January 24, 2022 and September 1, 2025. We refer to these as the *historic* and the *current* snapshot, respectively. We use only the historic snapshot to extract predictive features. The current snapshot is used to determine whether a startup turned out successful (i.e., received subsequent funding).

- **Startup websites:** For each startup, we retrieve a historic snapshot of its website through *Wayback Machine* from the latest available date before the announcement of the initial funding round.

- **Search results:** For the *Perplexity Sonar API*, we apply a time filter to return only those search results that originate from a time before the announcement of the initial funding round.

A.2 Data Preprocessing

To construct the dataset, we started with companies that were added to the VC's watchlist between January 1, 2021 and January 31, 2021 and systematically enriched this set with data from *Crunchbase*, startup websites, and search results, removing companies where such enrichments were not possible. The preprocessing followed four phases: (1) data cleaning, (2) entity matching, (3) label assignment, and (4) enrichment. Table 3 reports the number of companies retained after each step and the corresponding share of successful companies.

1. **Watchlist export:** The starting dataset contained 3,441 companies from the VC's watchlist that were between January 1, 2021 and January 31, 2021.

2. **Cleaning:** We cleaned the dataset by removing duplicate entries, *Missed Deals*, and companies that were considered for a founding round later than *seed*. We further updated all companies website URLs by sending HTTP requests to the domains listed in the watchlist export and recording the final redirect target as the current domain. Finally, we canonicalized company names and URLs (lowercasing, removing prefixes such as "www", unicode normalization) for later matching purposes. After cleaning, the dataset contained 3,357 companies.

3. **Entity matching:** To enrich the watchlist records with additional data from *Crunchbase*, we performed a left-join between the cleaned dataset and the current *Crunchbase* snapshot and then the historic *Crunchbase* snapshot. Matching followed a strict precedence: (1)

Source	Data Field	Description	Value As Of
Watchlist	*Name* F	The company name.	Some time in 2021
	Domain F	The company web domain.	Some time in 2021
	Date Added P	The date the company was added to the watchlist. Only companies added between January 1, 2021 and December 31, 2021 are considered.	Some time in 2021
	Status P	The last stage reached in the investment process (e.g., *Added to Watchlist*, *Initial Review*, or *Investment Made*). Used to determine whether the real VC invested in the company.	Some time in 2021
Crunchbase	*Funding Rounds* P	A list of all funding rounds, including round type (*pre-seed*, *seed*, *series A*, *IPO*, etc.), amount, and announcement date of the round.	September 1, 2025
	Current Name P	The company's current name.	September 1, 2025
	Current Domain P	The company's current web domain.	September 1, 2025
Crunchbase	*Short Description* F	Short description of the company.	January 24, 2022
	Long Description F	Long description of the company.	January 24, 2022
	Industries F	A list of industries the company is operating in.	January 24, 2022
	Team F	The names of the team members, their education, prior work experience, and current roles.	January 24, 2022
	Historic Name P	The company's former name.	January 24, 2022
	Historic Domain P	The company's former web domain.	January 24, 2022
	Historic Funding Rounds P	A list of all historic funding rounds, including round type, amount, and announcement date.	January 24, 2022
Startup Websites	*Website* F	The historic HTML content of the company's website (only the homepage).	Various dates
	Archived P	The date and time when the website was captured.	Various dates
Search Results	*Results* F	The list of search results for a given query, including title, content snippet, and URL.	Various dates

Table 2: An overview of the data sources and extracted data fields used for the dataset creation. Data fields marked with F are used as predictive features (independent variables), whereas data fields marked with P are only used during preprocessing (e.g., for merging data from different sources) and were **not** made available to the prediction method.

exact current domain match, (2) exact historic domain match, and (3) fuzzy name-and-domain match with a similarity threshold of 95%. After matching, the dataset contained 1,623 companies.

4. **Label assignment:** Success labels were constructed from the current *Crunchbase* snapshot. Companies that had raised a series A or later funding round by September 1, 2025 were labeled as *successful*, all others as *unsuccessful*.

5. **Enrichment:** We further enriched each startup's data with additional historic information from the startup's website, web search results, and historic *Crunchbase* snapshot to extract predictive features including long and short company descriptions, industry, team setup, website content, and information from online articles. We removed companies where no founding team information was available, leaving 637 companies.

B DIALECTIC Design Details

B.1 Seed Questions

To kick off DIALECTIC's question decomposition, we provide it with a set of seed questions Q_0. These are intended to guide DIALECTIC's fact gathering efforts by giving it a rough scaffolding of relevant fact categories. VC investors typically assess startups across the following dimensions: general company, market, product/service, entrepreneurial team, and funding (Retterath, 2020). While rich and reliable funding information is typically private and was not available to us for every startup in our dataset, we dedicate one seed question to each of the remaining four dimensions. Specifically, we use the following four seed questions:

1. **General Company:** "How do the company's sector, development stage, and geography align with the VC's investment strategy?"

2. **Team:** "Who are the key members of the founding team, and what relevant experience and track record do they have?"

3. **Product:** "What are the product's core features, underlying technology, and forms of protection?"

4. **Market:** "What is the current size, historical growth, and forecast growth of the target market, and which customer needs or market gaps does the company address?"

B.2 Generating Question Trees

In order to evaluate each startup in more detail, DIALECTIC decomposes each seed question into lower-level questions tailored to the specific industry of the startup. Together, the seed questions and all their lower-level questions are supposed to comprehensively cover the information required by the VC investor to make a decision. In order to derive and answer lower-level questions from the seed questions, we adapt the *Probabilistic Tree-of-Thought Prompting* (*ProbTree*) approach (Cao et al., 2023). ProbTree uses an LLM to create hierarchical question decomposition trees (HQDTs) and then answer the questions in a post-order traversal using three different answer strategies **Open Book** (retrieving information from online sources), **Closed Book** (asking an LLM for its internal knowledge), and **Child Aggregation** (deriving the answer to a higher-level question from the answers to its lower level questions). For each answer strategy, it calculates a confidence score and then probabilistically chooses the most confident answers for each question.

For DIALECTIC, we use a simplified adaptation of ProbTree. It first decomposes a given seed question into a HQDT in a single LLM prompt. Then it performs a post-order traversal through the tree to answer the questions from leaf nodes to the root node. Unlike ProbTree, we use a single answer prompt for each node and therefore forgo the confidence estimation. Each prompt includes a company summary (description, tagline, and team details including education and prior work experience) and optional web data that the LLM can obtain agentically, if it decides to do so, by using a web search tool. We only allow usage of the web search tool for leaf nodes.

For the web searches, we provide the LLM with access to the *Perplexity Sonar API*. We limit search results to five, each described by a title and content snippet. As described in Appendix A.1, we restrict search results to those originating from a time before the announcement date of the initial funding round to prevent look-ahead bias.

Preprocessing Stage	Companies	Success Rate (%)
Export from CRM system	3,441	—
Remove duplicates and missed deals	3,404	—
Remove series A investments by VC funds	3,401	—
Match with current funding data	2,192	21.6
Remove companies without cutoff date	1,715	23.4
Remove companies added post–seed announcement	587	18.7
Apply temporal cutoff (before February 28, 2023)	462	22.0
Match with historic Crunchbase snapshot	259	25.1

Table 3: Dataset size and success rate after successive preprocessing steps. The success rate refers to the share labeled *successful* at the corresponding stage.

B.3 Evaluating Arguments

DIALECTIC includes an evaluator agent that assigns a numeric quality score for each argument. We use an LLM judge (Zheng et al., 2023) to evaluate each argument and apply the taxonomy of argument quality proposed by Wachsmuth et al. (2017). Following the instruction design principles by Wachsmuth et al. (2024), we adapt the taxonomy criteria to the VC context. The revised framework explicitly defines the objective of argumentation (informing the investment decision), establishes domain-specific criteria for argument quality, specifies the intended audience (expert VC investors), and incorporates the surrounding decision context (high-stakes financial environments). Our argument quality evaluation scheme includes the following 14 questions:

1. **Local Acceptability:** Are the premises believable and factually plausible given the provided Q&A facts?
2. **Local Relevance:** Do the premises clearly contribute to supporting or rejecting the conclusion about investment?
3. **Local Sufficiency:** Do the premises provide enough support to justify the conclusion?
4. **Cogency:** Does the argument have premises that are acceptable, relevant, and sufficient to support the investment conclusion?
5. **Credibility:** Does the argument make the author appear credible and trustworthy to VC investors?
6. **Emotional Appeal:** Does the argument create emotions that make the VC investors more receptive?
7. **Clarity:** Does the argument use correct and widely unambiguous language as well as avoid deviation from the issue?
8. **Appropriateness:** Is the style of reasoning and language suitable for a professional VC investment discussion?
9. **Arrangement:** Is the argument well-structured, with a logical order of premises and conclusion?
10. **Effectiveness:** Does the argument succeed in persuading the VC investors toward or against investing?
11. **Global Acceptability:** Would most VCs consider it a valid and legitimate argument?
12. **Global Relevance:** Does the argument meaningfully contribute to resolving the overall investment question?
13. **Global Sufficiency:** Does the argument adequately anticipate and rebut the main counterarguments from the argument's stance?
14. **Reasonableness:** Does the argument resolve the issue in a way acceptable to the VC investors, balancing global acceptability, relevance, and sufficiency?

The LLM judge scores each argument across the above 14 criteria using a seven-point Likert scale from 1 (low) to 7 (high). To calculate the final argument quality score, we simply sum up all of the 14 individual scores. The judge also produces justifications explaining each score.

C Prompts

In the following, we report the prompts that we used for each of our LLM-based agents. These are:

- **Decomposer** prompt (Listing 1)
- **Answer Agent** prompt (Listing 2)
- **Generator** prompt for pro (Listing 3) and for contra arguments (Listing 4)
- **Critic** prompt for pro (Listing 5) and for contra arguments (Listing 6)
- **Evaluator** prompt (Listing 7)
- **Refiner** prompt for pro (Listing 8) and for contra arguments (Listing 9)
- **Input Output (IO) Prompting** baseline prompt (Listing 10)

Listing 1: Decomposer Prompt

```
SYSTEM: You are good at decomposing a complex
     question into a hierarchical question
     decomposition tree (HQDT).

USER: Please generate a hierarchical question
     decomposition tree (HQDT) with json format
     for a given question. In this tree, the root
      node is the original complex question, and
     each non-root node is a sub-question of its
     parent.

Q: How large is the company's market opportunity
      (TAM, SAM, SOM)?
A: {{
  "How large is the company's market opportunity
      (TAM, SAM, SOM)?": [
    "What is the Total Addressable Market (TAM)
    ?",
    "What is the Serviceable Available Market (
    SAM)?",
    "What is the Serviceable Obtainable Market (
    SOM)?"
  ],
  "What is the Total Addressable Market (TAM)?":
      [
    "What customer segments are included in the
    broadest market?",
    "What is the total number of potential
    customers?",
    "What is the total industry revenue across
    those segments?"
  ],
  "What is the Serviceable Available Market (SAM
    )?": [
    "Which subset of TAM does the company's
    product or service directly target?",
    "What portion of customers can realistically
     be reached given geography, regulations, or
     product scope?",
    "What is the annual spending of these
    customers?"
  ],
  "What is the Serviceable Obtainable Market (
    SOM)?": [
    "What portion of SAM can the company
    realistically capture in the next 3-5 years
    ?",
    "What customer acquisition assumptions
    support this share?",
    "What expected adoption rate drives this
    forecast?",
    "What annual revenue corresponds to this
    achievable market share?"
  ]
}}

Q: What is the competitive landscape, and how is
     the company positioned within it?
A: {{
  "What is the competitive landscape, and how is
     the company positioned within it?": [
    "What is the competitive landscape?",
    "How is the company positioned within the
    competitive landscape?"
  ],
  "What is the competitive landscape?": [
    "Who are the direct competitors?",
    "Who are the indirect competitors or
    substitutes?",
    "What are the major trends shaping
    competition in this space?"
  ],
  "How is the company positioned within the
    competitive landscape?": [
    "What is the company's relative pricing
    strategy?",
    "What is the company's market share or
    traction compared to peers?",
    "Does the company occupy a niche or broader
    category?",
    "What barriers to entry protect the company'
    s position?"
  ]
}}

Q: What is the company's product differentiation
     and value proposition?
A: {{
  "What is the company's product differentiation
     and value proposition?": [
    "What is the company's product
    differentiation?",
    "What is the company's value proposition?"
  ],
  "What is the company's product differentiation
    ?": [
    "What features or technologies distinguish
    the product?",
    "How is the product better than alternatives
    ?",
    "What intellectual property (e.g., patents,
    proprietary tech) supports defensibility?"
  ],
  "What is the company's value proposition?": [
    "What problem does the product solve for
    customers?",
    "What measurable benefits (e.g., cost
    savings, time savings, revenue uplift) does
    it deliver?",
    "Why would customers choose this company
    over competitors?"
```

```
  ]
}}

Here is the question to decompose:
Q: {question}

Generate its HQDT customized for a company in
     the {industry} industry.
```

Listing 2: Answer Agent Prompt

```
SYSTEM: Answer the question using company
     summary and sub Q&A if provided. Keep answer
      concise (<50 words) with data backing.
If unable to answer the question, use web_search
      for market data, trends, competitive
     analysis, funding info. Focus on industry-
     level searches, not specific companies. Use
     the tool only if necessary.
Make ONE tool call at a time.

USER: Question: {question}

Company summary: {company_summary}
{qa_pairs}
```

Listing 3: Generator Prompt (Pro Arguments)

```
SYSTEM: You are a very experienced investor at a
      top-tier VC fund. You are also a great
     storyteller and can tell a compelling story.

USER: Generate {n_pro_arguments} pro arguments
     why this company is a good investment
     opportunity.

Each argument should be concise (max. 100 words)
      and backed by specific data from the
     questions and answers.

A good argument provides a unique perspective on
      the investment opportunity that addresses
     the following criteria:
1. Local Acceptability - Are the premises
     believable and factually plausible given the
      provided Q&A facts?
2. Local Relevance - Do the premises clearly
     contribute to supporting or rejecting the
     conclusion about investment?
3. Local Sufficiency - Do the premises provide
     enough support to justify the conclusion?
4. Cogency - Does the argument have premises
     that are acceptable, relevant, and
     sufficient to support the investment
     conclusion?
5. Credibility - Does the argument make the
     author appear credible and trustworthy to VC
      investors?
6. Emotional Appeal - Does the argument create
     emotions that make the VC investors more
     receptive?
7. Clarity - Does the argument use correct and
     widely unambiguous language as well as avoid
      deviation from the issue?
8. Appropriateness - Is the style of reasoning
     and language suitable for a professional VC
     investment discussion?
9. Arrangement - Is the argument well-structured,
      with a logical order of premises and
     conclusion?
10. Effectiveness - Does the argument succeed in
      persuading the VC investors toward or
     against investing?
11. Global Acceptability - Would most VCs
     consider it a valid/legitimate argument?
12. Global Relevance - Does the argument
     meaningfully contribute to resolving the
     overall investment question?
13. Global Sufficiency - Does the argument
     adequately anticipate and rebut the main
     counterarguments from the argument's stance?
14. Reasonableness - Does the argument resolve
     the issue in a way acceptable to the VC
     investors, balancing global acceptability,
     relevance, and sufficiency?

Here are the questions and answers about the
     company:
{qa_pairs}

Provide the qa_indices that were used to
     generate the argument.
```

Listing 4: Generator Prompt (Contra Arguments)

```
SYSTEM: You are a very experienced investor at a
      top-tier VC fund. You are also a great
     storyteller and can tell a compelling story.

USER: Generate {n_contra_arguments} contra
     arguments why this company is a bad
     investment opportunity.

Each argument should be concise (2-3 sentences)
     and backed by specific data from the
     questions and answers.
Lack of data is not a good contra argument.

A good argument provides a unique perspective on
      the investment opportunity that addresses
     the following criteria:
1. Local Acceptability - Are the premises
     believable and factually plausible given the
      provided Q&A facts?
2. Local Relevance - Do the premises clearly
     contribute to supporting or rejecting the
     conclusion about investment?
3. Local Sufficiency - Do the premises provide
     enough support to justify the conclusion?
4. Cogency - Does the argument have premises
     that are acceptable, relevant, and
     sufficient to support the investment
     conclusion?
5. Credibility - Does the argument make the
     author appear credible and trustworthy to VC
      investors?
6. Emotional Appeal - Does the argument create
     emotions that make the VC investors more
     receptive?
7. Clarity - Does the argument use correct and
     widely unambiguous language as well as avoid
      deviation from the issue?
8. Appropriateness - Is the style of reasoning
     and language suitable for a professional VC
     investment discussion?
9. Arrangement - Is the argument well-structured,
      with a logical order of premises and
     conclusion?
10. Effectiveness - Does the argument succeed in
      persuading the VC investors toward or
```

```
    against investing?
11. Global Acceptability - Would most VCs
    consider it a valid/legitimate argument?
12. Global Relevance - Does the argument
    meaningfully contribute to resolving the
    overall investment question?
13. Global Sufficiency - Does the argument
    adequately anticipate and rebut the main
    counterarguments from the argument's stance?
14. Reasonableness - Does the argument resolve
    the issue in a way acceptable to the VC
    investors, balancing global acceptability,
    relevance, and sufficiency?

Here are the questions and answers about the
    company:
{qa_pairs}

Provide the qa_indices that were used to
    generate the argument.
```

Listing 5: Critic Prompt (Pro Arguments)

```
SYSTEM: You are a very experienced VC investor
    against investing in the company. However,
    your colleague thinks it is a good
    investment opportunity.
Your job is to criticize the pro argument given
    by your colleague using the questions and
    answers about the company and defend your
    position.
Be direct to persuade your colleague not to
    invest in the company.

USER: Here are the questions and answers about
    the company:
{qa_pairs}

Here is the argument you have to criticize to
    persuade the colleague not to invest in the
    company:
{argument}

Keep your critique concise in 3-4 sentences.
```

Listing 6: Critic Prompt (Contra Arguments)

```
SYSTEM:You are a very experienced VC investor in
     favor of investing in the company. However,
     your colleague thinks it is a bad
    investment opportunity.
Your job is to criticize the given contra
    argument given by your colleague using the
    questions and answers about the company and
    defend your position.
Be direct to persuade your colleague to invest
    in the company.

USER: Here are the questions and answers about
    the company:
{qa_pairs}

Here is the argument you have to criticize to
    persuade the colleague to invest in the
    company:
{argument}

Keep your critique concise in 3-4 sentences.
```

Listing 7: Evaluator Prompt

```
SYSTEM: You are an impartial LLM judge to
    evaluate the quality of an argument in the
    VC investment context. The goal of the
    argument is to support or reject a startup
    investment decision in a persuasive way.
The quality of an argument in the venture
    capital investment context should be
    evaluated along the following 14 dimensions.
     For each dimension, assign a score from 1 (
    Low) to 7 (High), and provide a short
    feedback (1 sentence) how to improve the
    score.

14 Dimensions:
1. Local Acceptability - Are the premises
    believable and factually plausible given the
     provided Q&A facts?
2. Local Relevance - Do the premises clearly
    contribute to supporting or rejecting the
    conclusion about investment?
3. Local Sufficiency - Do the premises provide
    enough support to justify the conclusion?
4. Cogency - Does the argument have premises
    that are acceptable, relevant, and
    sufficient to support the investment
    conclusion?
5. Credibility - Does the argument make the
    author appear credible and trustworthy to VC
     investors?
6. Emotional Appeal - Does the argument create
    emotions that make the VC investors more
    receptive?
7. Clarity - Does the argument use correct and
    widely unambiguous language as well as avoid
     deviation from the issue?
8. Appropriateness - Is the style of reasoning
    and language suitable for a professional VC
    investment discussion?
9. Arrangement - Is the argument well-structured,
     with a logical order of premises and
    conclusion?
10. Effectiveness - Does the argument succeed in
     persuading the VC investors toward or
    against investing?
11. Global Acceptability - Would most VCs
    consider it a valid/legitimate argument?
12. Global Relevance - Does the argument
    meaningfully contribute to resolving the
    overall investment question?
13. Global Sufficiency - Does the argument
    adequately anticipate and rebut the main
    counterarguments from the argument's stance?
14. Reasonableness - Does the argument resolve
    the issue in a way acceptable to the VC
    investors, balancing global acceptability,
    relevance, and sufficiency?

USER: Argument to evaluate:
{argument}
{critique}
...
```

Listing 8: Refiner Prompt (Pro Arguments)

```
SYSTEM: You are a very experienced investor at a
     top-tier VC fund. You are sure that the
    company is a good investment opportunity.
Your job is to revise your argument to reach
```

```
    better argument quality scores.

USER: Here are the Q&A facts about the company:
{qa_pairs}

Here is your previous argument:
{argument}

Here are the argument quality scores (1-7) to
    your previous argument:
{argument_feedback}

Refine your argument by improving argument
    quality scores.
```

Listing 9: Refiner Prompt (Contra Arguments)

```
SYSTEM: You are a very experienced investor at a
     top-tier VC fund. You are sure that the
    company is a bad investment opportunity.
Your job is to revise your argument to reach
    better argument quality scores.

USER: Here are the Q&A facts about the company:
{qa_pairs}

Here is your previous argument:
{argument}

Here are the argument quality scores (1-7) to
    your previous argument:
{argument_feedback}

Refine your argument by improving argument
    quality scores.
```

Listing 10: Input Output (IO) Prompting Baseline Prompt

```
SYSTEM: Assuming you are a venture capital
    investor, would you invest in the following
    company? Respond with only "Yes" or "No".

USER: Questions and Answers for the company:
{qa_pairs}
```

D Related Work

Table 4 provides an overview of related work, i.e., studies that propose startup success prediction methods based on machine learning. The table also shows the success criteria used by these studies, the accuracy, recall, and precision achieved by the best-performing models, as well as the interpretability approach taken, if any.

Work	Success Criterion	Accuracy	Recall	Precision	Interpretability
Arroyo et al. (2019)	First event in 3 yrs (AC = acquired, FR = funding round, IPO, CL = closed, NE = no event)	Global 82.2%	FR: 40%, AC: 3%, IPO: very low, NE: 95%	FR: 64%, AC: 33%, IPO: 44%, NE: 85%	Feature-based
Żbikowski and Antosiuk (2021)	AC, IPO, Series B	85%	34%	57%	Feature-based
Retterath (2020)	Follow-on round, trade sale, IPO	80%	80%	–	No mention
Antretter et al. (2019)	5-year survival	76%	86%	80%	Feature-based
Sharchilev et al. (2018)	Series A+ within 1 yr	–	–	62.6%	Feature-based
Gavrilenko et al. (2023)	Raise Series A+ within 1 yr	–	82.7%	74.4%	Feature-based
Maarouf et al. (2025)	IPO, AC, or funding	74.3%	78.3%	59.8%	Feature-based
Ozince and Ihlamur (2024)	IPO/AC/funding >$500M	66.7%	64.7%	68.8%	Persona-based
Xiong and Ihlamur (2023)	N/A		No backtesting		Pro/contra arguments
Xiong et al. (2024)	IPO/AC/funding >$500M	87.6%	27.1%	37.3%	No mention
Corea et al. (2021)	IPO, AC, or funding	–	–	–	Feature-based

Table 4: Comparison of startup success prediction studies: success criteria, predictive performance of the best model, and interpretability.