# Structured Reasoning Frameworks for LLM-Based Venture Capital Evaluation: Integrating Systems Engineering and Business Analysis

Dan Velarde, Stephen Gillespie
United States Military Academy
West Point, New York, USA
Emails: dan.velarde@westpoint.edu, stephen.gillespie@westpoint.edu

***Abstract*—Venture capital decision-making for early-stage startups involves high uncertainty and relies heavily on qualitative reasoning from product and business proposals called pitch decks. Traditional evaluation approaches depend on idiosyncratic heuristics that vary significantly across individual investors, limiting consistency and transparency in investment decisions. This study examines whether large language models (LLMs) augmented with structured knowledge frameworks can enhance predictive accuracy in early-stage startup evaluation. We designed four evaluation personas that are hosted on Gemini 2.5 Pro Preview: a baseline model, a venture capital-informed model, a systems engineering-guided model using established frameworks from IEEE, INCOSE, NASA, and MITRE, and a combined model integrating both frameworks. Each persona evaluated startups through a single-pass protocol that mirrors real venture capital constraints, where investors typically have only one opportunity to assess pitch materials. Testing on 14 historical pitch decks from companies including LinkedIn, Theranos, and WeWork, the combined persona achieved 85.7% accuracy in predicting startup outcomes, outperforming specialized and baseline configurations. The systems engineering rubric provided structured technical risk assessment across cost, schedule, and programmatic dimensions, while the venture capital knowledge stack enhanced market opportunity identification and business model evaluation. Results demonstrate that LLMs can function as structured reasoning engines when constrained by domain-specific frameworks, offering transparent and auditable decision support for high-uncertainty investment environments.**

***Index Terms*—Large language models, venture capital, systems engineering, artificial intelligence, decision support systems, startup evaluation, structured reasoning**

## I. Introduction

### A. Background and Motivation

Venture capital decision-making involves evaluating early-stage companies that often lack operational history, financial maturity, or technical validation. These high-uncertainty judgments rely on qualitative reasoning and expert interpretation of pitch decks. Systems engineering provides established methodologies for risk assessment, lifecycle evaluation, and technical feasibility analysis through frameworks codified by major organizations including the International Council on Systems Engineering (INCOSE), the Institute of Electrical and Electronics Engineers (IEEE), the National Aeronautics and Space Administration (NASA), and MITRE Corporation.

Recent advances in generative AI, particularly large language models (LLMs) [1], have introduced new possibilities for automating investment evaluations. LLMs can process unstructured data, extract context, and generate structured reasoning across domains including finance and engineering. Prior studies have demonstrated that language models can identify factors associated with startup performance [2], [3], and industry applications have explored AI-assisted pitch deck screening. However, prior work primarily relies on structured datasets from platforms like Crunchbase, focusing on numerical or categorical founder profiles and company metrics rather than the unstructured reasoning contained in pitch decks [2], [4], [5].

This study extends current research by testing whether LLMs can replicate both technical systems engineering analysis and business reasoning using only the information contained in pitch decks. While recent work has explored LLM-based founder evaluation [6] and prediction from free-form text descriptions [7], no prior research integrates structured systems engineering frameworks with venture capital reasoning for pitch deck analysis. By combining established systems engineering standards with venture capital knowledge stacks, we investigate whether structured reasoning can improve both accuracy and interpretability in startup evaluation.

### B. Research Significance

This research combines systems engineering, venture capital analysis, and artificial intelligence to test whether structured reasoning frameworks can improve predictive reliability in early-stage investment evaluation. From a systems engineering perspective, applying lifecycle evaluation frameworks and quantitative risk analysis to investment decisions introduces traceability and repeatability to typically subjective processes. Major systems engineering organizations including IEEE, INCOSE, NASA, and MITRE have developed comprehensive standards that provide fewer but more encompassing resources compared to fragmented venture capital literature. For AI research, the study demonstrates that LLMs can operate as structured reasoning engines rather than pure pattern generators when guided by domain-specific knowledge.

### C. Academic Gap

Despite growing interest in AI for venture analysis, key gaps remain. Prior work rarely uses unstructured pitch decks

as the primary data source, instead focusing on numerical or categorical datasets derived from structured databases [2], [4], [5]. Few studies combine engineering rigor with venture logic to examine how structured reasoning affects interpretability and justification.

This study addresses those gaps by incorporating a graded systems engineering rubric derived from established standards into LLM evaluation, enabling transparent and auditable reasoning. It examines not only the correctness of model outputs but the logic behind them through confidence ratings and rationale summaries, contributing to discussions on how AI can function as a structured reasoning partner in uncertain, high-stakes decision environments.

## II. Methodology

### A. Research Hypothesis

The central hypothesis of this study is that large language models equipped with structured systems engineering and venture capital knowledge frameworks will demonstrate higher accuracy, interpretability, and reasoning consistency in evaluating startup success than unstructured baseline models. The combined configuration, which merges the graded systems engineering rubric with venture capital reasoning, is expected to yield the most balanced performance by integrating both technical and market perspectives into a unified analytical process.

### B. Overview of Evaluation Framework

Figure 1 illustrates the complete evaluation framework. Historical pitch decks serve as input to the Gemini 2.5 Pro model operating within the MSTY.ai environment. Two knowledge stacks—Systems Engineering and Venture Capital—are indexed using BERT embeddings and made available for retrieval during inference. Four distinct personas process each pitch deck, generating investment decisions, confidence scores, and reasoning summaries. These outputs are then compared against historical company outcomes to compute accuracy metrics.

### C. Data Collection and Preparation

The dataset consisted of 14 historical startup pitch decks evenly divided between successful and failed ventures. The successful set included LinkedIn, Intercom, Coinbase, Monzo, Buffer, WeWork, and Tinder. The failed set included Bliss, Cardlife, Castle, FlowTab, LimeTree, Spartan, and Theranos. Each pitch deck was obtained from SlideShare [8]. Pitch decks were converted to text using optical character recognition and manually cleaned to ensure consistent formatting and readability. This balanced dataset design enabled fair evaluation across both successful and unsuccessful startup outcomes.

### D. Knowledge Stack Construction and Integration

Two distinct knowledge repositories were constructed to support domain-specific reasoning:

*1) Systems Engineering Knowledge Stack:* The Systems Engineering Knowledge Stack comprised authoritative materials from major standards organizations: the IEEE 15288 standard for systems and software engineering lifecycle processes [9], the INCOSE Systems Engineering Handbook Fourth Edition [10], the MITRE Systems Engineering Guide [11], and the NASA Systems Engineering Handbook [12]. These sources were selected for their comprehensive coverage of risk assessment frameworks, lifecycle evaluation methodologies, and technical maturity criteria. The materials provided structured rubrics for evaluating cost risk, schedule risk, technical risk, programmatic risk, and system quality attributes (ilities).

*2) Venture Capital Knowledge Stack:* The Venture Capital Knowledge Stack drew from investment theory, valuation frameworks, due diligence guidelines, and historical case studies [13]–[17]. These materials covered market opportunity assessment, founder evaluation heuristics, business model validation, competitive analysis, and scaling dynamics. The knowledge base emphasized the qualitative and market-oriented reasoning typical of early-stage investment decisions.

*3) Knowledge Indexing and Retrieval:* Both knowledge stacks were vectorized and indexed using BERT embeddings integrated into the MSTY.ai [18] desktop environment. This configuration enabled semantic retrieval during model inference, allowing relevant passages from the knowledge stacks to be dynamically retrieved based on the content of each pitch deck. The retrieval system operated locally without requiring external connectivity, ensuring consistent access to domain knowledge across all evaluations.

### E. Evaluation Environment and Model Configuration

All evaluations were conducted using Google Gemini 2.5 Pro [19] within the MSTY.ai [18] desktop environment hosted locally on a workstation equipped with an NVIDIA RTX 5080 graphics processor and an Intel Core i9 central processor. The model configuration was identical across all four personas to ensure comparability. The evaluation environment was configured with a maximum output token limit of 16,000 and a context length of 30 messages, with all other parameters maintained at default settings.

### F. Persona Configurations and Prompt Engineering

Four personas were designed to simulate distinct reasoning frameworks representative of real-world evaluators. Each persona was defined through carefully constructed system prompts that established role, analytical framework, and output format.

*1) Basic Persona:* The Basic Persona served as the experimental control. It received no domain-specific knowledge beyond the pitch deck text itself. The system prompt instructed the model to act as an independent early-stage investor evaluating the startup opportunity based solely on the information presented in the pitch deck. This configuration tested the baseline reasoning capability of the large language model without structured knowledge augmentation. Despite its simplicity, the Basic Persona achieved 78.6% accuracy, demonstrating that

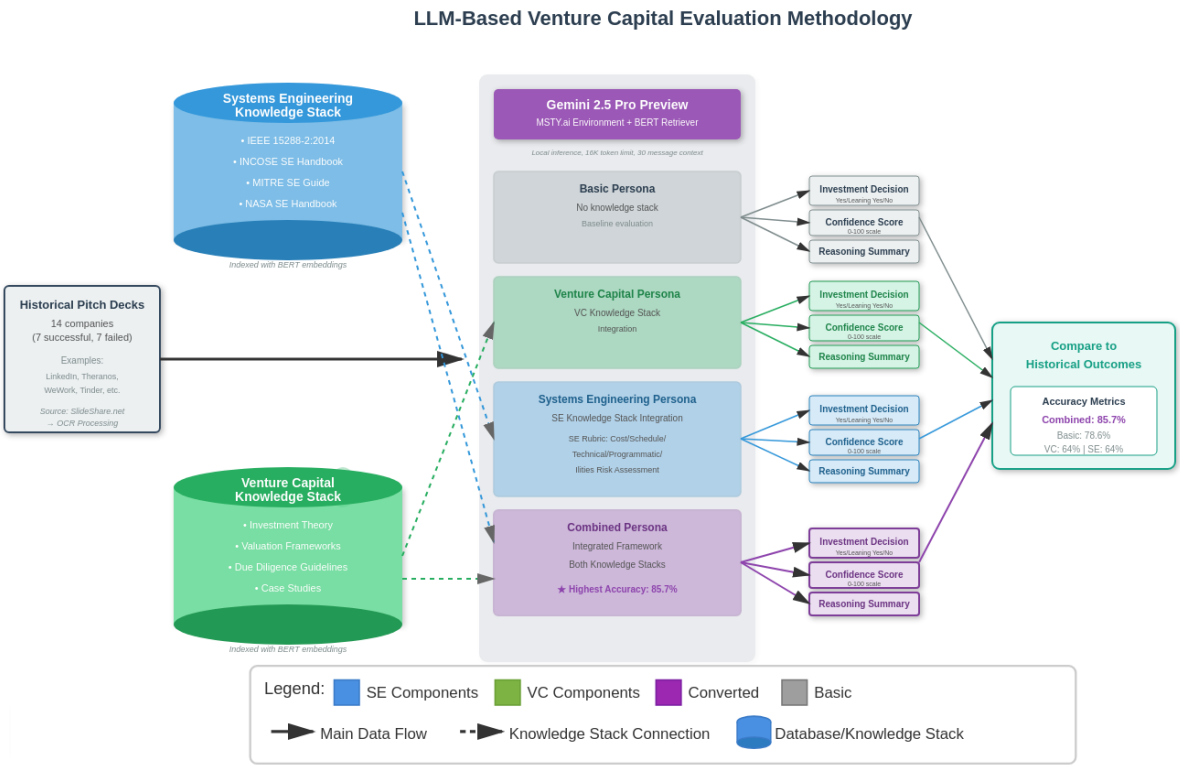


Fig. 1: Methodology workflow showing the evaluation framework with four personas (Basic, Venture Capital, Systems Engineering, Combined), knowledge stack integration, and evaluation pipeline from pitch deck input to investment decision output.

foundational LLM reasoning can provide meaningful signal even without specialized knowledge frameworks.

*2) Venture Capital Persona:* The Venture Capital Persona was equipped with access to the Venture Capital Knowledge Stack. The system prompt directed the model to evaluate pitch decks through the lens of investment heuristics, market opportunity assessment, and founder credibility—the qualitative factors that dominate early-stage investment decisions. The persona was instructed to prioritize scalability, market timing, competitive positioning, and team capabilities. During inference, relevant passages from venture capital literature were retrieved to inform the model's reasoning process.

*3) Systems Engineering Persona:* The Systems Engineering Persona was guided by the Systems Engineering Knowledge Stack. The system prompt instructed the model to evaluate pitch decks using structured risk assessment rubrics derived from IEEE, INCOSE, NASA, and MITRE standards. The persona applied quantitative criteria across five evaluation categories: Cost Risk, Schedule Risk, Technical Risk, Programmatic Risk, and System Ilities. Each category was subdivided into four subcriteria, scored between 0 and 25 points, for a total of 100 points per category. Individual category scores were averaged to produce a Project Risk Score representing overall risk exposure. This persona emphasized technical feasibility, integration complexity, and lifecycle maturity. A detailed example of this rubric applied to LinkedIn's pitch deck is presented in Section 3.4.

*4) Combined Persona:* The Combined Persona integrated both knowledge stacks into a unified analytical framework. The system prompt directed the model to balance technical credibility (from systems engineering) against market opportunity (from venture capital reasoning). This persona had access to both knowledge repositories during retrieval and was instructed to synthesize insights from both domains. The Combined Persona achieved the highest accuracy at 85.7%, validating the hypothesis that interdisciplinary reasoning improves predictive performance in startup evaluation.

### *G. Single-Pass Evaluation Protocol*

A critical design decision was the use of single-pass evaluation to simulate realistic venture capital conditions. In real investment environments, investors typically review pitch decks once during initial screening before deciding whether to advance a company to deeper due diligence. To replicate this constraint, each persona analyzed each pitch deck exactly once, generating a single set of outputs without iteration, refinement, or multiple sampling.

This protocol differed from common LLM evaluation practices that rely on ensemble methods, majority voting, or iterative refinement. By constraining the model to a single inference pass, the study tested whether structured knowledge frameworks could produce reliable judgments under conditions that mirror actual venture capital workflows. The strong performance of the Combined Persona (85.7%) under these constraints suggests

that knowledge-augmented LLMs can provide decision support even when limited to single-pass evaluation.

### H. Output Structure and Evaluation Metrics

Each persona produced three structured outputs per pitch deck:

1) **Investment Decision**: A categorical judgment (Hard Yes, Leaning Yes, Leaning No, No) indicating the strength of the recommendation.
2) **Confidence Score**: A numerical rating between 0 and 100 representing the model's confidence in its decision.
3) **Reasoning Summary**: A concise textual explanation of the rationale supporting the investment decision, highlighting key factors from the pitch deck.

All four personas evaluated all 14 pitch decks, generating 56 total evaluations. A classification was considered correct if the decision aligned with the company's historical outcome: for successful ventures, "Yes" or "Leaning Yes" counted as correct; for failed ventures, "No" or "Leaning No" counted as correct. Confidence values were retained for descriptive analysis but were not used to adjust accuracy scores. This binary classification approach reflected the fundamental venture capital question: should we invest or pass?

### I. Comparison to Historical Outcomes

Model predictions were evaluated against known company outcomes. Successful companies were defined as those that achieved significant market traction, sustained operations, or successful exits through acquisition or public offering. Failed companies were defined as those that ceased operations, filed for bankruptcy, or experienced fundamental business model failures. Ground truth labels were assigned based on publicly available information about each company's trajectory following the pitch deck creation date. Accuracy metrics were computed as the percentage of correct classifications across all evaluations for each persona.

## III. Results

### A. Descriptive Statistics

The Combined persona achieved the highest overall accuracy at 85.7%, followed by the Basic persona at 78.6%. The Venture Capital and Systems Engineering personas each achieved approximately 64.3% accuracy. The Combined persona demonstrated balanced reasoning between technical feasibility and market potential, resulting in the fewest misclassifications.

Confidence scores generally correlated with correctness. Evaluations above 70% confidence were more likely to align with historical outcomes, while low-confidence responses tended to appear on decks with incomplete or inconsistent data.

### B. Comparative Analysis

Analysis confirmed that structured reasoning frameworks improved both accuracy and interpretability. The Systems Engineering rubric enabled recognition of technical failure conditions, while the Venture Capital stack improved identification of scalable business opportunities. Confusion matrices revealed behavioral tendencies: the Venture Capital persona showed a higher rate of false positives, and the Systems Engineering persona was more conservative. The Combined persona minimized both tendencies.

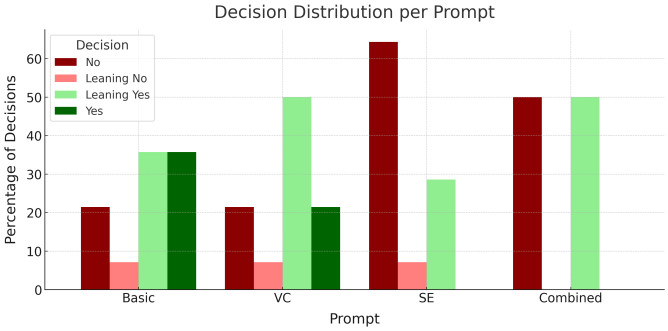


Fig. 2: Decision distribution per persona showing percent of each decision type (Yes, Leaning Yes, Leaning No, No).

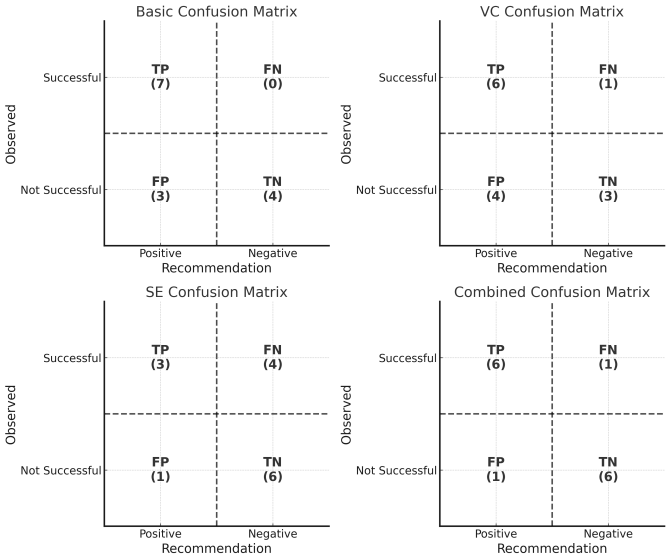


Fig. 3: Confusion matrix for the Combined persona: predictions versus actual outcomes (6 true positives, 6 true negatives, 1 false positive, 1 false negative).

Review of reasoning summaries revealed differences in cognitive style. The Basic persona produced brief heuristic justifications. The Venture Capital persona emphasized founder strength, traction, and market scale. The Systems Engineering persona cited integration complexity, schedule realism, and design maturity. The Combined persona synthesized these factors.

### C. Systems Engineering Rubric Application: LinkedIn Case Study

To illustrate the structured quantitative approach of the Systems Engineering persona, Table I presents the complete

rubric evaluation for LinkedIn's Series B pitch deck. This evaluation demonstrates how the SE framework systematically assesses risk across five major categories, each subdivided into four specific criteria. The rubric produces both numerical scores and qualitative rationales that trace back to specific evidence (or absence thereof) in the pitch deck.

The LinkedIn evaluation reveals several critical insights about how the SE rubric operates. First, it identifies specific documentation gaps that traditional VC analysis might overlook—the absence of verification plans, cost control systems, and dependency management. Second, it quantifies maturity disparities: while the core platform demonstrates high technical readiness (TRL 7/8), the revenue-generating features remain conceptual (TRL 3/4), creating integration risk. Third, the rubric captures positive signals that align with business success: stakeholder alignment scores 20/25, and usability receives 22/25 based on demonstrated viral growth.

Despite LinkedIn's eventual success, the SE persona correctly identified legitimate technical and programmatic risks present at the Series B stage. The average Project Risk Score of 35.4/100 reflects moderate-to-high risk, primarily driven by Cost Risk (15/100) and Schedule Risk (25/100). These scores influenced the SE persona's conservative evaluation stance, demonstrating that the rubric captures real uncertainty even for companies that ultimately succeed.

The structured format enables traceability: each score links to specific evidence (or its absence) in the pitch deck. This transparency contrasts with the more holistic, narrative-driven reasoning of the Venture Capital persona, which emphasized market opportunity and team strength without systematically assessing technical maturity or schedule feasibility.

### D. Confidence Calibration and Prediction Reliability

Confidence scores ranged from 40 to 100, with a mean of 82.5 across evaluations. The Combined persona showed the strongest confidence-accuracy correlation, with 91.7% of high-confidence predictions ($\geq$85) being correct. Confidence thresholds corresponded to meaningful differences in accuracy ($\geq$85: 84.6%; ¡75: 62.5%).

### E. Comparative Case Analysis

Three illustrative cases:

**Theranos:** VC and Basic personas recommended investment; SE and Combined correctly rejected due to unsubstantiated technical claims and missing regulatory pathway.

**Tinder:** SE and Combined rejected due to lack of technical detail; Basic correctly identified opportunity based on product novelty.

**LinkedIn:** All personas recommended investment; rationales varied across technical and market perspectives. The detailed SE rubric analysis (Table I) demonstrates how systematic risk assessment identified legitimate concerns even for an ultimately successful venture.

### F. Error Pattern Classification

Three primary error categories emerged:

- **Technical Conservatism (5 cases)**: SE persona produced false negatives for ventures lacking architecture documentation.
- **Narrative Bias (5 cases)**: VC and Basic personas produced false positives driven by founder story and traction claims.
- **Information Asymmetry (2 cases)**: Combined persona misclassified Bliss and Tinder where signals conflicted.

### G. Decision Distribution and Risk Tolerance

The Systems Engineering persona was most conservative (35.7% No), the Venture Capital persona most optimistic (35.7% Yes), and the Combined persona most balanced (42.9% Leaning Yes).

## IV. Discussion and Conclusion

### A. Summary of Findings

The study tested whether structured reasoning frameworks improve LLM evaluation of pitch decks. The Combined persona achieved the best predictive accuracy (85.7%) and the strongest confidence calibration, supporting the hypothesis that integrating systems engineering and venture capital frameworks enhances both accuracy and interpretability.

### B. Implications

LLMs can be guided by structured frameworks to emulate interdisciplinary reasoning. Constraining LLM reasoning with domain-specific rubrics produces outputs that are more traceable and auditable, suitable for decision-support contexts in high-uncertainty domains.

### C. Limitations and Threats to Validity

Key limitations include potential information leakage from LLM training data on high-profile companies, single-pass evaluation variability, lack of a human-expert benchmark, small sample size (14 decks), binary success/failure simplification, and hindsight bias from historical data.

### D. Future Work

Future work will expand the dataset (100+ decks), implement repeated evaluations with ensemble aggregation, benchmark against expert humans, explore alternative LLM architectures, refine confidence calibration, incorporate multimodal slide content, and investigate adaptive rubric weighting (e.g., via reinforcement learning) and ablation studies.

## Acknowledgments

The author acknowledges the use of Claude Sonnet 4.5 [20] (Anthropic, 2025) for assistance with data analysis, section development, and manuscript preparation, and ChatGPT [21] (OpenAI, 2025) for figure generation and visualization support.

TABLE I: Systems Engineering Rubric Evaluation Applied to LinkedIn Pitch Deck

| Category | Subcriteria | Evaluation Focus & Evidence | Score | Cat. Total |
|---|---|---|---|---|
| **Cost Risk** | Accuracy of Cost Estimates | Financial projections lack Basis of Estimate (BOE). Expenses listed without traceability to architecture or staffing plans. | 5/25 | 15/100 |
| | Funding Stability | Series B solicitation indicates current funds insufficient. Funding unstable until round closes. | 5/25 | |
| | Cost Control Measures | No cost-tracking systems (EVM, reserves) mentioned. Presents goal, not financial management plan. | 0/25 | |
| | Cost Risk Response Plan | Target is profitability by 2005. No documented mitigation for cost overruns. | 5/25 | |
| **Schedule Risk** | Schedule Realism | Revenue features scheduled Q4 2004/Q1 2005. Aggressive with no integrated master schedule or readiness assessment. | 10/25 | 25/100 |
| | Dependency Management | No identification or discussion of dependencies between core platform and revenue modules. | 0/25 | |
| | Resource Availability | Strong leadership team identified, but plan requires nearly doubling headcount in one year. | 15/25 | |
| | Schedule Recovery Plans | No schedule margin or recovery measures. Timeline appears best-case scenario. | 0/25 | |
| **Technical Risk** | Technology Maturity (TRLs) | Core networking platform operational with significant traction (TRL 7/8). Revenue subsystems undeveloped (TRL 3/4). | 15/25 | 35/100 |
| | Design Margin & Robustness | No system architecture, performance data, or scalability plans provided despite projected exponential growth. | 5/25 | |
| | Integration Complexity | New revenue features (ads, listings, subscriptions) must integrate with core network. Complexity undefined. | 10/25 | |
| | Verification & Validation | No V&V plan, test strategy, or quality assurance process mentioned for revenue features. | 5/25 | |
| **Programmatic Risk** | Stakeholder Alignment | Team and existing investors (Sequoia) well-aligned on "network first" strategy. Key strength. | 20/25 | 65/100 |
| | Policy/Regulatory Stability | Regulatory environment for social networking nascent in 2004. Low immediate compliance risk. | 20/25 | |
| | Supplier/Contractor Reliability | System developed in-house. Minimal external supplier risk for core technology. | 20/25 | |
| | External Event Preparedness | Plan focused on "happy path" with no contingency for competitive moves or market shifts. | 5/25 | |
| **System 'Ilities'** | Usability | Strong evidence through viral growth (930k+ users), high search volume, engagement. Product-market fit validated. | 22/25 | 37/100 |
| | Reliability & Availability | No data on uptime, latency, or failure rates. Current state and future capability unassessable. | 5/25 | |
| | Maintainability | No information on software architecture, modularity, or development practices provided. | 5/25 | |
| | Producibility (Scalability) | Largest technical risk. No architectural evidence of scaling from $\sim$1M to $\sim$10M users. | 5/25 | |
| **Average Project Risk Score** | | | **35.4/100** | |

## REFERENCES


[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[2] Y. Ozince and Y. Ihlamur, "Automating venture capital: Founder assessment using llm-powered segmentation, feature engineering and automated labeling techniques," in *arXiv preprint*, 2024, arXiv:2407.04885.

[3] L. Chen, M. Zhang, and Y. Wang, "A fused large language model for predicting startup success," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 2145–2158, 2024.

[4] J. Arroyo, F. Corea, G. Jimenez-Diaz, and J. A. Recio-Garcia, "Assessment of machine learning performance for decision support in venture capital investments," *IEEE Access*, vol. 7, pp. 124 233–124 243, 2019.

[5] B. Potanin, A. Scherbakov, and D. Muravev, "Machine learning for startup success prediction: A high-performance predictive model," *Journal of Business Venturing Insights*, vol. 19, p. e00364, 2023.

[6] X. Xiong and Y. Ihlamur, "Founder-gpt: A framework for evaluating founder-idea fit using large language models," *arXiv preprint arXiv:2310.xxxxx*, 2023.

[7] A. Maarouf, A. Alic, and E. H. Houssein, "Predicting startup success from free-form text descriptions using deep learning," *Expert Systems with Applications*, vol. 237, p. 121432, 2024.

[8] SlideShare, "Startup pitch decks," https://www.slideshare.net/, 2025.

[9] *ISO/IEC/IEEE 15288:2015 Systems and Software Engineering—System Life Cycle Processes*, IEEE Std. ISO/IEC/IEEE 15 288:2015, 2015.

[10] INCOSE, *Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities*, 4th ed. John Wiley & Sons, 2015.

[11] MITRE Corporation, "Systems engineering guide," MITRE Corporation, Tech. Rep., 2014. [Online]. Available: https://www.mitre.org/publications/systems-engineering-guide

[12] NASA, "Nasa systems engineering handbook," National Aeronautics and Space Administration, Tech. Rep. NASA/SP-2016-6105 Rev2, 2016.

[13] C. M. Christensen, *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press, 1997.

[14] P. A. Gompers and J. Lerner, "What drives venture capital fundraising?" *Brookings Papers on Economic Activity: Microeconomics*, pp. 149–204, 1998.

[15] Kauffman Foundation, "The anatomy of an entrepreneur: Family background and motivation," https://www.kauffman.org, 2009.

[16] OECD, "Financing smes and entrepreneurs 2024: An oecd scoreboard,"

Organisation for Economic Co-operation and Development, Tech. Rep., 2024. [Online]. Available: https://www.oecd.org/
[17] CB Insights, "The top 20 reasons startups fail," https://www.cbinsights.com/research/startup-failure-reasons-top/, 2019.
[18] MSTY, "Msty.ai desktop environment," https://msty.app, 2025.
[19] Google DeepMind, "Gemini 2.5 pro," https://deepmind.google/technologies/gemini/, 2025, accessed: 2025.
[20] Anthropic, "Claude sonnet 4.5," https://claude.ai, 2025.
[21] OpenAI, "Chatgpt," https://chat.openai.com, 2025.