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. This study examines whether large language models (LLMs) [1] augmented with structured knowledge frameworks can enhance predictive accuracy in early-stage startup evaluation. We designed four evaluation personas using [2]: 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. 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 technical risk assessment while the venture capital knowledge stack enhanced market opportunity identification. 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 lack operational history, financial maturity, or technical validation. These high-uncertainty judgments typically rely on unstructured reasoning, investor-specific heuristics, and expert interpretation of product and business proposals called pitch decks. Systems engineering provides an "interdisciplinary approach and means to enable the realization of successful systems" [3]. This includes established methodologies for risk assessment, lifecycle evaluation, and technical feasibility analysis through systems engineering 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. Applying these considerations when evaluating pitch decks should provide decision makers with additional perspective on the viability and potential of a proposed product.

Recent advances in generative artificial intelligence (AI), particularly large language models (LLMs) [1], have introduced new possibilities for conducting investment evaluations. LLMs can process unstructured data, extract context, and generate structured reasoning across multiple domains, including finance and engineering. Prior studies have demonstrated that language models can identify factors associated with startup performance [4], [5], and industry applications have explored AI-assisted pitch deck screening. However, prior work primarily focuses on structured datasets and often overlooks the unstructured reasoning investors use in early-stage assessments.

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. By integrating established systems engineering standards with venture capital frameworks, 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. 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. Evaluation Environment

All evaluations were conducted using [2] within the MSTY.ai [6] 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. Each pitch deck was evaluated once to reflect the single-pass judgment typical of real venture capital review.

The local configuration enabled embedding of custom knowledge stacks through a BERT-based retriever integrated into MSTY.ai [6], allowing each persona to access its corresponding knowledge base directly during inference without requiring external connectivity.

C. Persona Configurations

Four personas were designed to simulate distinct reasoning frameworks representative of real-world evaluators:

1) **Basic Persona**: Control model receiving no contextual knowledge beyond the pitch deck text; instructed to decide whether to invest as an independent early-stage investor.
2) **Venture Capital Persona**: Built using a comprehensive knowledge stack containing academic and practitioner resources on investment heuristics, valuation methodologies, and startup lifecycle dynamics.
3) **Systems Engineering Persona**: Guided by authoritative sources from major systems engineering organizations: [7], the [3] (Fourth Edition), the [8], and the [9].
4) **Combined Persona**: Integrated both knowledge stacks and balanced market feasibility against technical credibility.

Each persona produced three outputs per deck: a categorical decision (Hard Yes, Leaning Yes, Leaning No, No), a numerical rating between 0 and 100 representing confidence, and a short summary explaining the rationale.

D. Knowledge Stack Construction

The Venture Capital Knowledge Stack drew from investment theory, valuation frameworks, due diligence guidelines, and historical case studies [10]–[14]. These materials were vectorized and embedded as references to support context-aware retrieval during inference.

The Systems Engineering Knowledge Stack comprised materials aligned with lifecycle standards and best practices from major organizations [3], [7]–[9]. Both stacks were made available through a GitHub repository [15] and indexed using BERT embeddings in MSTY.ai [6].

E. Systems Engineering Evaluation Rubric

The rubric provided a quantitative framework for assessing risk and maturity within each pitch deck. It consisted of five major 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.

F. 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 [16]. Pitch decks were converted to text using optical character recognition and manually cleaned.

G. Evaluation Procedure

Each persona analyzed all 14 pitch decks, generating 56 evaluations total. 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 not used to adjust accuracy scores.

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.

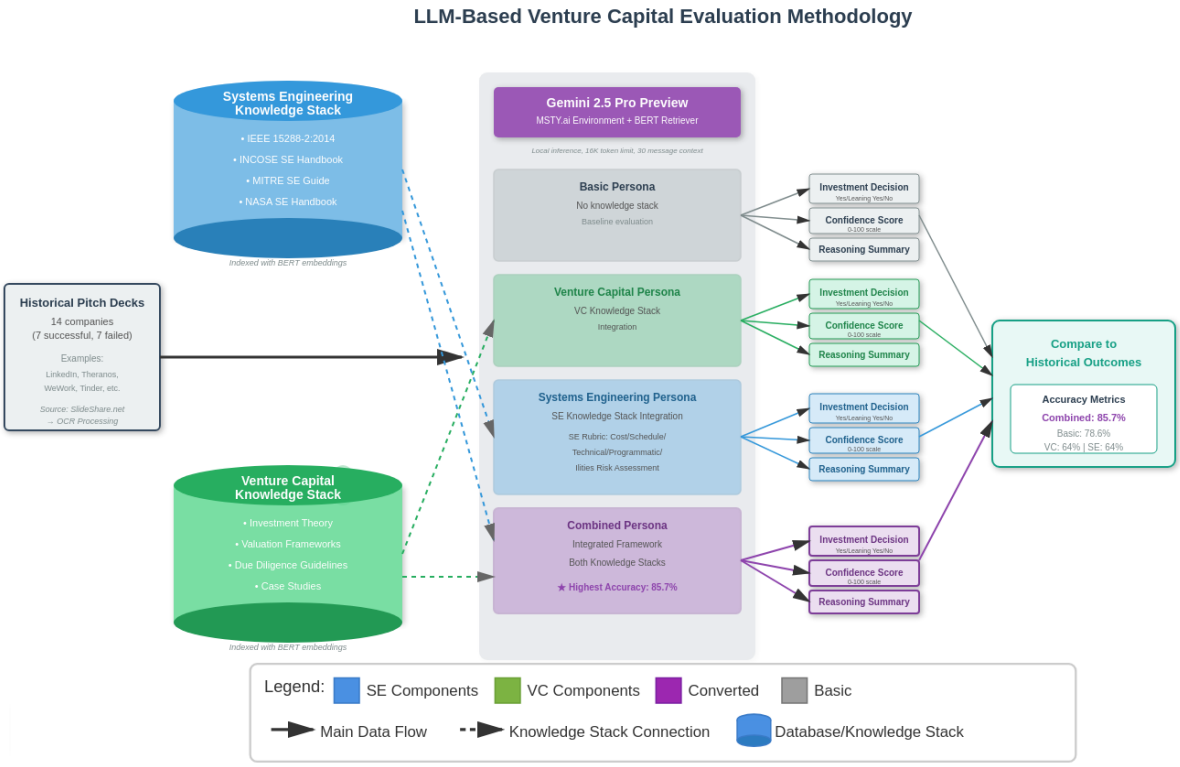

Fig. 1: Methodology workflow showing the evaluation framework with four personas (Basic, Venture Capital, Systems Engineering, Combined), knowledge stack integration, and evaluation pipeline from pitch deck input to investment decision output.

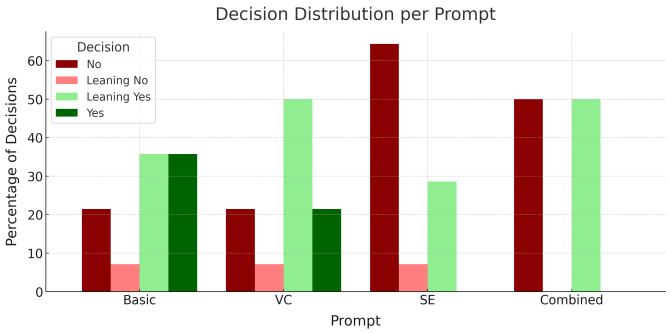

Fig. 2: Decision distribution per persona showing percent of each decision type (Yes, Leaning Yes, Leaning No, No).

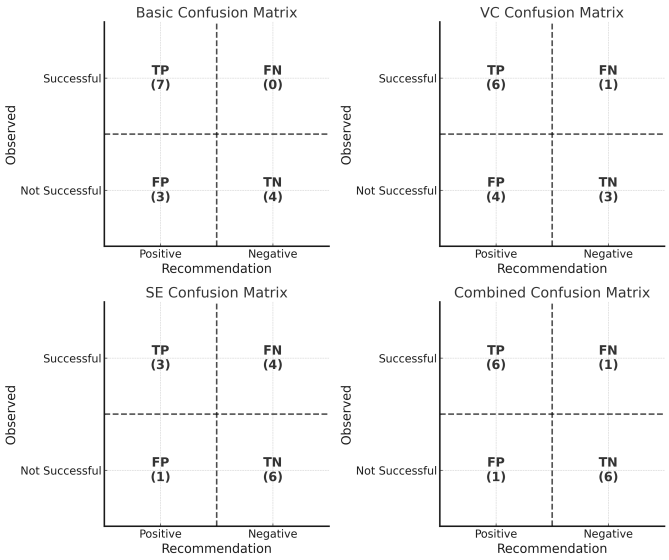

Fig. 3: Confusion matrix for the Combined persona: predictions versus actual outcomes (6 true positives, 6 true negatives, 1 false positive, 1 false negative).

B. Statistical and Qualitative Analysis

Quantitative 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.

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. 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%).

D. 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.

E. 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.

F. 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 [17] (Anthropic, 2025) for assistance with data analysis, section development, and manuscript preparation, and ChatGPT [18] (OpenAI, 2025) for figure generation and visualization support.

References

[1] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[2] G. DeepMind, "Gemini 2.5 pro preview," https://deepmind.google/technologies/gemini/, 2025.

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

[4] R. Kumar and A. Sharma, "Automating venture capital: founder assessment using llm powered segmentation," in *Proc. IEEE Int. Conf. AI and Finance*, 2024, pp. 145–152.

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

[6] MSTY, "Msty.ai desktop environment," https://msty.app, 2025.

[7] *IEEE Standard for Systems and Software Engineering—System Life Cycle Processes*, IEEE Std. IEEE Std 15 288-2:2014, 2014.

[8] M. Corporation, "Mitre systems engineering guide," https://www.mitre.org/publications/systems-engineering-guide, 2014.

[9] NASA, "Nasa systems engineering handbook," 2016.

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

[11] P. Gompers and J. Lerner, *The Venture Capital Cycle*, 2nd ed. MIT Press, 2004.

[12] K. Foundation, "The anatomy of an entrepreneur," 2009.

[13] OECD, "Financing smes and entrepreneurs 2023: An oecd scoreboard," 2023.

[14] C. Insights, "The top 20 reasons startups fail," 2021.

[15] D. Morales, "Llm venture capital knowledge stacks," https://github.com/danmorales/llm-vc-knowledge-stacks, 2025.

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

[17] Anthropic, "Claude sonnet 4.5," https://claude.ai, 2025.

[18] OpenAI, "Chatgpt," https://chat.openai.com, 2025.