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 typically rely on qualitative reasoning and expert interpretation of product and business proposals called pitch decks. This reasoning is highly dependent on the experience and perspective of individual investors, generally not explicitly structured or a repeatable process, limiting the efficacy and efficiency of any such evaluation. Systems engineering provides an "interdisciplinary approach and means to enable the realization of successful systems" [1]. This includes established methodologies for risk assessment, lifecycle evaluation, and technical feasibility analysis through frameworks codified by major engineering 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) [2], have introduced new possibilities for conducting 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 [3], [4], 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 [3], [5], [6].

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 [7] and prediction from free-form text descriptions [8], 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 early-stage investment evaluations by increasing the prediction accuracy and repeatability of such evaluations. This has obvious benefit for investors, but the ability to make consistent and accurate predictions in a way that accounts for both technical standards and business viability has applications for other applications such as technology

forecasting or concept development. Integrating established and well-defined engineering standards such as those by IEEE, INCOSE, and others with venture capital case-studies and AI to improve this prediction is therefore significant if it can be reasonably trusted.

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 [3], [5], [6]. Existing studies are also limited to analysis through a monolithic lens (e.g., only venture logic). By combining engineering rigor with venture logic enables one 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. Overview

The central hypothesis of this study is that LLMs equipped with structured systems engineering and venture capital knowledge frameworks will demonstrate higher accuracy, interpretability, and reasoning consistency in evaluating startup success based on a given pitch deck 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.

We assessed this hypothesis by following the evaluation framework illustrated in Figure 1. This framework follows three major sections. The first is preparing the input data from the pitch decks, systems engineering frameworks, and venture capital frameworks. This includes historical pitch decks and the two "knowledge stacks" of Systems Engineering and Venture Capital that are indexed using BERT embeddings and available to the LLM for retrieval during inference. Second is the development of different LLM personas. For the LLM we used the Gemini 2.5 Pro model operating within the MSTY.ai environment. We provided this with one of four personas, each either including or not including the "knowledge stacks" and prompted it to generate investment decisions, confidence scores, and reasoning summaries. Finally, we assessed the outputs of the model across a variety of metrics to assess These outputs are then compared against historical company outcomes to compute accuracy metrics.

B. Data and Preparation

The first major set of data for this study are the unstructured product and business proposals called pitch decks as represented by the grey square on the left side of Figure 1. These are 14 historical proposals from the website SlideShare [9] divided evenly between successful and unsuccessful ventures. The successful set included LinkedIn, Intercom, Coinbase, Monzo, Buffer, WeWork, and Tinder. The unsuccessful set included Bliss, Cardlife, Castle, FlowTab, LimeTree, Spartan, and Theranos.

Ventures were classified as successful or unsuccessful from the perspective of early-stage investors evaluating the pitch deck at the time of its creation. A venture was deemed successful if investors at that funding stage could have achieved a profitable exit through subsequent funding rounds, acquisition, or public offering—regardless of the company's ultimate long-term trajectory. For example, WeWork's 2014 Series D pitch deck is classified as successful because early investors achieved substantial returns through later funding rounds that valued the company at $47 billion, even though the company subsequently filed for bankruptcy in 2023. This classification reflects the practical reality of venture capital: investors seek returns within a typical 5-10 year fund lifecycle, not permanent company viability.. These slide decks are completely unstructured and contain a variety of images. We converted them to text using optical character recognition and manually cleaned them to ensure consistent formatting and readability. While this data does have limitations, for example, it does not include the associated oral presentation, it is a useful representation of the limited data available to decision makers for early stage ventures. This data set is balanced between successful and unsuccessful ventures enabling a fair evaluation of the LLM's ability to predict outcomes.

The second major set of data for this study are the knowledge repositories necessary to support domain-specific reasoning. This resulted in two "knowledge stacks":

1) Systems Engineering Knowledge Stack: The Systems Engineering Knowledge Stack, represented by the blue database in Figure 1, comprised authoritative materials from major standards organizations: the IEEE 15288 standard for systems and software engineering lifecycle processes [10], the INCOSE Systems Engineering Handbook Fourth Edition [1], 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, represented by the green database in Figure 1, 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. Detailed examples of how retrieved VC literature directly influenced model decisions are

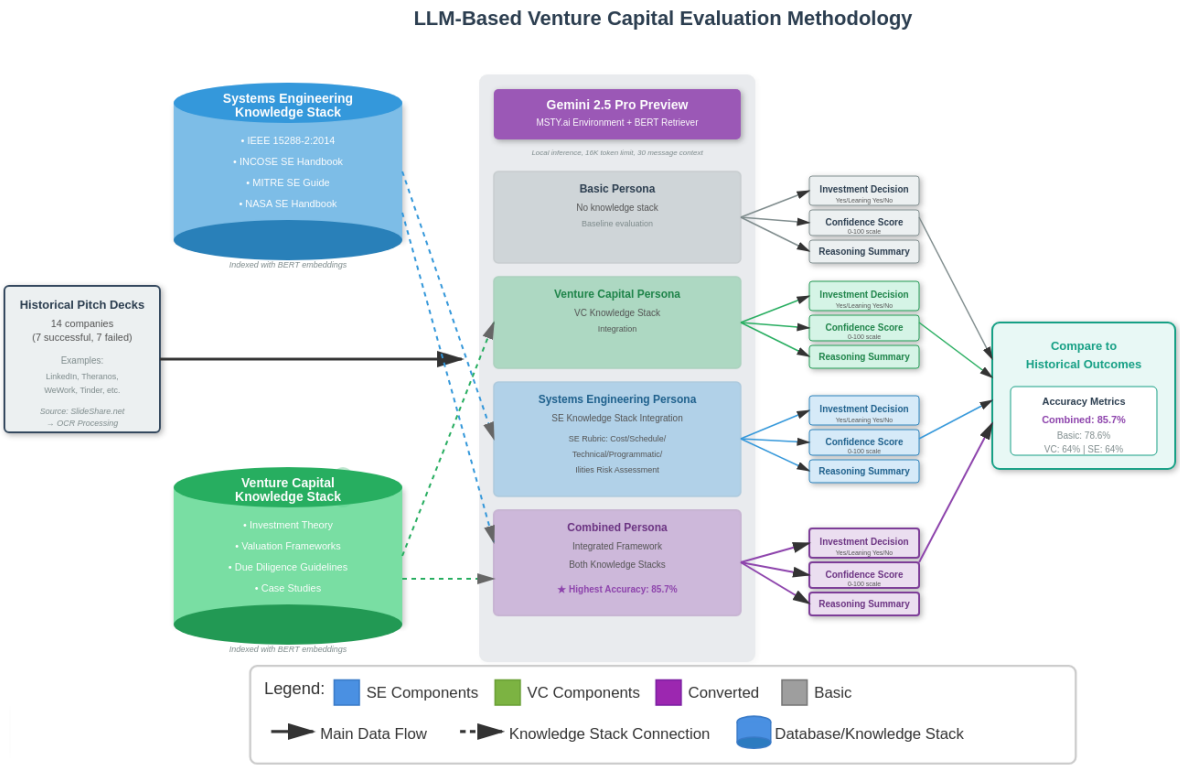

Fig. 1: Methodology workflow showing the evaluation framework with four personas (Basic, Venture Capital, Systems Engineering, Combined), knowledge stack integration, and evaluation pipeline from pitch deck input to investment decision output.

available in the supplementary materials.

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.

C. Proposal Evaluation

The second major aspect of the methodology is represented by the middle tier of Figure 1. In this section we configured the LLM with four different personas and prompted it to assess each of the 14 pitch decks. Each of the four personas assessed the 14 pitch decks for its investment decision, confidence score, and reasoning summary as shown in the grey, green, or purple outputs in Figure 1. While there are multiple ways to configure any particular LLM, we kept the baseline LLM configuration static and focused on varying the input personas.

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. The 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, as represented by the grey rectangle in Figure 1, 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.

2) Venture Capital Persona: The Venture Capital Persona, as represented by the green rectangle in Figure 1, 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, as represented by the blue rectangle in Figure 1, 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 3.3. The complete SE rubric (Table I) is designed as a generalizable artifact that can be applied independently of LLM evaluation—researchers and practitioners may adapt this framework for acquisition reviews, architecture trade studies, or technology readiness assessments in other domains.

4) Combined Persona: The Combined Persona, as represented by the purple rectangle in Figure 1, 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.

Finally, 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.

D. 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?

Model predictions were evaluated against historical investment outcomes. Successful companies were defined as those where early-stage investors could have achieved profitable exits through subsequent funding rounds, acquisitions, or public offerings within a typical venture capital fund lifecycle. Failed companies were defined as those that ceased operations, failed to secure follow-on funding, or experienced fundamental business model failures before providing exit opportunities for early investors. 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

In executing the methodology shown in Figure 1, we confirmed our hypothesis that 1) LLMs can make reasonable, interpretable assessments of pitch decks for venture capital decisions and 2) providing them with structured reasoning based on systems engineering and venture capital constructs improves their accuracy. This section details both the data analysis of the 56 iterations we ran and a qualitative assessment of the results.

A. Data Analysis

Figure 2 shows the distribution of decisions for each of the 14 pitch decks across the four different personas. In each case, the LLM output a decision that ranged from No, Leaning No, Leaning Yes, or Yes depending on the confidence of prediction. Recall that the proper classification should be 50-50 as half of the pitch decks were successful ventures and half were note. In this case we see that the Basic Persona and Venture Capital Persona are most aggressive recommending a lean yes or yes for 10/14 (71%) of cases, the SE Persona is most conservative with a no or leaning no for 10/14 (71%) of cases, and the Combined Persona achieves a 50-50 split. The raw percentage of predictions, in and of itself, is insufficient to assess the quality as there could be false-positives or false-negatives.

Figure 3 shows the confusion matrix for the results indicating the accuracy of the models. The Basic Persona at achieved a 78.6% accuracy indicating that baseline knowledge does enable some level of reasonable prediction. Surprisingly, the Venture Capital and Systems Engineering Personas each achieved a lower 64.3% accuracy, indicating that each one was potentially over-correcting for perceived business or technical weakness in the proposals. Finally, the Combined Persona achieved the highest overall accuracy at 85.7%. This demonstrates a balanced reasoning between technical feasibility and market potential is essential to have the fewest misclassifications.

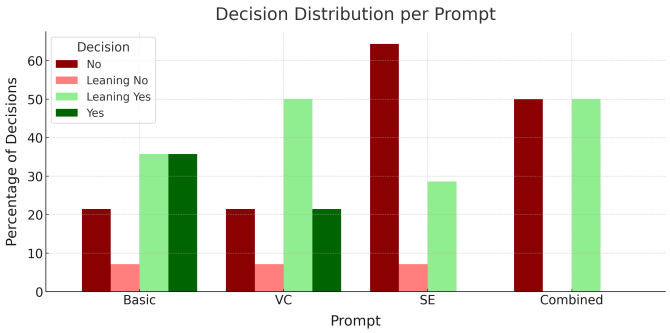

Fig. 2: Decision distribution per persona showing percent of each decision type (Yes, Leaning Yes, Leaning No, No).

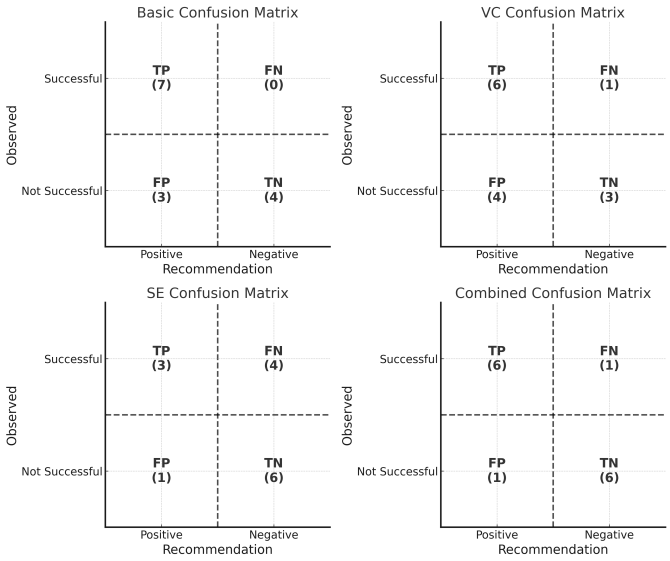

Fig. 3: Confusion matrix for the Combined persona: predictions versus actual outcomes (6 true positives, 6 true negatives, 1 false positive, 1 false negative).

The models output more than a simple yes or no, they also provide a numeric confidence score where a 100 was complete confidence and 0 was no confidence in their recommendation. These confidence scores ranged from 40 to 100, with a mean of 82.5 across evaluations. Confidence scores generally correlated with accuracy. 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. 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%).

B. Qualitative Assessment

The quantitative analysis confirmed that structured reasoning frameworks improved both accuracy and interpretability. Equally important to the analysis is not simply the accuracy of the recommendations, but also the reasoning provided for why recommendations were a positive or negative. This is best seen in the written explanation for the recommendation that the LLMs provided. These results generally followed how the personas assessed the data.

1) Basic Persona: The Basic Persona produced brief heuristic justifications with no common lexicon or foundational reasoning. It was somewhat susceptible to a narrative bias and the story and claims as presented in the pitch deck. This narrative bias occured in five of the cases.

2) Venture Capital Persona: The Venture Capital Persona was somewhat more aggressive with a higher rate of false positives. It emphasized founder strength, traction, and market scale, wtih an improved ability to identify scalable business opportunities. That stated, it was susceptible to the same narrative bias as the Basic Persona where it took the claims of the pitch deck on face value without doing a deeper assessment of the technical fundamentals.

3) Systems Engineering Persona: The Systems Engineering Persona was more conservative and tended to be wary of claims where it could not justify the technology behind the product. This led to it producing more false negatives for ventures that lacked necessary architectural data. Interestingly, none of the pitch decks specifically produced standard systems engineering architecture products, but it was capable of identifying the ones that did contain the relevant information.

4) Combined Persona: The Combined Persona was able to incorporate and balance both perspectives of the Venture Capital and Systems Engineering Personas to increase its accuracy, resulting in only 2 misclassifications (Bliss and Tinder) where technical and market signals conflicted.

While it is not possible to articulate the models' reasoning for all pitch decks, it is helpful to understand a few illustrative cases:

Theranos: Theranos was an unsuccessful venture that received significant initial investment due to the founder's claims of revolutionary blood-testing technology [20]. The Basic and Venture Capital Personas recommended investment, illustrating their susceptibility to the pitch deck's story and claims. The Systems Engineering and Combined Personas correctly rejected the pitch deck due to unsubstantiated claims and a missing regulatory pathway.

Tinder: Tinder is a successful venture that was one of the first applications to provide locality based date matching services. The Basic and Venture Capital Personas correctly classified this as successful based on product novelty and scalability. The Systems Engineering and Combined Personas rejected as the pitch deck lacked technical detail.

LinkedIn: All personas recommended investment, but for varying reasoning. The next section details how the model was able to apply a Systems Engineering risk assessment to the venture on relatively little information.

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.

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 of "pitch deck" assessments. These structured LLM-enabled assessments are repeatable, traceable, and auditable, suitable for decision support in high uncertainty contexts such as venture capital decisions.

B. Threats to Validity and Limitations

Several threats to validity warrant careful consideration when interpreting these results.

Information Leakage. High-profile companies in our dataset (LinkedIn, Theranos, WeWork) appear extensively in LLM training corpora, potentially enabling pattern matching rather than genuine reasoning. However, several observations suggest leakage effects are limited. First, the model's errors occurred on both well-known (Tinder) and obscure companies (Bliss), indicating no systematic advantage for famous cases. Second, the SE persona's structured rubric forces evaluation against specific technical criteria rather than reputation-based pattern matching—LinkedIn received only 35.4/100 on the risk rubric despite its known success. Third, less prominent companies in our dataset (Castle, FlowTab, LimeTree) showed similar accuracy patterns to high-profile cases. Future work should explicitly test anonymized decks to quantify leakage effects.

Hindsight Bias. Ground truth labels derive from known outcomes, potentially biasing interpretation toward post-hoc rationalization. The rubric's emphasis on contemporaneous evidence (what the pitch deck contained at the time) partially mitigates this concern, but cannot eliminate it entirely. The Theranos case illustrates this tension: the SE persona correctly rejected the pitch based on missing regulatory pathways and unsubstantiated technical claims—information available in 2014—yet this reasoning benefits from knowing which red flags proved consequential.

Confidence Calibration. While confidence scores correlated with accuracy, the calibration analysis remains descriptive rather than statistically validated. The confidence threshold of 85% achieving 84.6% accuracy suggests reasonable calibration, but the small sample size (n=56 evaluations) limits statistical power. Additionally, confidence scores reflect model self-assessment rather than calibrated probability estimates, and may not generalize across different prompt formulations or model versions.

Prompt Sensitivity. LLM outputs are known to vary with prompt wording, ordering, and formatting. While we held prompts constant across evaluations, we did not systematically test alternative phrasings. The personas' relative performance could shift under different prompt designs, particularly for the boundary between "Leaning Yes" and "Leaning No" classifications.

Generalizability. These results should be interpreted as methodological validation rather than predictive performance claims. The 85.7% accuracy demonstrates that structured frameworks improve LLM reasoning consistency, not that the system is ready for production deployment. The approach may transfer to other systems engineering decision domains—acquisition reviews, architecture trade studies, technology readiness assessments—where structured rubrics can constrain LLM reasoning, though such applications require domain-specific validation.

C. Future Work

Building on this methodological foundation, future work will expand the dataset to 100+ pitch decks with emphasis on lesser-known companies, implement repeated evaluations with ensemble aggregation to assess output stability, and benchmark against human expert judgments under identical single-pass constraints. Additional directions include exploring alternative LLM architectures, incorporating multimodal slide content analysis, and investigating adaptive rubric weighting through ablation studies to identify which framework components contribute most to predictive performance.

Acknowledgments

The author acknowledges the use of Claude Sonnet 4.5 [21] (Anthropic, 2025) for assistance with data analysis, section

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**	

development, and manuscript preparation, and ChatGPT [22] (OpenAI, 2025) for figure generation and visualization support.

Data Availability

All pitch decks, system prompts, knowledge stack contents, evaluation outputs, and analysis code used in this study are publicly available at https://github.com/DanVelarde00/LLM-Based-Venture-Capital-Evaluation to support reproducibility. The repository includes the complete SE rubric scoring guidance, persona prompt templates, VC knowledge stack examples showing how retrieved literature influenced model decisions, and raw model outputs for all 56 evaluations.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] SlideShare, "Startup pitch decks," https://www.slideshare.net/, 2025.
[10] *ISO/IEC/IEEE 15288:2015 Systems and Software Engineering—System Life Cycle Processes*, IEEE Std. ISO/IEC/IEEE 15 288:2015, 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] J. Carreyrou, *Bad Blood: Secrets and Lies in a Silicon Valley Startup*. Alfred A. Knopf, 2018.
[21] Anthropic, "Claude sonnet 4.5," https://claude.ai, 2025.
[22] OpenAI, "Chatgpt," https://chat.openai.com, 2025.