unibz
Fakultät für Ingenieurwesen
Facoltà di Ingegneria
Faculty of Engineering

Degree Course

BACHELOR IN INFORMATIK
CORSO DI LAUREA IN INFORMATICA
BACHELOR IN COMPUTER SCIENCE

Pista: A GenAI-Powered System for Automated Startup Pitch Evaluation

Student
Maxwell Aboagye

Berichterstatter / relatore / Supervisor
Prof. Xiaofeng Wang

Zweitbetreuer / correlatore / Second Supervisor
Triando

akademisches Jahr / anno accademico / academic Year
2024/2025

Abstract

Startup pitch evaluation faces challenges in providing consistent and accessible assessments across diverse entrepreneurial contexts. This thesis presents Pista, a GenAI-powered evaluation platform, and compares its performance with Winds2Ventures (W2V), another GenAI-based startup assessment system developed by a team collaborating with the thesis supervisor. A statistical analysis was conducted using 22 university startup pitches to understand how different GenAI evaluation approaches perform when assessing identical content.

Pista was developed as a full-stack web application using Next.js 15, Convex database, Clerk authentication, and GPT-4 integration. The system evaluates pitches across four weighted dimensions: Problem-Solution Fit (30%), Business Model & Market (30%), Team & Execution (25%), and Pitch Quality (15%). Pista supports text uploads, file processing, and audio transcription, allowing for real-time evaluation progress tracking. The platform provides structured feedback with specific improvement recommendations and is deployed at https://pista-app.vercel.app.

The comparative analysis reveals moderate agreement between the GenAI systems with a Cohen's kappa coefficient of 0.505 and 77.3% observed agreement when categorizing pitches as below average, average, or good. Pista scored an average of 5.36 compared to W2V's 5.20, showing a systematic 0.16-point difference that indicates distinct evaluation approaches rather than random variation. Both systems demonstrated consistent patterns, with Pista providing more optimistic assessments while W2V applied more conservative evaluation criteria.

Performance analysis shows that Pista delivers evaluations in 30-60 seconds at $0.10-0.15 per assessment with 24/7 availability. The system handles problem-solution Fit evaluation most effectively but shows limitations in assessing team capabilities and execution potential. The pitches in the technology sector showed the highest scoring differences between the systems, while the pitches in the healthcare sector demonstrated the highest agreement rates.

The research demonstrates that different GenAI evaluation systems bring distinct characteristics and perspectives to startup assessment. This supervisor-facilitated comparison between GenAI evaluation systems shows that Pista's consistent scoring approach works well for educational contexts and initial screening scenarios, while W2V's varied scoring patterns better reflect investment decision contexts. The findings suggest that multiple GenAI evaluation perspectives provide more comprehensive assessment than relying on single platforms.

This thesis contributes a documented statistical comparison of GenAI evaluation platforms using standardized metrics, a working proof-of-concept system demonstrating technical feasibility, and empirical evidence of GenAI evaluation capabilities and limitations. The results show GenAI evaluation systems can provide valuable startup assessment capabilities while maintaining practical advantages in speed, cost efficiency, and accessibility for underrepresented entrepreneurial communities.

Riassunto

La valutazione delle startup pitch affronta sfide nel fornire un'analisi coerente e accessibile in diversi contesti imprenditoriali. Questa tesi presenta Pista, una piattaforma di valutazione basata su GenAI, e confronta le sue prestazioni con Winds2Ventures (W2V), un altro sistema di valutazione GenAI sviluppato da un team di startup in collaborazione con il supervisore della tesi. È stata condotta un'analisi statistica utilizzando 22 pitch di startup universitarie per comprendere come diverse approcci di valutazione GenAI performano quando valutano contenuti identici.

Pista è stata sviluppata come applicazione web full-stack utilizzando Next.js 15, database Convex, autenticazione Clerk e integrazione GPT-4. Il sistema valuta i pitch attraverso quattro dimensioni ponderate: Problem-Solution Fit (30%), Business Model & Market (30%), Team & Execution (25%) e Pitch Quality (15%). Pista supporta caricamenti di testo, elaborazione di file e trascrizione audio con tracciamento del progresso di valutazione in tempo reale. La piattaforma fornisce feedback strutturato con raccomandazioni specifiche per il miglioramento ed è distribuita su https://pista-app.vercel.app.

L'analisi comparativa rivela un accordo moderato tra i sistemi GenAI con un coefficiente kappa di Cohen di 0.505 e 77.3% di accordo osservato quando si categorizzano i pitch come sotto la media, nella media o buoni. Pista ha ottenuto una media di 5.36 rispetto ai 5.20 di W2V, mostrando una differenza sistematica di 0.16 punti che indica approcci di valutazione distinti piuttosto che variazioni casuali. Entrambi i sistemi hanno dimostrato modelli coerenti, con Pista che fornisce valutazioni più ottimistiche mentre W2V applica criteri di valutazione più conservativi.

L'analisi delle prestazioni mostra che Pista fornisce valutazioni in 30-60 secondi a $0.10-0.15 per valutazione con disponibilità 24/7. Il sistema gestisce la valutazione del Problem-Solution Fit in modo più efficace ma mostra limitazioni nella valutazione delle capacità del team e del potenziale di esecuzione. I pitch nel settore tecnologico hanno mostrato le maggiori differenze di punteggio tra i sistemi, mentre i pitch nel settore sanitario hanno dimostrato i tassi di accordo più alti.

La ricerca dimostra che diversi sistemi di valutazione GenAI portano caratteristiche e prospettive distinte alla valutazione delle startup. Questo confronto facilitato dal supervisore tra sistemi di valutazione GenAI mostra che l'approccio di punteggio coerente di Pista funziona bene per contesti educativi e scenari di screening iniziale, mentre i modelli di punteggio variati di W2V riflettono meglio i contesti di decisione degli investimenti. I risultati suggeriscono che multiple prospettive di valutazione GenAI forniscono una valutazione più completa rispetto al fare affidamento su singole piattaforme.

Questa tesi contribuisce con un confronto statistico documentato di piattaforme di valutazione GenAI utilizzando metriche standardizzate, un sistema proof-of-concept funzionante che dimostra la fattibilità tecnica, e prove empiriche delle capacità e limitazioni della valutazione GenAI. I risultati mostrano che i sistemi di valutazione GenAI possono fornire capacità preziose di valutazione delle startup mantenendo vantaggi pratici in velocità, efficienza dei costi e accessibilità per le comunità imprenditoriali sottorappresentate.

Zusammenfassung

Die Bewertung von Startup-Pitches steht vor Herausforderungen bei der Bereitstellung konsistenter und zugänglicher Bewertungen in verschiedenen unternehmerischen Kontexten. Diese Arbeit stellt Pista vor, eine GenAI-gestützte Bewertungsplattform, und vergleicht ihre Leistung mit Winds2Ventures (W2V), einem anderen GenAI-Bewertungssystem für Startups, das von einem Startup-Team in Zusammenarbeit mit dem Betreuer der Arbeit entwickelt wurde. Eine statistische Analyse wurde mit 22 universitären Startup-Pitches durchgeführt, um zu verstehen, wie verschiedene GenAI-Bewertungsansätze bei der Bewertung identischer Inhalte abschneiden.

Pista wurde als vollständige Webanwendung unter Verwendung von Next.js 15, Convex-Datenbank, Clerk-Authentifizierung und GPT-4-Integration entwickelt. Das System bewertet Pitches anhand von vier gewichteten Dimensionen: Problem-Solution Fit (30%), Business Model & Market (30%), Team & Execution (25%) und Pitch Quality (15%). Pista unterstützt Text-Uploads, Dateiverarbeitung und Audio-Transkription mit Echtzeit-Fortschrittsverfolgung der Bewertung. Die Plattform liefert strukturiertes Feedback mit spezifischen Verbesserungsempfehlungen und ist unter https://pista-app.vercel.app bereitgestellt.

Die vergleichende Analyse zeigt, dass die GenAI-Systeme moderate Übereinstimmung aufweisen, mit einem Cohen's Kappa-Koeffizienten von 0.505 und 77.3% beobachteter Übereinstimmung bei der Kategorisierung von Pitches als unterdurchschnittlich, durchschnittlich oder gut. Pista erzielte einen Durchschnitt von 5.36 im Vergleich zu W2Vs 5.20, was eine systematische Differenz von 0.16 Punkten zeigt, die auf unterschiedliche Bewertungsansätze und nicht auf zufällige Variation hinweist. Beide Systeme zeigten konsistente Muster, wobei Pista optimistischere Bewertungen lieferte, während W2V konservativere Bewertungskriterien anwendete.

Die Leistungsanalyse zeigt, dass Pista Bewertungen mit 24/7-Verfügbarkeit in 30-60 Sekunden zu kosten von $0.10-0.15 pro Bewertung. Das System handhabt die Problem-Solution Fit-Bewertung am effektivsten, zeigt aber Einschränkungen bei der Bewertung von Teamfähigkeiten und Ausführungspotential. Die Pitches im Technologiesektor zeigten die größten Bewertungsunterschiede zwischen den Systemen, während die Pitches im Gesundheitssektor die höchsten Übereinstimmungsraten demonstrierten.

Die Forschung zeigt, dass verschiedene GenAI-Bewertungssysteme unterschiedliche Eigenschaften und Perspektiven zur Startup-Bewertung bringen; dieser Vergleich, unterstützt vom Betreuer, zeigt, dass Pistas konsistenter Bewertungsansatz gut für Bildungskontexte und erste Screening-Szenarien funktioniert, während W2Vs variierte Bewertungsmuster besser Investitionsentscheidungskontexte widerspiegeln. Die Ergebnisse deuten darauf hin, dass mehrere GenAI-Bewertungsperspektiven eine umfassendere Bewertung bieten als das Vertrauen auf einzelne Plattformen.

Diese Arbeit trägt einen dokumentierten statistischen Vergleich von GenAI-Bewertungsplattformen unter Verwendung standardisierter Metriken bei, ein funktionierendes Proof-of-Concept-System, das technische Machbarkeit demonstriert, und empirische Belege für GenAI-Bewertungsfähigkeiten und -einschränkungen. Die Ergebnisse zeigen, dass GenAI-Bewertungssysteme wertvolle Startup-Bewertungsfähigkeiten bieten können, während sie praktische Vorteile in Geschwindigkeit, Kosteneffizienz und Zugänglichkeit für unterrepräsentierte unternehmerische Gemeinschaften beibehalten.

Acknowledgements

I am incredibly grateful to my supervisor, Professor Xiaofeng Wang, for her expertise in entrepreneurship and technology, which greatly shaped both the technical development of Pista and the design of the research methodology. I am thankful for her facilitation of the comparative analysis with Winds2Ventures and for providing access to the startup pitch datasets, which were crucial for the statistical validation of this thesis. Additionally, I wish to acknowledge my thesis co-supervisor, Triando, for his steady support and thoughtful guidance, which were instrumental in navigating the project's challenges and motivated me to explore new research methods.

I am thankful to the Winds2Ventures team for providing access to their evaluation platform, which was instrumental in conducting the comparative research that significantly contributed to the depth and validity of the analysis.

In addition to my academic mentors, I owe heartfelt thanks to my family, whose consistent support and encouragement provided the foundation that carried me through the most challenging stages of this academic journey. I want to also acknowledge my girlfriend, who inspires me every day to be a better version of myself through her kindness, patience, and belief in me, which consistently gave me motivation and perspective during this journey.

Table of Contents

List of Tables

List of Figures

List of Listings

Chapter 1

Introduction

Communicating business ideas effectively can determine the success or failure of a startup. The startup pitch, as a primary medium for communicating visions, business models, and potential to investors and stakeholders, is traditionally evaluated through in-person presentations and expert judgment. Research shows that investment decisions are highly dependent on the quality of pitch presentations, which play an important role in securing financial support[1]. Many business functions have changed substantially through technological improvements. However, traditional pitch evaluation has not evolved much, still relying heavily on in-person presentations and individual expert judgment.

In recent years, GenAI, particularly large language models, has emerged as a powerful tool in business analysis. They have shown incredible abilities in natural language understanding, context, and structured evaluations in various domains [2]. Models such as GPT-4 represent the latest stage of large language models. They perform at levels comparable to humans in complex analytic tasks. This suggests a great potential to improve pitch evaluation processes through innovative approaches[3].

However, pitch evaluation in the startup community continues to face several obstacles. Many evaluation processes use inconsistent standards. Quality feedback is hard to access, and evaluations take too long [4, 5]. Current evaluation methods require significant time for each pitch and often yield vastly different interpretations among evaluators. These issues especially affect entrepreneurs from emerging markets and other underrepresented groups with limited access to expert feedback networks [6].

To address these challenges, this thesis presents Pista[1], a GenAI-based pitch evaluation system and compares it with Winds2Ventures (W2V)[2]; another GenAI evaluation system developed by a startup team collaborating with the thesis supervisor. The differences in their approaches are examined, and their performance is evaluated. The system processes pitches in multiple formats (text, files, audio recordings) and provides feedback across four dimensions: problem-solution fit, business model viability, team execution capability, and pitch quality. The comparison with Winds2Ventures reveals differences in scoring patterns and evaluation approaches. Through textual content analysis, it delivers fast and consistent evaluations that address accessibility barriers for entrepreneurs seeking feedback. When compared with Winds2Ventures, the system reveals distinct differences in scoring patterns and evaluation approaches. The comparison shows how different GenAI evaluation methods perform and where each approach works best [7].

[1] Project website: `https://pista-app.vercel.app` with source code at `https://github.com/Maxxy21/pista`
[2] `https://w2v.network/`

Chapter 2

Background and Related Work

This chapter examines existing research on startup pitch evaluation methods and GenAI-driven assessment systems. The problems with traditional pitch evaluation methods are first examined, followed by an exploration of how GenAI systems address these issues. The need for better evaluation methods is demonstrated, and current GenAI evaluation systems and their capabilities are examined.

2.1 Limitations of Traditional Pitch Evaluation Methods

Startup pitch evaluation has traditionally relied on human experts and standardized frameworks to assess entrepreneurial ventures. This section examines these established approaches and identifies their key limitations that motivate the development of automated evaluation systems.

Startup pitch evaluation is a systematic process of assessing entrepreneurial venture presentations to determine their viability, potential, and investment worthiness. Investors and stakeholders use these evaluations to make funding decisions by analyzing business models, market opportunities, team capabilities, and financial projections. For decades, this evaluation process has been dominated by two main approaches: expert panels that rely on human judgment and experience, and standardized frameworks that apply consistent criteria across different pitches.

Understanding these traditional methods is important because they form the foundation for most current investment decisions. However, recent research has identified three fundamental limitations that challenge their effectiveness: scalability constraints, consistency problems, and accessibility barriers. These limitations affect both the quality of evaluations and the fairness of access to funding opportunities.

2.1.1 Expert Panels

Expert panels represent the most traditional approach to startup pitch evaluation. These panels consist of industry professionals, experienced investors, and successful entrepreneurs who assess startup presentations based on their domain expertise and practical experience. These panels watch live presentations and give feedback based on specific criteria about market opportunity and team quality. Proponents of this approach argue that human insight is valuable for judging aspects that are difficult to measure computationally, such as founder passion, confidence levels, and presentation quality.

However, research shows that expert panels have consistency problems. Gius (2024) looked at data from 67 startup competitions and found that expert judges often disagree when evaluating the same pitches [8]. This disagreement is actually common and can even predict which startups will succeed. The variation in expert opinions raises questions about whether these evaluations are reliable and fair.

Bias is another serious problem with expert panels. Balachandra et al. (2019) found clear gender differences in how investors evaluate pitches [9]. They showed that the same business idea gets different treatment depending on who presents it. Female entrepreneurs often receive more questions about risks and potential problems, while male entrepreneurs receive inquires focused on growth opportunities. This indicates that even experienced evaluators have unconscious biases that affect their judgments.

2.1.2 Standardized Evaluation Frameworks

Standardized evaluation frameworks emerged as a response to the inconsistency issues inherent in expert panel assessments. These frameworks represent systematic methodologies that apply uniform evaluation criteria and scoring systems across all pitch assessments. The goal is to reduce subjective variability and improve evaluation reliability by providing structured assessment protocols that all evaluators follow. These frameworks usually examine four main areas:

- Market opportunity (market size, growth potential)
- Team capability (team experience, diversity)
- Product innovation (uniqueness, technical feasibility)
- Business model (revenue streams, scalability)

Structured scoring systems have been shown to provide better consistency between evaluators compared to unstructured approaches [10]. However, challenges still exist. A 2020 survey by Gompers et al.[11] found significant differences in how venture capital firms evaluate startups. There's no single evaluation framework that everyone in the industry uses. Standardization helps, but the main evaluation problems are still present.

These four areas are commonly emphasized in standardized frameworks due to their demonstrated predictive value for startup success [5].

2.1.3 Key Limitations

Despite their widespread use, traditional evaluation methods face three fundamental limitations that affect both the quality and fairness of startup assessments. These limitations represent inherent constraints in human-centered evaluation processes that create systematic barriers to effective pitch assessment:

1. **Scalability**: Human evaluators can only review a finite number of pitches before they become tired and make worse decisions. Research by Hirshleifer et al. (2019) showed that financial analysts become less accurate when [12] evaluating too many things in one day. When analysts get tired, they start making more basic mistakes. During busy periods, the quality of evaluations decreases because people struggle to manage the workload.

2. **Geographic Access Barriers**: Most expert evaluators reside in specific cities and regions, making it challenging for startups in other areas to receive good feedback. Colombo et al. (2019) studied how distance affects startup funding in Europe [13]. They found that startups far from major investment centers are less likely to seek funding because they feel like they cannot access the right evaluators. This clustering of experts in certain places creates problems for entrepreneurs in remote areas who want quality evaluation.

3. **Consistency Challenges**: Traditional evaluation methods struggle to balance subjective feelings with objective facts in a consistent manner. Different evaluators prioritize different levels of importance on personal impressions versus hard data when predicting which startups will succeed [10]. This inconsistency makes it difficult to determine what really matters in evaluation.

These problems disproportionately affect early-stage startups and underrepresented founders. They demonstrate the need for more systematic and fair ways to evaluate pitches. Traditional methods struggle to be both consistent and contextual at the same time. GenAI systems might help solve this problem because they can apply the same criteria consistently while providing detailed feedback. These limitations motivated the system design described in Chapter 3.

2.2 GenAI-Driven Evaluation Systems

This section reviews GenAI-based evaluation systems and compares their capabilities and limitations to traditional methods.

The problems with traditional evaluation methods have led to growing interest in GenAI-based systems. These new systems use large language models and automated text analysis to solve the scalability, consistency, and accessibility problems described above.

2.2.1 Current Approaches

The limitations of traditional evaluation methods have sparked growing interest in GenAI-based solutions, leading to diverse approaches that address specific challenges. These approaches can be categorized into several key types, each targeting different aspects of the evaluation process:

- **Machine Learning-Based Assessment**: Researchers have built systems that learn from data about past startups to make better evaluation decisions. Arroyo et al. [14] (2019) showed how machine learning can help venture capital firms make decisions . These systems examine patterns in successful and failed startups to figure out what makes startups succeed.
- **Text-Based Evaluation Systems**: These systems use automated text analysis to read and evaluate pitch content, business plans, and presentation materials. They can judge things like how clear the value proposition is, how well the team understands the market, and how they position themselves against competitors.
- **GenAI Evaluation Platforms**: Several GenAI-powered evaluation systems have emerged in this area. Winds2Ventures (W2V), developed within the same research group, provides startup evaluations based on multiple criteria and gives structured feedback [15]. This platform was made available for comparison through supervisor facilitation, enabling benchmarking against different GenAI evaluation approaches.
- **Large Language Model Integration**: Recent advances in large language models have made it possible to create more sophisticated systems that understand text better. These systems can read pitch content and provide detailed feedback on many different aspects of the pitch at once.

2.2.2 Performance Considerations

GenAI systems have several advantages compared to traditional evaluation methods:

- **Scalability**: GenAI systems can evaluate many pitches without getting tired like human evaluators do. This means they can maintain consistent quality even when processing large volumes of pitches.
- **Consistency**: Automated systems apply the same evaluation criteria every time, which reduces the variability that happens with human expert panels.
- **Accessibility**: GenAI-based evaluation can provide quality feedback to entrepreneurs anywhere in the world. This solves the access problems that come with expert panels that are only available in certain locations.
- **Structured Feedback**: GenAI systems provide detailed, organized feedback on multiple aspects of a pitch. This helps entrepreneurs understand exactly what they need to improve.

However, GenAI systems still have important limitations. Humans are still needed to check the context and make final investment decisions [16]. GenAI systems might have trouble understanding subtle aspects of entrepreneurship that require deep knowledge about specific industries or market dynamics that go beyond what they can learn from text.

2.2.3 Diversity in GenAI Evaluation Approaches

As GenAI evaluation systems develop, different platforms have appeared with different ways of assessing pitches. Understanding these differences positions the comparison research within the broader world of automated evaluation systems; GenAI systems differ in their evaluation criteria, algorithms, and feedback mechanisms. Some platforms focus on analyzing many different criteria at once, while others emphasize specific areas like market validation or team assessment. These differences create opportunities to study how different GenAI evaluation approaches compare when they look at the same pitches.

The growing diversity of GenAI evaluation systems creates both opportunities and challenges. This diversity motivates comparative studies that examine how different automated approaches agree or disagree in their assessments. This comparison guides the evaluation design in Chapter 4.

Chapter 3

System Design and Implementation

This chapter presents the design and implementation of the Pista system to address the challenges identified in startup pitch evaluation. The fundamental aspects of the system architecture are explained, along with the technical decisions made during development, and how each component works together to provide consistent, GenAI-powered startup evaluations.

The chapter is organized around the key technical elements that make Pista work. The chapter is structured to first explain the solution approach and design philosophy, followed by the system architecture, technical implementation, and finally, the performance characteristics and deployment approach

3.1 Solution Approach and Design Philosophy

The Pista system was designed to solve three main problems in startup pitch evaluation: inconsistency between different evaluators, limited access to expert-level assessment, and scalability challenges. These problems were addressed by creating a system that uses GenAI to provide standardized evaluations based on research-backed criteria.

The core design philosophy focuses on evidence-based assessment rather than subjective opinions. Traditional pitch evaluations often vary widely because different evaluators focus on different aspects or use different standards. Pista addresses this by using a fixed evaluation framework with clear criteria and requiring evidence to support higher scores.

The system was built around four key design principles. First, the evaluation framework uses standardized criteria derived from venture capital research to ensure assessments focus on factors that actually predict startup success. Second, the system requires concrete evidence for all scoring decisions to prevent unsupported claims from receiving high ratings. Third, the architecture prioritizes simplicity and reliability over complex features to maintain research data quality. Fourth, the user interface guides users through a consistent process while presenting results in a clear, structured format.

The system processes a startup pitch, which can be either text or audio. It generates targeted questions to fill information gaps, processes the content through a structured evaluation framework, and then returns both numeric scores and detailed qualitative feedback. This approach combines the speed and consistency of automated assessment with the depth and insight that entrepreneurs need for improvement.

3.2 System Architecture and Technical Implementation

This section explains how the technical architecture was built to support reliable GenAI-powered evaluations. The overall architecture design, evaluation framework implementation, GenAI integration, data models, and API structure are covered.

3.2.1 Architecture Overview

Modern web technologies were selected that work well together for rapid development while supporting the real-time features needed for research applications. The architecture uses Next.js 15 with React for the frontend, Convex for the database and backend functions, and OpenAI's API for GenAI processing.[1] This technol-

[1] The complete technology stack includes: Next.js 15 (React framework), TypeScript (type safety), Tailwind CSS (styling), Shadcn/ui (UI components), Convex (backend-as-a-service), Clerk (authentication), OpenAI API (GPT-4 and Whisper), Zustand (state management),

ogy stack was chosen after evaluating alternatives based on development speed, real-time capabilities, and research data reliability. The combination provides server-side rendering for fast page loads, real-time synchronization for evaluation updates, and robust API integrations for GenAI processing. TypeScript was used throughout the implementation to ensure type safety and reduce runtime errors, which is critical for maintaining research data integrity.

Figure 3.1 shows the complete user workflow and how data moves through the system. Users can upload either text pitches or audio files, which the system processes through transcription (if needed), question generation, evaluation, and results storage. The workflow was designed to handle both synchronous and asynchronous operations, with audio transcription running separately from evaluation processing to prevent interface blocking. Real-time progress updates are provided throughout each step using Convex's reactive query system, allowing users to monitor processing status without manual page refreshes. This architecture ensures consistent data flow regardless of input type while maintaining responsive user experience during computationally intensive operations.

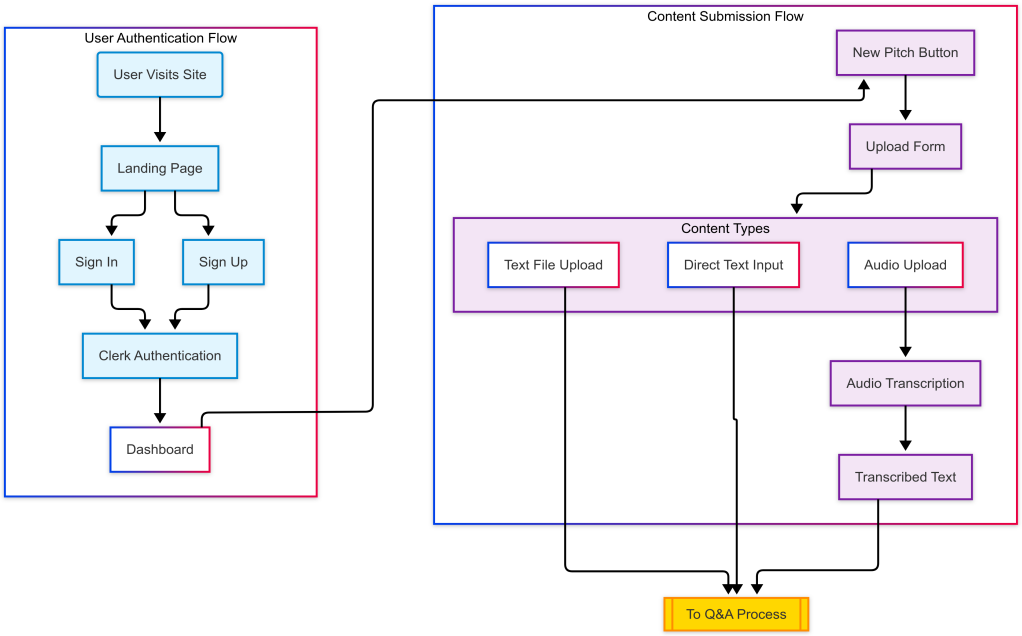

Figure 3.1: High level workflow showing user interaction and data processing

Next.js was chosen because it provides server-side rendering, built-in API routes, and excellent developer experience for building interactive applications. The App Router in Next.js 15 makes it easy to organize different parts of the application (dashboard, pitch pages, authentication) while maintaining clean code structure. This architecture supports the route groups used in the implementation, including '(dashboard)', '(pitch)', and '(auth)' directories that separate functional areas without affecting URL routing.

For the backend, Convex was selected after comparing it with traditional databases like PostgreSQL and other backend-as-a-service options. Convex provides real-time data synchronization, built-in schema validation, and server functions that run in the cloud. This eliminates the complexity of managing database connections, implementing real-time updates, or handling server deployment, allowing focus on the evaluation logic.

The authentication layer uses Clerk, which provides complete user management including organizations for multi-tenant data separation. This is important for research applications where different groups need to maintain separate data while using the same system. Clerk handles user registration, login, password management, and organization membership without requiring custom authentication logic.

For GenAI processing, OpenAI's GPT-4 API is used for all evaluation tasks and Whisper for audio transcription. This single-provider approach was chosen after testing multiple alternatives, as it provides the most consistent results while keeping the architecture simple. The system uses GPT-4 for structured evaluation, question generation, and answer processing, while Whisper handles audio file transcription with support for multiple audio formats.

Vitest (testing), and React Testing Library (component testing).

3.2.2 Evaluation Framework

The evaluation framework, developed by analyzing venture capital research and startup success factors, serves as the core mechanism for how Pista assesses startup pitches by identifying the most important evaluation criteria. The framework evaluates pitches systematically across multiple dimensions using standardized criteria and weighted scoring to ensure consistent assessments.

The framework evaluates pitches across four main dimensions, each with specific weights based on their importance for startup success:

- **Problem-Solution Fit**: 0.3
- **Business Model & Market**: 0.3
- **Team & Execution**: 0.25
- **Pitch Quality**: 0.15

The highest weights were assigned to Problem-Solution Fit and Business Model & Market because research consistently shows these factors are the strongest predictors of startup success. Team & Execution receives the next highest weight since execution capability is crucial for converting good ideas into successful businesses. Pitch Quality has the lowest weight because while presentation skills matter for raising capital, they are less predictive of actual business outcomes.

Each dimension evaluates five specific aspects, creating a comprehensive 20-aspect evaluation framework. For example, Problem-Solution Fit examines problem definition clarity, solution innovation, market understanding, competitive advantage, and value proposition. This systematic breakdown ensures thorough coverage of all critical elements.

Figure 3.2 illustrates how each pitch moves through the evaluation pipeline. The system analyzes the content against each dimension's criteria using GPT-4, then combines the results using the predetermined weights to calculate final scores.

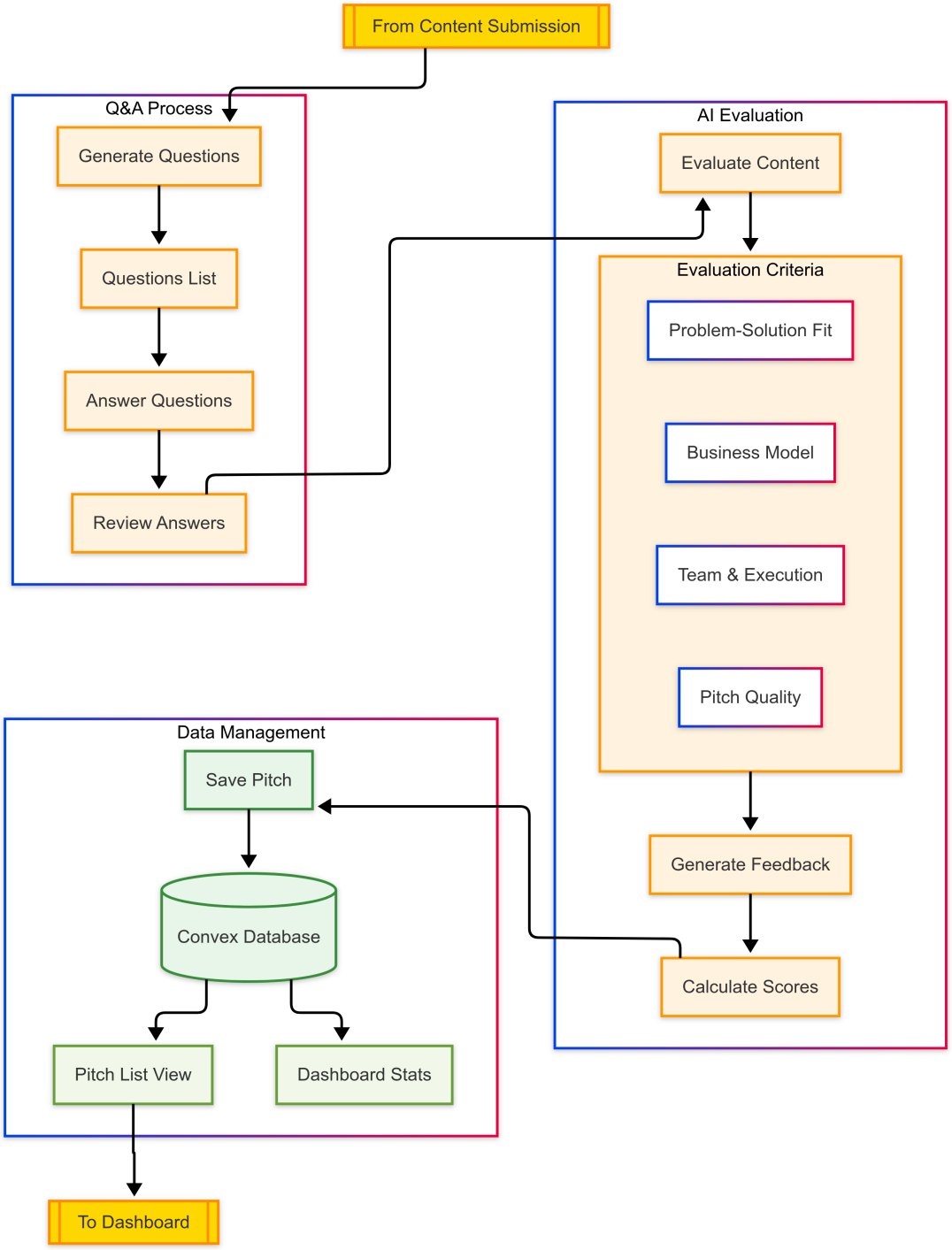

Figure 3.2: Evaluation and processing flow across four criteria

The scoring system uses a 1-10 scale with specific rubric anchors to ensure consistent evaluation; evidence-based scoring rules were implemented to prevent pitches from receiving high scores without supporting data. The rubric anchors define clear requirements for each score level:

1. **Scores 1–2**: No concrete evidence, claims only
2. **Scores 3–4**: Weak or indirect evidence, plan not validated
3. **Scores 5–6**: Mixed evidence, partial validation or unclear metrics
4. **Scores 7–8**: Strong evidence, credible metrics, risks addressed
5. **Scores 9–10**: Exceptional evidence, repeated traction, benchmarks exceeded

This evidence-gating approach addresses the central tendency problem where evaluations cluster around mid-scale scores without meaningful discrimination between pitch quality levels. The rubric was developed through iterative testing to ensure scores reflect actual business validation rather than presentation quality alone. These scoring rules prevent unsupported claims from receiving high ratings while rewarding pitches with concrete evidence of market traction and business validation.

Listing 3.1 shows how the evaluation criteria and weights were implemented as TypeScript constants to ensure consistency across all assessments.

```
const EVALUATION_CRITERIA = {
  problemSolution: {
    name: "Problem-Solution Fit",
    aspects: [
      "Problem Definition Clarity",
      "Solution Innovation",
      "Market Understanding",
      "Competitive Advantage",
      "Value Proposition",
    ],
  },
  businessModel: {
    name: "Business Model & Market",
    aspects: [
      "Revenue Model",
      "Market Size & Growth",
      "Go-to-Market Strategy",
      "Customer Acquisition",
      "Scalability Potential",
    ],
  },
  team: {
    name: "Team & Execution",
    aspects: [
      "Team Capability",
      "Domain Expertise",
      "Track Record",
      "Resource Management",
      "Implementation Plan",
    ],
  },
  presentation: {
    name: "Pitch Quality",
    aspects: [
      "Clarity & Structure",
      "Data & Evidence",
      "Story & Engagement",
      "Q&A Performance",
      "Overall Persuasiveness",
    ],
  },
} as const;

const WEIGHTS: Record<string, number> = {
  "Problem-Solution Fit": 0.3,
  "Business Model & Market": 0.3,
  "Team & Execution": 0.25,
  "Pitch Quality": 0.15,
};
```

Listing 3.1: Evaluation criteria and weights implementation

This implementation ensures every evaluation follows exactly the same criteria and weighting scheme. The criteria are defined as constants to prevent runtime modifications and maintain evaluation consistency. The TypeScript type system enforces proper usage of these constants throughout the codebase, preventing inadvertent changes to evaluation criteria during development.

3.2.3 GenAI Integration Strategy

GPT-4 was integrated as the core evaluation engine after systematic testing showed it provided the most consistent and high-quality assessments. The integration uses structured prompts that return constrained JSON responses, ensuring reliable data parsing and consistent formatting. All API interactions include strict JSON schema validation to prevent malformed responses from corrupting evaluation data.

The prompt engineering process involved several iterations to optimize evaluation quality. Early testing identified a central tendency problem, where evaluations clustered around 7 out of 10, failing to discriminate between pitches of different quality levels. This was solved by implementing evidence-based scoring rules and rubric anchors.

Listing 3.2 shows the main evaluation prompt template that was developed for structured assessment:

```
function buildStructuredPrompt(
  criteriaName: string,
  aspects: string[],
  fullContent: string
): string {
  return `
As an expert evaluator, score ONLY the criterion: ${criteriaName}.

Work from the content below without inventing facts. If evidence is
missing for an aspect, score that aspect <= 4 and note what is missing.

Aspects to score (each 1-10 with a one sentence rationale):
${aspects.map((aspect) => `- ${aspect}`).join("\n")}

Content to evaluate (pitch + Q&A):
${fullContent}

${RUBRIC_ANCHORS}

JSON response schema (valid JSON only):
{
  "score": 1-10, // criterion score = rounded average of aspectScores[].score
  "strengths": [{ "point": "...", "impact": "High"|"Medium"|"Low"}],
  "improvements": [{ "area": "...", "priority": "Critical"|"Important"|"Nice to Have
    ", "actionable": "..."}],
  "aspectScores": [
    { "aspect": "${aspects[0]}", "score": 1-10, "rationale": "..." },
    { "aspect": "${aspects[1]}", "score": 1-10, "rationale": "..." },
    { "aspect": "${aspects[2]}", "score": 1-10, "rationale": "..." },
    { "aspect": "${aspects[3]}", "score": 1-10, "rationale": "..." },
    { "aspect": "${aspects[4]}", "score": 1-10, "rationale": "..." }
  ],
  "summary": "2-3 sentence synthesis",
  "recommendations": ["actionable step 1", "actionable step 2"]
}

${SCORING_RULES}
`;
}
```

Listing 3.2: Main evaluation prompt template for structured assessment

This prompt template enforces evidence-based evaluation by explicitly instructing the model to avoid inventing facts and to cap scores at 4 when supporting evidence is missing. The structured JSON response ensures consistent data format across all evaluations.

A separate question generation system was also implemented to identify information gaps in pitch content. This system analyzes the pitch and generates up to 3 targeted questions that would materially improve evaluation quality if answered. The questions focus on specific evidence gaps that prevent confident evaluation rather than general business development advice.

Listing 3.3 shows the question generation prompt:

```
const prompt = [
  `Analyze the following pitch and identify the most important gaps that prevent a
    confident evaluation. Select up to 3 questions that, if answered, would
    materially improve the assessment.`,
  `Pitch:\n"${truncate(text, MAX_PROMPT_CHARS)}"`,
  `\nReturn valid JSON only with this schema:\n{\n  "items": [\n    {\n "dimension":
    "Problem-Solution Fit" | "Business Model & Market" | "Team & Execution" | "Pitch
    Quality",\n      "question": "one specific question, no multi-part prompts",\n
       "why_needed": "why this matters for evaluation",\n      "suggested_format": "
    how to answer: numbers, metrics, bullets, examples",\n      "priority": "Critical
    " | "Important"\n    }\n  ]\n}\n\nConstraints:\n- Ask 1-3 questions total.\n-
    Make each question specific and evidence-seeking.\n- Do not request sensitive
    data.\n- Avoid duplicates and multi-part questions.`,
].join("\n\n");
```

Listing 3.3: Question generation prompt for evidence gathering

This question generation approach systematically identifies evidence gaps before conducting the final assessment, improving evaluation accuracy while ensuring questions remain answerable within a typical pitch context. The system constrains questions to avoid requesting sensitive data or internal documentation that would not be available in standard pitch scenarios. Each generated question includes priority levels and suggested answer formats to guide users toward providing the most valuable additional information for evaluation improvement.

For reliability, exponential backoff retries were implemented for API interactions and strict JSON validation before data persistence. The system uses a low temperature setting (0.2) for all evaluation requests to reduce randomness while maintaining appropriate response variation. Error handling includes comprehensive logging for debugging while ensuring user-facing error messages remain clear and actionable without exposing system internals.

3.2.4 Data Models and Storage Architecture

The data architecture was designed to support both real-time user interactions and research data collection. Convex combines database functionality with built-in schema validation and real-time synchronization. Convex provides automatic schema enforcement and type safety that prevents data corruption while eliminating the need for separate database migration management.

The main data model centers around the `pitches` table, which stores essential pitch information including title, content, submission type, processing status, structured evaluation results, question-answer pairs, and organizational metadata. Each document is automatically scoped by user and organization identifiers, providing secure data isolation for multi-tenant use.

The database schema implements several indexes to optimize query performance: `by_org`, `by_user`, `by_user_org`, and `search_title`. The data model stores both structured evaluation results and complete audit trails for research reproducibility. Question-and-answer pairs are stored with each pitch and reused when building evaluation input, allowing systematic gathering of additional context that improves assessment quality. The system stores structured evaluation data with detailed scoring breakdowns and improvement recommendations for comprehensive analysis.

3.2.5 API Endpoints and Server Functions

The API architecture uses Next.js API routes for GenAI processing and Convex functions for data management. This separation allows each component to be optimized for its specific computational requirements. Next.js API routes handle external integrations with OpenAI, while Convex functions manage all database operations with built-in real-time synchronization.

The GenAI processing endpoints include:

- `/api/evaluate`: Processes pitch text using GPT-4 for comprehensive evaluation
- `/api/generate-questions`: Produces targeted follow-up questions using evidence-gap analysis
- `/api/transcribe`: Handles audio file transcription using OpenAI's Whisper model
- `/api/evaluate-answers`: Supports evaluation updates based on question responses

These endpoints run on Vercel's Edge runtime for improved performance and global distribution. All API responses are validated against strict schemas before data persistence to ensure research data integrity.

The Convex functions handle all data lifecycle operations including create, update, remove, favorite, unfavorite, and export operations. Query functions support diverse access patterns needed for both user interface functionality and research data analysis, including filtered searches, statistics, and CSV export for external analysis.

3.3 Authentication and Authorization Implementation

The authentication system uses Clerk to provide complete user management with multi-tenant organization support. This design ensures secure data isolation between different research groups while maintaining a simple security model. The implementation includes Next.js middleware that protects dashboard routes and redirects unauthenticated users to the sign-in page. Security headers are applied to all responses through the middleware layer, providing additional protection against common web vulnerabilities. The Clerk integration handles user registration, login, password management, and organization membership without requiring custom authentication logic, reducing security risks while maintaining a professional user experience.

3.4 User Interface Implementation

The user interface was built using Next.js 15 with React 18 and Tailwind CSS to provide a responsive, modern experience optimized for research applications. The application was organized using route groups to separate functional areas: dashboard for pitch management, individual pitch pages for detailed evaluation views, and authentication areas.

Figure 3.3 shows the main dashboard interface, which provides an overview of all pitch submissions with evaluation scores, search and filtering capabilities, and direct access to detailed results.

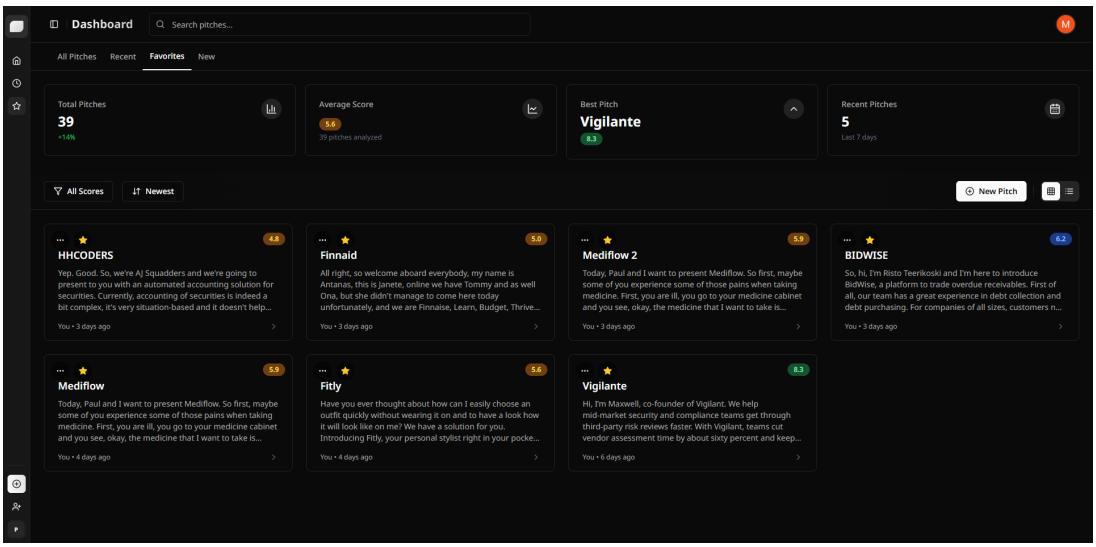

Figure 3.3: Main dashboard interface showing pitch management and evaluation scores

The component architecture separates shared UI primitives from feature-specific components to promote reusability and consistent design patterns. The dashboard implements comprehensive list management with user favorites, text-based search, and filtering by evaluation scores or submission dates. Components are organized into logical directories with 'src/components/ui/' containing base components and 'src/components/shared/' housing feature-specific components grouped by domain, including navigation, forms, modals, and authentication. The multi-step pitch upload process uses a dedicated component hierarchy in 'src/components/shared/forn pitches/steps/' that manages state progression, validation, and user guidance throughout the evaluation workflow. This architectural separation ensures consistent design language while enabling independent development and testing of complex features like the question generation and evaluation display components.

Figure 3.4 illustrates the pitch creation workflow that was designed to guide users through the evaluation process efficiently. The flow handles both text input and audio file uploads, manages question generation and response collection, and performs evaluation processing with complete metadata storage.

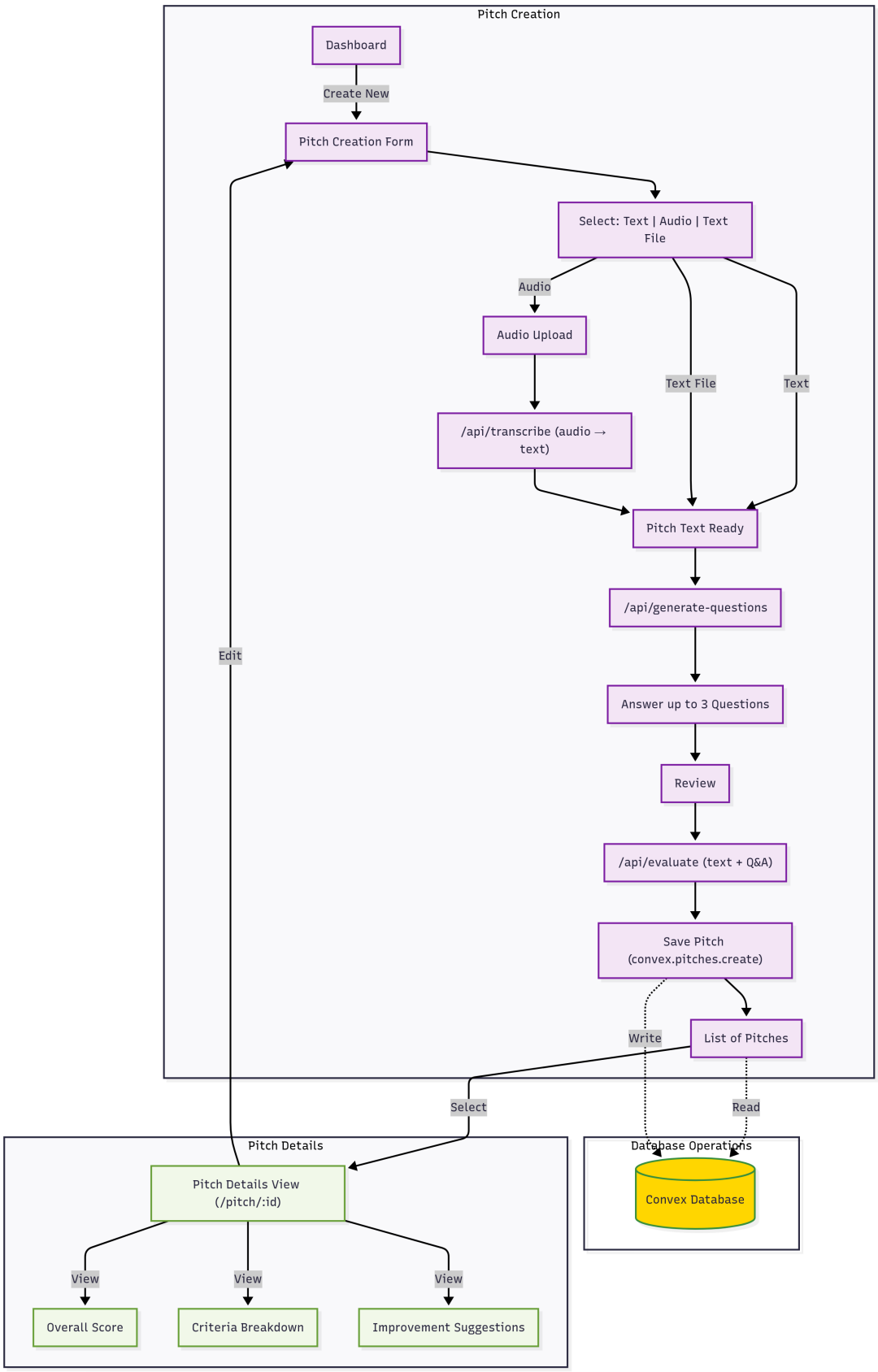

Figure 3.4: Pitch creation flow with Q&A generation and Convex storage

The question-and-answer step is optional but enabled by default based on testing that showed significant evaluation quality improvements. Users can skip this step for faster processing, though this may result in less

comprehensive evaluations when key information is missing from the original pitch. The questions component implements a multi-step interface that presents one question at a time with visual progress indicators, helping users maintain focus while providing targeted responses that address specific evaluation criteria gaps. The system dynamically generates up to three questions using GPT-4's analysis of the pitch content, with each question targeting critical missing information needed for accurate assessment. The component includes explanatory text stating "*These questions help improve your evaluation accuracy*" to guide user understanding of the purpose and value of providing additional context during the evaluation process.

The individual pitch view renders both numeric scores and qualitative feedback alongside generated follow-up questions, providing users with complete evaluation information in a clear, structured format optimized for review and improvement planning. The pitch page implementation uses lazy loading to optimize performance by loading components such as 'ScoreOverview', 'EvaluationSummary', 'DetailedAnalysis', and 'QuestionsSection' only when needed, while maintaining responsive skeleton placeholders during content loading. The system supports both legacy and structured evaluation formats through conditional rendering that automatically detects the evaluation data type and displays the appropriate analysis components. This modular architecture enables rich data visualization with transcript sections, comprehensive score breakdowns, and targeted improvement recommendations while maintaining fast initial page loads and smooth user interactions throughout the evaluation review process.

3.5 Performance, Deployment, and Reliability

System performance was optimized to handle the computational demands of GenAI processing while maintaining responsive user interactions. The implementation uses virtualized lists to maintain interface responsiveness with large datasets, code splitting to reduce initial page load times, and reactive queries that eliminate manual polling by automatically updating the interface when results become available.

The deployment architecture uses Vercel for hosting with environment variables for configuration, including database URLs, authentication keys, and API credentials. Critical API routes run on Vercel's Edge runtime for global distribution and reduced latency. The deployment pipeline ensures identical code paths across development and production environments to maintain evaluation result consistency.

For reliability, comprehensive error handling was implemented throughout the system. API routes return clear, actionable error messages that help users resolve problems without exposing sensitive system details. The system includes exponential backoff retries for external API interactions and strict JSON validation before data persistence to prevent corrupted evaluations from entering the research database.

3.6 Implementation Results and User Experience

The implemented system achieves the design goals of providing consistent, evidence-based startup pitch evaluations. Text-based pitch evaluations typically complete within 30 to 60 seconds, providing timely feedback without disrupting research workflows. Audio-based submissions require additional processing time for transcription but remain within acceptable bounds for practical use.

The user interface provides immediate visual feedback throughout the evaluation process with real-time updates that make data changes visible without manual page refreshes. The system maintains responsive performance during concurrent usage, supporting both individual use and classroom scenarios where multiple students submit pitches simultaneously.

The systematic development approach ensured all components integrate properly while maintaining research data integrity. The evidence-based evaluation framework successfully addresses the central tendency problem identified in initial testing, producing meaningful score distributions that discriminate between pitches of different quality levels.

The complete implementation represents a working system that combines GenAI capabilities with research-quality data collection, providing both immediate user value and a foundation for continued evaluation research.

Chapter 4

Evaluation and Results

This chapter evaluates how well the Pista system performs compared to Winds2Ventures (W2V). The research addresses a key question: how do different GenAI evaluation systems compare when rating the same startup pitches, and what can we learn from their agreement patterns?

To answer this question, both GenAI systems were compared on how they scored the same startup pitches. The analysis examined whether the two systems agreed on which pitches were good, average, or below average. The comparison focuses on the final overall scores that each system produces. Statistical measures of agreement were analyzed to understand how well different GenAI evaluation approaches align with each other.

Twenty-two startup pitches from university competitions were analyzed using statistical methods to measure agreement between the systems. The results show that the systems agree moderately well, but there are some systematic differences in their evaluation approaches. Pista tends to give slightly higher scores than W2V, which appears to use more conservative evaluation criteria. The Cohen's kappa coefficient indicates moderate agreement at 0.505, indicating the systems align much better than random chance.

This comparison helps us understand how different GenAI evaluation approaches perform and what we can learn from comparing automated assessment systems. The findings suggest that different GenAI evaluation systems can have distinct characteristics and evaluation patterns, even when assessing the same content. The results provide insights into the variety that exists within automated evaluation systems and how different algorithmic approaches can lead to different assessment outcomes. These findings are important for understanding the current diversity and future potential of GenAI-powered evaluation platforms.

4.1 Methodology

This section describes how the Pista system was tested against the Winds2Ventures platform. The final version of Pista was used, which utilizes GPT-4 with specialized prompts to evaluate startup pitches consistently. The testing approach focuses on comparing final scores from both systems using the same set of university competition pitches. Statistical methods were selected that account for chance agreement to provide meaningful comparisons.

4.1.1 Dataset Selection

Twenty-two startup pitches from university competitions were used as the test dataset for this evaluation. University pitches were chosen because they all follow the same format and contain similar information, which makes it easier to compare how the two systems evaluate them. University competition pitches provide a standardized structure that ensures both evaluation systems receive comparable input data. The pitches represent real student ventures that have been through competitive selection processes.

The pitches cover different areas such as technology, healthcare, agriculture, and consumer services. This variety demonstrates how both systems handle different types of businesses across multiple industry sectors. The diversity ensures that the evaluation results reflect how the systems perform across different business models and market contexts. Each industry sector presents unique challenges and opportunities that test different aspects of the evaluation frameworks.

To illustrate the dataset variety, here are representative examples from different sectors:

- **Coontent** (Marketing Technology): A GenAI-powered marketing automation platform that generates, publishes, and analyzes marketing campaigns. The team consists of software engineering students targeting marketing agencies in South Tyrol, with plans to expand beyond the region in 2025.

- **Serenity** (Digital Health): A mental wellbeing platform designed for young adults, offering GenAI-driven mood tracking, personalized wellness resources, and 24/7 chatbot support. The team targets the growing mental wellness app market valued at $139.9 billion.
- **Corptech** (Industrial Technology): A GenAI-powered methane leak detection system for industrial monitoring, addressing both financial losses and environmental concerns. The team already secured traction with distribution partners and interest from major gas companies.
- **Rideshare** (Transportation): A ridesharing platform focused on short-distance commutes in areas with limited public transportation, using a transaction-based model with 10% fees to help reduce traffic congestion.

Each pitch includes the standard parts you would expect: the problem they're solving, their solution, market analysis, team information, and how they plan to make money. Since all pitches follow the same format, both evaluation systems get the same type of information to work with.

The W2V evaluation results for these pitches were provided by the thesis supervisor to enable this comparative analysis. This supervisor-facilitated comparison allows for meaningful statistical analysis between different GenAI evaluation approaches using standardized pitch data.

4.1.2 Evaluation Frameworks

Both systems need to be explained since they use different ways to evaluate pitches. Understanding these differences helps explain why the scores sometimes vary between the systems. The frameworks use different criteria and weighting approaches, which can lead to different assessments of the same pitch. Comparing these frameworks helps identify the strengths and limitations of each evaluation approach.

Pista Evaluation Framework

Pista uses four main areas to evaluate pitches, with different weights for each:

1. **Problem-Solution Fit (30%)**: How well the startup understands the problem and how good their solution is
2. **Business Model & Market (30%)**: How they plan to make money and if the market is big enough
3. **Team & Execution (25%)**: How capable the team is and if they can actually build their solution
4. **Pitch Quality (15%)**: How well they present their idea and communicate their vision

Each area gets a score from 1 to 10, and then Pista calculates a final score using the weights above. The system also provides written feedback to explain the scores and reasoning behind each assessment. The weighted scoring approach ensures that the most important aspects of a startup pitch have greater influence on the final evaluation. This systematic approach helps maintain consistency across different pitches and evaluation sessions.

Winds2Ventures Evaluation Framework

W2V uses nine different criteria in its GenAI-powered evaluation system, focusing on aspects like problem clarity, solution viability, market analysis, team quality, and business model validation. The platform uses automated analysis trained on investment patterns and business assessment methodologies. The system appears to incorporate conservative evaluation approaches that reflect real-world investment considerations and startup challenges.

The key difference is that W2V uses a different algorithmic approach compared to Pista's evaluation methodology. This system tends to be more conservative in its assessments, applying stricter evaluation criteria when rating pitch quality. he evaluation algorithm weights risk factors and practical business challenges more heavily in its assessment process. The system's conservative approach reflects evaluation patterns that prioritize cautious assessment of startup viability and market potential.

4.1.3 Comparative Analysis Approach

Given the different methodologies employed by each system, a standardized comparison approach was required. The analysis focuses on the final overall scores that each system produces as the primary basis for comparison. This approach allows the systems' ultimate assessments to be compared regardless of their different internal processes. The comparison method focuses on whether both systems reach similar conclusions about pitch quality.

Both systems give final scores on a 1-10 scale, which makes it possible to compare them directly. Pista creates its final score by combining its four area scores with weights, while W2V gives an overall "Investibility" rating based on its nine criteria. Even though they calculate scores differently, the common 1-10 scale lets me compare how well they agree. The final scores from both systems were compared using statistical analysis. To calculate agreement levels, the following steps were taken:

1. Continuous scores were converted into three categories: *Below Average* (1.0–4.9), *Average* (5.0–6.9), and *Good* (7.0–10.0). By using this categorization, the focus was placed on meaningful quality differences rather than minor variations in scoring.

2. Cohen's kappa coefficient was employed as the primary agreement measure, since random chance agreement is accounted for by this statistic. The kappa coefficient is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

   where p_o represents observed agreement and p_e represents expected agreement by chance.

3. Descriptive statistics were calculated to provide insight into score distributions and to reveal potential systematic differences between the systems.

4.2 Results and Analysis

This section presents the comparison results between Pista and Winds2Ventures. The analysis shows moderate agreement between the systems with interesting patterns in how they evaluate startup pitches.

The following table shows the complete statistical analysis of the 22 pitches:

Table 4.1: Statistical Analysis: Pista vs Winds2Ventures Comparison

Agreement Metrics	**Value**
Cohen's Kappa	0.505
Interpretation	Moderate Agreement
Observed Agreement	77.3% (17/22)
Expected Agreement (Random)	54.1%
95% Confidence Interval	[0.144, 0.865]
Score Statistics	
Pista System	
Mean Score	5.36
Standard Deviation	0.71
Score Range	4.0 – 6.3
Winds2Ventures	
Mean Score	5.20
Standard Deviation	0.67
Score Range	4.3 – 6.4
System Comparison	
Mean Difference	+0.16 (Pista higher)
Correlation Coefficient	0.612

4.2.1 Statistical Interpretation

The Cohen's kappa coefficient of 0.505 indicates moderate agreement between Pista and W2V. This result shows that the systems agree more than they would by random chance, but there are still systematic differences in how they evaluate pitches. The moderate agreement level suggests that both systems capture similar underlying patterns in pitch quality. However, the differences indicate that each system brings unique perspectives to the evaluation process.

The systems agreed on 77.3% of the pitches (17 out of 22), meaning both systems categorized the pitches in the same way (below average, average, or good). This suggests that different GenAI evaluation systems can identify similar patterns in pitch quality assessment more than three-quarters of the time. The 95% confidence interval of [0.144, 0.865] indicates statistical significance for this agreement level.

The analysis shows that Pista gives slightly higher scores on average (0.22 points), suggesting that different GenAI evaluation systems can have distinct scoring tendencies, with some being more optimistic while others more conservative.

4.2.2 Score Distribution Analysis

Analysis of how the systems categorized the pitches reveals clear differences in their evaluation tendencies. W2V rated more pitches as "Below Average" (9 pitches) compared to Pista (6 pitches), while both systems rated all remaining pitches as "Average.". This pattern suggests that W2V applies stricter evaluation criteria when assessing pitch quality, being more likely to categorize pitches as below average. The differences reflect different algorithmic approaches to weighing risk factors and business viability in the evaluation process.

This pattern confirms that W2V uses more conservative evaluation criteria, which appears to be built into its assessment algorithm. The system seems designed to apply stricter standards when evaluating startup potential, possibly incorporating more cautious approaches to risk assessment and business viability analysis.

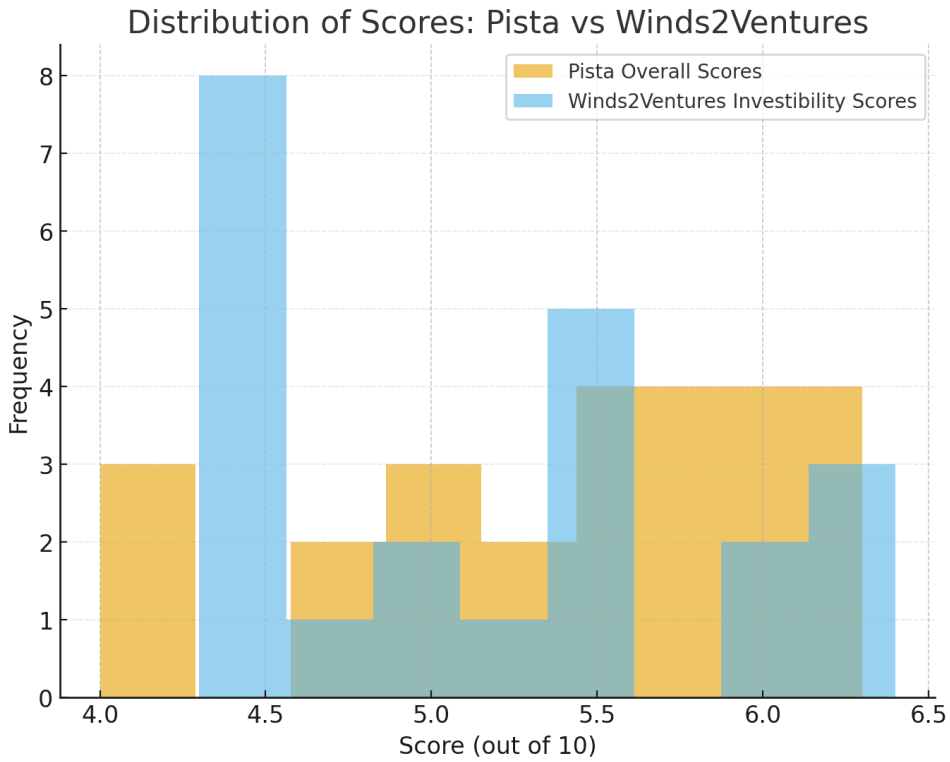

Figure 4.1: Histogram comparing the distribution of overall scores between Pista and Winds2Ventures.

Both systems put most pitches in the "Average" category (Pista: 16 pitches, W2V: 13 pitches), which is expected for university competition participants. These student ventures have potential but need more development before they are ready for investment. Notably, neither system awarded any pitch a "Good" rating, suggesting that both systems maintain relatively conservative evaluation standards for this dataset.

4.2.3 Discussion and Implications

These results show that different GenAI evaluation systems can produce meaningful assessments of startup pitches, with each system bringing distinct evaluation characteristics. The moderate agreement level demonstrates that GenAI systems can capture many important patterns in pitch evaluation consistently. However, the systematic differences suggest that different GenAI evaluation approaches can have unique characteristics and assessment tendencies. The findings highlight the diversity that exists within GenAI-powered evaluation systems and how different algorithmic approaches can complement each other.

The moderate agreement shows that Pista captures many of the same patterns that other GenAI evaluation systems identify when assessing startup pitches. This demonstrates the system's value in situations requiring consistent evaluation of multiple pitches, such as initial screening for competitions or standardized feedback provision to entrepreneurs. The GenAI system can process large volumes of pitches quickly while maintaining consistent evaluation criteria. These capabilities make GenAI evaluation particularly useful for high-volume assessment scenarios.

However, the observed scoring differences between GenAI systems indicate that multiple evaluation approaches may yield more comprehensive assessment perspectives. Different GenAI systems may weigh various factors differently, leading to more balanced overall evaluations. The combination of different algorithmic approaches can help identify both optimistic and conservative viewpoints in startup assessment. Understanding these differences helps users interpret GenAI evaluation results more effectively.

The optimal evaluation strategy may involve employing multiple GenAI systems to obtain diverse analytical perspectives on startup potential, combining their insights to create more comprehensive assessments. This multi-system approach gets the benefits of GenAI efficiency and consistency while capturing different evaluation methodologies and assessment criteria. Understanding how different GenAI systems approach evaluation helps users make more informed decisions about startup potential.

Chapter 5

Conclusion and Future Work

This thesis explored and implemented the development of Pista, a GenAI-powered startup pitch evaluation platform, and conducted a supervisor-facilitated statistical comparison with Winds2Ventures (W2V) to understand how different GenAI evaluation systems perform. Twenty-two university startup pitches were evaluated using both systems, and the agreement patterns were analyzed using Cohen's kappa statistics.

5.1 System Development and Key Findings

This section presents the system development outcomes and key research findings from this thesis work.

- **Complete Functional System**: Pista was developed and deployed as a full-stack web application using Next.js 15, Convex database, and Clerk authentication. The system integrates GPT-4 for GenAI analysis across four evaluation dimensions. Pista supports text uploads, file processing, and audio transcription with real-time evaluation progress. The platform is live at `https://pista-app.vercel.app` and provides structured feedback with specific improvement recommendations.
- **Statistical Comparison Results**: Moderate agreement was found between the two GenAI systems, with a Cohen's kappa coefficient of 0.505 and 77.3% observed agreement. The statistical analysis reveals that different GenAI systems bring distinct perspectives to startup evaluation, even when analyzing identical pitch content.
- **Practical Performance Insights**: Pista delivers evaluations in 30–60 seconds at $0.10–0.15 per assessment, compared to traditional manual evaluation processes. The system provides 24/7 availability and consistent evaluation criteria across all pitches. The research found that GenAI evaluation works best for problem-solution fit assessment but has limitations when evaluating team capabilities and execution potential.

5.2 Research Contributions

This section outlines the contributions of this thesis to understanding GenAI-powered startup evaluation.

This work contributed a functional GenAI evaluation system with empirical comparison data demonstrating how different GenAI approaches perform in practice. The statistical analysis provides the first documented comparison of GenAI evaluation platforms using standardized metrics. Specific patterns were identified where GenAI evaluation excels and areas where it has limitations. The research demonstrates that GenAI evaluation systems can provide valuable assessment capabilities while maintaining practical advantages like speed, cost efficiency, and accessibility.

The deployed Pista system serves as a proof-of-concept for how modern GenAI technologies can be applied to startup evaluation challenges. The research demonstrated that building such systems is technically feasible and can produce meaningful results when properly designed and implemented.

5.3 Practical Implications

This section discusses how these findings apply to real-world startup evaluation scenarios.

The results show that GenAI evaluation systems like Pista work well for initial screening and educational contexts where consistent feedback helps entrepreneurs improve their pitches. The system's speed and accessibility make it valuable for competition organizers and startup accelerators processing many applications. However, the moderate agreement between systems suggests that multiple GenAI perspectives provide more comprehensive assessment than relying on a single platform.

The research found that different GenAI systems serve distinct purposes based on their evaluation characteristics; Pista's consistent scoring standardizes feedback for learning environments, while W2V's varied scoring patterns better reflect investment decision contexts. Understanding these differences helps users choose appropriate GenAI tools for their specific evaluation needs.

5.4 Limitations and Future Work

This section outlines the current limitations of this research and potential directions for future work.

- **Current Limitations**: The evaluation used 22 university competition pitches, representing a limited sample size. The comparison involved only two GenAI platforms, so broader patterns across multiple GenAI systems remain unknown. University pitches may not reflect the complexity and variety found in professional startup environments. These limitations mean that the findings provide initial insights rather than a comprehensive understanding of GenAI evaluation system performance.

- **Technical Enhancement Opportunities**: Future versions of Pista could incorporate multimodal analysis, including video and audio evaluation capabilities. Real-time interactive features where the GenAI asks clarifying questions could improve assessment depth. Integration with pitch presentation tools could provide seamless evaluation workflows. Enhanced machine learning models trained on larger datasets could improve evaluation accuracy and reduce score clustering patterns.

- **Research Extensions**: Longitudinal studies tracking startup success rates could validate whether GenAI evaluation scores predict actual business outcomes. Comparative studies including more GenAI platforms would reveal broader patterns in automated evaluation approaches. Professional startup pitch datasets would test whether university pitch patterns apply to investment contexts. Human evaluator comparison studies could establish benchmarks for GenAI system performance against expert assessment.

5.5 Final Reflections

The development and evaluation of Pista demonstrate that creating functional GenAI evaluation systems is achievable with current technology and provides valuable insights into automated startup assessment. The statistical comparison with W2V reveals that GenAI evaluation systems bring distinct characteristics and perspectives to startup assessment, suggesting value in multiple-system approaches rather than relying on single platforms.

The moderate agreement between systems shows that GenAI evaluation captures meaningful patterns in pitch quality while maintaining practical advantages like speed, cost efficiency, and consistent availability. These capabilities make GenAI evaluation particularly valuable for initial screening, educational feedback, and accessibility in underrepresented regions where traditional evaluation resources may be limited.

This research shows that GenAI evaluation systems represent a practical tool for startup ecosystems, complementing rather than replacing human assessment. The key is understanding each system's strengths and limitations to use them effectively in appropriate contexts. As GenAI technology continues advancing, evaluation systems like Pista will likely become more sophisticated and valuable for supporting entrepreneurship development and startup assessment processes.

Bibliography

[1] Qubit, 2024. https://qubit.capital/blog/impress-investors-winning-presentation-skills.

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

[3] Actu.AI. Gpt-4 achieves human-level performance in analogical reasoning tasks, 2025. https://actu.ai/en/gpt-4-achieves-human-level-performance-in-analogical-reasoning-tasks-according-to-a-study-49284.html.

[4] FasterCapital. Startup evaluation and feedback: Evaluating startups: A comprehensive guide, 2025. https://fastercapital.com/content/Startup-Evaluation-and-Feedback--Evaluating-Startups--A-Comprehensive-Guide.html.

[5] Shruthi Garadi Kalvapalle, Nelson Phillips, and Joep Cornelissen. Entrepreneurial pitching: A critical review and integrative framework. *Academy of Management Annals*, 18(2):550–599, 2024.

[6] FasterCapital. Breaking barriers: How inclusive entrepreneurship is changing the game, 2025. https://fastercapital.com/content/Breaking-Barriers--How-Inclusive-Entrepreneurship-is-Changing-the-Game.html.

[7] GenQE. The future of ai evaluation: Exploring new benchmarks and challenges, 2025. https://genqe.ai/ai-blogs/2025/03/06/the-future-of-ai-evaluation-exploring-new-benchmarks-and-challenges/.

[8] L. Gius. Disagreement predicts startup success: Evidence from venture competitions. *Strategy Science*, 2024.

[9] L. Balachandra, T. Briggs, K. Eddleston, and C. Brush. Don't pitch like a girl!: How gender stereotypes influence investor decisions. *Entrepreneurship Theory and Practice*, 43(1):116–137, 2019.

[10] Chia-Jung Tsay. Visuals dominate investor decisions about entrepreneurial pitches. *Academy of Management Discoveries*, 7(3):343–366, 2021.

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

[12] David Hirshleifer, Yaron Levi, Ben Lourie, and Siew Hong Teoh. Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics*, 133(1):83–98, 2019.

[13] Massimo G. Colombo, Danilo D'Adda, and Annalisa Quas. The geography of venture capital and entrepreneurial ventures' demand for external equity. *Research Policy*, 48(5):1150–1170, 2019.

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

[15] Winds2Ventures. Winds2ventures startups, 2024. https://w2v.network.

[16] Mark Steyvers and Aakriti Kumar. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, 19(5):722–734, 2024.