Crypto Fund Due Diligence Automation: An AI-Driven TDSP Approach for Efficient and Scalable Risk Assessment

Chaima Mejri, Maram Mejri, Mariam Maammar, Oussema Bargougui, Firas Bouguerra, Houssam Kheder

Department of Data Science, ESPRIT University, Tunis, Tunisia
Integrated Project in Data Science Engineering in Collaboration with VALUE Company, Tunisia

Manuscript received January 20, 2025; revised May 19, 2025. Thesis defense scheduled for May 21, 2025. This research project was conducted as part of the final-year Integrated Engineering Project in Data Science at ESPRIT University. The project was carried out in collaboration with VALUE Company, Tunisia, and supervised by the academic faculty of the Department of Data Science and Artificial Intelligence. It follows the Team Data Science Process (TDSP) methodology to design and implement an AI-powered due diligence system for crypto funds, integrating document processing, risk assessment, and answer generation using machine learning and natural language processing.

Abstract: The rapid growth of the cryptocurrency industry has created an urgent need for scalable, intelligent tools that support effective due diligence. Traditional methods of evaluating crypto funds are time-consuming, manual, and ill-suited to the fragmented and fast-changing nature of decentralized markets. This paper presents DueXpert, an automated due diligence platform that leverages state-of-the-art AI components—including NLP, Graph-based Retrieval-Augmented Generation (GraphRAG), and LSTM-based forecasting—to assess legal, financial, and reputational risks associated with crypto investment opportunities. The system integrates document parsing, risk scoring, real-time question answering, and report generation into a unified workflow guided by the Team Data Science Process (TDSP). Through real examples and evaluations, we demonstrate how DueXpert enhances the speed, accuracy, and transparency of crypto fund analysis. The platform not only reduces operational overhead but also empowers investors and regulators to make informed decisions with greater confidence.

Index Terms: Crypto fund due diligence, GraphRAG, LLM evaluation, risk scoring, time-series forecasting, LSTM model, AI-driven compliance, NLP for investment analysis, automated report generation, unstructured document parsing.

1. Introduction

The rise of cryptocurrencies has transformed the financial landscape by introducing digital assets that operate on decentralized networks, without the need for traditional intermediaries like banks or governments. This decentralized nature offers many advantages, such as increased transparency, faster transaction times, and borderless transfers. However, it also introduces new challenges, particularly in how investors and regulators assess the risks and legitimacy of crypto-investments.

Due diligence is a critical process in traditional finance that involves thoroughly investigating and verifying an investment opportunity before committing funds. It helps ensure that investments are legitimate, comply with regulations, and align with the investor's risk tolerance and objectives. In the context of cryptocurrencies and crypto funds, due diligence becomes even more complex due to several factors:

- **Rapid Market Evolution:** The cryptocurrency market is highly volatile and evolves much faster than traditional financial markets. New tokens, protocols, and regulations frequently emerge, making it difficult to keep up with all relevant information.
- **Fragmented and Unstructured Data:** Information related to crypto funds is scattered across multiple sources, including regulatory filings, whitepapers, blockchain transaction data, news articles, and social media. Much of this data is unstructured or semi-structured, requiring significant effort to collect, clean, and interpret.
- **Lack of Standardization:** Unlike traditional assets, crypto investments often lack standardized reporting formats or universally accepted evaluation frameworks, which complicates direct comparisons and objective assessments.
- **Regulatory Uncertainty:** Cryptocurrency regulations vary widely by jurisdiction and continuously evolve, increasing the difficulty of ensuring compliance through manual processes.

These challenges mean that traditional due diligence methods, largely reliant on manual research and expert analysis, are increasingly insufficient. They are time-consuming, costly, and prone to human error, often taking weeks or months to complete for each investment. This creates barriers for both retail and institutional investors seeking to make timely informed decisions.

This research addresses these challenges by developing DueXpert, an automated platform that leverages artificial intelligence (AI), machine learning (ML), and natural language processing (NLP) to revolutionize crypto fund due diligence. DueXpert integrates data from diverse sources, applies AI-driven document parsing and analysis, provides tailored responses to the questions listed in the question bank, and produces comprehensive real-time risk reports. By automating key components of the due diligence process, the platform significantly reduces cost and time, while improving the accuracy and reliability of investment assessments.

The development and implementation of DueXpert are systematically guided by the Team Data Science Process (TDSP) methodology, which provides a structured framework from business understanding through deployment and evaluation. This ensures that the solution is not only technically robust, but also aligned with stakeholder needs, scalable, and maintainable over time.

This paper presents the full scope of the project, describing its motivation, technical design, data processing pipelines, AI models, evaluation results, and deployment architecture. Through DueXpert, our aim is to empower investors, regulators, and fund managers with a faster, smarter, and safer tool to navigate the complex and dynamic world of cryptocurrency investments.

2. Related Work

Several existing solutions have explored aspects of due diligence automation in the context of cryptocurrencies, but they typically remain limited in scope or technical depth.

Tools such as **ChainSecurity** and **MythX** focus on smart contract auditing by applying static and symbolic analysis methods to identify vulnerabilities in code. These tools are valuable in detecting programming risks but do not address fund-level concerns such as team transparency, legal compliance, or financial planning.

Platforms like **TokenInsight** and **ICOBench** offer high-level project scores and investment grades. However, these rely on shallow metrics, heuristics, and community voting, which lack methodological transparency and fail to explain the underlying reasoning or data sources used in the evaluation.

In the academic space, Liu et al. proposed a machine learning approach for classifying Initial Coin Offering (ICO) whitepapers using document features and shallow neural networks [1]. While effective for binary classification tasks, this method does not scale to nuanced, multi-label risk assessments or dynamic document querying. Similarly, Zhao et al. [2] investigated entity extraction in token-related documents using supervised models, which require significant manual annotation and are sensitive to domain shifts.

Unlike these works, **DueXpert** is designed as an end-to-end due diligence automation platform. It integrates GraphRAG-based document analysis, risk-specific question answering, and financial

forecasting into a unified pipeline. Our system goes beyond static rule-based checks and offers interpretability, flexibility, and dynamic analysis capabilities that better reflect the complexity of real-world crypto fund assessments.

3. Problem Definition and Business Understanding

3.1. Context and Challenges

The cryptocurrency market has seen rapid growth, attracting a wide range of investors, from individual retail investors to large institutional funds. However, this fast growth has brought several challenges, especially in evaluating investments in crypto funds. Traditional due diligence, which is essential to assess the legitimacy and risk of investments, faces significant difficulties in the crypto space.

One of the biggest problems is the cost and time involved in manual due diligence. Typically, creating a due diligence report on a crypto fund can cost anywhere from $15,000 to $50,000 and takes one to three months to complete. This makes it difficult for many investors, especially smaller ones, to conduct thorough evaluations. The process is slow because it requires experts to go through large amounts of data, such as white papers, financial reports, transaction histories, and regulatory documents, by hand.

In addition, the crypto industry has a high level of fraud and regulatory risk. In 2024, crypto fraud losses alone reached $5.6 billion, which shows the urgent need for better ways to detect fraud. These scams include everything from Ponzi schemes to money laundering, taking advantage of the complexity and lack of transparency in blockchain networks. In addition, cryptocurrency regulations are constantly changing and vary between countries. This makes it difficult to stay up to date and ensure compliance with the law, which increases the risks for investors.

Current due diligence platforms do not cover the full range of activities required for a comprehensive evaluation. Although some platforms provide isolated services, such as market data analysis or compliance checks, few offer end-to-end automation. This means they don't bring together all the necessary tasks, such as collecting data, analyzing documents, detecting fraud, and ensuring regulatory compliance, into a single real-time system. Without an integrated, automated solution, it's difficult for investors to get a complete picture of a fund's risk, leading to slower decision-making and higher costs.

3.2. Objectives

The primary goal of this project is to develop DueXpert, an AI-powered platform that will address the challenges faced in traditional crypto fund due diligence. DueXpert is designed with several key objectives that will streamline the due diligence process, making it faster, more affordable, and more accurate for investors and fund managers. These objectives are as follows:

- **Develop a Due Diligence Application:** The primary objective is to build a clear and structured application for conducting due diligence on cryptocurrency funds. This application ensures that each fund is evaluated based on consistent and well-defined criteria, such as market analysis, regulatory compliance, and financial health. It provides the foundation for assessing risks and opportunities in a systematic way, eliminating the subjectivity and inconsistency found in manual evaluations. The application also serves as the basis for automating the entire review process.
- **Identify Key Risk Factors:** DueXpert focuses on identifying the specific risks that are common and critical in the context of crypto assets. These include fraud, lack of regulatory compliance, market volatility, and security vulnerabilities. The system is designed to detect these issues through analysis of fund documents, financial data, and blockchain activity. It classifies risk into different categories and contributes to generating a risk score that reflects the overall health and reliability of the fund. This score provides a fast, reliable signal for investors and stakeholders.
- **Provide Automated Solutions:** Another key goal of the project is to reduce the time, cost,

and manual effort involved in the due diligence process by providing automated solutions. DueXpert automates data collection through APIs, web scraping, and document extraction. It processes and analyzes this data using artificial intelligence tools, including language models and vector search, to answer due diligence questions without requiring human intervention. This automation enables faster and more efficient analysis by reducing the time and manual effort required to assess individual crypto funds, while ensuring consistent and reliable results.

- **Generate Comprehensive Reports:** The final objective is to generate detailed, clear, and professional reports that summarize the findings of the due diligence process. These reports are automatically created in PowerPoint format and include the fund's profile, risk scores, regulatory observations, and other key insights. The report templates are designed to be visually appealing and easy to understand, allowing investors and decision-makers to review due diligence results quickly and make informed choices. This step also supports communication with external stakeholders, such as clients or regulatory bodies.

These four objectives-developing an application, detecting risk, automating the process, and generating reports are central to the value DueXpert brings to the due diligence process.

3.3. Stakeholders

DueXpert is designed to benefit several groups involved in the evaluation, regulation, and management of cryptocurrency investments. Each stakeholder has specific needs that the platform addresses through its features.

- **Retail and Institutional Investors:** Investors need to verify if a crypto fund is secure, profitable, and compliant. DueXpert helps by providing clear, structured reports and real-time answers to due diligence questions. These features reduce the effort and complexity of reviewing investment opportunities and help both small and large investors make better decisions, with less time and cost.
- **Crypto Fund Managers:** For fund managers, DueXpert offers a way to demonstrate transparency and credibility. By using the system to evaluate their own funds, managers can check compliance status, identify areas of improvement, and generate professional reports to share with investors. This enhances their ability to attract and retain investment by showing they meet key due diligence standards.
- **Regulators and Compliance Bodies:** Regulatory institutions benefit from DueXpert's ability to automate the review of fund documentation and detect signs of non-compliance. The platform can assist in monitoring whether crypto funds meet legal and financial reporting obligations, including those related to anti-money laundering (AML), know your customer (KYC), and tax reporting. This helps regulators oversee the ecosystem more effectively and take timely action when needed.
- **Traditional Financial Institutions:** Banks and other financial organizations entering the crypto space need reliable tools to assess potential risks. DueXpert provides a comprehensive analysis of crypto funds in a format that is structured and easy to integrate into internal review processes. It supports these institutions in making cautious, data-driven decisions about partnerships or investments.

DueXpert's features—including its structured framework, automated processing, risk detection, and report generation—are built around the needs of these core stakeholders. By simplifying and accelerating due diligence, it enables each of them to operate more securely and efficiently in the rapidly evolving world of digital assets.

4. Data Acquisition and Understanding

One of the foundational steps in the DueXpert project was the acquisition and exploration of data related to cryptocurrency funds. Given the decentralized nature of the crypto ecosystem and the diversity of data sources, this step required the collection of both structured and unstruc-

tured information from multiple channels. The goal was to ensure that all relevant aspects of a fund—financial, legal, technical, and operational—were included in the due diligence process.

4.1. Data Sources

The data acquisition strategy for DueXpert involved collecting information from various domains, including public APIs, official documentation, regulatory databases, and blockchain networks. The sources were chosen to align with the requirements of the due diligence application and to provide enough coverage for the identification of key risk factors.

- **APIs:** The platform integrated with trusted APIs such as CoinMarketCap, CoinGecko, Glassnode, CryptoCompare, and Etherscan to collect market data, on-chain analytics, and blockchain-related statistics. These sources provided insights into token prices, trading volumes, wallet movements, contract interactions, and other essential metrics.
- **Blockchain Nodes:** Ethereum and Bitcoin node connections were used to directly access and verify blockchain transaction data, offering an additional layer of trust and transparency for risk analysis.
- **Regulatory and Legal Sources:** Data from government repositories such as the SEC's EDGAR database was used to gather legal disclosures and regulatory filings. These documents helped determine whether a fund complied with existing financial laws and provided transparency on ownership and fund structure.
- **Web Scraping:** For information not available through structured APIs, tools like BeautifulSoup and Selenium were used to scrape content from crypto fund websites, legal portals, and news articles. This included whitepapers, investment brochures, compliance statements, and legal disclaimers.

This multi-layered data collection approach enabled DueXpert to cover the various dimensions of a crypto fund needed for effective due diligence.

4.2. Data Types and Content

The acquired data was both qualitative and quantitative in nature. Key categories included:

- **Descriptive Data:** Information about the fund's name, team, strategy, and jurisdiction.
- **Financial Data:** Historical token prices, fund returns, assets under management (AUM), liquidity metrics.
- **Regulatory Data:** Registration details, AML/KYC status, license verification, and any previous compliance violations.
- **Technical Data:** Smart contract details, GitHub activity, and contract audit reports.
- **Transactional Data:** Blockchain interactions, wallet balances, and transfer histories.

These datasets provided a comprehensive view of a fund's operations and risk exposure.

4.3. Data Understanding and Exploration

Before applying any AI models or automation techniques, the team conducted an exploratory data analysis (EDA) to understand the structure, distribution, and quality of the collected data. This step included:

- Verifying the completeness and consistency of the collected data.
- Identifying missing values and noisy entries in scraped text and PDFs.
- Analyzing token behavior over time to flag suspicious activities or abrupt changes in market patterns.

This preliminary phase was crucial in shaping the data engineering and modeling steps that followed. It ensured that only reliable and relevant information was passed to the due diligence pipeline, supporting the accuracy and credibility of the final outputs.

5. Data Preparation

After conducting exploratory data analysis, the project entered the data preparation phase, where both structured and unstructured data were cleaned, transformed, and organized for use in the DueXpert system. This step was critical to ensure the quality, consistency, and usability of the data before it was fed into the AI models, risk analysis components, and report generation tools.

5.1. Preparation of Unstructured Data

DueXpert relies heavily on unstructured documents such as fund whitepapers, legal filings, audit reports, and web-based content. These documents were collected via web scraping, PDF extraction tools (e.g., PDFMiner, PyMuPDF), and regulatory sources.

Once collected, the documents underwent text cleaning and normalization. This included removing unwanted characters, correcting encoding issues, and ensuring sentence segmentation. The cleaned text was then processed using natural language processing (NLP) tools to extract important entities such as project names, dates, jurisdictions, wallet addresses, compliance terms, and team members. These entities were stored in structured formats for downstream analysis and used for semantic search and compliance verification.

Large text documents were also chunked semantically into smaller, meaningful sections to support the GenAI-powered question-answering module. These chunks were vectorized using sentence embedding models like nomic-embed-text, and indexed using tools like FAISS to allow fast and relevant document retrieval.

5.2. Preparation of structured Data

In addition to unstructured text, the project also used structured time-series data, particularly token price histories and transaction volume data. This data was collected from APIs such as CoinMarketCap, CoinGecko, and Etherscan, and included metrics like:

- Token price over time
- Market volume
- Wallet activity
- On-chain transaction counts

After collection, this structured data was cleaned by handling missing values, ensuring consistent date formats, and verifying numeric ranges. It was then formatted into time series datasets. This structured data was prepared specifically for modeling with an LSTM (Long Short-Term Memory) neural network, which was used later in the pipeline to forecast token behavior and enrich the fund's risk profile.

5.3. Data Storage and Organization

Once cleaned and processed, the data was stored in appropriate systems:

- **MongoDB** stored semi-structured content and document metadata.
- **FAISS** was used for indexing document embeddings to support fast retrieval in the Q&A system.

This clear separation of data types ensured smooth integration between the data preparation pipeline and the later AI models and applications.

6. System Architecture Overview

To provide a high-level understanding of the platform's components and data flow, we present the overall architecture of DueXpert in Fig. 1. The system is organized into multiple layers, each responsible for a distinct part of the pipeline: data collection, storage, processing, modeling, deployment, and user interaction. These layers interact seamlessly to enable end-to-end automation of the crypto fund due diligence process. This modular design supports scalability, maintainability, and efficient integration of AI components such as LLM-based risk scoring and LSTM forecasting.

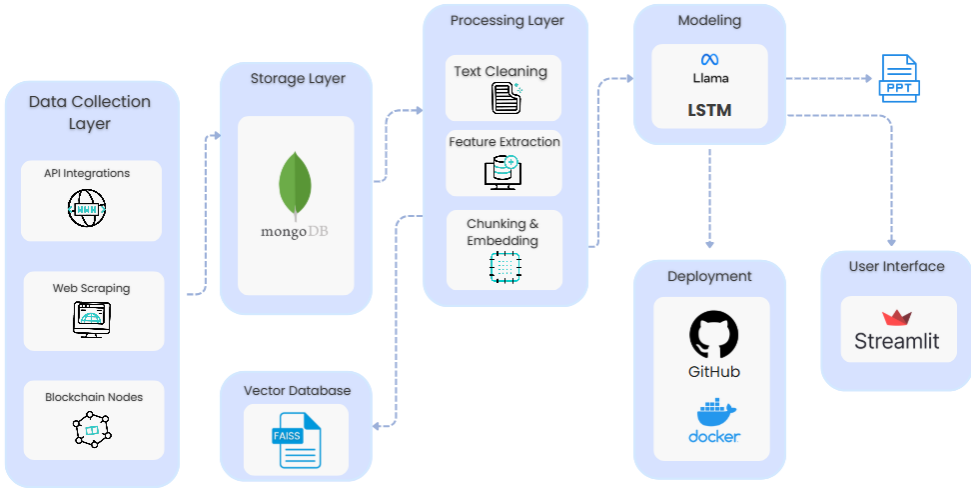

Fig. 1: System Architecture Overview of the DueXpert Platform.

The architecture illustrates the pipeline from data acquisition and processing to modeling, deployment, and interface delivery.

In the following sections, we detail the key AI modules illustrated in the architecture, including NLP-driven document analysis, the GraphRAG question-answering engine, and the LSTM-based financial forecasting model.

7. Modeling and Algorithm Development

The modeling phase of the DueXpert project focused on building intelligent, scalable components that could automate due diligence tasks such as document analysis, risk detection, dynamic question answering, and report generation. The modeling approach was carefully designed to work with the types of data collected during the earlier stages—both unstructured (e.g., whitepapers, audits) and structured (e.g., token prices, transaction metrics). Instead of training or fine-tuning models from scratch, the project relied on pretrained large language models and GraphRAG (Graph-based Retrieval-Augmented Generation) to extract and interpret knowledge, while a time-series LSTM model was used for risk-aware forecasting.

8. Modeling and Algorithm Development

The modeling phase of the DueXpert project integrates natural language processing components to automate the due diligence process. This phase begins with document ingestion and ends with the generation of reliable, context-aware answers to risk-based questions using a Large Language Model (LLM).

8.1. Document Ingestion and Parsing

Users upload unstructured documents such as crypto fund whitepapers, compliance reports, or scanned disclosures. These documents are parsed using a combination of OCR-based tools (such as PyMuPDF and PDFMiner) and natural language processing (NLP) pipelines. The goal is to extract clean, structured text from a variety of formats, including PDFs, images, and scanned documents.

8.2. Chunking and Embedding

Once the text is extracted and cleaned, it is semantically segmented into coherent units (or ”chunks”). These chunks are then converted into vector representations using a pretrained nomic-embed-text embedding model. The resulting embeddings are stored in a FAISS-based vector database for fast and accurate similarity retrieval. In parallel, the extracted and processed chunks are saved in MongoDB.

8.3. Graph-Based Retrieval with GraphRAG

To enhance context awareness and reduce hallucinations in LLM responses, the platform uses a Graph-based Retrieval-Augmented Generation (GraphRAG) system. In this approach, each text chunk becomes a node in a semantic graph. Edges between nodes represent topic similarity, shared entities, or logical continuity. When a due diligence question is posed, the query is embedded and matched to a relevant subgraph, rather than retrieving isolated chunks. This ensures that the response is based on semantically connected evidence.

8.4. Question Answering with LLaMA 3.1

Once the relevant subgraph is retrieved, it is passed as context to a Large Language Model—specifically, LLaMA 3.1. The model uses this grounded knowledge to answer predefined due diligence questions from the system's question bank. These questions cover legal, financial, technical, and governance-related risks. The answers generated by the LLM are then forwarded to the evaluation pipeline for risk scoring.

8.5. Risk Scoring and Rule-Based Evaluation

The risk scoring module in DueXpert is designed to deliver a transparent, explainable, and fully automated risk evaluation of a crypto fund. The system does not rely on probabilistic machine learning models for classification. Instead, it uses a structured pipeline powered by a language model acting under strict instructions—combined with a mathematical aggregation framework to generate the final score.

Each due diligence question is first posed to the GraphRAG-based retrieval system, which returns a context grounded in the fund's documentation. The context is then passed to a pretrained Large Language Model (Llama3.1). To ensure reliability and consistency, the model is given a specific role prompt:

”You are a Senior Investment Risk Analyst. Your task is to review the answer and classify the risk based on its content. Return one and only one of the following labels: 'positive', 'partial', or 'missing'.”

This approach eliminates ambiguity and hallucination by forcing the model to return only one of the four allowed labels per answer.

Each label is then mapped to a numerical risk score using the following logic:

- If the evaluation is ”positive”, the risk score is set to 0.0, indicating low risk.
- If the evaluation is ”partial”, the score is 0.5, representing a moderate risk.
- For evaluations labeled ”negative” or ”missing”, the score is assigned a value of 1.0, corresponding to high or critical risk.

These individual scores are treated as **Question Risk %**.

8.5.0.a. Example – Risk Classification Mapping
LLM Evaluation: partial
Mapped Score: 0.5
Risk Category: Use of Funds $\rightarrow$ Financial Risk

Each due diligence category (e.g., Legal, Financial, Governance, Technical, etc.) contains a defined number of critical questions. The **Category Risk %** is computed as the average of the

individual question scores within that category:

$$\text{Category Risk \%} = \frac{\sum \text{Question Risk \%}}{\text{Number of Critical Questions in Category}} \tag{1}$$

This provides a normalized score per category, where 0% is ideal and 100% is high risk.

The platform groups categories into two main dimensions:

- **Core Risk %:** Represents operational and compliance-related risk.
- **Strategic Risk %:** Reflects reputational, governance, and future-oriented concerns.

These are calculated using weighted averages:

$$\text{Core Risk \%} = \sum (\text{Category Risk \%} \times \text{Category Weight}) \tag{2}$$

$$\text{Strategic Risk \%} = \sum (\text{Category Risk \%} \times \text{Normalized Category Weight}) \tag{3}$$

Category weights are pre-assigned based on business relevance, and normalization ensures fair aggregation.

The final **Total Risk %** for a fund is the sum of the core and strategic scores:

$$\text{Total Risk \%} = \text{Core Risk \%} + \text{Strategic Risk \%} \tag{4}$$

This single percentage value reflects the fund's overall risk and can be used to drive reporting, decision-making, and flagging of high-risk projects.

8.6. LSTM-Based Market Behavior Forecasting

While traditional due diligence focuses on reviewing past and present information, it is equally important to assess how a fund's associated cryptocurrency may behave in the near future. For this reason, DueXpert includes a forecasting module that adds a forward-looking financial dimension to the risk evaluation. This component uses a deep learning model—called a **Long Short-Term Memory (LSTM)** network—to analyze historical price data and predict how the price of a fund's token might evolve over the next month.

Why Forecasting Matters
In the crypto market, prices are highly volatile and can shift quickly due to external events, internal project changes, or investor sentiment. Forecasting helps anticipate such movements. If the model predicts a strong drop or a period of instability, this is flagged as a potential risk. Even if a project looks reliable on paper, such signals may suggest underlying concerns that require deeper investigation.

What the Model Does
The model is trained to look at the past 60 days of a token's activity—including its price, volume, and daily changes—and use this information to forecast the next 30 days. The goal is not to predict exact prices but to detect unusual patterns, such as:

- A sudden expected decline
- Strong fluctuations (volatility)
- Signs of market instability

These patterns, once detected, are passed along to the report and risk scoring system.

How It Works
The LSTM model is a type of deep learning models for sequential data that is especially good

at learning patterns over time, like recognizing trends in a graph. It was trained using historical token data from trusted sources such as CoinGecko and blockchain explorers.

The model learns how token prices usually behave, and when it sees new data, it can estimate where the price might go next. It doesn't work by guessing, but by identifying statistical trends and comparing them to previous behaviors.

To ensure that the model is accurate and useful, it was tested using common evaluation techniques. The results showed that its forecasts were highly reliable, with small errors and a strong ability to match actual price movements.

How Forecasting Fits into Due Diligence

The insights produced by the forecasting model do not replace document analysis—they add an extra layer of financial awareness. For example, if a token's documentation looks solid, but the model forecasts instability, DueXpert highlights this in the final report. It's a way of catching risks that might not be visible in the whitepaper or legal materials.

By combining both static and dynamic analysis, DueXpert offers a more complete picture of the fund's risk profile—helping users not just evaluate the present, but also prepare for what might come next.

Input: 60 days of token data (price, open, high, low, volume, percent change)
Model Output: Predicted prices for 30 future days show a consistent downward trend with high volatility.
Interpretation: DueXpert flags this token as financially unstable over the short term, despite strong documentation.

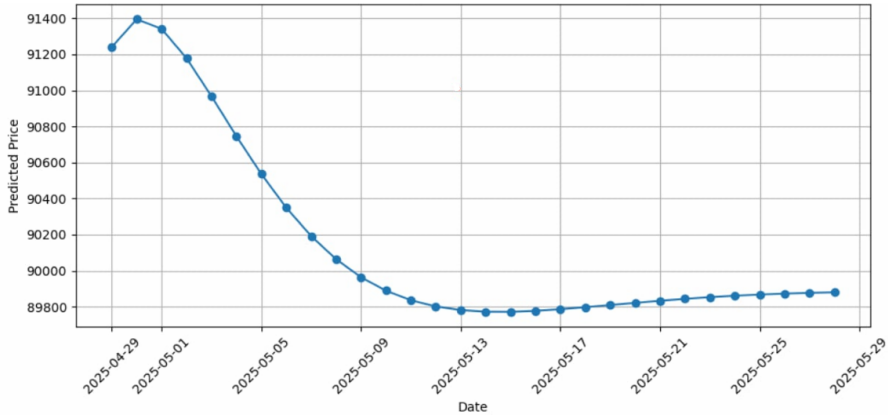

Fig. 2: Predicted Prices from BTC_USD Bitfinex

Forecasted token price trend showing a sharp decline followed by a flattening pattern.

8.7. Automated Report Generation

One of the final steps in the DueXpert pipeline is the automatic generation of a professional due diligence report, designed to summarize all the insights gathered during the analysis in a clear, structured, and investor-friendly format. This feature is especially important because, while the internal system performs advanced AI-driven processing, the final output must be understandable, actionable, and ready to share with decision-makers who may not be technically trained.

Purpose of the Report

The primary goal of the report is to translate complex data outputs into a structured summary that highlights the key findings about a crypto fund. It includes sections that cover the fund's

background, regulatory compliance, team transparency, technical infrastructure, token behavior, and risk forecasts. This report enables investors, analysts, or regulators to review a fund's status without needing to manually read through whitepapers, legal disclosures, or blockchain records.

Report Format and Structure
The report is generated in Microsoft PowerPoint (.pptx) format, using an automated process powered by the python-pptx library. A branded and pre-designed template is used to ensure consistency across different reports. Each report includes the following key sections:

- **Executive Summary:** A concise overview of the fund and its overall risk score
- **Risk Score Breakdown:** Visual representations of the scores in each category (e.g., legal, technical, financial)
- **Compliance Review:** Key findings related to KYC, jurisdiction, and regulatory disclosures
- **Forecast Analysis:** Graphs showing predicted token price behavior
- **Detected Red Flags:** Warnings or missing elements automatically flagged by the system
- **Recommendations:** Optional suggestions or next steps based on the findings

The use of PowerPoint format makes the report easy to read, share, and present in meetings or internal review boards.

The report is generated automatically once the analysis pipeline completes. There is no need for human editing or formatting. Users can download the report with one click through the interface, ensuring fast delivery of insights.

9. Evaluation

9.1. Evaluation of the Risk Classification (LLM-Based)

To ensure the reliability of the LLM-based classification system used for scoring due diligence answers, the team carefully designed prompts that instructed the LLM to rely only on the provided source materials. This was done to prevent hallucinations and ensure that the model did not generate answers beyond the given content

prompts:" "

9.2. Evaluation of the LSTM Forecasting Model

To assess the forecasting model's ability to anticipate crypto token trends, the LSTM was tested on unseen time-series data. It was evaluated using standard regression metrics:

TABLE I: Evaluation metrics for LSTM model

Metric	Value
Mean Absolute Error (MAE)	2,117.38
Root Mean Squared Error (RMSE)	2,886.87
Mean Absolute Percentage Error (MAPE)	3.73%
R^2 Score	0.986

These results show that the model was able to predict price movements with high accuracy and minimal error. The strong R^2 score indicates that the model was able to explain most of the variation in the data, confirming its usefulness as an added layer of risk awareness.

10. Deployment

11. Conclusions

Acknowledgements

The authors would like to express their sincere gratitude to the academic faculty of the Department of Data Science at ESPRIT School of Engineering and Technology for their continued support, guidance, and mentorship throughout the course of this project.

We also extend our appreciation to VALUE Company for their valuable collaboration, real-world insights, and technical engagement, which played a key role in shaping the practical aspects of our solution.

This project was developed as part of the 4th-year Integrated Project in Data Science Engineering, and benefited from the combined contributions of all team members: Chaima Mejri, Maram Mejri, Mariam Maammar, Oussema Bargougui, Firas Bouguerra, and Houssam Kheder.

References

[1] X. Liu and Q. Wang, "A machine learning approach for evaluating cryptocurrency projects through ico whitepapers," *Expert Systems with Applications*, vol. 175, 2021.

[2] M. Zhao and L. Chen, "Entity recognition and risk classification in blockchain project documents," in *Proceedings of the 2020 International Conference on Blockchain and Trustworthy Systems*, 2020, pp. 123–134.