|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Graduation Project Thesis**  **To obtain**  **Engineering diploma**  **Computer Engineering**  **Class** **2024 – 2025**   |  | | --- | | **Optimization of Resumes Selection Process Using Generative AI and Predictive Analysis** |   A blue text on a black background  AI-generated content may be incorrect.  **Mr. /Ms. Mohamed Amin El Metni**  **Defended on:**   |  |  | | --- | --- | | **Jury Members:** |  | | **Mr./Ms. Sara Bayour**  **Mr./Ms. Yasser Mesmoudi**  **Mr./Ms.**  **Mr./Ms.** | **Company’s supervisor**  **ENSATé’s supervisor**  **Professor, ENSATé (Chair)**  **Professor, ENSATé (Examiner)** |   **June 25, 2025** |

A logo with text on it

AI-generated content may be incorrect.

**Academic year: 2024-2025**

# Dedication

**To my dear mother,**

My mother, who has worked tirelessly for my success through her love, support, countless sacrifices, and invaluable advice, for all her help and her constant presence in my life, please accept through this work, however modest it may be, the expression of my deep feelings and eternal gratitude.

**To my dear father,**

Who can be proud and see in this work the result of many years of sacrifice and hardship to help me move forward in life. May God allow this work to bear fruit. Thank you for the noble values, the education, and the unwavering support you have given me.

**To my dear brothers,**

Hamza, Yassir and Oussama,  
Who have always been examples of perseverance, courage, and generosity to me.

**To all my friends and family,**

Thank you for believing in me.

# Acknowledgements

It is both a pleasure and an honor to extend my heartfelt thanks to all those who have contributed, directly or indirectly, to the development of this project — through their support, guidance, or encouragement — and helped make this work possible.

I would like to express my sincere gratitude to my academic supervisor at the National School of Applied Sciences of Tetouan (ENSATé), Professor **Yassir Mesmoudi**, for his invaluable guidance, continuous support, and insightful advice throughout this project. His expertise and encouragement have been a constant source of motivation.

I am deeply thankful to **Sara Bayour**, my supervisor at NTT DATA Morocco, for offering me the opportunity to work on such an enriching and stimulating project. Her time, dedication, and guidance played a crucial role in shaping my experience and helping me grow both professionally and personally.

I would also like to express my appreciation to **Taha El Yakoubi** for his outstanding support in facilitating my integration into the NTT DATA Morocco team. His constant availability and kindness made my journey within the company much smoother and more enjoyable.

A heartfelt thank you goes to my project teammate, **Khadija El Kammas**, for being an exceptional collaborator. Her dedication, team spirit, and unwavering support were vital to the successful completion of this project.

I would also like to extend my gratitude to all the staff at NTT DATA Morocco for their support and generous willingness to share their knowledge and insights.

Finally, I sincerely thank the entire teaching staff of ENSATé for their commitment to our training, and for the time, attention, and energy they have invested in us, all within a pleasant environment of respect and mutual understanding.

# Résumé

Ce rapport présente le travail réalisé dans le cadre du projet de fin d'études effectué au sein de la société **NTT DATA Maroc**.

Ce projet de fin d'études vise à optimiser le processus de sélection de CV dans le cadre du recrutement en développant un système automatisé reposant sur l'intelligence artificielle générative et l'analyse prédictive. Les méthodes traditionnelles de recrutement nécessitent un travail manuel important de la part des ressources humaines, ce qui engendre des pertes de temps considérables et une baisse potentielle de la précision face au grand volume de candidatures à traiter. Le système proposé automatise l'extraction, la classification, l’évaluation et le classement des CV afin d’identifier les profils les plus adaptés selon les besoins de l’entreprise.

Ce travail repose sur des technologies avancées telles que les modèles transformeurs (Mistral pour l’extraction, LED pour la synthèse, et un BERT ajusté pour la catégorisation et l’évaluation des compétences). La méthodologie inclut l’analyse des CV via des modèles IA, l’enrichissement avec des projets GitHub, la catégorisation des profils, et l’évaluation des compétences selon le contexte professionnel et les expériences décrites. Des techniques prédictives telles que l’embedding sémantique, la similarité cosinus et un système de scoring pondéré assurent des recommandations pertinentes.

Développée avec Python, ReactJS, Spring Boot, MongoDB et PostgreSQL, la solution a démontré une nette amélioration des performances par rapport à la méthode manuelle, en fournissant des résultats rapides et fiables. Bien qu’aucune métrique formelle n’ait été calculée, les résultats indiquent une cohérence logique entre les scores attribués et les parcours des candidats. Ce projet met en lumière le potentiel de l’IA dans la transformation du processus de recrutement, tout en ouvrant la voie à des améliorations futures telles que l’intégration d’entretiens automatisés par chatbot et une modélisation plus fine du scoring.

# Abstract

This report summarizes the work carried out as part of the graduation project completed at **NTT DATA Morocco**.

This graduation project addresses the challenge of optimizing the resume selection process in recruitment by developing an automated system powered by Generative AI and predictive analytics. Classic hiring methods often require significant manual effort from HR personnel, leading to time inefficiencies and reduced accuracy when screening large volumes of applications. The proposed system automates the extraction, classification, scoring, and ranking of resumes to identify the best profiles based on organizational needs.

The solution leverages a combination of advanced technologies, including transformer-based models such as Mistral for data extraction, LED for summarization, and fine-tuned BERT for categorization and skill scoring. The methodology involves parsing resumes using AI models, enriching data with GitHub projects, categorizing resumes into relevant domains, and scoring individual skills based on contextual experience and project history. Predictive techniques such as semantic embedding, cosine similarity, and rule-based scoring ensure accurate and relevant profile recommendations.

Developed using Python, ReactJS, Spring Boot, MongoDB, and PostgreSQL, the system demonstrated significant performance improvements compared to manual methods, providing fast and reliable recommendations. Although no formal metrics have been computed, the results qualitatively confirm that the scoring aligns well with candidate experiences and qualifications. The project showcases how AI can revolutionize hiring by enhancing decision-making speed and accuracy while laying the groundwork for future improvements such as AI-driven interviews and enriched scoring models.

# List of Abbreviations

Table 1 : List of Abbreviations

|  |  |
| --- | --- |
| **Abbreviation** | **Designation** |
| **NTT** | Nippon Telegraph and Telephone |
| **ENSATé** | École Nationale des Sciences Appliquées de Tétouan |
| **AI** | Artificial Intelligence |
| **LLMs** | Large Language Models |
| **BERT** | Bidirectional Encoder Representations from Transformers |
| **PDF** | Portable Document Format |
| **OCR** | Optical Character Recognition |
| **JSON** | JavaScript Object Notation |
| **HR** | Human Resources |
| **LED** | Longformer-Encoder-Decoder |
| **DL** | Deep Learning |
| **ML** | Machine Learning |
| **IT** | Information Technology |
| **LLC** | Limited Liability Company |
| **SARL** | Société à Responsabilité Limitée |
| **MAD** | Moroccan Dirham |
| **S&P** | Standard & Poor’s |
| **SAP** | Systems, Applications & Products in Data Processing |
| **Pvt Ltd** | Private Limited Company |
| **GenAI** | Generative AI |
| **FY** | Fiscal Year |
| **ATS** | Applicant Tracking System |
| **ERP** | Enterprise Resource Planning |

# List of Figures

[Figure 1 : NTT DATA's Global Distribution of Employees by Region 20](#_Toc200527314)

[Figure 2 : Strategic Global Delivery Centers and Delivery Capabilities 21](#_Toc200527315)

[Figure 3 : NTT DATA’s Ecosystem Integration in the Automotive Industry 22](#_Toc200527316)

[Figure 4 : The Growth Timeline of NTT DATA Business Solutions 23](#_Toc200527317)

[Figure 5 : Organizational Structure of NTT DATA 24](#_Toc200527318)

[Figure 6 : Generative AI Applications Across Business Functions in NTT DATA 26](#_Toc200527319)

[Figure 7 : The Three Major Problems of Manual Resume Screening 28](#_Toc200527320)

[Figure 8 : Agile Method Steps 31](#_Toc200527321)

[Figure 9 : SCRUM Methodology 32](#_Toc200527322)

[Figure 10 : Gantt Chart 34](#_Toc200527323)

[Figure 11 : Key Differences Between Narrow AI and General AI 38](#_Toc200527324)

[Figure 12 : General Working Mechanism of AI Systems 39](#_Toc200527325)

[Figure 13 : Typical Lifecycle of an AI Project 40](#_Toc200527326)

[Figure 14 : Key NLP Roles in HireLens 43](#_Toc200527327)

[Figure 15 : High-level Overview of the HireLens Pipeline 48](#_Toc200527328)

# List of Tables

[Table 1 : List of Abbreviations 9](#_Toc200527375)

[Table 2 : Corporate Fact Sheet of NTT DATA Morocco, Tetouan 24](#_Toc200527376)

[Table 3 : Numbers of New Hires Through Fiscal Years 27](#_Toc200527377)

[Table 4 : NLP and LLM Role Distinction in HireLens 44](#_Toc200527378)

[Table 5 : Benefits of NLP/LLM Integration 45](#_Toc200527379)

[Table 6 : Challenges of NLP/LLM Integration 45](#_Toc200527380)

# Table of Contents

[Dedication 3](#_Toc200706495)

[Acknowledgements 5](#_Toc200706496)

[Résumé 6](#_Toc200706497)

[Abstract 7](#_Toc200706498)

[List of Abbreviations 9](#_Toc200706499)

[List of Figures 10](#_Toc200706500)

[List of Tables 11](#_Toc200706501)

[Table of Contents 12](#_Toc200706502)

[General Introduction 15](#_Toc200706503)

[Chapter 1: General Context of The Project 18](#_Toc200706504)

[1. Host Organization Overview 19](#_Toc200706505)

[1.1. Presentation of NTT DATA 19](#_Toc200706506)

[1.2. Geographic Presence 20](#_Toc200706507)

[1.3. Sectors of Activity 21](#_Toc200706508)

[1.4. NTT DATA Growth Story 22](#_Toc200706509)

[1.5. Organizational Structure of NTT DATA 23](#_Toc200706510)

[1.2. NTT DATA Morocco Profile 24](#_Toc200706511)

[1.6. NTT DATA Morocco Services 25](#_Toc200706512)

[1.7. The Evolution of Generative AI at NTT DATA 25](#_Toc200706513)

[2. Project Overview 26](#_Toc200706514)

[2.1. Background and Context 26](#_Toc200706515)

[2.2. Problem Statement 27](#_Toc200706516)

[2.3. Project Objectives 28](#_Toc200706517)

[2.4. Project Boundaries and Constraints 29](#_Toc200706518)

[2.5. Stakeholders and Users 29](#_Toc200706519)

[3. Project Management 30](#_Toc200706520)

[3.1. Introduction 30](#_Toc200706521)

[3.2. SCRUM Methodological Framework 31](#_Toc200706522)

[3.3. Internship Planning (Gantt Chart) 32](#_Toc200706523)

[Chapter 2: Background and System Methodology 36](#_Toc200706524)

[1. Background 37](#_Toc200706525)

[1.1. Artificial intelligence (AI) 37](#_Toc200706526)

[1.1.1. General Overview 37](#_Toc200706527)

[1.1.2. How it works 38](#_Toc200706528)

[1.1.3. Ai Project Lifecycle 39](#_Toc200706529)

[1.2. Analysis in the Context of *HireLens* 40](#_Toc200706530)

[1.2.1. Natural Language Processing (NLP) and Large Language Models (LLMs) 41](#_Toc200706531)

[1.2.2. NLP Foundations in *HireLens* 42](#_Toc200706532)

[1.2.3. Integration of Large Language Models (LLMs) in *HireLens* 43](#_Toc200706533)

[1.2.4. NLP vs LLMs: Roles and Synergies 44](#_Toc200706534)

[1.2.5. Benefits and Challenges of NLP/LLM Integration 45](#_Toc200706535)

[2. Workflow Methodology 46](#_Toc200706536)

[2.1. Data Extraction and Preprocessing 48](#_Toc200706537)

[2.3. Categorization Model Training and Evaluation Process 49](#_Toc200706538)

[2.3. Specialized Scoring System Components 50](#_Toc200706539)

[3. Explored Approach During the Research Phase 53](#_Toc200706540)

[3.1. Initially Considered Solutions 53](#_Toc200706541)

[3.2. Reasons for Rejecting This Approach 53](#_Toc200706542)

[3.3. Comparison with the Final Adopted Solution 54](#_Toc200706543)

[Chapter 3: Implementation 57](#_Toc200706544)

# General Introduction

In recent years, Artificial Intelligence (AI) has evolved from a promising innovation to a transformative force reshaping various industries. Within the domain of computer science, AI and its subfields, Machine Learning (ML), Deep Learning (DL), and more recently, Large Language Models (LLMs), have opened new frontiers of automation, efficiency, and intelligence. These technologies are becoming increasingly essential in automating complex, time-consuming processes that previously required significant human effort and expertise.

In the context of recruitment and human resource management, the classic manual screening of resumes poses a significant challenge, especially in environments dealing with high volumes of applications. This process is not only time-consuming but also prone to inconsistencies and bias. The need for optimization is evident, and AI offers a powerful solution: automating the resume selection process while maintaining high accuracy and efficiency.

This thesis is carried out as part of my end-of-studies internship to fulfill the requirements for obtaining the Computer Engineering diploma at the National School of Applied Sciences of Tetouan (ENSATé). The objective of this internship is to bridge the gap between academic learning and industrial practices by applying theoretical knowledge to solve a real-world problem within a professional environment.

The project aims to design and develop an AI-powered system capable of automatically extracting, analyzing, and scoring resumes based on company-specific criteria. By leveraging generative AI and predictive analysis, the system is intended to identify the most relevant candidates for a given role or shift, thus optimizing the selection process.

The scope of this project extends beyond simple resume parsing. It encompasses the full pipeline of recruitment automation—from data extraction and semantic classification to skill-based scoring and intelligent profile matching. Although the system is primarily developed within the IT sector, its design is flexible enough to be adapted to other industries.

The solution supports both internal use cases (employee shift or position changes) and external recruitment needs. It integrates various data sources including GitHub and LinkedIn, uses advanced AI models for classification and scoring, and offers a user-friendly interface to visualize and manage candidate profiles.

To provide a comprehensive understanding of the work accomplished, this thesis is structured as follows:

* The first chapter **“Internship Environment & Project Context”** presents the host organization, NTT DATA Morocco, and outlines the strategic relevance of the project, the problem addressed, and the management methodology adopted.
* The second chapter **“Scientific and Technical Foundations”** discusses the theoretical background of AI technologies, the methodology employed, and the different AI approaches explored during the development phase.
* The third chapter **“System Implementation and Application Development”** details the technical architecture, tools, and technologies used, along with the integration of AI models into a functional backend and frontend system.
* The fifth chapter **“Results, Testing, and Evaluation”** presents the performance metrics, user testing feedback, and evaluation of the system's accuracy, efficiency, and practical utility.
* The sixth chapter **“Innovation & Future Perspectives”** highlights the innovative elements of the solution, discusses its limitations, and proposes future enhancements to improve accuracy and scalability.
* **Lastly the “Conclusion”** summarizes the key accomplishments of the project, personal learnings, and the broader impact of the work on recruitment automation.

# Chapter 1: General Context of The Project

Introduction

In this chapter, I will present the environment in which my internship took place, along with the broader context of the project we developed. The aim is to provide a clear understanding of the professional setting, the motivations behind the project, and the methods we followed to ensure its successful implementation.

I will begin by introducing NTT DATA, the host organization, highlighting its global services, strategic partnerships, and its local presence in Morocco. Then, I will give an overview of the project, including the background, the core problem it addresses, its strategic importance, and the expected results. Finally, I will explain the project management approach adopted during the internship, including the tools used and the timeline we followed to structure our work.

This chapter sets the stage for the rest of the thesis by framing the project within its real-world context and showing how it aligns with both the company’s goals and my academic objectives.

## Host Organization Overview

### 1.1. Presentation of NTT DATA

NTT DATA is a global leader in IT services and consulting, headquartered in Tokyo, Japan. It is a part of the NTT Group, one of the largest telecommunications companies in the world. NTT DATA offers a wide range of services including system integration, application development, cloud services, digital transformation, and business process outsourcing. The company supports its clients in their efforts to modernize their information systems and embrace technological innovations, notably Artificial Intelligence, Big Data, and the Internet of Things.

With its customer-centric approach and strong emphasis on innovation and sustainability, NTT DATA contributes to building smarter and more resilient digital societies.

### 1.2. Geographic Presence

Operating in more than 50 countries, NTT DATA boasts a workforce of over 200,000 employees. This global reach allows the company to offer services tailored to local markets while maintaining a unified strategy across regions.

A map of the world with people and numbers

AI-generated content may be incorrect.

Figure 1 : NTT DATA's Global Distribution of Employees by Region

To operationalize its presence, NTT DATA has established several global delivery centers strategically located across continents, including North Africa (e.g., Morocco), Eastern Europe, India, and the Americas. These centers support 24/7 service delivery, digital transformation initiatives, and global project execution with high process maturity.

A map of the world with different colored labels

AI-generated content may be incorrect.

Figure 2 : Strategic Global Delivery Centers and Delivery Capabilities

### 1.3. Sectors of Activity

NTT DATA serves a wide array of industries through its diversified portfolio. The company is especially active in sectors such as:

* Banking & Financial Services
* Healthcare & Life Sciences
* Automotive & Manufacturing
* Public Sector & Government
* Energy & Utilities
* Retail & Consumer Goods
* Telecommunications & Media

A prominent example of its cross-sector capabilities is seen in its Ecosystem Integration approach within the Automotive Industry, where digital solutions are linked to domains like insurance, logistics, smart cities, and healthcare.

A diagram of an industry

AI-generated content may be incorrect.

Figure 3 : NTT DATA’s Ecosystem Integration in the Automotive Industry

### 1.4. NTT DATA Growth Story

Founded in 1989 as S&P Consult in Bielefeld, Germany, NTT DATA Business Solutions has grown into a global leader in digital transformation services, with a strong emphasis on SAP solutions. Over the years, the company has demonstrated steady growth through continuous innovation, international expansion, and strategic acquisitions, such as the integration into the NTT DATA Group in 2007 and the recent acquisition of ProvenTech Pvt Ltd in 2024. From its early focus on mid-market SAP clients to becoming a global digital powerhouse, NTT DATA Business Solutions has consistently evolved to meet the dynamic needs of businesses worldwide. The timeline below (Figure 4) visually summarizes the key milestones in this journey of transformation and global impact.

A road with numbers and words

AI-generated content may be incorrect.

Figure 4 : The Growth Timeline of NTT DATA Business Solutions

### 1.5. Organizational Structure of NTT DATA

The organizational structure of NTT DATA reflects a robust and multidimensional approach to corporate governance and technological innovation. With its headquarters in Japan, the company operates through a network of regional entities, specialized departments, and global innovation hubs. The structure is divided into corporate staff functions, cost centers focusing on technology and quality assurance, and executive leadership bodies including the Board of Directors and Corporate Management Committee. This alignment facilitates agility, strategic development, and global coordination — supporting NTT DATA’s mission of delivering cutting-edge digital solutions worldwide.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5 : Organizational Structure of NTT DATA

### 1.2. NTT DATA Morocco Profile

Table 2 : Corporate Fact Sheet of NTT DATA Morocco, Tetouan

|  |  |
| --- | --- |
| Corporate Fact sheet of NTT DATA Morocco, Tetouan | |
| Trade Register Number | 19687 |
| Legal business name | NTT DATA Morocco Centers |
| Year Established | 22 October 2017 |
| Court of Registration | Tetouan |
| Headquarters Address | Parc De Tetouanshore Route De Cabo Negro Commune De Martil, Tetouan, Maroc |
| Legal Structure | Limited Liability Company (LLC | SARL) |
| Primary Activity | IT Development |
| Capital | 7,700,000 MAD |

### 1.6. NTT DATA Morocco Services

NTT DATA Morocco delivers a wide array of advanced IT services that align with the global standards of the group. These include Application Services, Consulting, Cloud Solutions, and Cybersecurity, along with emerging areas such as Data & Artificial Intelligence (AI), CX and Digital Products, and Sustainability Services. The company also provides robust Network Services, Digital Workplace solutions, and Tech Solutions & Integration, supported by a global network of Data Centers and Edge Computing capabilities. This comprehensive portfolio enables NTT DATA Morocco to support end-to-end digital transformation initiatives for clients across various industries.

### 1.7. The Evolution of Generative AI at NTT DATA

NTT DATA has strategically embraced generative AI (GenAI) to drive innovation and enhance operational efficiency across its global operations. In June 2023, the company established the Global Generative AI Laboratory, aiming to integrate GenAI into all facets of software development, from requirement definition to testing, and to standardize its application across a workforce of approximately 195,000 employees. To support this initiative, NTT DATA introduced a Generative AI Talent Development Framework in October 2024, designed to train around 200,000 employees globally and develop 30,000 GenAI experts by fiscal year 2026.

In January 2025, NTT DATA launched the Smart AI Agent™, a next-generation AI tool aimed at accelerating GenAI adoption, with a projected revenue impact of $2 billion by 2027. This tool enhances operational efficiency by automating complex workflows and enabling multi-agent collaboration. These developments underscore NTT DATA's commitment to leveraging GenAI for transformative business solutions and maintaining a competitive edge in the rapidly evolving digital landscape.

A diagram of a company

AI-generated content may be incorrect.

Figure 6 : Generative AI Applications Across Business Functions in NTT DATA

## Project Overview

### 2.1. Background and Context

In the current competitive and fast-paced job market, organizations are under increasing pressure to attract, evaluate, and hire the appropriate talent quickly and accurately. Classic hiring techniques —often manual, biased, and time-consuming— are becoming increasingly insufficient for managing the growing complexity and amount of applicant data.

Given NTT DATA, a worldwide leader in IT services and digital transformation, where hiring operations have expanded dramatically in recent years, this difficulty is especially pertinent. Rising from 802 in FY2020 to 1,249 in FY2023, the overall number of hires has consistently climbed year after year, as the table below shows. Moreover, many current NTT DATA staff members often move between positions and departments, which adds even another level of difficulty to internal personnel management.

Table 3 : Numbers of New Hires Through Fiscal Years

|  |  |  |  |
| --- | --- | --- | --- |
| Fiscal Year | Newly Recruited Graduates | Mid-Career Hires | Total Hires |
| FY2020 | 477 | 325 | 802 |
| FY2021 | 511 | 314 | 825 |
| FY2022 | 548 | 507 | 1,055 |
| FY2023 | 674 | 565 | 1,239 |

We created "*HireLens*" in response to this increasing demand, a project meant to maximize internal mobility and recruiting procedures through Generative AI and predictive analytics. *HireLens* aims to simplify resume reviews, provide consistent, statistically based applicant rankings, and assist HR teams with smart suggestions thus quickening and enhancing decision-making.

This project fits exactly the strategy orientation of NTT DATA. The corporation is progressively funding Generative AI to enable value across many business divisions as part of its larger innovative path. The *HireLens* project not only represents this ambition but also provides a useful model of how artificial intelligence may transform HR operations within a big, active company such as NTT DATA.

### 2.2. Problem Statement

While the rapid growth of hiring at NTT DATA and the rise of Generative AI form a promising backdrop, the core challenge remains unchanged: the recruitment process itself is still burdened by inefficiencies.

Conventional hiring procedures mostly depend on subjective judgment and manual resume screening. Consequently, this strategy raises three major problems:

* **Slowness:** Recruiters labor for hours sorting hundreds of resumes, therefore slowing down the recruiting process and postponing important project staffing.
* **Bias:** Human screening brings unconscious bias that runs the danger of biased assessments and loses chances for competent applicants.
* **Lack of insight:** Manual processes rarely utilize historical hiring data or candidate performance trends, leaving decisions unsupported by meaningful analytics.

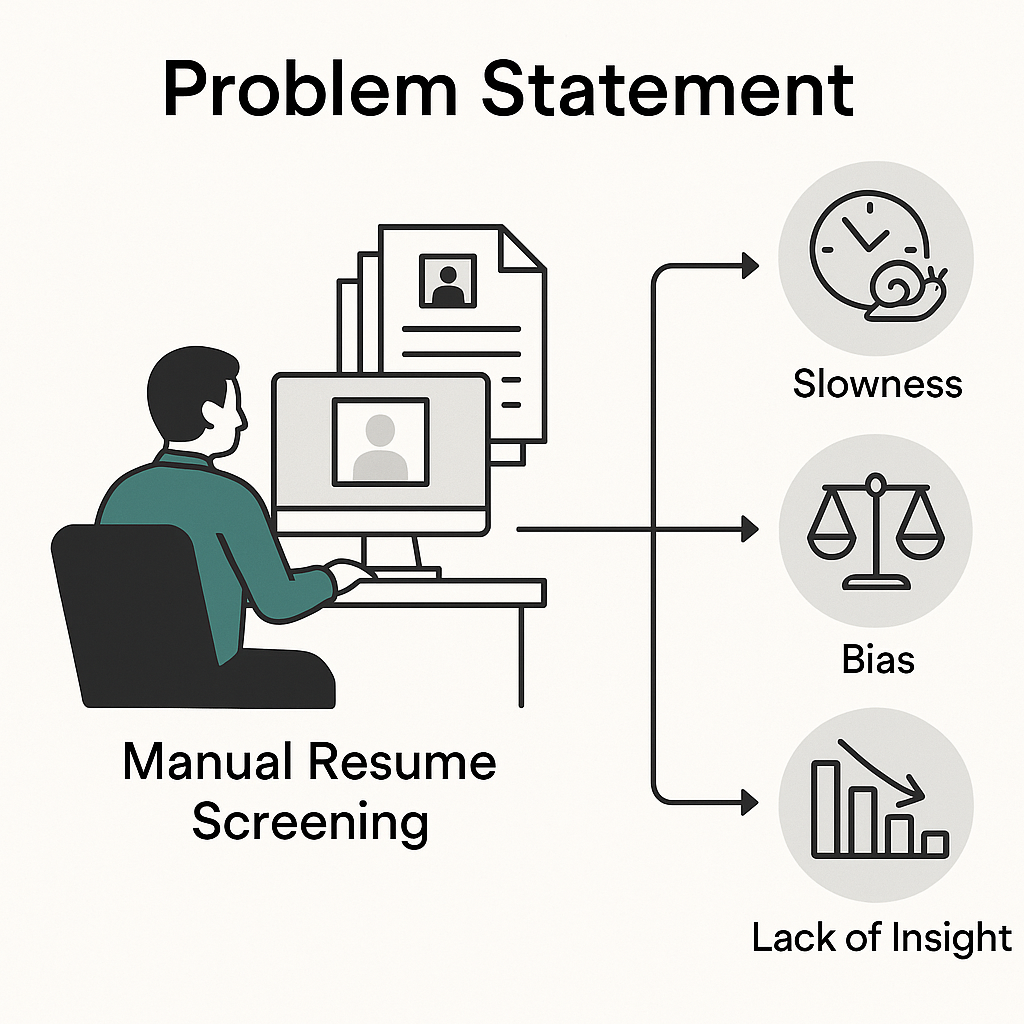


Figure 7 : The Three Major Problems of Manual Resume Screening

Companies therefore find it difficult to objectively assess, rate, and choose candidates, especially in view of rising application quantities and job complexity. This issue affects internal job mobility as well as external recruiting, since assessing current talent for new positions is often similarly unstructured. Dealing with these issues calls for a change toward data-driven, intelligent recruiting solutions that simplify assessment, lower prejudice, and support better decision-making precisely what *HireLens* seeks to do.

### 2.3. Project Objectives

The primary objective of this project is to develop an intelligent resume evaluation system, *HireLens*, designed to assist HR professionals in identifying the best-matching candidates for a given job role or internal shift. The system should analyze resumes and suggest top-fit profiles based on specific criteria such as required skills, languages, years of experience, and other job-related parameters defined by the user.

The solution aims to achieve two key technical objectives:

* **High scoring accuracy**: The system should deliver a skills scoring accuracy of over 90%, taking into account a comprehensive set of factors such as candidate experience, personal projects, certifications, and education. This ensures meaningful differentiation between applicants.
* **Efficient performance**: It should maintain acceptable processing times, enabling quick and responsive filtering and ranking of resumes even with large candidate pools.

In addition to these core goals, the project also emphasizes the importance of delivering a **user-friendly and secure interface** that supports the following features:

* Secure login with multi-user access.
* Advanced filters (e.g., skills, experience, resume categories).
* Resume sorting based on defined criteria.
* A prompt field allowing users to input job or shift descriptions for contextual matching.
* Functionality to like, comment on, and save resumes into personal collections.
* Viewing resume content and downloading original files.

*HireLens* aims to combine the power of AI with an intuitive interface to streamline candidate screening and enhance the quality of hiring decisions.

### 2.4. Project Boundaries and Constraints

While *HireLens* offers a powerful AI-driven approach to resume evaluation, its scope remains focused on the screening and ranking phase of the recruitment process. The system is designed to match candidate profiles with job or shift descriptions based on extracted textual information from resumes. It does not address the following areas:

* Post-hiring activities such as onboarding, performance evaluation, or employee development.
* Multilingual resume parsing beyond English and French (initial release).
* Integration with third-party Applicant Tracking Systems (ATS) or Enterprise Resource Planning (ERP) tools.
* Legal and ethical review of AI recommendations (currently outside project scope but to be considered in future iterations).

These constraints are defined to maintain a manageable and technically feasible project timeline, ensuring that core functionalities are robust and effective before future expansion.

### 2.5. Stakeholders and Users

The success of *HireLens* relies on addressing the needs of its various stakeholders, both direct and indirect. These include:

* **Primary Users:**
  + HR Managers and Recruiters: Direct users who initiate resume evaluations and use AI-generated rankings for decision-making.
  + Team Leads / Project Managers: May participate in defining role requirements and reviewing candidate lists for internal mobility.
* **Indirect Beneficiaries:**
  + Job Candidates: Internal or external applicants who benefit from a fairer and faster selection process.
  + NTT DATA Organization: Gains improved hiring efficiency, reduced time-to-hire, and data-driven HR insights.
* **Project Contributors:**
  + My colleague Khadija El Kamas and me (Interns/Developers): Responsible for system design, development, and maintenance.
  + Internship Supervisor Sara Bayour: Provide technical and strategic guidance throughout the project lifecycle.

This multi-stakeholder approach ensures that *HireLens* delivers both functional and strategic value, addressing operational needs while aligning with broader organizational goals in AI adoption and innovation.

## Project Management

### 3.1. Introduction

In the development of an IT solution, choosing an appropriate project management methodology is crucial to ensure the project's success. In an environment where requirements change rapidly and adaptability is a constant demand, we opted for an Agile approach, specifically the Scrum framework. Scrum is known for its ability to efficiently manage complex projects through iterative organization, strong stakeholder communication, and a focus on continuous improvement. This approach is widely used in the IT sector, especially at NTT DATA, as it enables the delivery of functional product increments at regular intervals, thereby facilitating frequent user feedback and better alignment of the product with actual needs.

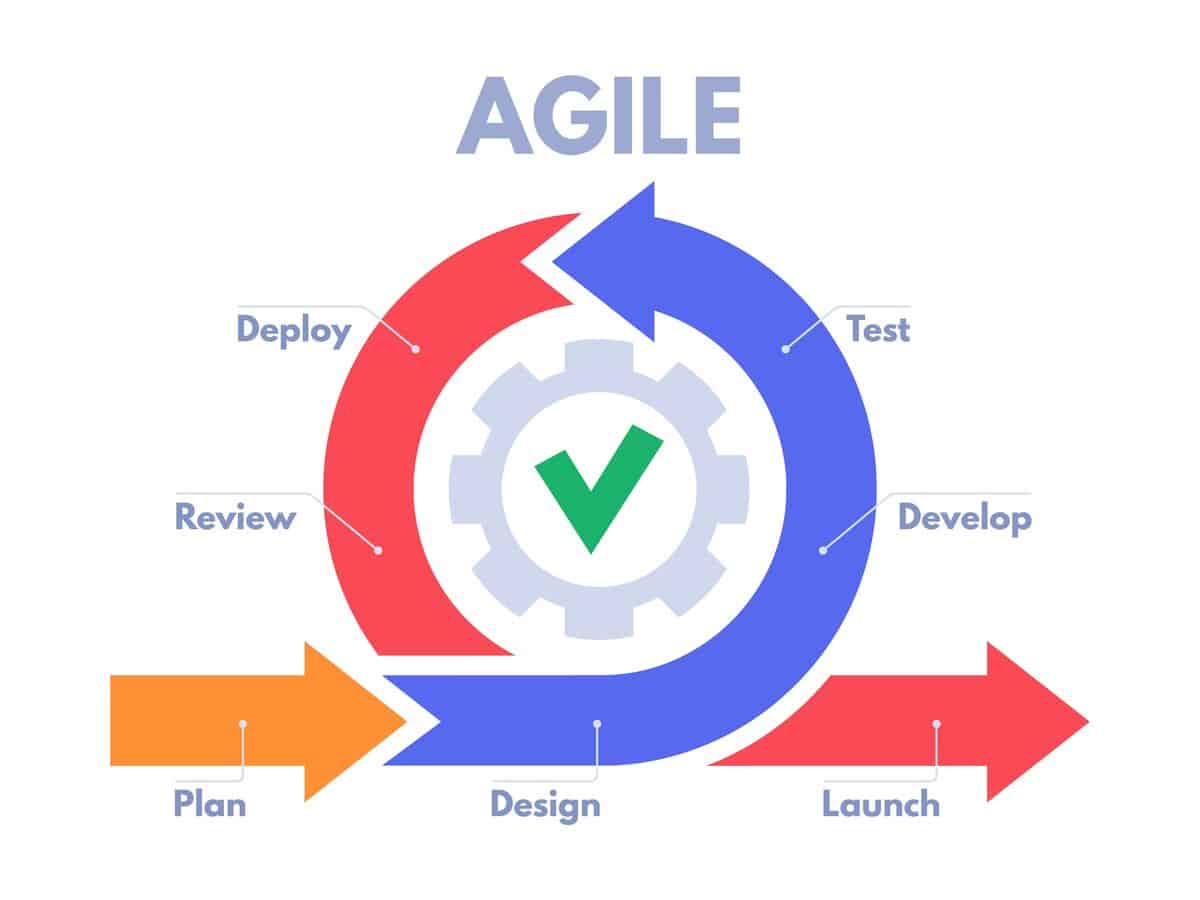


Figure 8 : Agile Method Steps

### 3.2. SCRUM Methodological Framework

Scrum is an Agile methodological framework designed to improve the management of complex projects by ensuring flexibility, transparency, and fast execution. It is based on a series of sprints, which are fixed-duration iterations (often 2 to 4 weeks) during which the team develops a set of defined features. Each sprint ends with a deliverable increment — a potentially usable version of the product. Scrum emphasizes high interaction among stakeholders, daily team communication, and regular assessment of results, allowing for quick adjustments and continuous steering. The method offers a high degree of autonomy to the team while structuring work organization to ensure goal achievement.

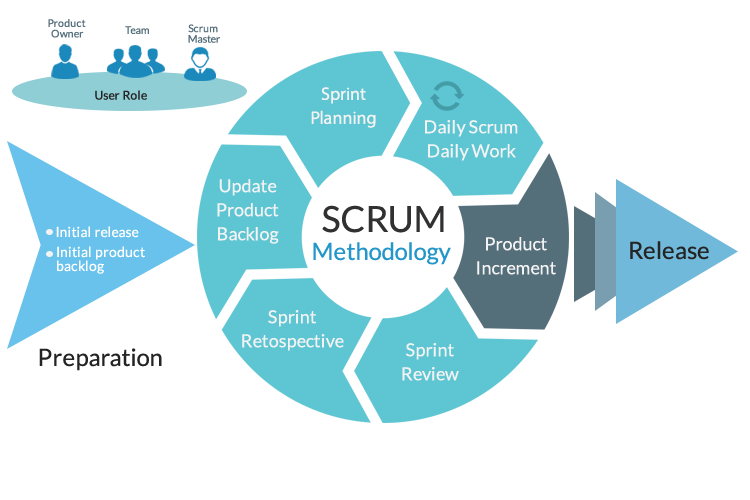


Figure 9 : SCRUM Methodology

### 3.3. Internship Planning (Gantt Chart)

To ensure effective time and task management during the internship, we created a Gantt chart. This chart outlines the various phases of the project, grouped by major development areas, along with the corresponding timelines. We structured *HireLens* into five key phases:

**a-** Resume Parsing & Analysis

We began by setting up the development environment and connecting to the necessary data sources. Then, we focused on extracting and structuring resume data using AI-driven techniques.

b- Resume Categorization

In this phase, we fine-tuned the AI model to automatically categorize resumes based on relevant job fields and skills.

c- Resume Scoring System Implementation

Next, we implemented the scoring system by developing classification models, setting up the database, and ensuring the system's stability. We also integrated similarity analysis features and visualized the results for better clarity.

d- Automation via API

We automated several processes, including the execution of “Jupyter” notebooks via API, skill and category extraction from resumes, and real-time change detection in the PostgreSQL database.

e- Application Development

Finally, we built the core application: this involved designing the PostgreSQL database tables, developing the backend logic through RESTful APIs, and creating a functional and user-friendly frontend.

The Gantt chart below shows the timeline of these tasks, highlighting how various stages overlapped and depended on each other.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 10 : Gantt Chart

Conclusion

In this first chapter, I provided an overview of the environment in which my internship took place, highlighting NTT DATA’s global presence, its investment in emerging technologies such as Generative AI, and how these align with the HireLens project.

I introduced the project by outlining its objectives, scope, stakeholders, and strategic relevance to the company. I also explained the use of the SCRUM methodology and tools like Jira to manage the project efficiently, and I concluded with a Gantt chart summarizing the internship timeline.

This chapter laid the foundation for understanding the professional context and project structure. In the next chapter, I will present the scientific and technical foundations that guided the development of HireLens.

# Chapter 2: Background and System Methodology

Introduction

The rapid evolution of artificial intelligence (AI) and its subfields has paved the way for highly intelligent and automated systems across various industries. In the context of talent acquisition and resume analysis, cutting-edge techniques such as machine learning, deep learning, natural language processing (NLP), and large language models (LLMs) offer transformative capabilities. This chapter explores the scientific and technical foundations that underpin the development of *HireLens*, our AI-driven platform for optimizing the candidate selection process.

The chapter begins with a comprehensive overview of artificial intelligence, emphasizing its lifecycle, fundamental principles, and distinctions between machine learning and deep learning. This theoretical background sets the stage for understanding the technologies leveraged in *HireLens*. Following this, we delve into the specialized domain of NLP and LLMs, key pillars of the system responsible for semantic resume analysis, categorization, and information extraction.

Subsequently, the chapter transitions into a detailed presentation of the system's workflow methodology. This includes data extraction from resumes, enrichment through GitHub integration, fine-tuning of categorization models, and the skill scoring mechanisms that form the backbone of *HireLens*' intelligence. The aim is to bridge foundational AI concepts with their concrete implementation within the project, demonstrating how theoretical knowledge translates into practical innovation.

## 1. Background

### 1.1. Artificial intelligence (AI)

#### 1.1.1. General Overview

Artificial Intelligence (AI) refers to the field of computer science that focuses on creating systems capable of performing tasks that typically require human intelligence. These tasks include reasoning, learning, problem-solving, perception, and language understanding. AI systems aim to simulate human cognitive processes to enable machines to make autonomous decisions, adapt to new inputs, and perform complex operations.

AI can be broadly categorized into two main types:

* **Narrow AI (Weak AI):** Designed to perform a specific task (e.g., email filtering, facial recognition). These systems operate under a limited set of constraints and are the most common form of AI today.
* **General AI (Strong AI):** A theoretical form of AI that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks—similar to human intelligence. This type of AI remains largely conceptual and is not yet realized in practice.

A comparison between a person and a robot

AI-generated content may be incorrect.

Figure 11 : Key Differences Between Narrow AI and General AI

#### 1.1.2. How it works

AI operates through a multi-layered process that involves learning patterns from data, making predictions, and ultimately performing intelligent tasks. These processes are often realized using various subfields of AI such as Machine Learning, Neural Networks, and Expert Systems.

A diagram of a machine learning

AI-generated content may be incorrect.

Figure 12 : General Working Mechanism of AI Systems

As illustrated in Figure 1, data is the fuel for AI systems. Once provided with input data, AI models learn patterns (through ML and DL techniques), apply reasoning or classification, and perform actions such as making recommendations or automating decisions. In the *HireLens* project, this concept translates directly into the automated screening of resumes, where the system learns to identify relevant qualifications, skills, and experience by analyzing textual data.

#### 1.1.3. Ai Project Lifecycle

Developing a functional AI-powered application involves several sequential and iterative steps, from data collection to deployment. The lifecycle of an AI project is outlined in this Figure 2.

A diagram of a lifecycle of an ai project

AI-generated content may be incorrect.

Figure 13 : Typical Lifecycle of an AI Project

This lifecycle maps closely to the development flow adopted in *HireLens*:

* **Data Collection & Preprocessing:** Thousands of resumes and job descriptions were collected and cleaned.
* **Feature Engineering:** Relevant features like education level, years of experience, added value in professional experiences, and keyword similarity were extracted.
* **Model Development & Evaluation:** Machine Learning and Deep Learning models were trained to predict candidate-job fit.
* **Deployment:** The final model was integrated into a RESTful API to serve predictions in real-time within the web interface.

This structure ensures the AI component in *HireLens* is not just a black box, but a well-integrated, maintainable, and explainable system aligned with real-world requirements.

### 1.2. Analysis in the Context of *HireLens*

In the development of *HireLens*, we primarily relied on deep learning techniques tailored to the nature of each task, while leveraging key machine learning concepts such as supervised learning. For company classification, we used a Long Short-Term Memory (LSTM) model trained on a dataset containing company names labeled with a binary indicator (0 or 1) denoting whether a company is a startup or not. This approach enabled the model to capture sequential patterns in textual data related to company names, facilitating accurate classification despite the relatively small and structured dataset.

For resume categorization, which involves unstructured and semantically rich natural language text, we fine-tuned transformer-based models like BERT. This enabled us to classify resumes into multiple relevant job categories based on deep contextual understanding of the text. While deep learning is a subset of machine learning, our approach specifically emphasized deep architectures to leverage their superior performance on complex language tasks.

Thus, *HireLens* integrates deep learning models that embody supervised learning principles, capitalizing on their strengths for different data types and problem complexities. This strategy maximizes both model performance and interpretability where applicable.

#### 1.2.1. Natural Language Processing (NLP) and Large Language Models (LLMs)

Natural Language Processing (NLP) and Large Language Models (LLMs) represent a specialized yet increasingly dominant branch within the broader fields of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). While AI encompasses the overall ambition of replicating human intelligence in machines, ML and DL focus on enabling systems to learn from data. NLP specifically addresses the challenge of enabling machines to understand, process, and generate human language—a critical component of intelligent behavior.

NLP began with rule-based and statistical approaches but has advanced significantly with the introduction of deep learning techniques and, more recently, the emergence of LLMs. These models, such as BERT, GPT, and Mixtral, leverage transformer architectures to model the complex dependencies and semantics in natural language. They allow machines not only to parse and extract information from unstructured text but also to reason, summarize, classify, and interact in a more human-like way.

In the context of the *HireLens* project, NLP and LLMs form a foundational layer of intelligence. From extracting structured information from resumes to classifying candidate profiles and summarizing GitHub repositories, the system relies on a combination of traditional NLP tools (e.g., regex, tokenization) and cutting-edge LLMs fine-tuned for domain-specific tasks. These components enable *HireLens* to automate and enhance the recruitment process with semantic understanding, contextual relevance, and scalable processing of large volumes of textual data.

#### 1.2.2. NLP Foundations in *HireLens*

Natural Language Processing (NLP) plays a foundational role in the *HireLens* platform, enabling the transformation of unstructured textual content such as resumes, GitHub repositories, and company names into structured, actionable insights.

At its core, NLP in *HireLens* serves five main goals:

* **Extracting and Cleaning Text**: Parsing resume documents and project data into clean, normalized text.
* **Understanding and Structuring Data**: Leveraging pre-trained language models to identify and organize key elements (e.g., skills, experience, education).
* **Text Classification**: Categorizing resumes and companies into relevant labels using machine learning and deep learning models.
* **Summarization and Representation**: Compressing lengthy project documentation into meaningful summaries.
* **Semantic Analysis**: Generating embeddings to assess similarity between skills, job descriptions, and candidate profiles.

These tasks blend traditional NLP methods (e.g., tokenization, regex-based preprocessing) with modern transformer-based techniques, reflecting the dual evolution of the field.



Figure 14 : Key NLP Roles in HireLens

Further implementation details and model-specific pipelines will be covered in the Methodology section.

#### 1.2.3. Integration of Large Language Models (LLMs) in *HireLens*

LLMs serve as the core intelligence layer for extracting meaning, generating structured data, and performing high-level reasoning over textual content. These models are particularly effective for understanding the nuanced language of resumes, job descriptions, and technical documentation, allowing the system to automate complex tasks such as:

* **Resume Structuring**: Parsing raw resume text into structured fields (skills, education, experience) via prompt-based generation.
* **Text Summarization**: Compressing lengthy project descriptions (e.g., GitHub READMEs) into concise summaries for rapid evaluation.
* **Multi-label Classification**: Categorizing resumes across multiple domains using fine-tuned transformer models.
* **Semantic Similarity**: Embedding skills and CV data into high-dimensional space to enable similarity scoring and recommendation.

These applications illustrate how LLMs are not simply passive tools, but active agents in decision-making and representation—bridging the gap between unstructured inputs and structured intelligence.

#### 1.2.4. NLP vs LLMs: Roles and Synergies

While often mentioned together, Natural Language Processing (NLP) and Large Language Models (LLMs) represent two layers of language technology, complementary yet distinct. NLP refers to the entire field of computational techniques used to process and analyze human language, ranging from basic text preprocessing to complex linguistic modeling. LLMs, on the other hand, are specific deep learning models within NLP that excel at understanding and generating text in a context-aware manner.

Table 4 : NLP and LLM Role Distinction in HireLens

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Traditional NLP** | **Large Language Models (LLMs)** |
| **Purpose** | Preprocessing, parsing, basic feature extraction | Contextual understanding, classification, semantic reasoning |
| **Examples in *HireLens*** | Text extraction from PDFs, regex-based cleaning, tokenization | Resume field generation (Mixtral), project summarization (LED), category prediction (BERT) |
| **Strengths** | Speed, simplicity, rule-based precision | Flexibility, generalization, handling ambiguity |
| **Limitations** | Limited semantic depth, hard-coded logic | Computational cost, reliance on high-quality prompts or fine-tuning |

The success of *HireLens* relies on the synergy between classical NLP and LLMs. Traditional techniques are used as foundational steps (e.g., extracting raw text, cleaning HTML, standardizing formats), ensuring that inputs fed into LLMs are clean and meaningful. Once processed, LLMs take over to perform high-level understanding and inference.

#### 1.2.5. Benefits and Challenges of NLP/LLM Integration

Integrating traditional Natural Language Processing (NLP) with Large Language Models (LLMs) offers significant advantages for intelligent systems like *HireLens*, but it also introduces practical challenges. Understanding this balance is essential for designing robust, scalable, and effective AI-driven applications.

a- Benefits of Integration

Table 5 : Benefits of NLP/LLM Integration

|  |  |  |
| --- | --- | --- |
| **Benefit** | **Description** | **Example in *HireLens*** |
| **Automation** | Reduces manual feature engineering and data structuring tasks. | Auto-conversion of unstructured resume text to structured JSON fields via LLMs. |
| **Semantic Richness** | LLMs capture deep contextual meaning beyond keyword matching. | Summarization of GitHub READMEs and resume categorization. |
| **Flexibility** | Easily adapt to various formats and styles of input text. | Handling diverse resume templates and job titles. |
| **Scalability** | Once trained or prompted well, models can process large volumes with consistency. | Batch processing of resumes for large recruitment databases. |

b- Challenges of Integration

Table 6 : Challenges of NLP/LLM Integration

|  |  |  |
| --- | --- | --- |
| **Challenge** | **Impact** | **How *HireLens* Addressed It** |
| **High Inference Cost** | Running LLMs can be computationally expensive. | Used quantized version of Mixtral-8x7B (4-bit) for resume extraction. |
| **Latency** | Longer processing time for large inputs. | Combined lightweight NLP preprocessing to reduce token count before LLM inference. |
| **Hallucinations** | LLMs may generate plausible but incorrect outputs. | Prompt engineering + post-processing checks (e.g., JSON schema validation). |
| **Prompt Sensitivity** | Model output can vary depending on prompt phrasing. | Developed and iteratively refined domain-specific prompts. |
| **Fine-tuning Complexity** | Requires labeled data and compute resources. | Fine-tuned BERT models specifically on resume categories and company types. |

## 2. Workflow Methodology

The *HireLens* system is designed as an end-to-end intelligent pipeline for automated resume understanding, categorization, and scoring. It leverages modern techniques from Natural Language Processing (NLP), Large Language Models (LLMs), and Deep Learning (DL) to transform unstructured resume documents into structured, enriched, and scored profiles.

At a high level, the system is composed of five sequential stages:

1. **Resume Extraction and Structuring**: Raw resume documents in PDF format are processed using a hybrid OCR and parsing approach to extract text. This text is then cleaned and structured using a prompt-engineered LLM, resulting in a normalized JSON representation stored in a MongoDB collection.
2. **GitHub Project Enrichment**: Candidate GitHub profiles are linked to resume entries, and their repositories are analyzed. NLP and LLM techniques summarize README files and extract programming languages, enriching the resume with concrete, verifiable projects.
3. **Resume Categorization**: A fine-tuned BERT model is used to assign multiple professional categories to each resume based on its content (skills, experience, and education). This enables recruiters to filter or match profiles more effectively to job domains.
4. **Company Classification**: Company entities mentioned in resumes are classified as startups or large enterprises using a sequence model, providing additional context for work experience relevance.
5. **Skill Scoring and Similarity**: Each resume is vectorized through a BERT-based skill embedding model, allowing the system to compute semantic similarity scores between resume skills and desired job skills. This enables ranking resumes based on technical relevance.

This modular workflow allows the system to be both extensible and scalable. Each component is loosely coupled, enabling independent improvement (e.g., replacing the summarization model or upgrading the classifier).

The overall system architecture after all can be structured around three main phases: extraction and preprocessing of data from various sources, training and evaluation of the categorization model, and finally the implementation of scoring algorithms with data integration.

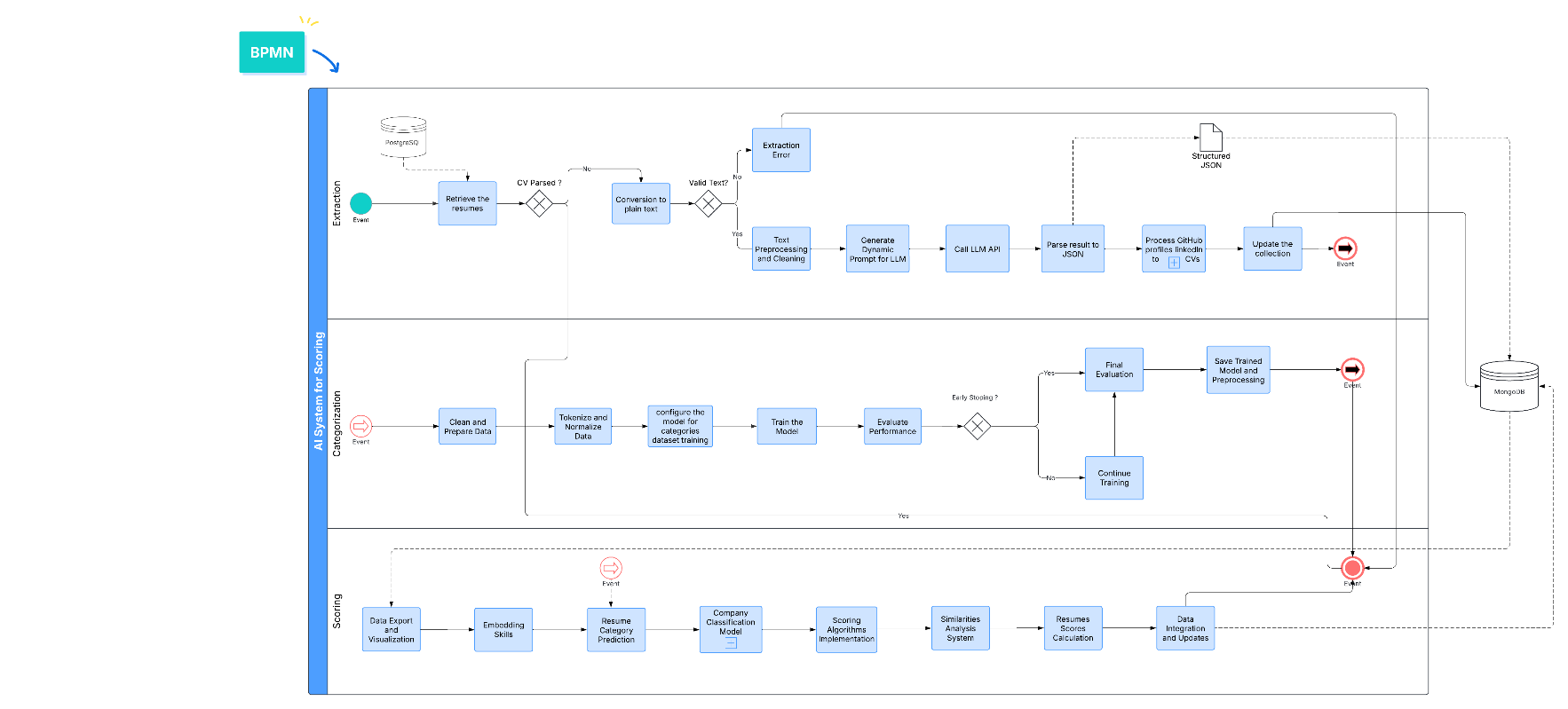


Figure 15 : BPMN for HireLens System

### 2.1. Data Extraction and Preprocessing

The process begins with the retrieval of resumes. The system performs preliminary verification to determine if the resume is already in an analyzable format “CV Parsed?”. Otherwise, automatic conversion to plain text is performed “Conversion to plain text” to normalize the data input.

Rigorous validation of the extracted text is then carried out “Valid Text?”. If the text is not valid or exploitable, the system triggers an extraction error stored in PostgreSQL that redirects correction or reprocessing mechanisms. This critical step ensures the quality of input data before processing by artificial intelligence algorithms.

Once validation is successful, the system proceeds with comprehensive text preprocessing and cleaning “Text Preprocessing and Cleaning”. This step includes removing unwanted characters, standardizing formats, and structuring extracted information.

The system then dynamically generates personalized prompts for a large language model “Generate Dynamic Prompt for LLM”. This dynamic generation allows adapting queries to the specificities of each resume and optimizing information extraction precision. The LLM API call is made with these optimized prompts “Call LLM API”, and results are systematically parsed to a structured JSON format “Parse result to JSON, Structured JSON Event” to facilitate subsequent processing.

* Subprocess: GitHub Links Extraction

The process intelligently integrates GitHub profiles associated with. This integration considerably enriches the data available for evaluation, enabling more comprehensive analysis of technical skills and professional experience of candidates.

A diagram of a diagram

AI-generated content may be incorrect.

Figure 16 : GitHub Projects Extraction Subprocess

The process begins with extracting the GitHub URL from CV or candidate profile data. Once the URL is identified, the system makes a call to the GitHub API “Call the GitHub API” to retrieve detailed information about the profile and associated projects. This retrieval includes repository metadata, contribution statistics, and information about technologies used.

The README content of GitHub projects is then cleaned “Clean the README Markdown content” to eliminate superfluous Markdown formatting while preserving essential information. Automatic README summarization is performed “Summarize the README” to extract key points and main technologies of each project.

After that, the system builds an organized structure of GitHub projects “Build a GitHub projects structure”, creating a hierarchical and exploitable representation of technical skills and development experience of the candidate.

Finally, the system automatically updates the data collection with this enriched information “Update the collection”, creating a complete candidate profile that combines classical resume data with dynamic one from professional platforms stored in MongoDB.

### 2.3. Categorization Model Training and Evaluation Process

The system proceeds with meticulous cleaning and preparation of data intended for model training “Clean and Prepare Data”. This phase includes tokenization and data normalization “Tokenize and Normalize Data”, essential steps for optimizing artificial intelligence model performance.

Model configuration adapts specifically to the categories dataset “configure the model for categories dataset”, allowing fine personalization of training parameters according to specific application domain needs.

Model training is performed in an iterative cycle “training” including continuous performance evaluation “Evaluate Performance”. The system implements an early stopping mechanism “Early Stopping?” that monitors performance metrics and interrupts training when improvement becomes marginal, thus avoiding overfitting.

If early stopping is not triggered, the system continues training by dynamically adjusting parameters to optimize results. Once training is completed, final evaluation is performed “Final Evaluation” and the trained model along with preprocessing parameters are saved for future use.

### 2.3. Specialized Scoring System Components

The system integrates several specialized components including skill embedding “Embedding Skills”, which transforms textual skills into vectorial representations to facilitate automated comparisons. CV category prediction “Resume Category Prediction” uses pretrained categorization model to automatically classify profiles according to predefined criteria.

* Subprocess: Company Classification Model

A diagram of a diagram

AI-generated content may be incorrect.

Figure 17 : Company Classification Model Subprocess

This dedicated subprocess develops a specialized model for company classification, allowing CV scoring adaptation according to specific organizational context.

The process begins with cleaning and preparing company data “Clean and Prepare Data”, including normalization of sectoral information, standardization of company sizes, and categorization of organization types. Tokenization and normalization of company data “Tokenize and Normalize Data” ensure consistency in information processing.

Model configuration adapts specifically to the company dataset “Configure the model for company dataset”, optimizing parameters to capture sectoral and organizational nuances. Model training “Train the Model” is performed with continuous validation (Validation) to ensure prediction robustness.

The early stopping mechanism “Early Stopping?” monitors validation metrics to avoid overfitting, while the training continuation option “Continue Training” allows fine optimization of performance. Once optimal training is achieved, final evaluation “Final Evaluation” and trained model saving “Save Trained Model” complete this specialized subprocess

The company classification model “Company Classification Model” allows adapting scoring according to organizational context, while the implementation of sophisticated scoring algorithms “Scoring Algorithms Implementation” ensures precise and objective evaluation of applications.

The similarity analysis system “Similarities Analysis System” compares candidate profiles with job requirements, generating precise compatibility scores. CV score calculation is performed by combining multiple criteria weighted according to the relative importance of each factor.

Data integration and updates ensure overall system consistency, while data export and visualization “Data Export and Visualization” allow end users to easily interpret results generated by the artificial intelligence system.

The integration of the main process and specialized subprocesses creates a comprehensive evaluation ecosystem where each component enriches the overall system capabilities. The GitHub extraction process provides detailed technical data, while the company classification model adapts evaluation criteria according to organizational context.

This modular approach allows advanced scoring personalization where different data sources (CV, GitHub) are combined with contextual company classification, thus optimizing evaluation relevance for each specific situation.

This BPMN architecture represents a complete and sophisticated automated CV evaluation system, integrating advanced artificial intelligence technologies with robust data processes. The modular approach with specialized subprocesses (GitHub links extraction and company classification model) ensures optimal flexibility and maintainability of the overall system, while providing precise and contextualized evaluation capabilities adapted to modern recruitment needs.

## 3. Explored Approach During the Research Phase

### 3.1. Initially Considered Solutions

During the research phase, we explored an alternative solution that was more interactive than the one ultimately adopted. The main idea was to create a personalized matching system in which the user enters their criteria from the frontend (e.g., desired skills, years of experience, target company type, etc.), and then an intelligent backend calculates relevance scores for each CV based on these criteria. This approach aimed at providing dynamic and personalized experience for recruiters.

The backend of this experimental solution combined several natural language processing (NLP) techniques:

* TF-IDF with n-grams (from 1 to 4 words) to capture key expressions.
* Semantic embeddings via Sentence-BERT to understand the context of skills.
* Detection of experience indicators (duration, etc.).
* Weighted scoring system based on the context of mention (experience, project, education, etc.).
* Summary generation via Gemini-type APIs to assist in quick CV reading.

The architecture included:

* A text preprocessing.
* A semantic matching engine.
* A dynamic weighting system based on user criteria.
* An automated CV summarization component.

### 3.2. Reasons for Rejecting This Approach

This approach, although functionally rich, was discarded due to several major limitations. First, the execution time turned out to be particularly long, mainly due to the use of BERT embeddings and deep contextual analysis, which require significant computing resources and greatly slow down the overall processing. Next, the results were sometimes inconsistent, due to difficulties in precisely calibrating the weightings between the various analysis methods, which made the generated scores difficult to interpret and therefore hard to exploit. Finally, the approach posed scalability issues: the massive storage of embeddings and the high latency when processing large volumes of CVs compromised the system’s performance in a production environment.

### 3.3. Comparison with the Final Adopted Solution

The final adopted solution stands in clear contrast to the initially explored approaches. While the latter relied on an interactive frontend interface involving user input, the current solution favors a fully automated backend process, ensuring greater fluidity and autonomy. From an NLP techniques standpoint, the early trials employed methods such as TF-IDF, Sentence-BERT, and heavy contextual analysis, whereas the adopted version relies on optimized contextual extraction with a more efficient weighting system. This evolution significantly reduced execution time: processes that were previously long due to embeddings and AI-generated summaries are now fast thanks to a lightweight, optimized model. In terms of maintenance, the complex architecture of the rejected approaches has been replaced by a modular and more maintainable solution. Additionally, dependency on external APIs and models has been greatly reduced, with the new solution being largely internal. Lastly, while the results of the earlier approaches were often variable and hard to interpret, the current solution delivers more stable, interpretable results that are well-suited for large-scale processing.

The adopted solution is based on a fully automated scoring mechanism on the backend, without requiring interaction through a frontend interface. This system evaluates CVs based on several key criteria: extracted skills using natural language processing (NLP) techniques, the duration and nature of professional experiences, the relevance of completed projects, online presence through platforms like GitHub, and the type of companies worked for, whether large corporations or startups. A global score is then calculated in a weighted manner, considering the recurrence of skills across different contexts such as experience, projects, education, and certifications. This score is further enriched by the notoriety of companies, assessed through a classifier, and adjusted using a rarity factor designed to reward specialized and less common technical skills.

Conclusion

This chapter provided a thorough examination of the theoretical and technical building blocks that enable the *HireLens* system. Through an exploration of artificial intelligence principles, including both machine learning and deep learning, we established the necessary groundwork for understanding how intelligent systems learn from data. We then explored the essential role of NLP and LLMs, highlighting their integration into the system for tasks such as resume classification, skill extraction, and semantic analysis.

The second half of the chapter detailed the step-by-step workflow that transforms raw resume data into structured, enriched, and actionable information. From data ingestion and GitHub enrichment to the fine-tuning of categorization models and advanced skill scoring techniques, each stage was designed to maximize precision, scalability, and relevance in candidate evaluation.

Together, these foundational concepts and methodologies form the technological core of *HireLens*, allowing it to intelligently process and evaluate resumes at scale. With this foundation laid, the next chapter will focus on the implementation details, user interface design, and system integration aspects that bring the solution to life in a production environment.

# Chapter 3: Implementation

Introduction

## 1. Technologies and Tools

This section presents the key technologies and tools selected for the development of the HireLens project. Each technology was carefully chosen to meet specific requirements such as performance, maintainability, scalability, and ease of collaboration. Together, they form modern, robust, and extensible.

### Development Environment

#### 1.1.1. Cloud Services:

* Google Colab

A logo with orange and yellow circles

AI-generated content may be incorrect.

Figure 18 : Google Colab's logo

Google Colaboratory, commonly known as Google Colab, is a free cloud-based service provided by Google that enables the execution of Jupyter Notebooks in a remote environment. It offers developers, researchers, and students the ability to run Python code in a preconfigured setting, without the need to install dependencies locally.

As part of the HireLens project, Colab played a central role during the data exploration, preprocessing, and model prototyping phases. Its ease of use allowed for a quick start, without the technical constraints of hardware setup.

* Kaggle

A blue text on a black background

AI-generated content may be incorrect.

Figure 19 : Kaggle's Logo

Kaggle, a Google-owned platform, is much more than a simple data science competition site — it serves as a large-scale experimentation lab. It provides a free, ready-to-use Jupyter Notebook environment with GPU and TPU support, a vast collection of public datasets, a dynamic community of experts, and infrastructure that ensures experiment reproducibility.

For our project, Kaggle was used as a platform for training and validating our models on real-world datasets. This approach allowed us to:

* Validate the robustness of our models using authentic and diverse datasets.
* Efficiently leverage rich public datasets covering various domains (NLP, computer vision, time series, etc.) without the need for manual downloads or configurations.
* Benefit from a scalable infrastructure for fast and reproducible experimentation.

#### 1.1.2. Local Development Tools:

* PyCharm

A logo with colorful squares

AI-generated content may be incorrect.

Figure 20 : Pycharm's Logo

PyCharm, developed by JetBrains, served as one of the primary Integrated Development Environments (IDEs) for backend development in our project, particularly for Python-based services such as FastAPI applications, data processing scripts, and automation workflows. Designed specifically for Python, PyCharm provides a robust, scalable, and feature-rich environment ideal for building medium- to large-scale software systems.

Due to its rich plugin ecosystem, intelligent refactoring tools, and advanced Python-specific optimizations, PyCharm played a central role in delivering a fast, maintainable, and scalable backend codebase.

* Visual Studio Code

A blue ribbon with a cross

AI-generated content may be incorrect.

Figure 21 : VS Code Logo

Visual Studio Code (VS Code), developed by Microsoft, served as the primary code editor for frontend development, particularly with React and Tailwind CSS. Its modular design and lightweight architecture made it an ideal tool for maintaining high productivity across diverse development contexts. VS Code’s flexibility was further enhanced through a rich ecosystem of extensions, including support for React, Prettier, ESLint, Tailwind CSS IntelliSense, Docker, and GitLens.

#### 1.1.3. Project Management Tools:

* GitHub

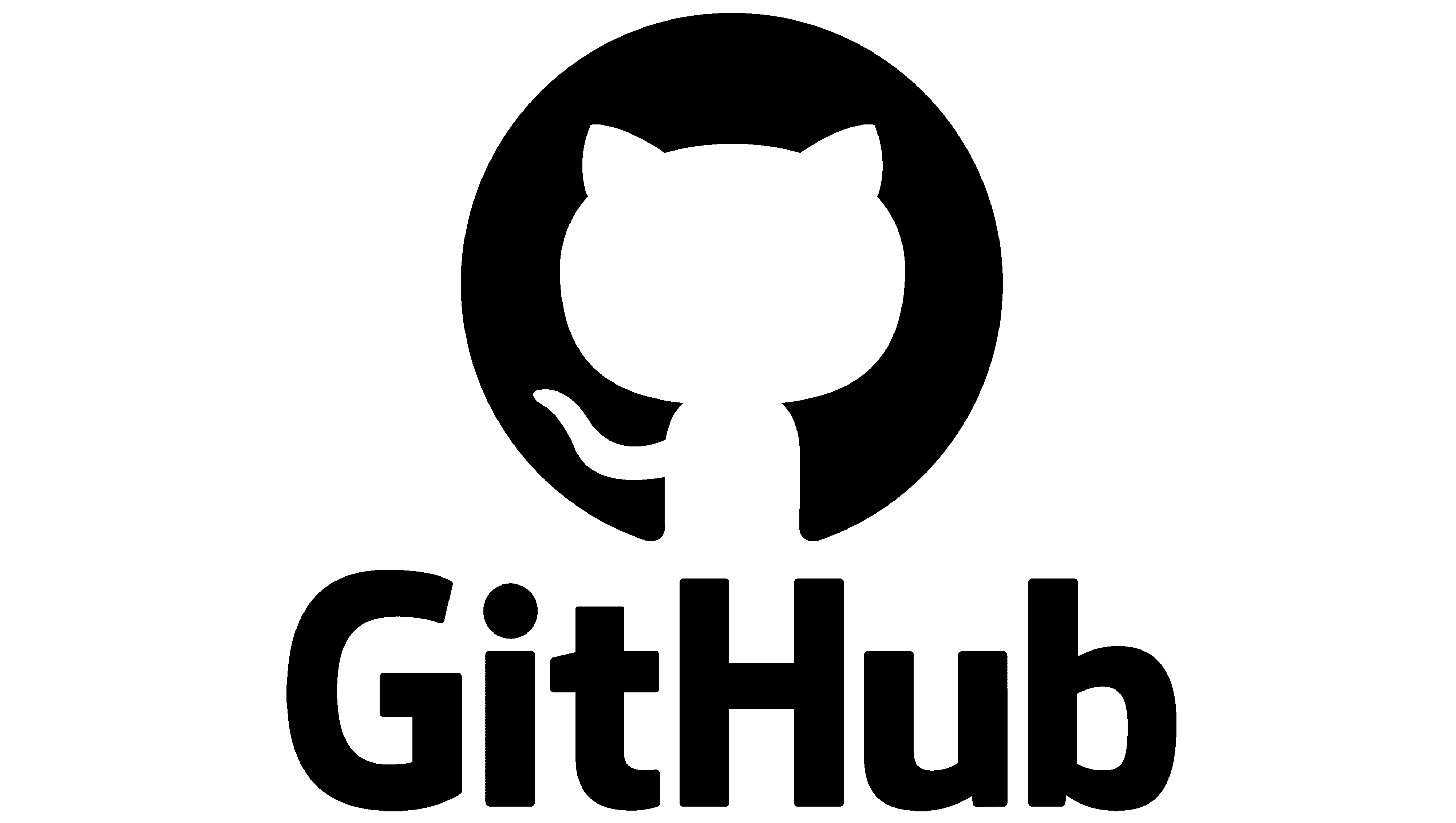


Figure 22 : GitHub's Logo

GitHub played a central role throughout the entire lifecycle of the project. Far more than just a code hosting platform, it served as a foundational and indispensable tool at every stage of development. GitHub not only enabled reliable source code management but also facilitated seamless collaboration, automated testing workflows, and the maintenance of dynamic, accessible documentation.

Thanks to its Git-based version control system, every code change was traceable, well-documented, reviewed, and cleanly integrated into the project. This versioning mechanism ensured full traceability, rigorous evolution tracking, and effective conflict resolution, an essential feature in collaborative development environments.

* Jira

A blue and black logo

AI-generated content may be incorrect.

Figure 23 : Jira's Logo

Jira is a comprehensive project management platform developed by Atlassian, widely used in software development and beyond. It is especially renowned for its support of Agile methodologies such as Scrum and Kanban. Jira enables effective planning, tracking, and management of all phases in a project’s lifecycle, from initial design to final delivery.

Thanks to its flexibility, Jira can be customized to meet the specific needs of each team, offering advanced features such as creating custom workflows, managing dependencies, automatic notifications, and detailed reporting. Its intuitive web interface promotes collaboration among team members, transparency, and individual accountability.

In our project, *HireLens*, we adopted a Scrum approach, and Jira proved to be an indispensable ally to structure and guide our teamwork.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 24 : Provided Epics in Jira Environment

Here is a detailed description of how we used it:

* **Product Backlog Management:**

The Product Owner uses Jira to list all features to be developed as user stories that describe business needs. Each user story is enriched with a clear description, acceptance criteria, and an estimation in complexity points. Stories are prioritized, providing a clear vision of the most urgent or high-impact items.

* **Sprint Organization via Sprint Backlog:**

Before each sprint, the team holds a planning meeting (Sprint Planning) during which the highest-priority user stories from the Product Backlog are selected for implementation. These user stories are then broken down into technical tasks, estimated, and added to the Sprint Backlog in Jira, with a deadline corresponding to the sprint duration (usually two to four weeks).

* **Daily Tracking with the Scrum Board:**

Jira’s Scrum Board displays tasks as cards moved across columns ("To Do", "In Progress", "Done"). Each team member moves the tasks they are working on, enabling real-time visibility of progress. This transparency facilitates coordination, rapid identification of blockers, and workload redistribution if necessary.

* **Assignment and Collaboration:**

Tasks are assigned individually, and each member can add comments, attach files, or create subtasks to organize work in detail. Jira sends automatic notifications to alert relevant members about changes or new assignments, thus enhancing communication.

* **Incident and Improvement Management:**

Alongside developing new features, Jira also serves as a platform to report and track bugs, incidents, and improvement requests. Each issue is logged with a priority level, detailed description, and evolving status until resolution.

* **Progress Analysis Using the Burndown Chart:**

The Burndown Chart in Jira shows the evolution of the amount of work remaining during the sprint, typically expressed in effort points. This chart helps the Scrum Master evaluate the team’s work pace, identify potential delays, and take corrective actions such as reprioritizing tasks or reallocating resources.

* **Facilitation of Agile Ceremonies:**

During daily meetings (Daily Scrum), team members consult Jira to quickly review their tasks and report obstacles. Sprint reviews (Sprint Review) rely on Jira to present delivered items. Finally, during retrospectives, the team analyzes data collected in Jira to suggest process improvements.

Each user story described below was created in Jira with a unique identifier, a clear title, and detailed acceptance criteria. For each story, related technical tasks were added as sub-items to be completed during the sprint. These tasks were estimated in story points, assigned to team members, and tracked daily through the Scrum Board. This clear structure helped ensure smooth sprint execution and full traceability from business needs to implementation.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 25 : Jira Stories and Tasks

### 1.2. Development Technologies

#### 1.2.1. Backend

* Python

A black and grey text

AI-generated content may be incorrect.

Figure 26 : Python Logo

Python is a high-level interpreted programming language known for its clear syntax, versatility, and extensive standard library. It is widely used in many fields, including web development, automation, data analysis, and scientific computing.

The Python language was specifically chosen for implementing the Artificial Intelligence (AI) part of the project. Thanks to its simplicity, flexibility, and vast ecosystem of scientific and machine learning libraries, Python has become the essential reference in the field of machine learning and data analysis.

In our project, Python was not used merely as a general-purpose programming language but as the central tool for building, training, evaluating, and integrating artificial intelligence models into the overall architecture.

* FastAPI

A close up of a logo

AI-generated content may be incorrect.

Figure 27 : FastAPI's Logo

The FastAPI framework was chosen to develop the project's backend API, enabling fast, smooth, and efficient integration between the artificial intelligence component (developed in Python) and other system components, particularly the user interface.

Designed specifically for modern applications, FastAPI is built on Python 3.7+, using Pydantic for data validation and Starlette as the asynchronous web server.

FastAPI was selected for this project due to its modern features, speed, and ease of integration with machine learning components. The following table highlights the key advantages that make FastAPI an excellent choice for developing AI-powered backend services.

* Spring Boot

A green logo with black background

AI-generated content may be incorrect.

Figure 28 : Spring Boot's Logo

Spring Boot is an open-source framework based on Spring, designed to simplify the creation of standalone, productive, and production-ready Java applications. It facilitates automatic configuration, quick startup, and integration of many essential modules (security, databases, web services) while significantly reducing the amount of code required. Thanks to its opinionated approach, Spring Boot allows developers to focus on business logic without getting lost in configuration details.

In the context of the HireLens project, Spring Boot was used to build certain critical modules of the web application, especially those requiring a robust architecture, fine-grained user access management, and reliable interaction with the PostgreSQL relational database. This solution proved ideal for providing a solid and secure backend foundation while enabling gradual scalability.

#### 1.2.2. Databases

In the development of our application, the choice of databases was a key factor in meeting the specific requirements for data management and structuring. We opted for a hybrid approach using both MongoDB, a document-oriented NoSQL database, and PostgreSQL, a powerful and robust relational database. This combination allows us to leverage the complementary strengths of each technology.

* MongoDB

A green leaf with brown text

AI-generated content may be incorrect.

Figure 29 : MongoDB's Logo

MongoDB is a widely adopted NoSQL database designed to store data in flexible, JSON-like documents called BSON (Binary JSON). This document-oriented approach allows for dynamic schemas, making it especially well-suited for handling semi-structured and rapidly evolving data. Its flexible data model supports complex nested structures and arrays, enabling developers to represent real-world entities more naturally than traditional relational databases. MongoDB is optimized for high-performance read and write operations, making it ideal for applications that demand low latency and fast throughput.

In HireLens, MongoDB is used to store unstructured or semi-structured data, specifically AI processing results and logs. This allows efficient handling and quick access to the varied data generated during AI model execution and monitoring.

* PosgreSQL

A blue elephant logo with black text

AI-generated content may be incorrect.

Figure 30 : PostgreSQL's Logo

PostgreSQL is a widely acclaimed open-source relational database known for its robustness, full compliance with SQL standards, and a rich set of advanced features. It supports complex transactions, data integrity through constraints, and allows the creation of custom data types to suit diverse application needs. Its extensibility and strong support for concurrency make it ideal for handling large-scale, mission-critical applications requiring reliable and consistent data storage.

In the HireLens project, PostgreSQL is the main database used to store structured data such as candidate profiles, CVs, and administrator information. It also handles the management of CVs and supports the tracking of various application features, ensuring reliable and consistent data organization throughout the system.

#### 1.2.3. Frontend

* ReactJS

A blue and black logo

AI-generated content may be incorrect.

Figure 31 : React's Logo

ReactJS is a powerful JavaScript library developed by Facebook, designed to build dynamic and high-performance user interfaces. It has quickly become a preferred choice for modern frontend development due to its component-based architecture, which promotes reusability and modularity. React’s virtual DOM enables efficient rendering and fast user interactions, enhancing the overall responsiveness of web applications. Additionally, its large and active community provides extensive resources, tools, and third-party libraries that accelerate development and foster innovation.

This combination makes ReactJS ideal for creating scalable, maintainable, and engaging user experiences.

* Tailwind CSS

A blue waves on a black background

AI-generated content may be incorrect.

Figure 32 : Tailwind's Logo

To complement React, we integrated Tailwind CSS, a utility-first CSS framework that enables rapid and efficient development of elegant, responsive user interfaces. Its rich set of predefined utility classes allows for seamless styling directly within component structures, promoting both design consistency and faster UI prototyping without writing custom CSS.

The combination of these tools and technologies results in complete, modern, and scalable architecture. Each technological choice is driven by well-defined project requirements including performance, maintainability, collaboration, and scalability.