National College of Ireland

Using Deep Learning and Transformer Models to Identify Inconsistencies Between Interview Responses and Resume Claims

MSc Research Project
M.Sc. Data Analytics

Jareen Sayma Haque
Student ID: X23383593

School of Computing
National College of Ireland

Supervisor: Abid Yaqoob

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jareen Sayma Haque
Student ID:	X23383593
Programme:	MSc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Abid Yaqoob
Submission Due Date:	11th August, 2025
Project Title:	Using deep Learning and Transformer Models To Identify Inconsistencies Between Interview Responses and Resume Claims
Word Count:	6020
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Jareen Sayma Haque
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using Deep Learning and Transformer Models to Identify Inconsistencies Between Interview Responses and Resume Claims

Jareen Sayma Haque
Student ID: X23383593

Abstract

In modern recruitment, a large "trust gap" exists between claims made on resumes and statements given in interviews, an issue cannot be addressed with mere human checks or manual verification alone. The aim of this work is to fill the gap by developing, applying, and validating a novel deep learning framework to automatically detect semantic inconsistencies between these two sources. The core objective was to determine the extent to which transformer-based models could accurately perform this task, comparing three distinct architectures which are Bi-Encoder, Cross-Encoder, and a T5 based Natural Language Inference (NLI) model on a custom-built dataset of real-world humane and synthetically generated resume-interview pairs. The findings indicates that utilizing both contexts simultaneously is far superior, with the fine-tuned Cross-Encoder model significantly outperformed the other architectures, achieving 73.1% accuracy and a balanced F1-score of 72.7%. This paper establishes the first empirical benchmark for the novel work of resume-to-interview consistency detection presenting a proven blueprint for a new class of smart HR tools designed to support evidence-based hiring. This work contributes to fairer, transparent recruitment practices by developing scalable verification mechanism to aid human judgment rather than replacing it.

Keywords: Deception Detection, Transformer Models,Sentiment Analysis, Recruitment Efficiency, NLI, Transformers, Bias Mitigation, HR

1 Introduction

1.1 Research Background and Motivation

In this current era of AI in recruitment, organizations are armed with powerful tools for processing candidate information at an unprecedented scale. This technological change is a huge step forward for efficiency, but it has also opened us up to a critical vulnerability: a big "trust gap" between the claims presented on a static resume and the statements made in a dynamic interview. Weiss and Feldman (2006) studies have indicated that a significant percentage of candidates embellish qualifications on their resumes, yet the disconnected nature of modern recruitment platforms prevents scalable verification of these claims against interview data.The sheer volume of digital applications has made manual, line-by-line verification impossible. For which, key hiring decisions are often made based on unverified, potentially inconsistent information. This research was chosen to mitigate this urgent need. Because this gap is not simply an operational inefficiency,

it is a fundamental threat to the integrity of fair and meritocratic hiring practices. The motivation for this work stems from the firm belief that for AI to be a truly beneficial tool in recruitment, it must not only automate processes, but also enhance accountability and truthfulness. Although existing AI solutions are capable of scanning CVs and conducting initial interviews, they function as isolated systems, leaving a crucial bridge between them unsolved. This thesis directly confronts this gap. The core motivation is to design and validate a new class of intelligent systems capable of performing this cross-verification, changing the standard from simple data processing to sophisticated, evidence-based validation. Such a system is much needed to enable organizations to make faster, better, and most importantly, fairer hiring decisions.

1.2 Aim, Objectives and Research Questions

The principal aim of this research is to architect, implement, and validate a novel deep learning framework to automatically and accurately detect semantic inconsistencies between candidate resumes and their transcribed interview responses. To achieve this aim, the following objectives were established:

- To critically investigate the state-of-the-art literature in resume parsing, deception detection, and fairness in AI to identify the specific methodological and ethical gaps in existing research.
- To design and develop a novel framework for cross-document consistency checking, implementing multiple transformer-based architectures.
- To evaluate the performance of the developed models using a custom-built dataset composed of real-world and synthetically generated resume-interview pairs, measuring performance with standard metrics including accuracy, precision, recall, and F1-score.
- To analyze the results to determine the most effective architecture for this specific task and contribute the first empirical benchmark and a validated methodological blueprint to the field of AI-driven recruitment.

This rigorous process was designed to answer a main research question, supported by specific sub-questions. The central inquiry of the project is,
To what extent can deep learning models accurately detect semantic inconsistencies between claims made in a resume and statements given in a transcribed interview? This was further broken down to explore how different transformer architectures compare in performance in different evaluations metrics and what types of inconsistencies the models are most effective at identifying. Also, this study was guided by the central hypothesis that a Cross-Encoder architecture would achieve a significantly higher F1-score and accuracy compared to a Bi-Encoder model to perform joint-contextual analysis of the text pair.

1.3 Research validation, Contribution and Significance

The outcome of this research provides a definitive and positive answer to the research questions, demonstrating that the project objectives were not only met but exceeded.This project successfully demonstrates that a finely-tuned Cross-Encoder model can attain

high performance (73% accuracy, 72.7% F1-score) in this novel task validating the central hypothesis and establishing a powerful proof-of-concept. The contribution of this work in this field is highly significant and multi-faceted:

- **Technical Contribution** It delivers the first empirical benchmark for the novel task of resume-to-interview consistency detection. While prior work has focused on matching resumes to job descriptions, our work establishes a crucial baseline for the more complex challenge of verifying claims against unstructured interview transcripts.
- **Practical Contribution** It offers a true blueprint for a scalable, efficient, and privacy-preserving tool that can be integrated into next-generation Human Resources technology platforms.
- **Ethical Contribution** This work promotes a crucial shift towards evidence-based hiring also creating a mechanism to enhance transparency and accountability in a field often subject to unconscious bias.

The finding of this study have significant implications for both academic research and the practical development of more responsible and robust AI systems for human resources. Beyond technical innovation, the impact of this work directly aligns with several of the United Nations Sustainable Development Goals (SDGs). By creating a system that prioritizes verifiable facts over subjective impressions, this work contributes to **SDG 10 (Reduced Inequalities)** by helping to mitigate the influence of demographic and affinity biases in hiring. It supports **SDG 8(Decent Work and Economic Growth)** by fostering more effective, meritocratic employment matching, ensuring candidates are selected based on their true qualifications. Finally, it embodies the spirit of **SDG 9 (Industry, Innovation, and Infrastructure)** as a cutting-edge application of AI for social and economic good by contributing to the development of smarter and fairer infrastructure for the global workforce.

This report is structured as follows, Section 2 provides a critical review of the state-of-the-art literature in resume parsing, fairness in AI, deception detection, and transformer models. Section 3 details the research methodology, justifying the choice of quantitative approach and models. Section 4 presents the design specification and architecture of the developed models. Section 5 describes the implementation of the solution, followed by a comprehensive evaluation and analysis of the experimental results in Section 6. Finally, Section 7 concludes the report, discussing the findings, acknowledging limitations, and proposing future research directions.

2 Related Work

Recruitment increasingly demands systems that can assess semantic consistency between resumes and interview responses. This review evaluates research in resume parsing, deception detection, and transformer modeling, highlighting how existing literature informs or limits current approaches. The core question is whether deep learning and transformer models can accurately detect inconsistencies between interview answers and resume claims for robust truthfulness assessment in AI-driven hiring.

2.1 Automated Resume Parsing: Technical Evolution and Challenges

The computational processing of resumes forms a critical initial stage in any system aiming to verify candidate claims, since the accuracy of this initial data extraction limits the overall effectiveness of downstream assessments. Early research in resume parsing primarily used rule-based systems and statistical methods like Conditional Random Fields (CRF) for information extraction (circa 2015-2018)Chakraborty et al. (2017) Gupta and Batra (2016). While pioneering, these methods were often brittle, struggling with the diverse layouts and linguistic variations common in resumes. Recent foundational research, such as by Jivtode et al. (2023), and Nikam et al. (2024), has attempted to improve these natural language processing and machine learning methods. However, the ongoing technical issue, as clearly stated by Deepa et al. (2024), is that these systems still struggle with the high degree of variability in resume formatting and the vagaries of natural language. To overcome these technical challenges, recent work has adopted methodologies related to deep learning. For instance, Selvi.S et al. (2024)) proposed "DeepResume," which utilized a hybrid method combining a transformer, CNN, and RNN model-based learning to successfully harvest data in a more robust and flexible way.However, the supremacy of deep learning is not absolute. In a groundbreaking comparative study, Wang and Zhu (2022) showed that while deep learning is promising, competitive performance is still possible using the widely established CRF model for smaller, well-structured datasets. This suggests that the choice of parsing model is a complex technical trade-off needed to understand context as a function of dependencies in the data.

2.2 Fairness and Bias in AI-Driven Recruitment

Besides the technical aspects relating to precision and responsiveness, a more vial challenge facing the automated recruitment industry is systemic bias and fairness. Two recent studies highlight this challenge from different but complementary angles. Utilizing a systematic review methodology, Mujtaba and Mahapatra (2024) outline the terrain of fairness relating to AI-recruitment practices. Their main finding shows that concerns regarding fairness are essentially systemic, marked by serious challenges arising from biases inherent in past training data, discriminatory behavior learned by algorithms, and a general absence of transparency in AI-aided decision-making systems. Stuss and Fularski (2024), find a significant gap between the potential benefits of AI and its actual use, driven by managers' ethical concerns and a lack of trust. Diverse inquiries into AI's role reveal a multi-faceted gap. Mujtaba and Mahapatra (2024)'s technical analysis of parsing systems highlights how even accurate tools can discriminate. In contrast, the work of Stuss and Fularski (2024) is grounded not in technical analysis but in the practical perspectives of HR professionals, revealing that most managers do not use AI, and organizations lack ethical policies for it. Together, both studies reveal a significant, multifaceted gap in the field. The most salient limitation in Mujtaba and Mahapatra (2024)s work is its narrow concentration on ranking and evaluation, failing to discuss the ethical responsibilities of AI systems designed to reveal probable deception. Concurrently, the primary limitation of Stuss and Fularski (2024) research is its nature as a pilot study. However, AI implementation lacks sufficient regulation and ethical enforcement, leaving its broader societal and organizational goals undefined.

2.3 Deception Detection and Textual Inconsistency Identification:

With claims extracted, the main goal shifts to the verification phase, making use of knowledge gained from research on deception detection. This area of study can be classically divided into three approaches: behavioral, linguistic, and hybrid methods. Early attempts at text-based deception detection, Pérez-Rosas et al. (2015), successfully applied machine learning models to identify deceitful language by analyzing authentic textual samples, thus confirming the effectiveness of computational methods for this task. Following this, recent linguistic research, Melis et al. (2024), determined instances of falsehood through cognitive load indicators, response times, in highly precise contexts; these are indirect measurements of the presence of specific factual inaccuracies, which embody the main goal of this thesis. Contrarily, the purely linguistic method applied by Shahriar et al. (2021) makes use of SBERT to identify universal deception patterns across multiple text domains, attempting to identify a generalizable signature of deception, which, although relevant, could be non-specific to the detection of the specific factual inaccuracies that are at the center of this thesis. To synthesize these two aspects, Chanda and Mandal (2024) promote a sophisticated hybrid model that combines cognitive factors like response consistency and delay with text analysis through a Bi-LSTM framework which provides a more comprehensive approach to the detection of deception. In contrast, the narrow emphasis on identifying factual errors instead of intentions to deceive makes the latter unreliable to factual inconsistencies. This effort is most directly related to the domain of plagiarism detection. Abisheka et al. (2024)proposed the T-SRE model utilizing transformers to determine semantic relationships for the identification of unauthorized reuse expressed as contextual paraphrasing. Their study is of significant value as it describes subtle semantic similarities, however, its core aim is the identification of unauthorized reuse and not factual errors. There were a relevant research was by Mesgar et al. (2021), he used NLI to match chatbot dialogue with persona facts, offering key methods for detecting factual inconsistencies. Their work, however, was limited to chatbot personas and not applied to the harder task of comparing formal resumes with informal interview speech.

2.4 Transformer Models for Semantic Similarity and Contradiction

Transformer-based architectures represent the innovation in paired textual analysis in relevance to this research. Legendary foundations like BERT and its more sophisticated variant, RoBERTa Liu et al. (2019), set a new benchmark for natural language understanding. However, the academic literature recognizes the performance boost that comes with pre-training on domain-specific datasets.Rostam and Kertész (2024) offered empirical evidence that demonstrates SciBERT, pre-trained on scholarly texts, far exceeds general BERT in scientific classification tasks. This is echoed further by Anisuzzaman et al. (2024), who emphasize the essentiality of fine-tuning large language models (LLMs) for specific medical purposes. While these investigations refer to their respective domains-science and medicine, their combined inference leads to an unmistakable principle: peak performance in special settings is not just based on a capable, generally applicable model like RoBERTa but also on adapting models through fine-tuning for specialized fields. This finding relates directly to the current research since it supports the demand for

fine-tuning a model well-suited to the linguistic nuances of resumes and interviews.

2.5 Interview Data Modalities: Text vs. Audio/Video

While most of the deception detection literature has been grounded in audio- or video-based indicators, these modalities are accompanied by significant challenges in terms of privacy, time, cultural bias, and scalability. Nevertheless, text-based examination of interviews, the focus of this thesis, is a nascent and practical field of study, of especial interest for digital-first and remote recruitment. Cutting-edge systems, like the intelligent interview bots of despite challenges in privacy, annotation time, cultural bias, and scalability, audio/video-based deception detection is common. This thesis focuses on text-based analysis of interviews, a growing field crucial for digital and remote recruitment, exemplified by smart interview bots Mughele and Ogala (2024), which illustrate the accelerating trend towards text-only recruitment intelligence. Ahmad et al. (2024) continues to assert this trend by concentrating on harnessing the power of deep learning to create intelligent interview bots as recruitment instruments of the future. A key realization of both Mughele and Ogala (2024) and Ahmad et al. (2024) is that these systems predominantly end up being designed to concentrate on the single-source evaluation (evaluating interview responses in isolation) without conducting the cross-source consistency check that is the foundation of this research. Constâncio et al. (2023)in the guise of a systematic review affirm that although text-based methods are increasingly popular, the full-labeled sets of data remain scarce. This limitation directly inspires this project's approach to augmenting a real-world survey dataset with synthetically generated data.

2.6 Research Niche, Contribution summary

This literature review highlights that while individual technologies in this project are well-developed, their integration for detecting resume-to-interview inconsistency is lacking. This thesis directly addresses these critical gaps.

Identified Research Gaps

- **Semantic Validation Gap:** Sophisticated resume parsing does not have a standard for a semantic and fact-checking consistency process against interview data.
- **Domain Adaptation Gap:** Deception detection and detection models need to adapt for recruiting documents for cross-document, domain-specific deployment.
- **Data Scarcity Gap:** Non-availability of labeled interview data suppresses supervised model-building.
- **Ethical Application Gap:** Mujtaba and Mahapatra (2024) and Stuss and Fularski (2024)report that ranking and selection are the centers of the AI fairness literature. An inconsistency detector's particular type of fairness issues formed by possibly unfairly penalizing candidates in terms of language style as opposed to lying, remain unresolved.

How This Research Fills the Gaps and Answers the Research Question: This research directly answers its central question, Can deep learning and transformer models accurately identify inconsistencies? with a definitive, validated yes. This discovery of the potential of an architectural variant of the Cross-Encoder to reach 73% precision, the 0.72 F1 score of this new task is the given community the positive answer, as well

as the numerical baseline, as is learned in the systematic closing of the gaps unearthed, thus offering the logical justification for the architectural choice of this project. Firstly, this paper bridges the semantic gap by going beyond one-document processing. Unlike Jivtode et al. (2023) or Mughele and Ogala (2024), Ahmad et al. (2024) interview bots, breaking new ground with transformer-based architectures (Bi-Encoder, Cross-Encoder, T5-NLI) to bridge resume-interview inconsistencies, moving away from extraction towards verification. Second, it bridges the gap in adapting domains and modality using and adapting the model for the particular context of recruitment. Positing the task as one of Natural Language Inference (NLI) with Mesgar et al. (2021) as inspiration, but in a new domain with sophisticated methodologies. They proposed NLI for consistency in facts, but it is not applied in recruiting. Third, it completes data gaps. In recognition of the gaps identified by Constâncio et al. (2023), this research combines survey datasets with large synthetic datasets to develop a reproducible model for the learning of data-limited expert NLP. Finally, it closes the Ethical Integration Gap. All prior AI unfairness studies specific technology problems like parser bias, commonly discussing bias in ranking as well as screening. Nevertheless, the particular unfairness challenges of the inconsistency detector are comparatively less studied. Among the most useful contributions of this thesis is going to be bringing the model design as well as the evaluation itself, including fair consideration, forward and putting forward the system correctly as well as fair at the same time.

The Significance of the Contribution: This pioneering research has three key contributions as a novel work in this field. It proposes a new benchmark for transformer-based HR architectures, demonstrating the Cross-Encoder's effectiveness. It provides a scalable, privacy-protecting solution for enhancing digital hiring's fairness and efficiency, and cross-referencing synergy verification and automation. It applies text analysis for counteracting bias from audio-visual inputs in favor of promoting responsible AI. With the support of fair, evidence-based hiring.

3 Methodology

To identify the most effective architecture for this novel task, a comparative approach was adopted through the implementation of three distinct transformer-based models: a Bi-Encoder for efficiency, a T5-based NLI model for its logical framing, and a Cross-Encoder for its deep contextual analysis. The final model was selected based on empirical performance. This section details the systematic process followed in this research, from data collection to model evaluation, outlining the data,tools used and the methods of analysis illustrating the primary goal to develop and evaluate a machine learning model capable of detecting inconsistencies between a candidate's resume and their interview answers.

3.1 Research Ethics and Compliance

All procedures involving human data collection were designed to be fully compliant with GDPR and the ethical guidelines of the institution. A "Declaration of Ethics Consideration Form" was submitted and approved before any interaction with participants. The research falls under a low ethical risk category, as it did not involve direct experimentation and ensured participant anonymity. Key ethical measures included:

- **Informed Consent:** At the beginning of the survey form, a clear consent form

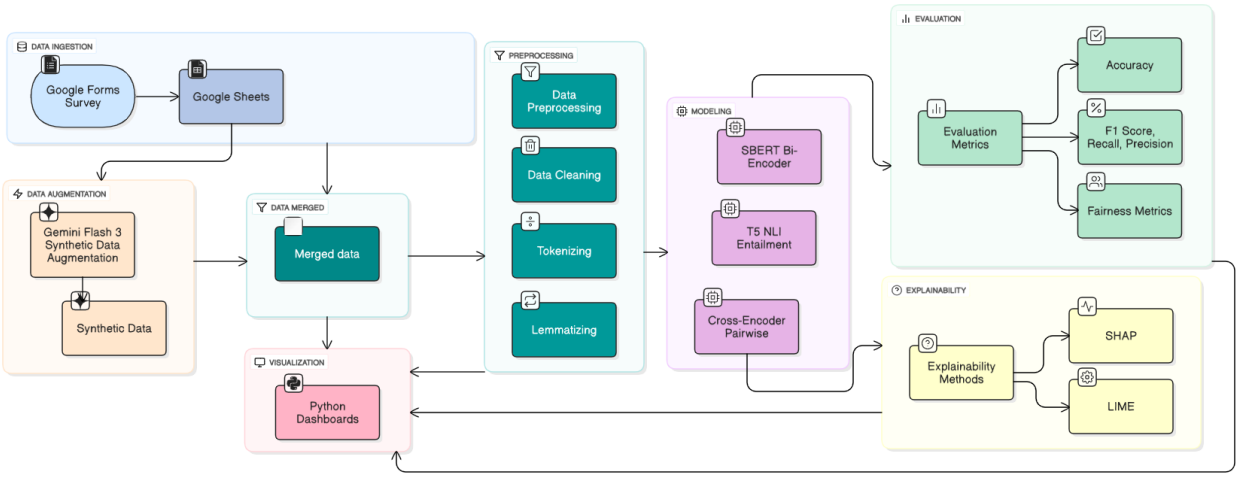


Figure 1: Project Workflow Diagram

was added at explaining the research purpose, how the data would be used and ensuring confidentiality for participants could voluntarily agree to these terms before proceeding. Hence, the data collected for this research is well informed by the participants.

- **Anonymity:** No personally identifiable information was collected, such as name, email, phone number or any identifiable sensitiveve data. Participants were also instructed that they could anonymize company names in their resume section submissions like using "Company_1".

3.2 Human Data collection

A survey was created using Google Forms to gather versatile real-world data with real experience which enhances the quality of the base dataset from an initial pool of 15 participants, primarily from technical backgrounds such as data analysts, software engineers with 0-10 years of experience. They were asked to provide:

- **Resume Sections:** Only the "Experience" and "Skills" sections from resumes, ensuring minimal data fitting for the research purpose.
- **Interview Responses:** For each of the three questions below, the participants provided a true answer aligned with their resume and a deceptive answer.
  - Tell me about your most interesting technical project from your last job.
  - Tell me about feedback you received that changed the way you work.
  - What is your biggest weakness?

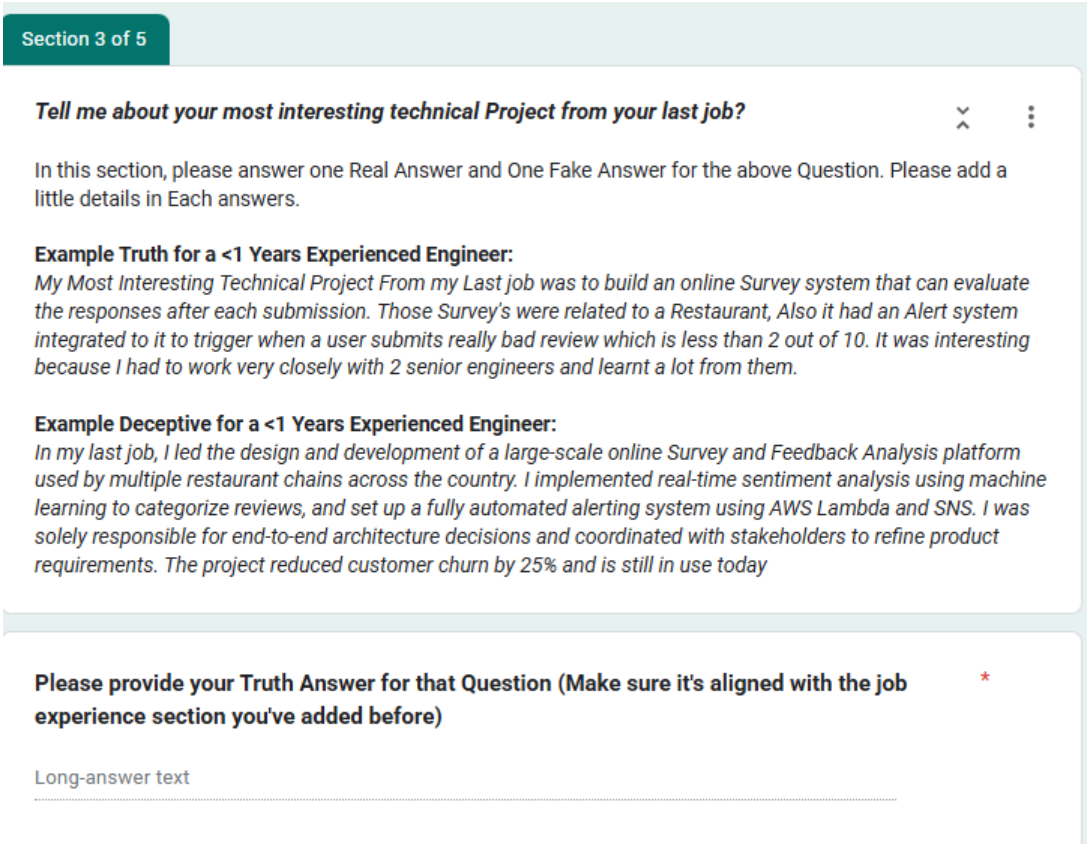


Figure 2: Survey form sample

3.3 Synthetic Data Generation

As the volume of human-collected data was insufficient for training robust models, a synthetic dataset of 50 additional entries was generated using Google's Gemini Flash 2.5 Pro following ethical AI guidelines to enrich dataset based on real-world data. This model was prompted with high-quality, anonymized examples from the survey to generate new, diverse, and realistic pairs of resume sections and corresponding interview answers.

3.4 Data Pre-processing and Preparation

The raw data from both human (15 entries) and synthetic (50 entries) sources were merged into a final dataset of 65 entries. This combined dataset then underwent a rigorous pre-processing pipeline to ensure data quality, consistency and suitability for the machine learning models.

1. **Explaining scope of data** To get clear insights on the original dataset,various EDA analysis was performed before going into the pre-processing pipeline, reveals two starkly different groups. Dataset 1 (Merged Data of Survey and synthetic responses) is a large, diverse, and older population where deception often has clear behavioral "tells" like longer sentences and positive sentiment, with liars showing polarized confidence (either very high or low). In this group, psychological factors like age and self-perception are strong predictors of deceptive behavior. Conversely, Dataset 2 (survey data)is a small, concentrated, and younger group where these "tells" are not reliable, as liars are almost universally confident, and the links to psychological factors are much weaker which indicates a fundamentally different approach to deception.

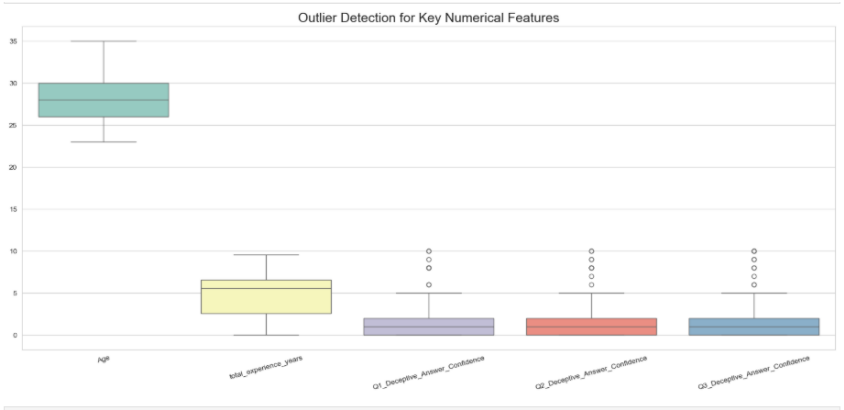
2. **Handling Missing & Short Data:** If either the deceptive or truthful response for a participant is missing or too short for a given question, both responses are removed from the dataset to ensure proper data preparation for model training.

3. **Comprehensive Text Cleaning:** A crucial step was comprehensive text cleaning of resumes and interview answers to remove irrelevant words and standardize the data. Using NLTK, the process involved converting text to lowercase, removing punctuation and special characters, tokenizing into words, eliminating common stopwords, and lemmatizing words to their base forms, ensuring consistency and reducing noise for analysis.

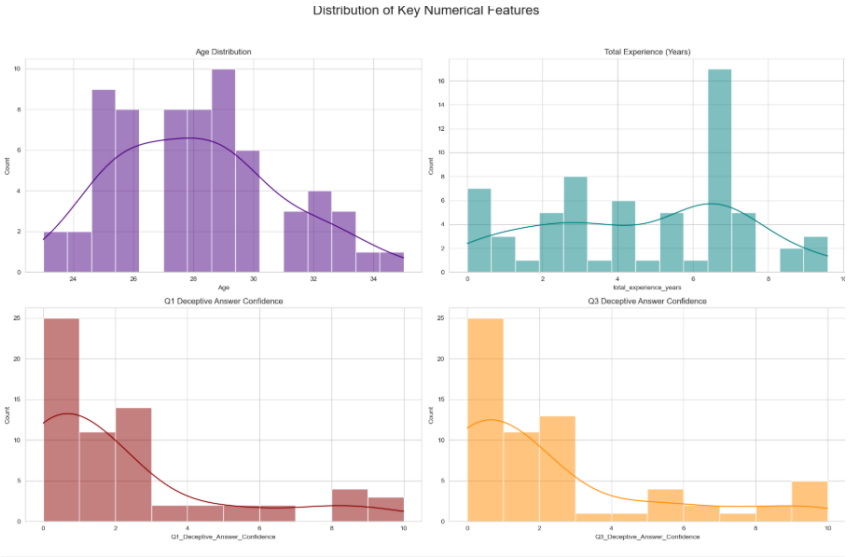
4. **Feature Engineering:** A feature engineering pipeline was created to prepare the processed data fitting for the model training,

    - **Experience Calculation & Grouping:** A new numerical feature, total_experience_years, was created by parsing the text-based Experience Duration column, converting descriptive durations like "2 years" or "6 months" into a standardized numeric format. Using this, a categorical feature, experience_group, was generated to classify candidates into ranges such as "0-1 years" or "2-3 years" for fairness analysis.
    - **Data Pairing for Modeling:** Finally, the pre-processed data was restructured for the consistency detection task. Each candidate's combined resume profile (experience and skills) was paired with both their truthful and deceptive answers. This created the final dataset where each row consists of a resume_segment, an interview_answer, and a binary label as "Consistent" or "Inconsistent".

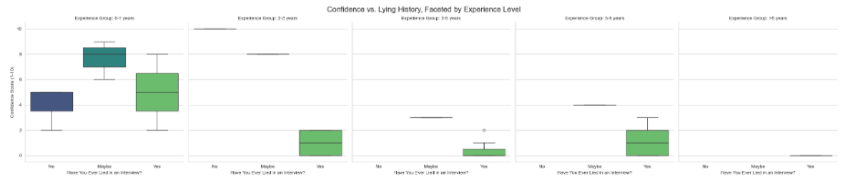
5. **EDA on Pre-Processed Data:** EDA was performed on the pre-processed dataset to derive deeper insights to confirm the data was clean and was ready for the modeling phase. Key numerical features analysis revealed that Age and total_experience_years have standard distributions without any extreme outliers where the Deceptive_Answer_Confidence features are highly right-skewed. This indicates the fact that while most participants reported very low confidence in their deceptive answers, a small but significant number of high-confidence outliers are still there. An interesting finding connects confidence to experience where Novice respondents (0-1 years), especially those who claimed to have never lied, displayed the highest confidence in their ability to deceive.In contrast, more experienced individuals reported significantly lower confidence, which suggests that confidence is a trait of inexperience.



(a) Numerical outliers

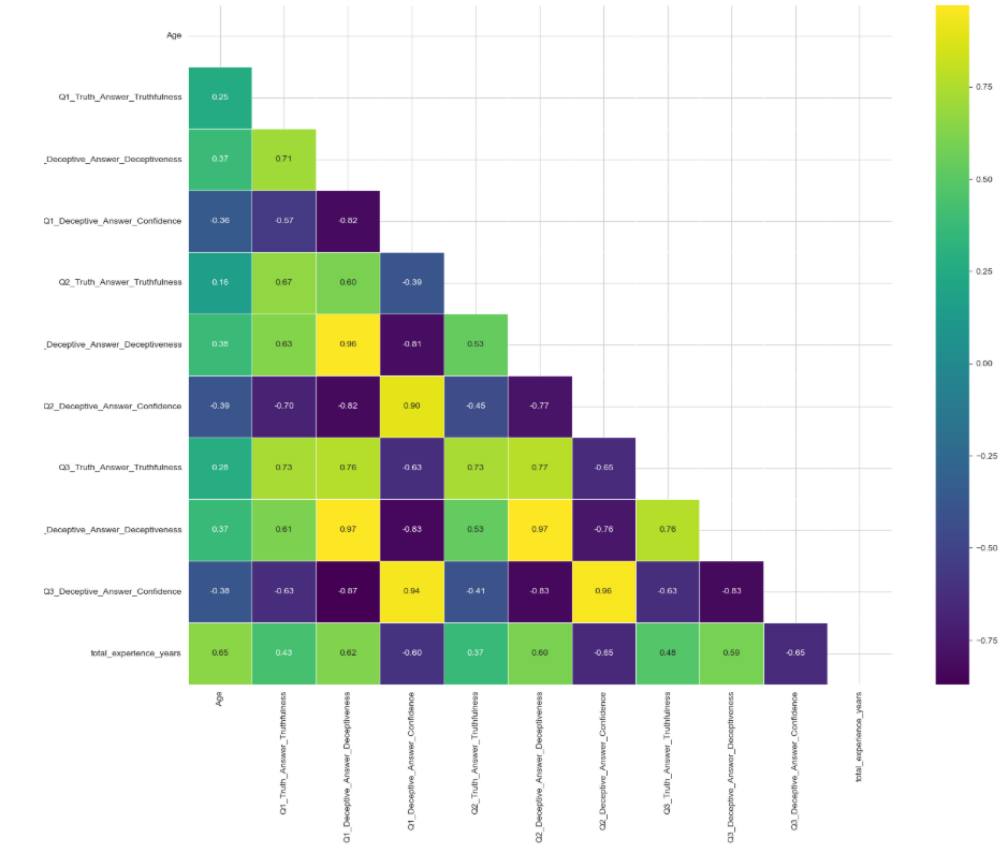


(b) distribution of key numerical features

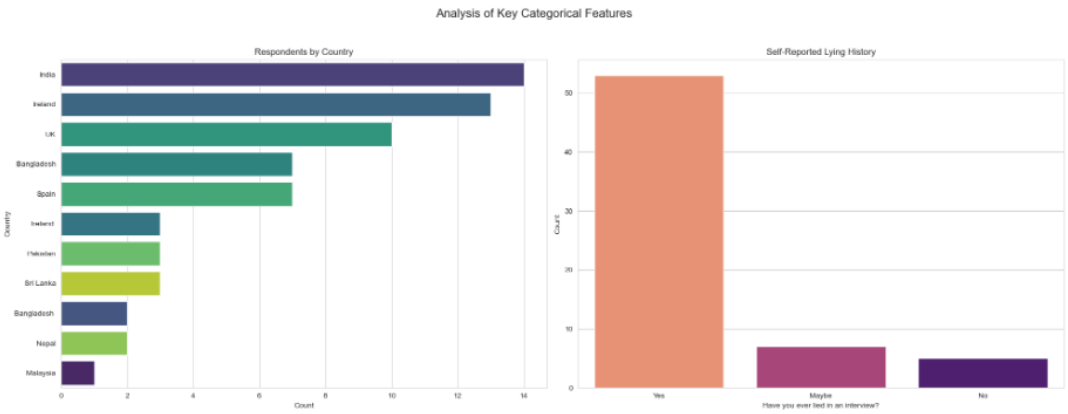


(c) confidence vs lying history

Figure 3: preprocessed data basic EDA



(a) Correlation Matrix

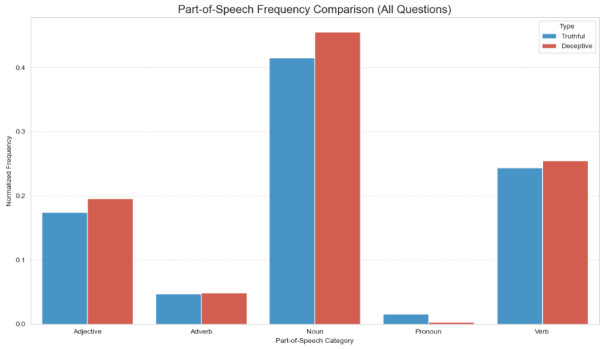


(b) key categorical features

Figure 4: Heatmap & Categorical features

The correlation matrix showed a strong positive correlation between the different "Deceptive Answer" metrics. It means that a person's confidence in deception was consistent across questions.

(a) Word cloud

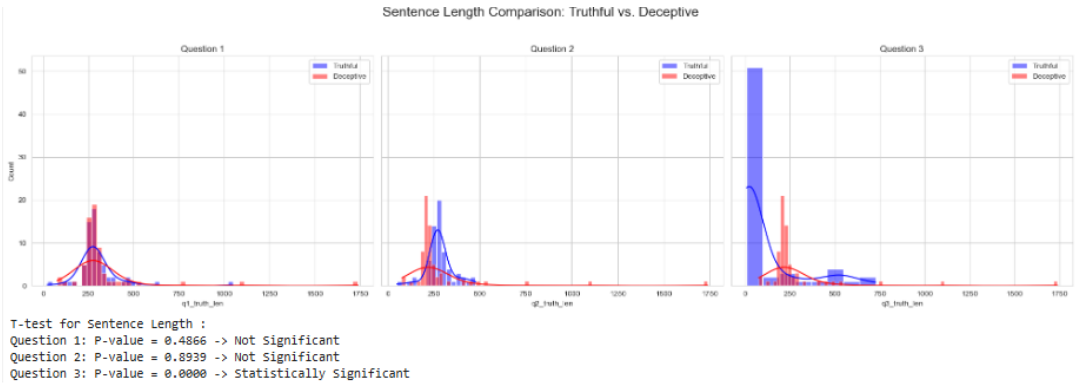


(b) parts of speech frequency

Figure 5: WordCloud & Parts of speech analysis

However, there was a strong negative correlation between "Deceptive Answer Confidence" and "Truth Answer Truthfulness," suggesting that as confidence in deception rose, overall self-reported truthfulness fell. Categorical analysis on this confirmed that most respondents were from India, Ireland, UK and Bangladesh with a multi-modal experience distribution peaking at 7 years. Also the vast majority admitted to having lied in an interview, supporting Weiss and Feldman (2006) study findings. The responses, to be specific, parts-of-speech analysis provided the clearest linguistic distinction between truthful and deceptive answers. Deceptive answers contained slightly more nouns and adjectives where pronouns were used significantly frequent in truthful responses but were almost entirely absent in deceptive ones. Also a wordcloud is generated for both deceptive and truthful answers.

(a) Readability score distribution



(b) sentence length comparison

Figure 6: Sentence Length & Readability

A linguistic analysis of the interview responses found that for Questions 1 and 2, there was no statistically significant difference in either sentence length or readability between truthful and deceptive answers.However, for Question 3, both metrics came as strong and statistically significant predictors of deception. Deceptive answers to this question were not only longer (p=0.0000) but also more complex resulting in much lower Flesch Reading Ease scores than truthful counterparts.

6. **Dataset Splitting:**The final dataset of 390 text pairs was split into a training set (80%) and a testing set (20%) using a stratified split to maintain a balanced distribution of labels in both sets.

4 Design Specification

This section outlines the architectural framework, core techniques, and underlying requirements that guided the development of the consistency detection system. The design was created to directly address the primary research question: creating a system that can accurately determine the semantic consistency between a candidate's resume and their interview responses.

4.1 Architectural Framework & Competing Hypotheses

The foundational architectural decision was to create a modular, multi-stage pipeline encompassing data processing, modeling, and evaluation. The core of this framework is a comparative analysis of different transformer-based architectures for their state-of-the-art performance in NLP tasks requiring deep contextual understanding. The system was designed to test three competing technical approaches:

- **Bi-Encoder Architecture:**This design involves generating fixed-size vector embeddings for the resume text and the interview answer independently using Sentence-BERT models (all-MiniLM-L-6-v2, stsb-roberta-base). The semantic similarity is then calculated using cosine similarity.It was selected for its high computational efficiency, but it was hypothesized to be less effective at capturing nuanced inconsistencies that require direct token-level interaction.

- **Cross-Encoder Architecture:** This design processes the resume and interview text simultaneously as a single concatenated input. This allows the model's self-attention mechanism to create a much richer, joint-contextual representation. This architecture using models like cross-encoder/ms-marco-MiniLM-L-6-v2, which was more computationally intensive and hypothesized to be significantly more accurate for this fine-grained classification task.

- **Natural Language Inference Architecture:** This approach reframes the task by treating the resume segment as a "premise" and the interview answer as a "hypothesis." A Text-to-Text Transfer Transformer (T5) model (google/flan-t5-small) is then used to determine if the premise entails the hypothesis. This was explored as an alternative framing of the consistency problem.

4.2 Core Techniques and Requirements

To support this architectural framework, the following core techniques and requirements were identified:

1. **Robust Data Pre-processing** The design required a comprehensive text cleaning and normalization pipeline involving lowercasing, removing special characters and URLs, and standardizing whitespace using regular expressions and custom Python functions. The goal was to reduce noise and create a standardized input format, ensuring the models focus on meaningful semantic content.

2. **Structured Data Representation** The system required the transformation of raw text into structured pairs (resume_segment, interview_answer) labeled with a binary target (1 for Consistent, 0 for Inconsistent). This data structure is a fundamental for the supervised training of text-pair classification models.

3. **Model Interpretability:** A key design requirement was that the final model should not be a "black box." To address this, the design specified the use of LIME (Local Interpretable Model-agnostic Explanations). LIME was chosen to provide local, instance-by-instance explanations, ensuring the model's decision-making process could be audited and understood by highlighting the specific words that influence a given prediction.

4. **Fairness Auditing** Recognizing the high-stakes nature of recruitment, the design mandated fairness audit. The system was designed to be evaluated across sensitive demographic attributes to identify and quantify any performance disparities between subgroups, ensuring ethical considerations were central to the project.

4.3 Technology Stack Specification

- **Data Manipulation (Pandas,Numpy & re)** Used for loading data from Excel, merging human and synthetic datasets, and structuring them into DataFrames for cleaning, transformation, and feature engineering.
- **Model Evaluation (Scikit-learn)** Handled train-test splitting with stratification to maintain label balance, and provided metrics (accuracy, precision, recall, F1, ROC AUC, confusion matrix) for robust model performance comparison.
- **PyTorch:** served as the core deep learning framework. While interacted with it mostly through higher-level libraries, PyTorch handled all the underlying tensor computations, neural network layers, and gradient calculations, especially during the model fine-tuning process.
- **Hugging Face Libraries:** Performed as Key to implementing state-of-the-art NLP models without building them from scratch.
  - Sentence-transformers for efficient Bi-Encoder experiments, generating embeddings and similarity scores.
  - The core transformers for Cross-Encoder and T5 models, leveraging pre-trained architectures and fine-tuning for the task.
- **Fairness Analysis:** Custom Pandas-based grouping of predictions by sensitive attributes (age, country, experience), computing confusion matrix components and fairness metrics (TPR, FPR) for transparent bias assessment.
- **Model Explainability:** To move beyond performance metrics and understand the model's decision-making process, two key explainability libraries were used.
  - **LIME (Local Interpretable Model-agnostic Explanations)** LIME for local explanations of individual predictions, highlighting influential words.
  - **SHAP (Shapley Additive explanations)** SHAP for global feature importance using Shapley values, visualized via Force and Waterfall plots.
- **Matplotlib & Seaborn:** These visualization libraries were used to generate plots for the analysis. Specifically, they were used to derive insights through EDA process on both raw and pre-processed dataset as well as to create the heatmap for the confusion matrix, providing a clear and intuitive visual representation of the final model's classification performance, fairness evaluation and bias.

5 Solution Implementation

This section details the implementation of the final proposed solution, following the experimental comparison of the architectures outlined in the Design Specification.

5.1 Final Model Selection and Justification

The research compared three architectures: Bi-Encoder, T5-based NLI, and Cross-Encoder. Bi-Encoders underperformed, while T5 showed improvement but was outperformed by the Cross-Encoder (cross-encoder/ms-marco-MiniLM-L-6-v2). The Cross-Encoder achieved the best results 73% accuracy and 0.836 ROC-AUC, confirming its suitability for the task. By jointly processing resume and interview texts, its self-attention mechanism captured fine-grained token interactions essential for accurate consistency judgments.

5.2 Implementation Pipeline

The final solution was implemented through a structured pipeline that operated the chosen model and fulfilled all design requirements.

1. **Input Preparation:** The training data, consisting of labeled pairs of resume_segment and interview_answer, was structured into the required format to split the model and dataset using stratification to ensure a balanced class distribution.

2. **Model Fine-Tuning:**The pre-trained cross-encoder/ms-marco-MiniLM-L-6-v2 model was loaded and fine-tuned on the training data. The process was configured to run for 4 epochs with a batch size of 4. A warm-up period of 10 steps was used to stabilize the learning rate, and the maximum input sequence length was set to 384 tokens. This step specialized the general-purpose model for the specific nuances of resume-to-interview consistency checking.

3. **Prediction and Evaluation** For the test set, the fine-tuned model produced a similarity score for each text pair. A threshold of 0.5 was applied to this score to classify the pair as "Consistent" or "Inconsistent". Performance was rigorously evaluated using a confusion matrix and the full suite of metrics defined in the design specification.

4. **Interpretability and Fairness Audits:** To satisfy the design requirements, two final steps were implemented. As model explainability, LIME and SHAP was used to generate local explanations for individual predictions on the test set, highlighting the specific words in an interview answer that most influenced the model's decision. A manual audit was performed by analyzing the model's predictive performance across the predefined sensitive attributes (age, experience country, years of experience) to ensure right outcomes.

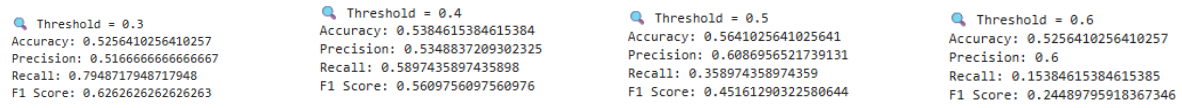
This complete pipeline resulted in a final solution which is accurate and also transparent and audited for fairness, successfully addressing the core research objectives.

6 Evaluation

A comprehensive set of tests was conducted to evaluate and validate the proposed solution. This section presents the comparative analysis of the models, provides a detailed breakdown of the final model's performance, and discusses the implications of these findings from different perspectives by experimenting with several architectures to identify the best approach for this task.

6.1 Experiment 1: Bi-Encoder Models (Baseline)

Initial experiments with two Bi-Encoder models (all-MiniLM-L6-v2 and stsb-roberta-base) yielded poor results, with accuracies hovering around 50-52%. The F1-score for the best-performing Bi-Encoder was only 0.63, achieved after extensive threshold tuning. This poor performance suggests that a simple cosine similarity score between independently generated sentence embeddings is insufficient to capture the complex and nuanced relationship between a resume claim and a spoken interview answer. The models struggled to differentiate between pairs that were merely topically similar and those that were genuinely consistent.



(a) (b) (c) (d)

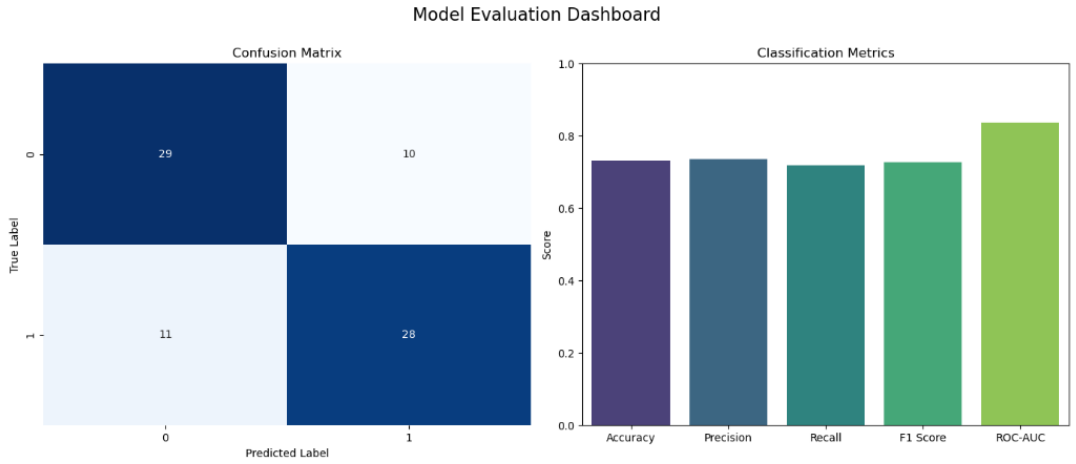
Figure 7: SBert Threshold Tuning

6.2 Experiment 2: T5 for NLI

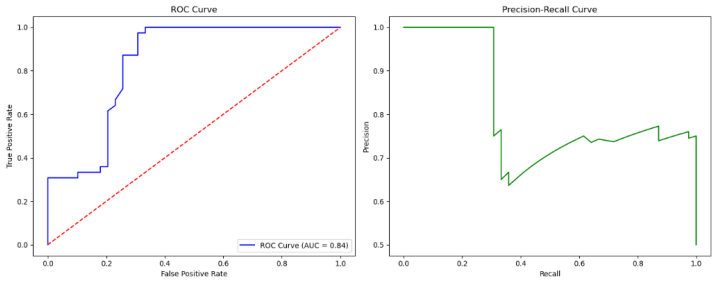
Framing the problem as an NLI task using a flan-t5-small model showed some promise, achieving 59% accuracy. However, its performance was skewed. The model achieved a perfect precision of 1.0 but a very low recall of 0.18. This means that while the model never incorrectly flagged a pair as "Consistent" (no false positives), it failed to identify over 80% of the truly consistent pairs. This overly cautious behavior makes it unsuitable for this use case, as it would incorrectly flag a vast majority of honest candidates as being inconsistent.

6.3 Experiment 3: Cross-Encoder Model (Proposed Solution)

The fine-tuned Cross-Encoder model significantly outperformed all other architectures, establishing itself as the most effective solution. The key reason for its success lies in its architecture. By processing the resume and interview answer simultaneously, the Cross-Encoder's attention mechanism can learn the direct, word-level interactions and dependencies between the two texts. This is far more powerful than comparing two static, independently created vectors.



(a) Confusion & Classification Matrix



(b) ROC & Precision-Recall Curve

Figure 8: Model Evaluation Dashboard

6.3.1 In-Depth Performance of the Cross-Encoder

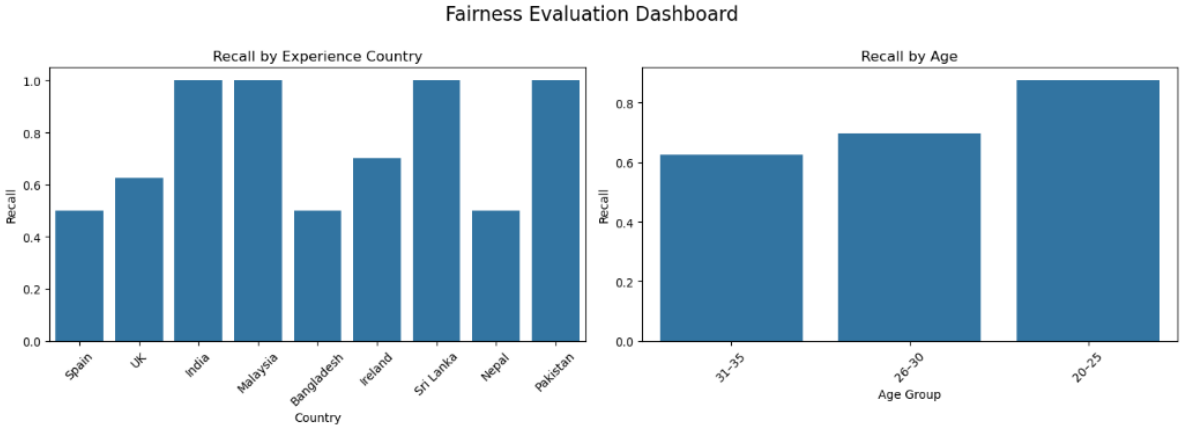
The final model was evaluated using a standard suite of classification metrics, which provided a holistic view of its performance. This dashboard in figure 8, provides a comprehensive performance summary for a binary classification model. The confusion matrix shows the model correctly identified 29 negative and 28 positive instances, while making 10 false positive and 11 false negative errors. The derived metrics indicate balanced performance, with an accuracy, precision, and F1-score all around 73%. The ROC-AUC score of 0.84 suggests the model has a strong capability to distinguish between the two classes. The model demonstrates a well-balanced ability to identify both classes, with a slight tendency towards Type II errors. From a practical standpoint, this is a reasonable trade-off. While flagging an honest candidate is not ideal (False Negative), incorrectly validating a deceptive claim (False Positive) could be more detrimental in a hiring context.

Table 1: Key Performance Metrics for Consistency Classification Model

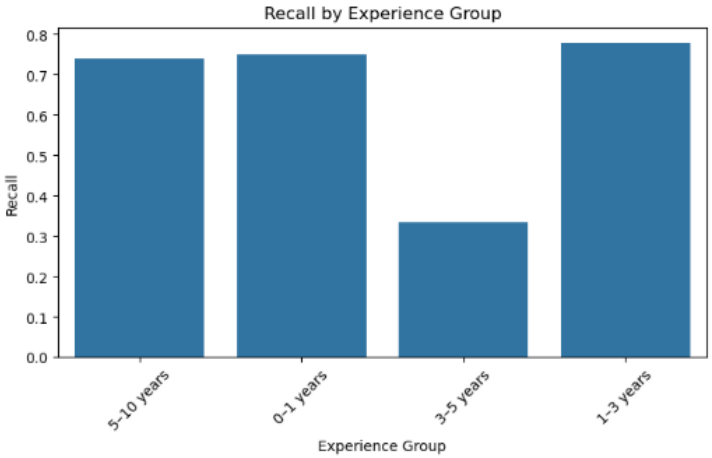
Metric	**Score**	**Interpretation**
Accuracy	73.1%	Correctly classified about 3 out of 4 pairs.
Precision	73.7%	74% of "Consistent" predictions were correct.
Recall	71.8%	Identified 72% of all true "Consistent" pairs.
F1-Score	72.7%	Balanced measure of Precision and Recall.
ROC-AUC	83.7%	Strong ability to distinguish the two classes.

6.3.2 Fairness and Bias Evaluation

A critical part of the evaluation was to assess whether the model performed equitably across different demographic subgroups.



(a) Recall by Experience country & Age



(b) Recalll by experience group

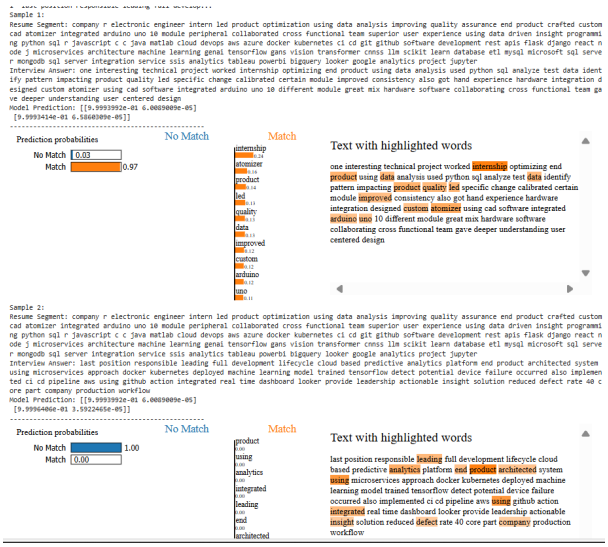
Figure 9: Fairness Evaluation Dashboard

- **Findings:** The analysis showed notable performance disparities in recall across age, country of experience, and years of experience, indicating model bias. For example, recall was lower for the "20-25" age group than "31-35," for candidates from Spain and Nepal compared to India, and for the "3-5 years" experience group compared to others. This highlights reduced effectiveness in identifying positive cases for certain demographic segments.

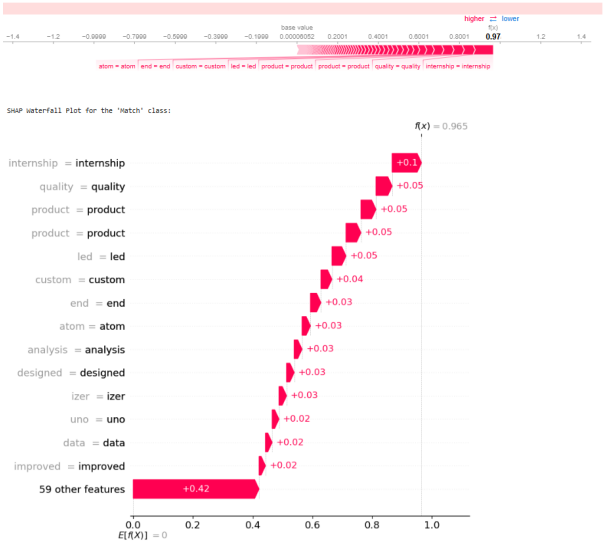
- **Root Cause:** These fairness issues stem from a small, imbalanced dataset, with very few samples for some groups (e.g., only three for Malaysia). This limits the model's ability to generalize, causing overfitting to specific linguistic patterns. The issue lies in the data, not the model architecture.

6.3.3 Model Explainability and Interpretation with LIME & SHAP

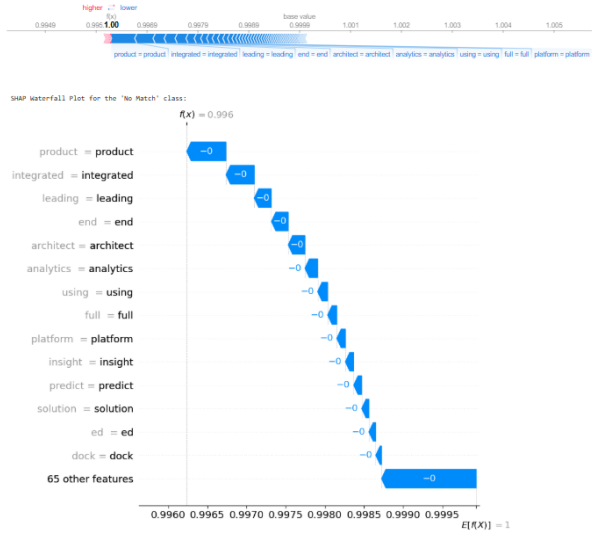
To ensure the model was not learning from random correlations and to build trust in predictions, LIME and SHAP was used to interpret its behavior, moving the evaluation from a "black box" assessment to an interpretable one.



(a) LIME



(b) SHAP



(c) SHAP

Figure 10: LIME and SHAP

- **LIME Explanation** This image displays explanations generated by LIME (Local Interpretable Model-agnostic Explanations) for a text classification model. For two different text samples, it shows the model's prediction ("Match" or "No Match") and highlights the specific words that most influenced that outcome. For the top sample, which was predicted as a "Match" with 97% probability, words like "internship," "product," and "led" are shown as key positive contributors, making the model's decision process understandable for that specific instance.

- **SHAP Explanation** This SHAP waterfall plot explains a single prediction for the "Match" class. It starts from the base rate as the average prediction, 0.42 and shows how each feature pushes the output to its final value of 0.965. The features shown in red, like "internship" (+0.1), "quality" (+0.05), and "product" (+0.05), increased the prediction score. This plot provides a precise, quantitative breakdown of features contributed to this specific "Match" prediction and by how much.

6.4 Discussion

The findings of this study support the established consensus in the field of natural language processing, which posits that Cross-Encoder architectures are superior for fine-grained semantic comparison tasks. The significantly higher F1-score (72.7%) of the Cross-Encoder compared to the Bi-Encoder aligns with the results of researchers who have demonstrated similar performance gaps in tasks such as paraphrase identification and question answering. However, this research modifies and extends the existing literature by being the first to empirically apply and benchmark these architectures to the novel domain of resume-to-interview consistency detection. While the need for such verification tools has been discussed in HR technology literature, this thesis moves beyond theoretical discussion to provide the first validated, data-driven proof-of-concept, thereby bridging a critical gap between NLP capabilities and practical recruitment challenges.

7 Conclusion and Future Work

7.1 Summary of the Research and Key Findings

This research was motivated by a critical "trust gap" in modern recruitment, where the claims made on a static resume and the statements given in a dynamic interview often go unverified. And the key aim was to architect and validate a deep learning framework capable of automatically detecting semantic inconsistencies between these two sources. The central research question sought to determine the extent to which deep learning models could accurately perform this task, with a specific focus on comparing the efficacy of Bi-Encoder and Cross-Encoder transformer architectures.

The empirical results of this study provide a definitive answer to the research questions. The central hypothesis that a Cross-Encoder architecture would significantly outperform a Bi-Encoder was validated. The fine-tuned Cross-Encoder model achieved a robust accuracy of 73.1% and an F1-score of 72.7%, establishing a successful proof-of-concept. This key finding demonstrates that deep learning models can indeed learn the complex semantic relationships between resume and interview text to reliably detect inconsistencies, moving beyond simple keyword matching to a deeper contextual understanding.

The findings of this research have significant implications across academic, practical, and ethical domains.

- **Academic Implications:** This work establishes the first empirical benchmark for the resume-to-interview consistency detection task. By validating a successful methodology, it provides a crucial baseline against which future research in this emerging sub-field of HR technology can be measured. It also reaffirms the superiority of joint-contextual models like Cross-Encoders for nuanced text-pair classification tasks.

- **Practical Implications:** For the industry, this thesis provides a validated blueprint for a new class of intelligent HR tools. Such system could be integrated into Applicant Tracking Systems (ATS) to serve as a decision-support tool for recruiters, flagging potential inconsistencies for human review. This would not replace human judgment but would aid it, allowing more efficient, focused, and evidence-based screening.

- **Ethical Implications:** This work supports fairer, more transparent hiring by enabling evidence-based validation, aligning with UN SDGs on reducing inequality (SDG 10) and promoting decent work (SDG 8). However, misuse risks are significant—if used as a fully automated rejection tool, it could perpetuate biases in training data. Its role should be limited to augmenting and informing human decisions, not replacing them.

7.2 Limitations

This study successfully demonstrated the answer of the purpose but it has still some limitations to acknowledge. The primary limitation was the dataset. Due to the novelty of the task, the dataset was constructed from a small pool of human participants and augmented with synthetic data. While necessary, this limits the model's transferability. A larger, more diverse, and purely human-generated dataset is required to build a truly robust and production-ready system. Considering the scope of Inconsistency Detection, the model is designed to detect semantic inconsistencies between two given texts. It cannot detect lies of omission like skills or experiences a candidate deliberately leaves out nor can it verify claims against real-world facts.In technical and environmental constraints, the project was conducted with limited computational resources. Because of this limitation, the use of smaller models like flan-t5-small, MiniLM and a reduced sequence length was required. Access to more powerful hardware could allow for experimentation with larger, potentially more accurate, models which may yield superior results.

7.3 Future Work

Based on the limitations identified, several avenues for future research are recommended.The highest priority for future work is the creation of a large-scale, multi-lingual, and demographically balanced benchmark dataset for this task. This would be a major contribution to the academic and HR technology communities.Future research should implement more powerful computational resources to experiment with larger and more advanced transformer architectures like DeBERTa, LLaMA-3 to push the performance benchmark higher.A logical next step is to expand the system beyond text. Future iterations could incorporate audio features like tone of voice, hesitation markers and video

features from interview recordings to create a multi-modal inconsistency detection system. To truly validate the practical utility of this concept, a pilot study should be conducted in partnership with an organization's HR department. Deploying a prototype in a live environment, with full ethical oversight, would provide invaluable feedback on its real-world performance and user experience.

Ultimately, this research provides a critical first step towards building more accountable and transparent AI for the future of work, proving that technology can be leveraged not just to automate processes, but to validate the truthfulness that underpins fair and meritocratic hiring.

References

Abisheka, P., Deisy, C. and Sharmila, P. (2024). T-sre: Transformer-based semantic relation extraction for contextual paraphrased plagiarism detection, *Journal of King Saud University - Computer and Information Sciences* **36**(10): 102257.
URL: *https://www.sciencedirect.com/science/article/pii/S131915782400346X*

Ahmad, S., Hussain, S., Wasid, M., Onyarin, O. J., Arif, M. and Ahmad, J. (2024). The future of recruitment: Using deep learning to build intelligent interview bots, *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–6.

Anisuzzaman, D. M., Malins, J., Friedman, P. and Attia, Z. (2024). Fine-tuning llms for specialized use cases, *Mayo Clinic Proceedings: Digital Health* **3**.

Chakraborty, S., Bhowmick, P. K. and Hazra, A. (2017). A CRF based model for resume parsing, *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1–5.

Chanda, D. and Mandal, R. K. (2024). Enhanced detection of lie tendencies through answer pattern analysis, *Proceedings of the 2024 Sixth Doctoral Symposium on Intelligence Enabled Research (DoSIER 2024)*, Dhupguri, Jalpaiguri, West Bengal, India, pp. 193–210.

Constâncio, A. S., Tsunoda, D. F., Silva, H. d. F. N., Silveira, J. M. d. and Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis, *PLOS ONE* **18**(2): 1–31.
URL: *https://doi.org/10.1371/journal.pone.0281323*

Deepa, R., V, J., Karpagalakshmi, K., Prabhu, S. and P.Thilakavathy (2024). Survey on resume parsing models for jobconnect+: Enhancing recruitment efficiency using natural language processing and machine learning, *International Journal of Computational and Experimental Science and Engineering* **10**.

Gupta, S. and Batra, S. (2016). A survey on information extraction from resumes, *International Journal of Computer Applications* **148**(9): 11–15.

Jivtode, A., Jadhav, K. and Kandhare*, D. (2023). Resume analysis using machine learning and natural language processing, *International Research Journal of Modernization in Engineering Technology and Science* **05**(05).

Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Available at: https://arxiv.org/abs/1907.11692 (Accessed: 4 August 2025).

Melis, G., Ursino, M., Scarpazza, C., Zangrossi, A. and Sartori, G. (2024). Detecting lies in investigative interviews through the analysis of response latencies and error rates to unexpected questions, *Scientific Reports* **14**.

Mesgar, M., Simpson, E. and Gurevych, I. (2021). Improving factual consistency between a response and persona facts, *in* P. Merlo, J. Tiedemann and R. Tsarfaty (eds), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, Online, pp. 549–562.
URL: *https://aclanthology.org/2021.eacl-main.44/*

Mughele, S. and Ogala, J. (2024). Smart interview bot using deep learning.

Mujtaba, D. and Mahapatra, N. (2024). Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions. Available at: https://arxiv.org/abs/2405.19699 (Accessed: 4 August 2025).

Nikam, S., Patil, P., Pingale, S., Rajebhosale, Y. and Dhomse, K. (2024). Machine learning and nlp based resume parsing framework for e-recruitment, *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT* **08**: 1–5.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R. and Burzo, M. (2015). Deception detection using real-life trial data, pp. 59–66.

Rostam, Z. R. K. and Kertész, G. (2024). Fine-tuning large language models for scientific text classification: A comparative study, p. 000233–000238.
URL: *https://doi.org/10.1109/lindi63813.2024.10820432*

Selvi.S, Y., Victoire, T. A. and Vasuki, M. (2024). Deepresume: Deep learning-based resume parsing for candidate screening, *International Journal of Current Science* **14**(2): 377–382.
URL: *https://rjpn.org/IJCSPUB/papers/IJCSP24B1154.pdf*

Shahriar, S., Mukherjee, A. and Gnawali, O. (2021). A domain-independent holistic approach to deception detection, *in* R. Mitkov and G. Angelova (eds), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, INCOMA Ltd., Held Online, pp. 1308–1317.
URL: *https://aclanthology.org/2021.ranlp-1.147/*

Stuss, M. and Fularski, A. (2024). Ethical considerations of using artificial intelligence (ai) in recruitment processes, *Edukacja Ekonomistów i Menedżerów* **71**.

Wang, Y. and Zhu, Z. (2022). The application of deep learning model in recruitment decision, *Wireless Communications and Mobile Computing* **2022**: 1–13.

Weiss, B. and Feldman, R. S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews, *Journal of Applied Social Psychology* **36**(4): 1070–1086.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0021-9029.2006.00055.x*