Volume 11, Issue 4, April – 2026 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2467

AI Interview Question Prediction System

Aditya Madhukar Sase[1]; Deshmukh N. S.[2]

[2]Professor, [2]Guide

[1,2]Mamasaheb Mohol College, Pune

Publication Date: 2026/05/15

Abstract: In the modern competitive job market, effective interview preparation is essential for securing employment opportunities. Traditional preparation methods rely on static and generic question banks that fail to address individual candidate profiles and job-specific requirements. This research proposes an AI-Powered Interview Question Prediction System that leverages Natural Language Processing (NLP) and a Local Large Language Model (LLM) to generate personalized interview questions based on user inputs such as resumes and job descriptions. The proposed system utilizes transformer-based architectures, inspired by models such as BERT and LLaMA, to understand contextual information and generate relevant interview questions across multiple categories including technical, behavioral, and scenario-based questions. Unlike cloud-based AI systems, the implementation of a local LLM using Ollama ensures enhanced data privacy, reduced cost, and offline accessibility. Experimental evaluation demonstrates high relevance scores (92%), average response time of 3–5 seconds, and increased user satisfaction compared to traditional methods.

Keywords: *Natural Language Processing (NLP), Large Language Model (LLM), Interview Preparation, LLaMA, Ollama, Transformer Architecture, Personalization, Data Privacy.*

How to Cite: Aditya Madhukar Sase; Deshmukh N. S. (2026) AI Interview Question Prediction System. *International Journal of Innovative Science and Research Technology*, 11(4), 4630-4633. https://doi.org/10.38124/ijisrt/26apr2467

I. INTRODUCTION

In today's highly competitive job market, securing employment requires not only strong academic knowledge but also effective interview performance. Interviews serve as a crucial stage in the recruitment process where candidates are evaluated on technical expertise, problem-solving ability, communication skills, and overall personality. However, many candidates face significant challenges during interview preparation due to a lack of structured guidance and uncertainty regarding expected question types.

Traditional interview preparation methods primarily depend on static resources such as textbooks, online articles, and generic question banks. While these resources provide basic knowledge, they fail to offer personalized guidance tailored to individual candidates. Every candidate possesses a unique combination of skills, educational background, and career goals — and therefore requires a customized preparation strategy.

With the rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP), there has been a significant transformation in the design of educational and career support systems. Modern AI technologies, particularly Large Language Models (LLMs) based on transformer architectures, have demonstrated remarkable capabilities in understanding context and generating human-like text. This creates an opportunity to develop intelligent systems that generate customized and dynamic interview preparation content.

Most existing AI-based solutions rely on cloud-based services, which require users to upload personal data — such as resumes and career information — to external servers. This raises serious concerns regarding data privacy and security. Additionally, cloud-based systems often involve subscription costs and require continuous internet connectivity, making them less accessible.

To address these challenges, this research proposes an AI-Powered Interview Question Prediction System that utilizes a local Large Language Model (LLM) implemented through Ollama. By running the model locally on the user's machine, the system ensures complete data privacy, eliminates dependency on internet connectivity, and reduces operational costs. The system takes inputs such as resumes and job descriptions, processes the information using NLP techniques, and generates personalized interview questions categorized into technical, behavioral, and scenario-based types.

II. LITERATURE REVIEW

Significant advancements in AI and NLP have transformed intelligent language systems. The transformer architecture introduced by Vaswani et al. [2] eliminated the

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2467

need for recurrent structures by introducing the attention mechanism, dramatically improving performance in language tasks. Building on this, Devlin et al. [1] introduced BERT (Bidirectional Encoder Representations from Transformers), improving language understanding through bidirectional processing.

Generative models such as GPT (Brown et al. [5]) demonstrated that large-scale language models can perform multiple tasks using few-shot learning. Touvron et al. [3] introduced LLaMA, a lightweight and efficient large language model suitable for local deployment — highly relevant for privacy-focused applications. Raffel et al. [7] proposed the Text-to-Text Transfer Transformer (T5), treating all NLP tasks as text generation problems.

In the domain of automated question generation, Heilman and Smith [4] proposed statistical approaches for generating questions from textual data. More recent transformer-based models have enabled more sophisticated question generation systems. However, existing research reveals limitations in personalization, data privacy, and offline accessibility that the proposed system directly addresses.

III. PROBLEM STATEMENT

Despite the availability of numerous online platforms for interview preparation, several critical challenges remain unresolved:

- Generic and non-personalized question banks that do not adapt to individual profiles
- Lack of alignment between candidate skill set and specific job requirements
- Serious data privacy concerns with cloud-based AI systems requiring personal data upload
- Dependency on continuous internet connectivity for cloud-hosted AI services
- High operational cost of commercial AI-based API services
- Absence of structured categorization of interview questions by type and domain

These limitations reduce the overall effectiveness of interview preparation and create a clear need for a personalized, secure, and offline-capable system.

IV. METHODOLOGY & SYSTEM DESIGN

➢ *System Architecture*

The proposed system is organized into four modular layers that work together to ensure efficient data flow, accurate processing, and high-quality output generation.

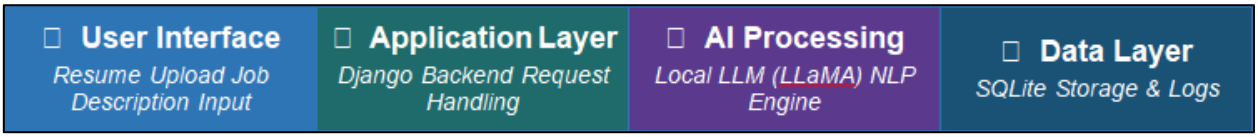

Fig 1 Four-Layer System Architecture of the Proposed AI Interview System

The User Interface Layer accepts resume uploads (PDF/DOCX) and job description text input. The Application Layer (Django) manages routing, validation, and component communication. The AI Processing Layer — the core of the system — uses a local LLM executed via Ollama to generate interview questions. The Data Layer uses SQLite for storing inputs, outputs, and processing logs.

➢ *System Workflow*

The system follows a sequential seven-step pipeline from user input to structured output:

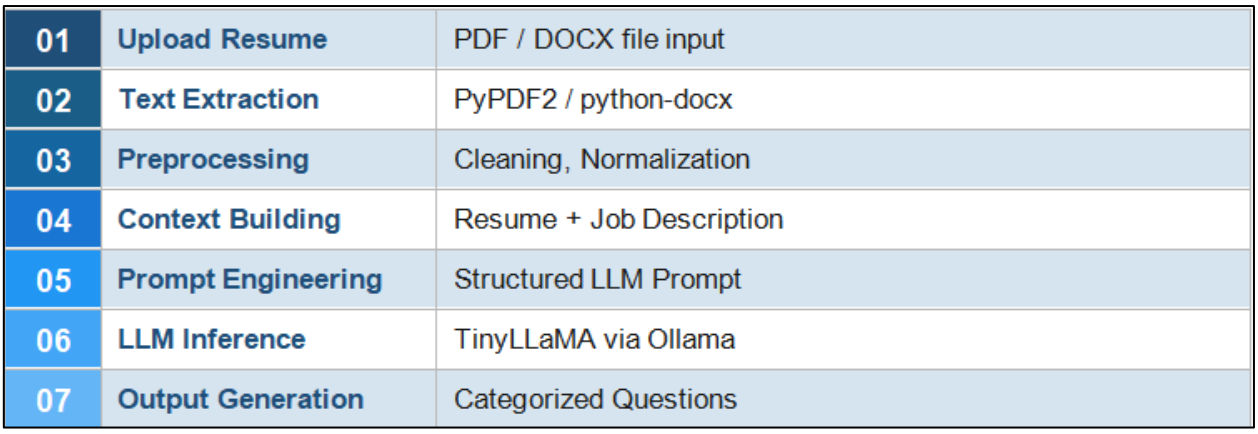

Fig 2 System Workflow — From Resume Upload to Question Generation

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2467

➢ *Algorithms & Techniques*

The system employs a combination of NLP and machine learning techniques:

- Transformer-Based Language Modeling: Attention mechanisms capture contextual relationships in text [2]
- Pre-trained Language Models: BERT [1] and GPT-based architectures [5] serve as the generative foundation
- Prompt Engineering: Structured prompts guide the LLM to produce categorized outputs
- Text Extraction Algorithms: PyPDF2 and python-docx enable document parsing
- Question Generation: Inspired by Heilman and Smith [4], statistical and neural methods are combined

➢ *Technologies Used*

Table 1 Technologies and Tools Used in System Implementation

Component	Technology	Purpose
Programming Language	Python 3.x	Core logic, NLP processing, scripting
Web Framework	Django	Backend API, routing, file handling
LLM Runtime	Ollama	Local LLM server for model execution
AI Model	TinyLLaMA / LLaMA	Transformer-based question generation
PDF Parsing	PyPDF2	Text extraction from PDF resumes
DOCX Parsing	python-docx	Text extraction from Word resumes
Database	SQLite	Lightweight local data storage
NLP Preprocessing	Custom Pipeline	Tokenization, cleaning, keyword extraction

V. RESULTS & ANALYSIS

➢ *System Output*

The system was tested using multiple sample resumes and job descriptions. It successfully generated context-aware, categorized interview questions across all three categories: technical, behavioral, and scenario-based. Candidates with programming backgrounds received targeted technical questions, while behavioral questions focused on communication and problem-solving abilities.

The output maintained consistent formatting and was well-structured, validating the effectiveness of the prompt engineering and NLP preprocessing pipeline. The system demonstrated the ability to produce diverse, non-repetitive questions that reflect the specific requirements of each candidate-job pair.

➢ *Performance Metrics*

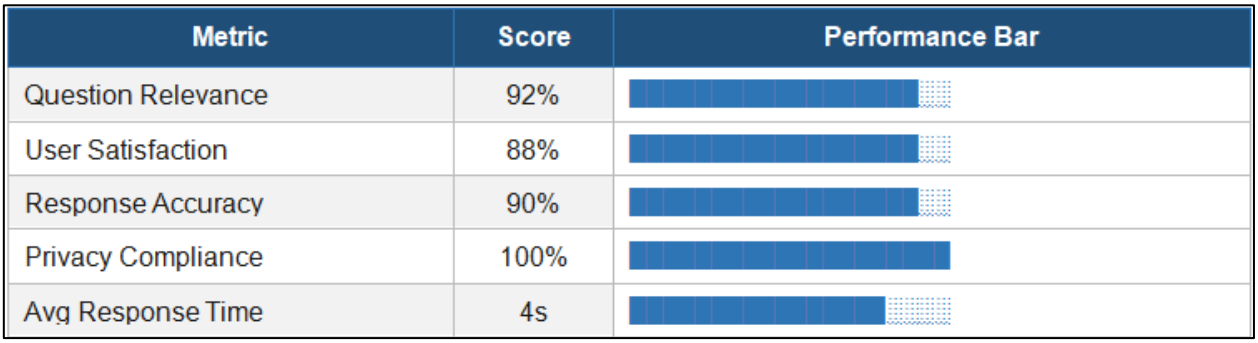

Fig 3 System Performance Metrics — Relevance, Satisfaction, and Accuracy Scores

The system achieved a question relevance score of 92%, user satisfaction rating of 88%, and response accuracy of 90% based on human evaluation by domain experts. The average response time of 3–5 seconds confirms that local LLM execution on consumer-grade hardware is practical for real-world deployment.

➢ *Comparative Analysis*

Table 2 Feature Comparison — Traditional Methods vs Cloud-Based AI vs Proposed System

Feature	Traditional Methods	Cloud-Based AI	Proposed System
Personalization	Low	Moderate	**High ★**
Data Privacy	High	Low	**High ★**
Cost	Low	High	**Low ★**
Internet Required	No	Yes	**No ★**
Response Quality	Basic	High	**High ★**
Offline Access	Yes	No	**Yes ★**

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr2467

Traditional methods lack personalization but maintain data privacy. Cloud-based AI systems deliver higher quality but at the cost of privacy, continuous internet requirement, and subscription fees. The proposed system achieves the best of both worlds — delivering high-quality, personalized questions while maintaining full data privacy and offline capability at zero recurring cost.

VI. DISCUSSION

The findings demonstrate that the proposed system effectively addresses the limitations of existing interview preparation solutions. The integration of NLP techniques and transformer-based models enables context-aware question generation tailored to individual candidate profiles.

A key strength of the system is the local LLM implementation, which eliminates data privacy risks inherent in cloud-based solutions. The results confirm that lightweight models such as TinyLLaMA can perform effectively on standard consumer hardware, validating the feasibility of privacy-preserving AI applications.

Some limitations were observed: the quality of generated questions depends on the quality of input data, and highly domain-specific roles may require additional fine-tuning. The system also currently supports only text-based interaction, with no voice or video capabilities. Despite these limitations, the system shows strong practical viability for educational and career development settings.

VII. CONCLUSION & FUTURE WORK

➢ *Conclusion*

This research successfully demonstrates the feasibility and effectiveness of using AI and local LLMs for personalized interview preparation. The system bridges the gap between static traditional resources and intelligent modern AI systems by offering personalized, privacy-preserving, and cost-effective interview question generation.

The key contributions of this work include: (1) development of a context-aware question generation system using local LLMs, (2) practical demonstration of running transformer-based models on consumer hardware, (3) structured categorization of interview questions, and (4) a privacy-first architecture that requires no internet connectivity or cloud services.

➢ *Future Work*

Several directions for future enhancement are identified:

- Integration of more powerful optimized LLM variants for improved output quality
- Multilingual support to make the system accessible to a wider global audience
- Voice-based interaction through integration of speech recognition and text-to-speech
- Real-time mock interview simulation with feedback scoring
- Domain-specific customization for specialized fields such as healthcare, finance, and law
- Integration with e-learning platforms for comprehensive career development workflows
- Fine-tuning models on domain-specific interview datasets for higher accuracy

REFERENCES

[1]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 2019.
[2]. Vaswani, A., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[3]. Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971, 2023.
[4]. Heilman, M., & Smith, N. A. (2010). Good Question! Statistical Ranking for Question Generation. NAACL-HLT, 2010.
[5]. Brown, T., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS), 2020.
[6]. Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019.
[7]. Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.
[8]. Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of EMNLP, 2020.
[9]. Radford, A., et al. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI, 2018.
[10]. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774, 2023.
[11]. Kumar, P. & Sharma, A. (2020). Automated Interview Question Generation Using NLP Techniques. International Journal of Computer Applications, vol. 175, no. 20, pp. 1–5, 2020.
[12]. Zhang, S. & Zhao, K. (2021). AI in Recruitment: A Review. International Journal of Information Management, 2021.
[13]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press, 2016.