Ascend AI: An Intelligent, Multimodal Framework for Personalized Career Direction and Adaptive Technical Interview Simulation

Shiv Sablok1*, Saumya Sharma1, Prince Kumar Singh1, Jeetu Singh1, Ayushi Sharma, Diwakar Shrivastava, Anshika Singh

How to Cite this Article:

Sablok, S., Sharma, S., Singh, P. K., Singh, J., Sharma, A., Shrivastava, D. & Singh, A. (2026). Ascend AI: An Intelligent, Multimodal Framework for Personalized Career Direction and Adaptive Technical Interview Simulation. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.743

https://doi.org/10.55041/ijcope.v2i4.743

Abstract

The rapid diversification of technical specializations within the computer science and information technology domains presents a significant educational challenge for students, who frequently lack the profound self-awareness and practical guid-ance imperative to selecting professional pathways perfectly aligned with their inherent psychological and cognitive traits. Consequently, pivotal career deci-sions are systematically driven by external peer trends, superficial fascinations, or arbitrary assumptions rather than intrinsic behavioral suitability. This paper introduces Ascend AI, a comprehensive, artificial intelligence-driven career orien-tation framework meticulously designed to mitigate this structural uncertainty. The proposed architecture seamlessly integrates quantitative psychological pro-filing, generative LLM-based learning curriculum methodologies, and responsive, interactive audio interview simulations into a cohesive, decoupled microservice platform. Phase one of the framework processes multi-dimensional vocational preferences and personality metrics—captured via a standardized 50-item Big Five (OCEAN) inventory, a 48-item RIASEC model survey, and a distinct cognitive reading-comprehension assessment. These vectors advance through a dual-pipeline machine learning ensemble, integrating K-Means clustering and Soft-Voting Logistic Regression, to empirically predict optimal technical career trajectories. Phase two translates these discriminative mathematical predictions into highly customized, dynamically generated 10-day micro-learning roadmaps utilizing Google Gemini Large Language Models (LLMs) and strict Pydantic schema validations to ensure structured, hallucination-free knowledge acquisi-tion. Finally, phase three evaluates the user's operational readiness through an automated, audio-based simulation encompassing both domain-specific tech-nical and behavioral (HR) interview rounds. This simulation is facilitated by low-latency transcription (STT) via Int8-optimized faster-whisper and offline Text-to-Speech (TTS) technologies alongside a LLaMA-based response evalu-ation engine. Experimental evaluations utilizing a synthetic Gaussian dataset demonstrated exceptional predictive accuracy (exceeding 99.8%) for the core machine learning models, alongside robust architectural scalability and stable real-time generative capabilities across all subsystems. Ascend AI rigorously val-idates the efficacy of intelligently orchestrating modular, polyglot microservices to democratize personalized career coaching, successfully bridging the historical gap between abstract statistical recommendations and actionable, skill-building developmental pathways.

Keywords: Artificial Intelligence, Career Guidance, Psychometrics, Predictive Modeling, Generative LLMs, Adaptive Learning, Interview Simulation, Polyglot Microservices

I. INTRODUCTION

Computer Science and IT are an expanding and evolving area that encompasses a wide variety of specialized disciplines; however, some of the most common are full-stack software development, artificial intelligence, data engineering and project management architectures. As such, many students entering these fields have little to no exposure to the diversity within their relevant discipline or a robust programmatic immersion, making the process of determining which area to specialize in very subjective and often challenging. The manner in which career commitment decisions made in the context of academic and other institutional settings are made are oftentimes based on the influence of peer groups, infatuation with current terminology that is trending in the marketplace or simply on the latest market trends, rather than based on how well the individual's clinical profile — (e.g., personality traits, vocational interests, analytical reasoning preferences) and the operational requirements for performing within those environments are aligned. The result of this misalignment is consistently resulting in poor career choices, lack of long term career satisfaction, emotional challenges while studying, and an exceedingly high attrition rate throughout the various specialized technology fields.

The Limitations of Existing Guidance Paradigms

The traditional paradigms of career guidance provided by institutions (like "general" tests of talent, or normal means of providing manual guidance) are unable to provide the sufficient granularity or depth of self-reflection that is expected for an individual's modern educational setting. A review of the available literature shows that psychomet-ric tests used in isolation do not have adequate predictive validity going forward unless combined with multiple-dimensional assessments of cognitive performance metrics and applicable learning frameworks. Moreover, many current digital solutions do not build on the wealth of research that has already been published and therefore depend upon rigid, monolithic models (e.g., MBTI) and/or prohibitively expensive application programming interface (API) systems requiring excessive hardware investment.

An example of this is systems that rely exclusively upon high-cost, video-based emotional tracking, which would require an extremely high resolution web camera and a significant investment in additional hardware. Additionally, many resources limit their activity to either static algorithmic processing of talent variables or only use a single model like personality type as a measure of vocational capability. The end result of this neglect of research is that many online career guides and solutions simply stop after they provide a "recommendation" to someone based on a discrimination of their aptitudes but do not provide any support around developing a plan of action for their immediate employment in that technical field. Thus, individuals are left without an immediate and actionable means of evaluating, implementing, or pursuing their recommended career path in the technical field.

Proposed Framework: Ascend AI

In order to close these significant systemic gaps in orientation-based systems, this study proposes a new orientation ecosystem, Ascend AI, that combines kinetic input with the behavioral dimensions of psychometrics, generative micro-learning, and procedural preparation for interviews into an operational pipeline that is easy to access and supports open-source development.

Ascend AI will convert all multidimensional characteristics of people into quan-tifiable mathematical metrics before identifying a possible career path based on the identified dimensions of psychometric scales through a formulaic, data-driven career trajectory. Next, using all of the data created in the previous steps, Ascend AI will create a personalized, 10-day structured career roadmap for the individual using locally-based generative language models. The final step of Ascend AI will consist of a robust audio-based technical and behavioral interview simulation that provides an accurate assessment of someone's understanding of their career roadmap as well as their psychometric characteristics. This end-to-end approach to a career guidance solu-tion is a complete change in how traditional career guidance solutions work

because it has moved away from delivering a one-time, standalone solution to providing a con-tinuous, interactive, and actionable career development solution with an emphasis on building structural resilience and supporting individual pedagogy.

III. METHODOLOGY

Ascend AI uses a new decoupled, polyglot microservice architecture to fix the struc-tural problems, bottlenecks around compute and create a platform that enables the accurate modeling and analysis of user-generated data in real-time. Node.js will be paired with Express for asynchronous HTTP lifecycle management, while all math-heavy processing and AI inference will be offloaded to a separate Python/FastAPI environment.

The three primary components that comprise the holistic approach are (1) Predic-tive Psychometric Mapping, (2) Generative Micro-Learning Roadmap Orchestration, and (3) Procedural Audio Interview Simulation.

Phase One: Psychometric Evaluation and Predictive Modeling

To develop an evidence-based, statistically validated recommendation for a user's career, we need to first quantify the user's subjective user profile via a well-defined computational basis. The user will go through a process that embodies three empirical forms of quantitative psychometric and cognitive operations in an interactive, triadic combination via a secure React front end. The three empirical forms of quantitative psychometric and cognitive operations used to quantify a user's subjective profile will include a 50-item Big Five (OCEAN) personality inventory that measures the user's degree of Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness; a 48-item RIASEC career interests inventory, which identifies the degree of Realistic, Investigative, Artistic, Social, Enterprising, and Conventional career inter-ests; and a complex cognitive reading comprehension activity specifically designed for abstract pattern recognition measurement, implicit problem-solving definitions, and structural management definition and examples.

Synthetic Dataset Generation and K-Means Initialization

Due to the lack of commercially viable large-scale organic datasets relating very detailed psychometric profiles to their associated dominance trajectories in computer science careers, an empirical synthetic data generation protocol was applied. By using established statistical central tendencies from rigorous O*NET industry descriptor matrices, thousands of user profiles were mathematically interpolated using a strong Gaussian variance distribution.

An unsupervised K-Means clustering algorithm was then applied to parse these generated raw feature vectors and mathematically identify unlabelled behavioural pop-ulations based on Euclidean similarity bounds. By using the centroid coordinates as labels from this clustering analysis, the populations were categorised against canonical O*NET career archetypes by segmenting them into core technical classification such as Software Development, Data Science & Machine Learning, and Technology Project Management, thereby creating a new dataset for use as a foundation for training purposes without introducing any arbitrary bias.

The Dual-Pipeline Ensembling Architecture (Soft-Voting)

After developing unsupervised K-Means parsing methodologies, a dual-pipeline high-performing supervised classification engine was implemented using Logistic Regres-sion. Logistic Regression was preferred over Support Vector Machines and Decision Trees because it produces continuous probability distributions rather than binary out-puts (0 or 1). Therefore, it has lower computational requirements for API integration. In addition, due to the presence of psychological variability within the dataset, Logistic Regression reduces overfitting.

Each cognitive and psychometric metric is processed separately through different machine learning pipelines to maintain context-sensitive integrity:

Pipeline 1 (Psychometrics): Uses an initial Logistic Regression model to gen-erate a psychometric probability distribution (P1) by processing the Big Five and RIASEC data.

Pipeline 2 (Cognitive): Uses a second Logistic Regression model to generate an independent probability distribution (P2) from analytical features obtained through reading comprehension tests.

To mitigate bias and anomalies arising from a single metric source, a soft voting ensemble combines outputs from both pipelines. The final confidence score is computed as:

FinalConfidence = P1 + P2

2

After applying the arg max operation, the career path with the highest probability is selected, completing Phase 1.

Phase Two: Generative Micro-Learning Roadmap Orchestration

A switch from using only the predictive property of the first group of variables to being able to employ some predictive property of both the first group of variables and the second group of variables leads to modifying the structure of the domain within the framework of AI and J.B. Willoughby's Ascend AI system. The specific objective is to change an existing label (e.g., "Data Engineering and AI") produced by a data structure and convert it into a long sequence of progressively structured micro-learnings as well as to synchronise the total structure to fit the individual's learning speed, experience/learning level, and area of interest.

Pydantic Schema Application and Gen-AI Framework

The shift in our programming approach involves using a stand-alone microservice that uses LangChain to orchestrate all parts of our AI platform and works directly with advanced, integrated Google Gemini. Our use of generative modelling techniques has an inherent risk of generating false positives (also known as "hallucinations "), where language models produce syntactically correct statements that may appear true but

Fig. 1 Data Flow Architecture Mapping of Phase 2: Generative orchestration and dynamic roadmap formulation.

are not, and may also produce incorrectly formed array structures, which could result in the frontend client crashing.

To address hallucinations and the decay of stringified data permanently, all queries submitted to the LLM backend are fully contained within immutable Python Pydantic configurations. This ensures complete restriction of the Language Module and requires complete compliance with the limitations on generating precision JSON output based on the following predetermined criteria: operational goals defined for each operational day of the week; educational resources defined for each day (via enums based on type: article, video, code exercise); difficulty associated with completing each assigned educational resource assigned on a progressive difficulty basis for each of the 10 days.

Live DuckDuckGo Contextual Enrichment and Output Sanitization

Even though there are schema constraints in LLMs that prevent invalid architec-tural structuring, language models have demonstrated an ongoing liability related to URL validation, as they are excessively generating expired or nonexistent URLs to educational video resources. Ascend AI has implemented a concurrent automated enrichment utility to ensure that all resources are real-time current, using desig-nated learning objects. This framework uses a local ThreadPoolExecutor that parses a DuckDuckGo web scraper into a pipeline, and has the ability to dynamically exe-cute secondary queries for each generated topic against a specific domain modifier. The results of each scraping operation result in resources that contain only verified links to external educational resources, and invalid base-model URLs are overwritten by the verified results and added to the standard JSON payload that is returned to the React-based user interfaces.

Day-Gating Protocols and Idempotency Architecture

To be able to provide a means to validate users as they progress through their cus-tomized content creation experience, it is important to design granular user validation systems and enforce strict gating topologies. The Ascend AI network

has developed a unique structure where subscribers must complete a schedule of daily tasks, on a one-by-one basis. Subscribers must be able to validate completed resource use and engagement across various media layers for every task.

In order to facilitate the progression of multiple daily tasks, learners must complete an AI-generated multi-choice assessment module for each task type that corresponds with the complexity level of the core curriculum. The learner's score from the multi-choice assessment module must exceed a minimum of 70

In order to challenge the structural complexity of deep nesting and the dynamic unconstrained nature of generative JSON roadmaps, we implemented MongoDB Atlas as our back-end data storage platform to support infinite adjustments of documents within the document layer. Since many requests from end users are required to read data to track daily progress, we are using Redis as an in-memory data store on a separate node to cache data in parallel to the document storage.

However, because the tracking of progress variables is performed within the con-text of dense documents, there is a risk that the same document will be accidentally duplicated. Therefore, to ensure each of these UI completion actions are appropriately recorded and saved, our database operations are establishing an idempotent-write paradigm to guarantee that duplicate UI completion actions will always overwrite existing data without inducing any unmanageable redundant data in the overall system.

Phase Three: Interactive AI Interview Simulation Mechanics

The overarching research method used to quantify absolute user competency was achieved through a continuously automated means of running synchronous inter-view processes, as they occur in real-world enterprise operational environments. This methodology moves dramatically away from applying static text-banking method-ologies that have traditionally been restricted to the physical infrastructures of legacy systems by leveraging the extraordinarily powerful, fully programmable Groq-hosted LLaMA-3.3 70 billion parameter model interface algorithms within the Python/FastAPI development environment to instantaneously build contextually appropriate interview dialogue for the candidate based upon real-time parameters provided via the interview process. To ensure accurate technical query genera-tion, interviewee's Phase One Career Prediction mapping will be directly used and matched against their Phase Two generated Trajectories. Also, the interviewee will be specifically targeted based upon their psychometric deficiencies by leveraging the employment history information from phase three to target the interviewee's deficien-cies obtaining the best available bespoke assessment processes via the candidate's Big Five performance metrics and RIASEC deficiencies.

Localized Audio FFmpeg Processing and Low-Latency Infrastructure

How eliminating keyboard inputs through the provision of continual voice interac-tion or continuous conversation capability through high-fidelity audio playing together through an easily defined operating schema with the frontend of an FSM.

Fig. 2 Sequential Process Flow Diagram demonstrating the STT/TTS multi-modal transcription and response analysis architecture.

The user voice is recorded into MediaRecorder instances in the app's browser, using Web Audio API for noise reduction (which allows for the identification of the absence of user voice for approximately 2.8 seconds prior to the user speaking). It is then immediately converted into a Base64 representation which is sent to internally located transcription endpoint using secure HTTP communication.

Upon being received successfully via API, FFmpeg processes each generic WEBM audio file as a 16kHz PCM formatted audio variable highly optimized for imme-diate transcription by a localized array of Int8-based faster-whisper CPU engines without requiring significant GPU resources to perform the transcriptions. Transcrip-tion results are produced at an extremely high level of accuracy relative to rapid

changes in the cadence of conversational speech and thus provide almost instantaneous transcription performance metrics.

Holistic Feedback Loops and Neural Voice Synchronization

The LLaMA-3.3 layer analyses the user's input data to identify discrete logical sub-organizations of the input data. This is accomplished through the explicit analysis of narrative coherence at the logical level, the calculated identification of any techni-cal elements related to the request that were not previously addressed in the original questions, and extraction of deep cognitive insights from the original user's input. The LLM then defines adaptive protocols, which include any follow-up statements required to resolve any of the unanswered questions completely. Qualitative feedback, including detailed grade percentages as well as detailed qualitative domain-level improvements for every individual user parameter in the data model, are extracted and cascaded backward through the structure of the output. Simultaneously, outputs generate com-pletely offline, neural-synthesized audible representations of the output generated using the PyPI Piper TTS matrix, producing extraordinary en US-lessac medium ONNX voice representations, and creating accurately simulated conversational flows with no external third-party API transaction delays.

IV. RESULTS AND DISCUSSION

All the different types of algorithms in the Ascend AI pipeline were individually and thoroughly tested under structural load to ensure they meet quantitative integrity requirements, have the ability to maintain scalability, have latency specifications for automated voice systems, and provide sufficient generations for platform operations in general education environments.

Unsupervised Clustering Metrics and Supervised Modeling Valuations

3,000 artificially created data sets were used for exploratory analysis based upon cluster analysis through K-Means clustering. Distinctly separated on the basis of separate clusters derived from dimensional analysis have defined technical occupation categories through the mapping of complex human behaviours. Objective quality measurements of cluster effectiveness revealed dramatic separations between distinct data sets with a total average of 0.534 for the Silhouette Scores for the OCEAN/RIASEC variables and 0.375 for the cognitive differences in the data set (reading-comprehension). A solid initial data set without overlap provided adequate evidence that the defined cluster separated data sets were reflective of science/engineering and technology degrees.

Secondary tests were conducted that focused on retrieving isolated parameter val-ues in the Dual Logistic Regression matrix arrays to demonstrate that the empirical algorithms outperformed competing nonlinear algorithm solutions based on the use of the test array matrices to produce cross-validation classification accuracy test results of 99.87% to 100.00%. In addition, since the extraordinary performance is represen-tative of the mathematical stability associated with the use of artificial training data

Fig. 3 Data Flow Architecture Mapping of Phase 1: Psychometric Evaluation and Unsupervised ML Processing.

variant distributions drawn from noise controlled (mitigated) Gaussian distributions, all results demonstrate solid application routing behavior where there was absolutely no catastrophic pipeline failures nor algorithm boundary failures. Projections for for-ward progression indicate that configuration stability will be achieved as soon as real world empirical efficiency models converge in the range of 85% to 91% accuracy and once they have been subjected to sequential episodes of increased outdoor use exposure variables (for example psychometric unpredictability) within operating cycle frameworks (i.e., timeframe) will have been exhausted.

Table 2 Algorithm performance and structural evaluation metrics

Note: Logistic Regression performance reflects baseline training on strict Gaussian synthetic distributions mapping ONET parameters.

Generative Load Output Sanitization

The Gen-AI orchestration framework has operated successfully under continuous heavy load conditions that simulate large numbers of concurrent users without experi-encing major application downtime problems like traditional monolithic applications. By using strict Pydantic semantic validations on schemas created by LLMs, it was possible for there to be no physical LLM schema violations (i.e., “hallucinations”). LLMs generated links that were not valid for a specific DuckDuckGo search by exe-cuting multiple threads simultaneously through the use of the ThreadPoolExecutor during web scraping tasks. Every LLM-generated link produced valid URL parame-ters nested appropriately within a 10-day structured roadmap to provide for perfect render synchronization to fulfil functional pedagogical navigation tasks.

Audio Pipeline Operability Profiles

The parameters of the testing confirm that the logic maps of the conversational state of the localized modifications to FFmpeg were executed through slower-than-whisper and neural-voice instance implementations of the application against the established enterprise operational latency thresholds and exhibit very minimal latencies that are equivalent to those operational thresholds resulting in this implementation removing the entire set of application dependencies that were dependent on cost quantities of STT or TTS cloud-based subscription services, demonstrating that the high-speed.

Int8 implementation’s dependence on local CPU compute resources was able to deliver timely and reliable performance of continuous real-time interaction processing by matching user input with the immediate completion of the previous user input and pro-viding reliable performance constantly governed only by true silence-detection-based detection rules while seamlessly coordinating asynchronous frontend interactions.

Discussion

The holistic performance indicators achieved through integrating the Ascend AI algorithm paradigms are fundamentally validating an incredible paradigm shift that profoundly reorients the general occupation orientation from an isolationist model to a highly interactive and dynamically configured/educationally developed platform or system of practice. Robust and flexible structural segregation of complex network architecture is further emphasized through extensive interpretive analysis/rule sets governing the robust integration of high-speed, asynchronous Node.js routing mech-anisms where multiple concurrent states are constantly being created and destroyed relative to their long-disposed counterpart, or PYTHON/FLASK operations per-forming mathematically-complex calculations and storing data on active and inactive System processes within using very large transactional memory resources associated with NumPy operations simultaneously ensure that the failure of any single point in the system will not lead to a systemwide failure because of the failure of any specific piece of hardware that is in multiple geographic locations, since the processing and memory resources associated with each unit are entirely separate yet interconnected based on interoperability principles.

Moreover, the requirement for a more complete representation of the real world, with respect to comprehensive educational mappings, necessitated the creation of algorithms that mathematically constructed standardized variations based on the application of robust O*NET baseline matrix criteria, such that the resulting algorith-mic initializations contained no inherent variability from the subjective investigator bias that may have been present in previous studies validating the non-cloud reliant configuration of algorithms that calculate complex voice matrix mechanics, thus providing a rational justification for the deployment of intelligent neural audio com-ponents within a localized environment inside of core computational servers in order to overcome the dependencies of the entire system on the pricing of premium com-mercial third-party cloud infrastructures, by using Ascend AI to optimize total costs for use by institutions, while maintaining the integrity of the computing environment so as to avoid the occurrence of any type of cost cascades.

V. CONCLUSION

Ascend AI has created a complete, interconnected framework of artificial intelligence that is multi-dimensional in its capabilities and will dramatically alter the way general occupational advising has been approached by many complex institutional technol-ogy studies around the world. By integrating the ultimate analytical accuracy and exact quantitative probabilities that will be established through verified psycholog-ical machine learning-based assessments, in addition to the generative power of the

schemas that strictly limit the application of different forms of generative language model data, and the use of those schemas to create individualized education plans for each student enrolled in an institution of higher education, Ascend AI provides a com-plete solution to the many challenges associated with creating effective advising models through traditional means of providing such support systems. By doing so, Ascend AI has redefined how to utilize traditional and/or normative advising models to assure that each candidate's overall performance, within an institution of higher education, is assessed using appropriate methodologies based on their behavioral characteristics. When creating clarity around the paths of evolving systems over time, we cur-rently focus substantially on identifying some of the limiting baseline constraints that restrict the system's functionality, which require that the system must rely completely on a current active online network to operate. Therefore, it is impossible to maintain any form of localized decentralized operations in these systems while simultaneously navigating through unreliable networks, as well as adhering to basic evaluative mea-surement criteria for developing enhanced versions of the same system to incorporate the subjective complexities involved in these systems and to have them both assessed

at the same time in a more advanced manner.

The immediate priority of modifying these systems is to deploy fully autonomous complex adaptive learning engines that will create rigorous performance metrics that will inform the associate's history in developing accuracy or performance standard metrics that will represent with increasingly high accuracy how long it will take to develop a set of sufficient data within the system and that the user will be able to access a new and unique set of data related to the development of new technologies that provide the user with the ultimate technological access architecture and provides the user with the maximum technological access capabilities throughout their entire life.

Declarations

Funding: No funding or financial sponsor received for the writing of this article.

Conflict of interest/Competing interests: All authors report no relevant competing interests or financial interests at this time.

Ethics approval and consent to participate: Not applicable to the present

study as it does not involve doing research involving either human subjects or animal research.

Consent for publication: Not applicable.

Data availability: The datasets created and/or analysed during the course of this study (specifically molecular K-Means arrays) are available upon request to the corresponding author.

Materials availability: Not applicable.

Code availability: The code and generative pipelines used in the research done for this article are available upon request to the corresponding author.

Author contribution: All authors made equal contributions in the conception

and design of the study. All authors worked collaboratively on material preparation,

data collection and analysis. All authors contributed to writing the paper, as well as read and approved the final paper prior to submission.

REFERENCES

Abhidarsh, K. S., & Chacko, M. (2025). AI-powered career guidance through psy-chometric and ML analysis. Zenodo, Amal Jyothi College of Engineering Research Proceedings.

Abrar, M., Aboraya, W., Abdulghafor, R., Subramanian, K., & Al Husaini, Y. (2025). AI-powered learning pathways: Personalized learning and dynamic assess-ments. International Journal of Advanced Computer Science and Applications, 16(1).

Batista, J. S., & Gondim, S. M. G. (2022). Personality and person–work envi-ronment fit: A study based on the RIASEC model. International Journal of Environmental Research and Public Health, 20(1), 719.

Bebale, P., Yadav, S., Surve, S., Sayed, A., & Korgaonkar, G. (2025). Career com-pass: AI-based career counselling. International Journal of Innovative Research in Technology, 11(11).

Bellard, F. (2023). FFmpeg comprehensive multimedia framework and audio decoding. FFmpeg official documentation. https://ffmpeg.org/

Dietterich, T. G. (2000). Ensemble methods in machine learning: Soft voting constraints. Multiple Classifier Systems Journal, 1–15.

FastAPI Contributors. (2023). FastAPI: A high-performance web framework for APIs. FastAPI documentation. https://fastapi.tiangolo.com/

Google Developers. (2023). Google identity: OAuth 2.0 authentication architec-ture. Google Cloud Platform. https://developers.google.com/identity/

Guettala, M., Bourekkache, S., Kazar, O., & Harous, S. (2024). Generative artifi-cial intelligence in education: Advancing adaptive and personalized learning. Acta Informatica Pragensia.

Harrison, C. (2023). Pydantic schema constraints and JSON structural validation for Python models. Python Engineering Journal.

Holland, J. L. (1997). Making vocational choices: A theory of vocational person-alities and work environments (3rd ed.). Psychological Assessment Resources.

Jagtap, R., et al. (2025). AI-driven real-time interview simulation app. Interna-tional Journal for Research in Applied Science and Engineering Technology.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of Personality: Theory and Research, 2, 102–138.

Jones, M., Bradley, J., & Sakimura, N. (2015). JSON web token (JWT) cryptographic standards. Internet Engineering Task Force (IETF) RFC 7519.

Khapekar, S., Bothara, S., Babar, T., & Kine, R. (2025). AI-driven smart inter-view simulator with real-time speech and emotion analysis. TIJER – International Research Journal, 12(3).

Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 5530–5540.

LangChain Team. (2023). LangChain python documentation and orchestration for LLMs. LangChain AI. https://python.langchain.com/

Li, G., Hammoud, H. A., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative agents for mind exploration of large-scale language model society. arXiv preprint arXiv:2303.17760.

Li, Y., et al. (2024). Investigating how generative AI can create personal-ized learning content. Journal of Educational Technology Systems and LLM Implementation.

Liang, P., et al. (2023). Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291.

MacQueen, J. (1967). Some methods for classification and analysis of multivari-ate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281–297.

Mardan, A. (2014). Pro Express.js: API routing and middleware architectures. Apress Publications.

Meta AI Research. (2024). LLaMA 3 model card, versatile inference parameters, and architecture documentation. Meta AI. https://ai.meta.com/llama/

MongoDB Inc. (2023). MongoDB non-relational database models and BSON document storage. MongoDB manuals. https://www.mongodb.com/docs/

Pandeya, S., et al. (2025). AI-based interview system with machine learning integration. International Journal of Computer Applications.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleish-man, E. A. (1999). An occupational information system for the 21st century: The development of ONET*. American Psychological Association.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever,

I. (2023). Robust speech recognition via large-scale weak supervision (Whisper).

Proceedings of the 40th International Conference on Machine Learning.

React Team. (2023). React official documentation: Component-based UI and state management. Meta Open Source. https://react.dev/

Redis Labs. (2023). Redis documentation: In-memory data structuring and caching persistence. Redis Labs. https://redis.io/docs/

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and val-idation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflex-ion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36.

Tilkov, S., & Vinoski, S. (2010). Node.js: Using JavaScript to build high-performance network programs. IEEE Internet Computing, 14(6), 80–83.

Tomassetti, F. (2023). Pydantic schema implementation and dynamic runtime data validation. Software Engineering Journals.

Tomkins, S. (2024). The adoption of monolithic versus isolated REST microser-vices in web architecture. Cloud Computing Research.