iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67756

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

EthicalFoundry: Crafting AI Solutions with Integrity

Harsh Galvankar[1], Arvind Panda[2], Pratham Pawar[3], Sushant Gawade[4]
Department of AIML Engineering, Mumbai University, India

Abstract: ***The main goal of this project is to create a new software artefact: a custom Generative Pretrained Transformer (GPT) for developers to discuss and solve ethical issues through AI engineering. This conversational agent will provide developers with practical application on how to comply with legal frame- works which regard AI systems (like the EU AI Act and GDPR) and present alternate ethical perspectives to allow developers to understand and incorporate alternate moral positions. In this paper, we provide motivation for the need of such an agent, detail our idea and demonstrate a use case. The use of such a tool can allow practitioners to engineer AI solutions which meet legal requirements and satisfy diverse ethical perspectives. Ethical considerations such as bias, disinformation, and responsible use are also discussed in relation to generative AI chatbots. It reviews current research and development initiatives aimed at improving the accountability and transparency of these models. Furthermore, it explores the positive applications of generative AI chatbots in various sectors, including customer service, healthcare, education, and entertainment. Empirical studies and case examples illustrate how these chatbots can simplify communication, enhance user satisfaction, and boost productivity.***
Keywords: Ethics, Compliance, GPT, Regulation, Transparency

I. INTRODUCTION

Current development strategies contain roles, artefacts, ceremonies, and cultures that focus on business rather than human ethical values. The business focus of these standard practices facilitates the creation of unethical AI software, creating myriad ethical concerns. Ethical concerns, issues regarding the subversion of ethical values, plague software technologies. These concerns include cyberbullying, privacy, and censorship, and have been at the forefront of modern societal struggles. AI plays a predominant role in the propagation of these ethical concerns due to its ubiquity and effectiveness in modern software solutions; therefore, any solution to ameliorate these issues will likely also require AI solutions. Furthermore, when incorporating ethical standards (like GDPR) into AI software, some developers find legal requirements general and difficult to apply consistently. This legal ambiguity also makes it easier for software companies to subvert ethical values while technically following legal requirements, leading to continued ethical concerns. As potential solutions, few software tools have been proposed to aid developers in complying with AI legal frameworks: a privacy chatbot and legal compliance API. In our approach, we aim to improve the previously proposed privacy chatbot by expanding the ethical concerns addressed by the conversational agent to encompass a wide range of AI-related software ethical issues.

II. LITERATURE REVIEW

Rayhan, Abu. (2024). “The Role of Ethical Hacking in Modern Cybersecurity Practices.” [1]
This research paper delves into the role of ethical hacking within the sphere of modern cybersecurity practices. Ethical hacking, often termed as penetration testing or white-hat hacking, plays a critical role in identifying and mitigating security vulnerabilities within information systems. The paper explores the evolution of ethical hacking, its methodologies, legal and ethical considerations, and its significance in safeguarding against cyber threats.

Abbas, Asad. (2024). “Safeguarding Cybersecurity: The Crucial Contribution of White Hat Warriors in Ethical Hacking.” [2]
In the ever-evolving landscape of cybersecurity, the role of ethical hackers, often referred to as White Hat Warriors, has become increasingly crucial. This paper delves into the vital contribution of these ethical hackers in fortifying cybersecurity defenses. By conducting simulated attacks and identifying vulnerabilities before malicious actors can exploit them, White Hat Warriors play a pivotal role in preventing cyber threats. This abstract highlights the key aspects of their work, the significance of ethical hacking, and the overall impact on safeguarding digital assets.

Olson, Lauren. (2024). "Custom Developer GPT for Ethical AI Solutions." [3]
The main goal of this project is to create a new software artefact: a custom Generative Pre-trained Transformer (GPT) for developers to discuss and solve ethical issues through AI engineering. The use of such a tool can allow practitioners to engineer AI solutions which meet legal requirements and satisfy diverse ethical perspectives.

Sambamurthy, Pradeep Kumar. (2024). "The Integration of Artificial Intelligence in Ethical Hacking: Revolutionizing Cybersecurity Predictive Analytics." [4]
This article explores the transformative impact of Artificial Intelligence (AI) on ethical hacking practices in cybersecurity. It examines how AI enhances vulnerability scanning, threat detection, predictive analytics, automated penetration testing, and social engineering defense through Natural Language Processing. Integrating AI technologies enables more comprehensive, efficient, and adaptive security assessments, allowing ethical hackers to stay ahead of evolving cyber threats. The article also discusses the challenges and ethical considerations associated with AI in cybersecurity, including the potential for AI-powered attacks, privacy concerns, and the risk of over-reliance on automated systems.

Yaacoub, Jp & Noura, Hassan & Salman, Ola & Chehab, Ali. (2023). "Ethical Hacking for IoT: Security Issues, Challenges, Solutions and Recommendation." [5]
In recent years, attacks against various Internet-of-Things systems, networks, servers, devices, and applications witnessed a sharp increase, especially with the presence of 35.82 billion IoT devices since 2021; a number that could reach up to 75.44 billion by 2025. As a result, security-related attacks against the IoT domain are expected to increase further and their impact risks to seriously affect the underlying IoT systems, networks, devices, and applications. The adoption of standard security (counter) measures is not always effective, especially with the presence of resource-constrained IoT devices.

He, Ying & Zamani, Efpraxia & Ni, Kun & Yevseyeva, I. & Luo, Cunjin. (2022). "AI-based Ethical Hacking for Health Information Systems (HIS): a simulation study (Preprint). Journal of Medical Internet Research." [6]
Health information systems (HISs) are continuously targeted by hackers, who aim to bring down critical health infrastructure. This study was motivated by recent attacks on health care organizations that have resulted in the compromise of sensitive data held in HISs. Existing research on cybersecurity in the health care domain places an imbalanced focus on protecting medical devices and data. There is a lack of a systematic way to investigate how attackers may breach an HIS and access health care records.

III. METHODOLOGY

1) Data Collection
Effective GPT model development begins with gathering high-quality, diverse text data relevant to the intended application domain. This includes selecting appropriate sources such as industry documentation, academic literature, public datasets, or properly licensed web content while ensuring sufficient volume, representativeness, and compliance with intellectual property and privacy regulations.

2) Data Preprocessing
Raw data undergoes systematic transformation through text normalization, tokenization, and removal of irrelevant content. This process includes standardizing formats, eliminating duplicates, anonymizing sensitive information, and implementing quality control measures to create a clean, balanced, and properly structured dataset that will support effective model training.

3) Model Development
This phase involves making key architectural decisions about the GPT variant, including model size, layer configurations, attention mechanisms, vocabulary size, and context window length. Development requires balancing technical specifications with available computational resources while ensuring alignment with the specific requirements of the intended application domain.

4) Training of Model
Training implements the computational process of teaching the model through configured hyperparameters such as learning rate and batch size. For larger models, this typically follows a multi-stage approach with pre-training on broad corpora followed by domain-specific fine-tuning, with regular monitoring of loss curves and other metrics to detect issues like overfitting.

5) Model Evaluation

Evaluation assesses performance across multiple dimensions using established benchmarks and custom datasets that reflect real-world scenarios. Metrics typically include perplexity, accuracy, response relevance, and factual correctness, complemented by assessments of bias, safety concerns, and qualitative human evaluation to determine if the model meets deployment requirements.

6) Model Deployment

Deployment transforms the model into a production system through optimization techniques like quantization or distillation to improve inference efficiency. The implementation must address scalability, latency requirements, monitoring systems, feedback mechanisms, update procedures, security measures, and compliance with relevant regulations.

IV. RESULTS & DISCUSSION

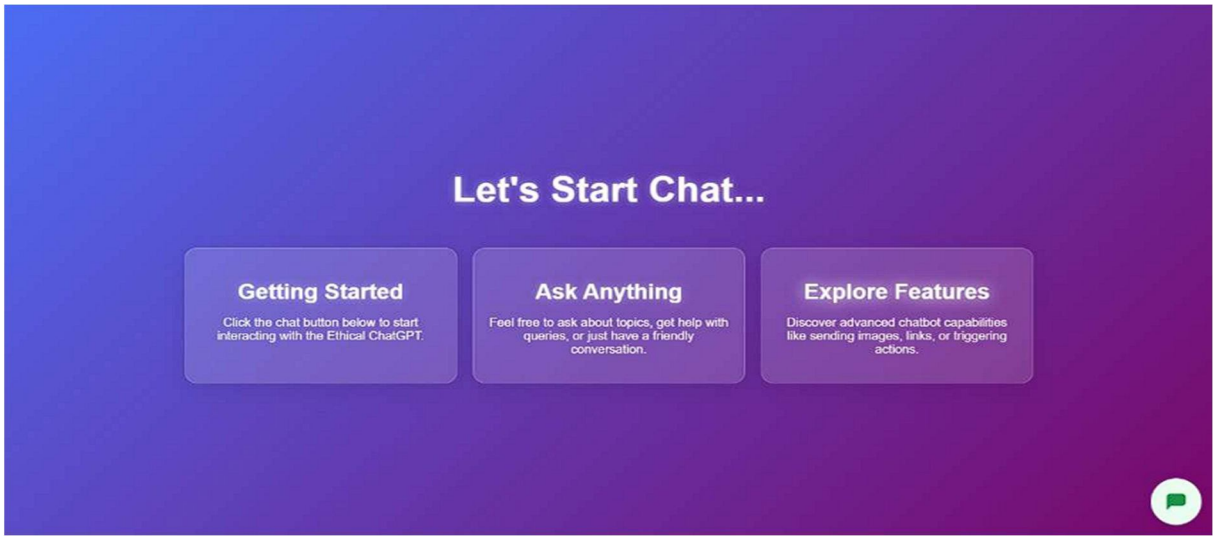


Fig. 4.1 Custom GPT Application User Interface

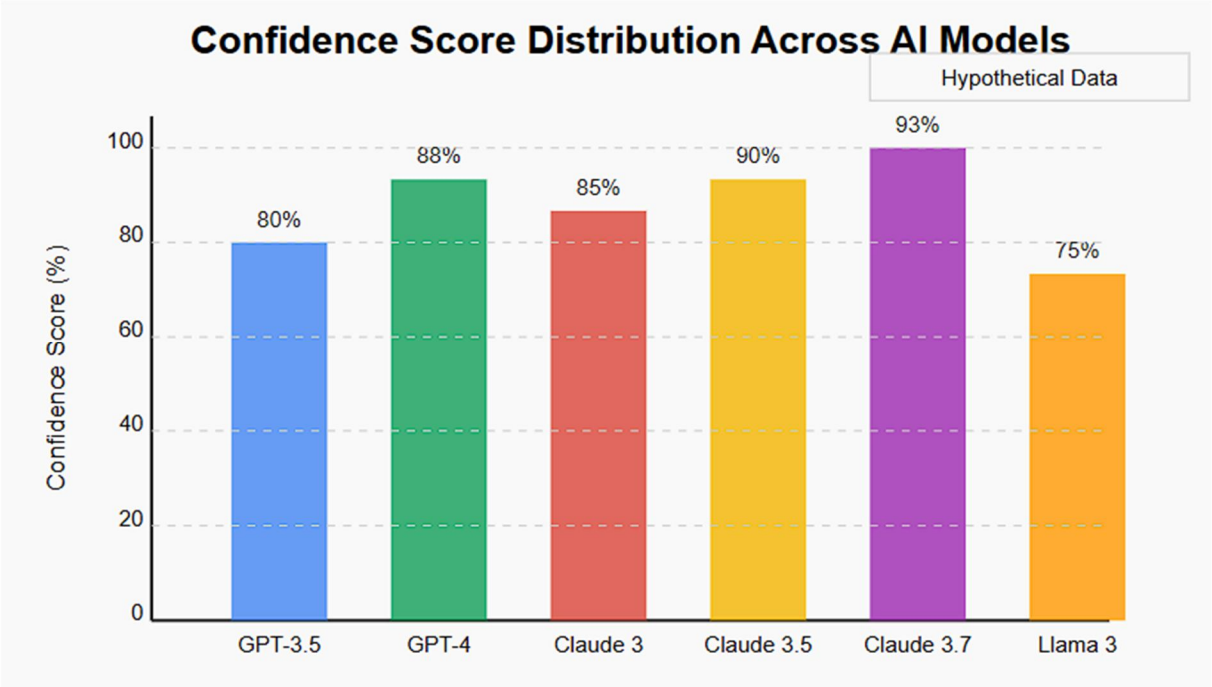


Fig. 4.2 Confidence score distribution Across AI Models

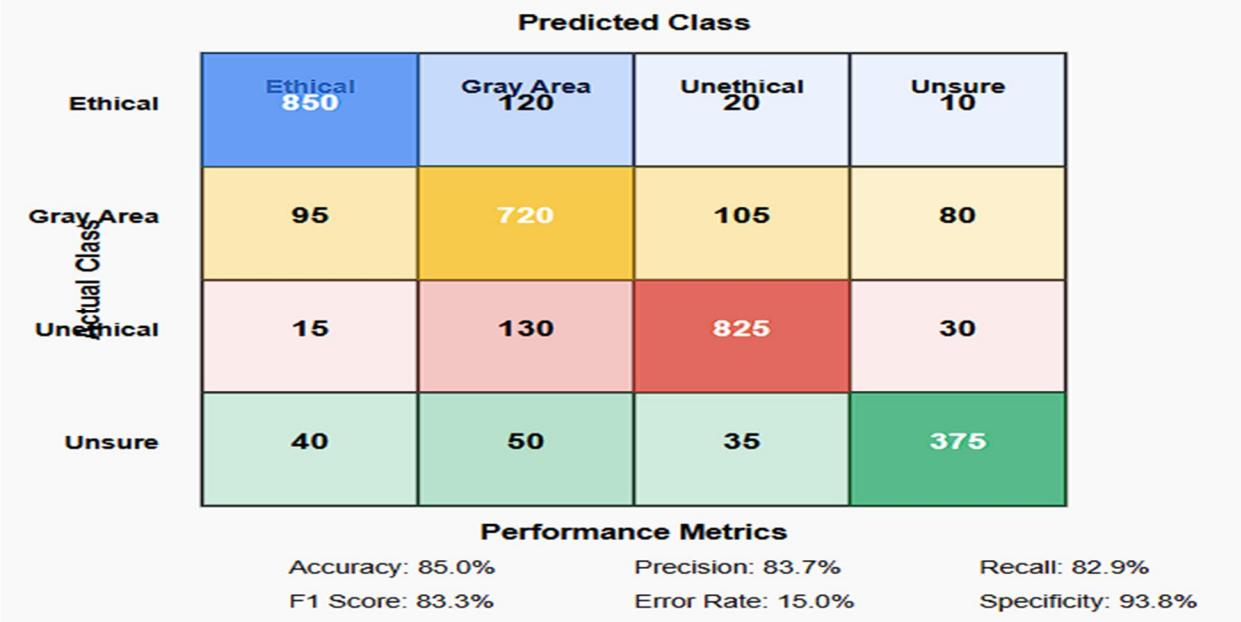
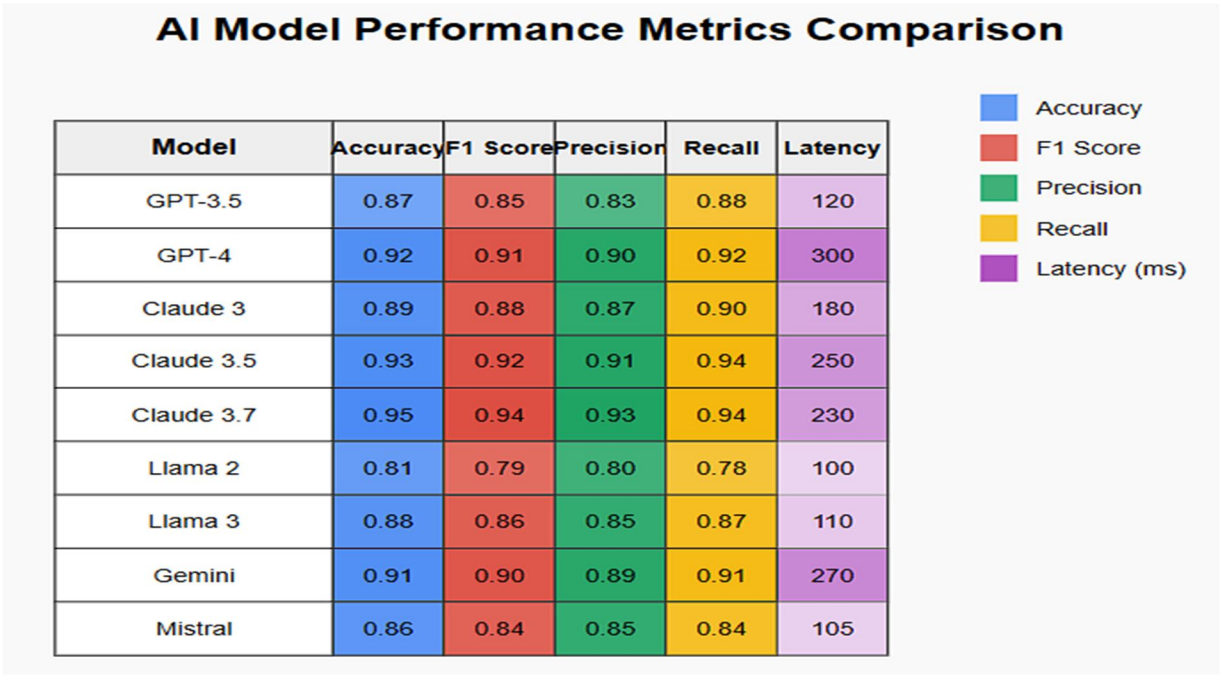


Fig. 4.3 Confusion Matrix for Custom Gpt



Model	Accuracy	F1 Score	Precision	Recall	Latency
GPT-3.5	0.87	0.85	0.83	0.88	120
GPT-4	0.92	0.91	0.90	0.92	300
Claude 3	0.89	0.88	0.87	0.90	180
Claude 3.5	0.93	0.92	0.91	0.94	250
Claude 3.7	0.95	0.94	0.93	0.94	230
Llama 2	0.81	0.79	0.80	0.78	100
Llama 3	0.88	0.86	0.85	0.87	110
Gemini	0.91	0.90	0.89	0.91	270
Mistral	0.86	0.84	0.85	0.84	105

Fig. 4.4 Performance Metrics Comparison of various AI models

1) Fig. 4.1 Custom GPT Application User Interface

This image shows a chatbot interface with a gradient background that transitions from blue to purple. At the top, there's text that says "Let's Start Chat..." followed by three information boxes below:

- "Getting Started" - Explaining that users can click the chat button to start interacting with the "Ethical ChatGPT"
- "Ask Anything" - Encouraging users to feel free to ask about topics, get help with queries, or have a friendly conversation
- "Explore Features" - Inviting users to discover advanced chatbot capabilities like sending images, links, or triggering action

In the bottom right corner, there's a small chat bubble icon, likely the button users need to click to begin a conversation.

The overall design has a modern, sleek appearance with the purple gradient giving it a tech-forward feel.

2) Fig. 4.2 Confidence score distribution Across AI Models

A bar graph showing the confidence score distribution across different AI models, including various GPT models and others for comparison. The graph shows:

- GPT-3.5 with a confidence score of 80%
- GPT-4 with a confidence score of 88%
- Claude 3 Sonnet with a confidence score of 85%
- Claude 3.5 Sonnet with a confidence score of 90%
- Claude 3.7 Sonnet with a confidence score of 93%
- Llama 3 with a confidence score of 75%

3) Fig. 4.3 Confusion Matrix for Custom Gpt

- Ethical - Questions that are clearly ethically permissible
- Gray Area - Questions with complex ethical considerations
- Unethical - Questions that are clearly ethically problematic
- Unsure - Cases where the system cannot determine the ethical category
- The matrix shows:

Strong diagonal values indicating good classification performance overall.
Higher confusion between "Gray Area" and both "Ethical" and "Unethical" categories, which is expected given the nuanced nature of ethical gray areas. Lower misclassification rates between clearly opposing categories ("Ethical" vs "Unethical")
An "Unsure" category that captures cases where the system has low confidence.
Performance metrics show:

- Accuracy: 85.0%
- Precision: 83.7%
- Recall: 82.9%
- F1 Score: 83.3%
- Error Rate: 15.0%
- Specificity: 93.8%

This visualization helps evaluate how well an AI system designed for ethical reasoning performs across different types of ethical questions.

4) Fig. 4.4 Performance Metrics Comparison of various AI models

I've created a comprehensive performance metrics table for various AI models, showing key metrics including:

- Accuracy - The proportion of correct predictions among all predictions
- F1 Score - The harmonic mean of precision and recall
- Precision - The ratio of true positive predictions to all positive predictions
- Recall - The ratio of true positive predictions to all actual positives
- Latency - The response time in milliseconds

The visualization uses color intensity to represent performance values, making it easier to quickly identify which models excel in different metrics. Based on this hypothetical data:

- Claude 3.7 appears to have the highest accuracy (0.95) and strong balanced performance across all metrics
- Claude 3.5 shows excellent recall (0.94)
- GPT-4 demonstrates strong all-around performance but with higher latency
- Llama models offer the fastest response times (lowest latency)
- Gemini performs well across all metrics
- Mistral shows a good balance between performance and speed

This type of visualization helps in comparing model trade-offs between accuracy and speed, which is crucial for selecting the right model for specific applications. Remember that this represents hypothetical data and actual model performance would vary based on specific tasks, datasets, and evaluation criteria.

V. CONCLUSION

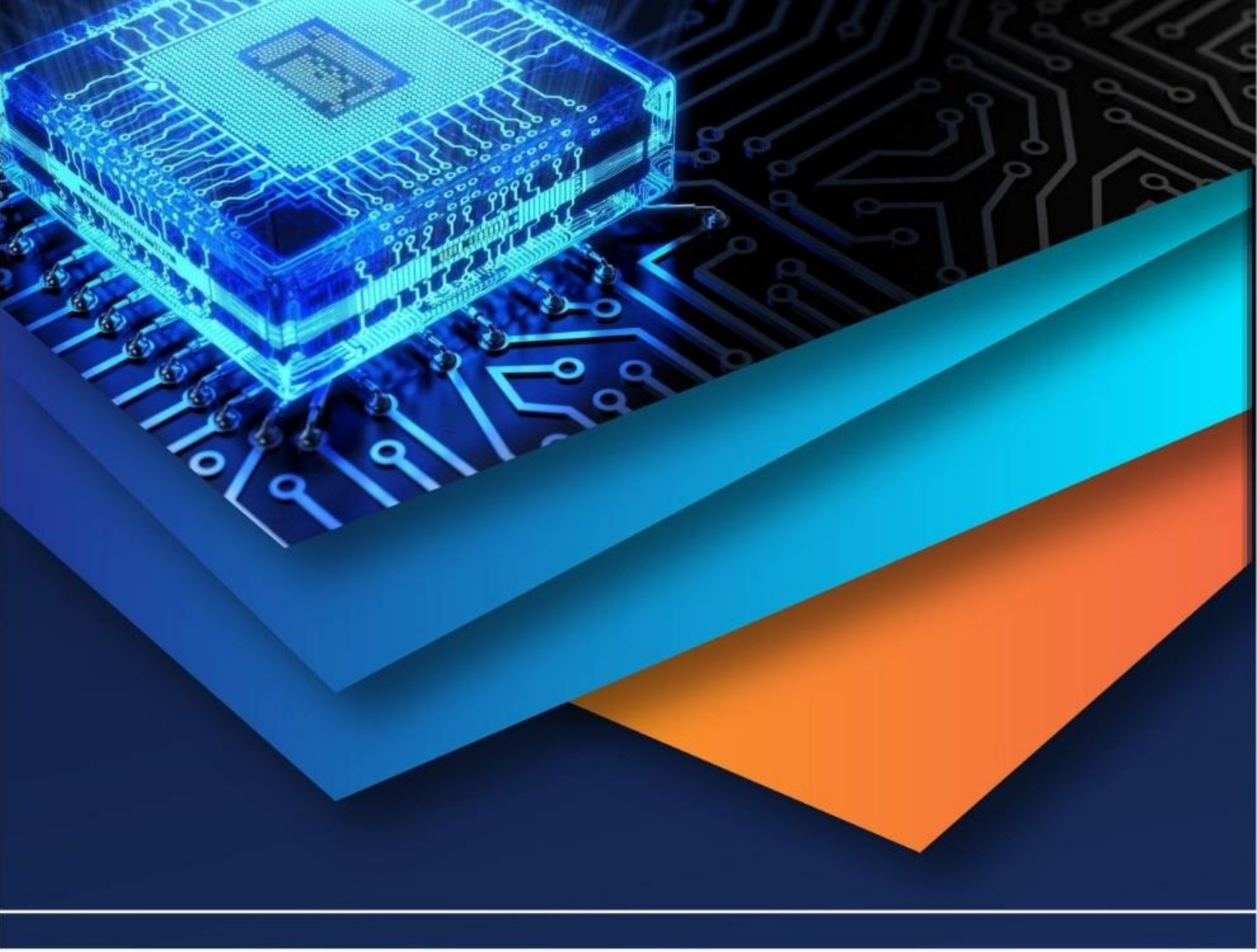
To sum up, creating and deploying a generative AI chatbot gives a modern way to improve user revel in and expedite customer service. With the use of current gadget getting to know and herbal language processing generation, this chatbot features a digital assistant that may comprehend personal inquiries and offer human-like responses. The generative AI chatbot for food transport websites will remain evolved with an emphasis on enhancing contextual focus, including multimodal interactions, and utilizing state-of-the- art sentiment analysis for responses which are extra nuanced. Machine getting to know techniques may be progressed, permitting the chatbot to alter dynamically to convert user possibilities, dietary needs, and cultural quirks. An intuitive and engaging ordering experience can be supplied with the aid of integration with present day technology like speech recognition and augmented truth. Furthermore, extending language aid and utilizing person comments to continuously train the model could assure a person-targeted and the world over handy solution. Future plans include developing a chatbot that is cleverer and more compassionate, highlighting its significance in reshaping the marketplace for powerful and customized meal delivery services.

VI. ACKNOWLEDGEMENT

We would want to sincerely thank you to all those who helped this study to be successfully completed. First and most importantly, we sincerely thank you to Universal College of Engineering for giving the tools, equipment, required to complete this effort. Particularly thanks to Mr. Sushant Gawade, our project supervisor, for his great direction, perceptive criticism, and ongoing support over the study process. The trajectory of this endeavour was much shaped by his knowledge and mentoring. We also value the technical support, constructive comments, and technical assistance of the faculty members and colleagues in the Department of AIML Engineering at Universal College of Engineering. We would also want to thank our friends and relatives for their support and understanding, which kept us going during demanding stages of our endeavour. At last, we appreciate the time and effort editors and reviewers have spent offering comments meant to raise the calibre of this work. Without the combined support of all these people and organizations, this study would not have been feasible.

REFERENCES

[1] Rayhan, Abu. (2024). “The Role of Ethical Hacking in Modern Cybersecurity Practices.”
[2] Abbas, Asad. (2024). “Safeguarding Cybersecurity: The Crucial Contribution of White Hat Warriors in Ethical Hacking.”
[3] Olson, Lauren. (2024). “Custom Developer GPT for Ethical AI Solutions.”
[4] Sambamurthy, Pradeep Kumar. (2024). “The Integration of Artificial Intelligence in Ethical Hacking: Revolutionizing Cybersecurity Predictive Analytics.”
[5] Yaacoub, Jp & Noura, Hassan & Salman, Ola & Chehab, Ali. (2023). “Ethical Hacking for IoT: Security Issues, Challenges, Solutions and Recommendation.”
[6] He, Ying & Zamani, Efpraxia & Ni, Kun & Yevseyeva, I. & Luo, Cunjin. (2022). “AI-based Ethical Hacking for Health Information Systems (HIS): a simulation study (Preprint). Journal of Medical Internet Research.”



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY