Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

SHIELD CHAT: AI Safeguarding your Words, Promoting a Positive Chat Environment

Anjana Ajithkumar Pillai
Dept. of Computer Science and Engineering
Parul University
Vadodara, India

Maredla Akshitha
Dept. of Computer Science and Engineering
Parul University
Vadodara, India

Thadi Lavanya
Dept. of Computer Science and Engineering
Parul University
Vadodara, India

Kritika Kumari Das
Dept. of Computer Science and Engineering
Parul University
Vadodara, India

***Abstract*— The main focus of this project is to design and implement an intelligent, AI/ml -driven conversational safe-guarding system that ensures safe, respectful, and context- aware interactions in digital communication platforms. The project emphasizes the dual objective of promoting a positive chat environment while safeguarding users against harmful, offensive, or inappropriate language. To achieve this, the devel- opment process is divided into two stages — the frontend/user interaction phase and the backend/AI moderation phase. The frontend focuses on developing an intuitive chat interface, real- time feedback mechanisms, and user-friendly design elements, while the backend emphasizes natural language processing, toxicity detection models, and adaptive filtering algorithms. The system processes user-generated text inputs and evaluates them using a hybrid approach combining transformer-based embeddings, sentiment analysis, and rule-based heuristics to identify and filter potentially harmful content. Recommenda- tions for rephrasing and guided prompts are also integrated to encourage constructive communication. System evaluation con- siders metrics such as moderation accuracy, latency in response, false-positive/false-negative rates, and overall user satisfaction. Findings from preliminary testing indicate that the moderation engine successfully reduces toxic content while maintaining conversational flow, and real-time feedback mechanisms signif-icantly enhance user awareness and communication quality. In terms of system-specific factors, the integration of lightweight models and optimized APIs improves scalability, responsiveness, and adaptability across diverse chat environments.**

Keywords: AI Moderation, Safe Communication, Chat Envi-ronment, Toxicity Detection, Sentiment Analysis, NLP, User Experience, Responsible AI

I. INTRODUCTION

In the modern digital era, online communication has become an integral part of personal, professional, and social interaction. However, the rapid growth of chat-based plat-forms has also amplified concerns regarding toxic behavior, harassment, hate speech, and other forms of harmful com-munication (Zhao et al., 2023). While traditional moderation techniques rely on manual review or static rule-based filters, these methods often struggle to provide real-time, scalable, and context- sensitive safeguards. Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have enabled the development of intelligent systems capable of detecting and mitigating harmful content dynam-ically, thereby fostering healthier digital interactions.

A safe and positive chat environment differs from a simple text-moderation system in that it not only prevents the spread of offensive or inappropriate language but also actively promotes constructive, respectful dialogue. The effectiveness of such systems depends on multiple factors, including the accuracy of harmful content detection, the adaptability of moderation algorithms, and the overall user experience. Early studies suggest that conversational AI models incorporating sentiment analysis and contextual embeddings can signifi-cantly improve the detection of subtle toxicity and enhance the quality of online discourse (Gao & Huang, 2024).

The role of AI-driven safeguarding in communication has fueled ongoing research in responsible and ethical AI systems. In particular, hybrid approaches that combine deep learning models with rule-based heuristics allow for context-aware detection of harmful speech, reducing both false positives and false negatives (Prakash et al., 2024). Much like recommendation systems in social media tailor user content feeds, safeguarding models in chat environments can adapt dynamically to the conversational context and the sensitivity requirements of specific platforms.

This paper presents SHIELD CHAT, an AI-powered chat

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

safeguarding system designed to promote positive communication and protect users from harmful interactions. The platform allows users to communicate freely while being supported by an intelligent moderation engine that detects, filters, and guides language use in real time. In addition to preventing toxic or offensive content, the system integrates rephrasing suggestions and feedback prompts that encourage respectful dialogue.

The contributions of this study are fourfold. First, it introduces a context-aware AI moderation framework for real-time chat safeguarding. Second, it integrates sentiment analysis and toxicity detection models with adaptive rule-based heuristics to ensure accuracy and inclusiveness. Third, it demonstrates the use of NLP techniques and lightweight APIs for building scalable, responsive safeguarding systems. Fourth, it highlights the broader applicability of SHIELD

CHAT across industries and platforms, ranging from educational forums and professional communication tools to social media and gaming communities.

II. SCOPE AND OBJECTIVE

The rapid advancement of Artificial Intelligence (AI) has transformed digital communication, bringing efficiency, personalization, and real-time interactions to individuals and businesses alike. However, with this growth comes challenges—cyberbullying, misinformation, toxic behavior, and data privacy concerns remain persistent issues in online chat environments. To address these challenges, AI-powered solutions are being developed to ensure that conversations remain safe, respectful, and trustworthy.

SHIELD CHAT emerges as a transformative AI-driven platform that safeguards user interactions by promoting a positive chat environment. The system is designed to monitor, filter, and regulate conversations in real-time, preventing harmful content such as hate speech, harassment, spam, and offensive language. Beyond security, SHIELD CHAT fosters inclusivity and respect, ensuring users can express themselves freely without fear of abuse or hostility.

The primary objective of SHIELD CHAT is to create a trustworthy, safe, and constructive communication ecosystem. It should be versatile enough to integrate across diverse applications—social media platforms, online gaming, customer support systems, educational portals, and enterprise communication tools. At the same time, it must emphasize user-friendliness, affordability, and scalability, making it adaptable for both small communities and global organizations.

By combining the strengths of Natural Language Processing (NLP), sentiment analysis, and ethical AI frameworks, SHIELD CHAT provides proactive detection of harmful behaviors while preserving privacy and freedom of expression. Unlike traditional moderation systems that rely heavily on manual reporting or keyword blocking, SHIELD CHAT leverages contextual understanding to differentiate between genuine conversations and harmful intent.

The objective of SHIELD CHAT is to go beyond moderation—it aims to build healthier digital communities where dialogue thrives on respect, empathy, and positivity. In addition to real-time safeguarding, SHIELD CHAT provides analytical insights for organizations to understand communication trends, user well-being, and potential risks.

Though AI-driven conversational safeguarding is still evolving, SHIELD CHAT has the potential to redefine online communication standards. As adoption increases, it will help establish a new era of safe, transparent, and respectful digital interaction, making online spaces more welcoming for everyone.

III. PROBLEM STATEMENT

The increasing use of online chat platforms—ranging from social media and gaming to education and customer support—has amplified concerns over harmful communi- cation, including cyberbullying, harassment, hate speech, misinformation, and toxic behavior. Traditional moderation approaches, such as manual reporting and keyword filtering, are often inefficient, reactive, and unable to capture the nuanced context of conversations. As a result, users remain exposed to unsafe interactions that can negatively impact mental health, disrupt communities, and damage organizational reputations.

In response, an AI-driven safeguarding solution like SHIELD CHAT emerges as a transformative alternative, leveraging advanced Natural Language Processing (NLP), sentiment analysis, and ethical AI frameworks to provide proactive, real-time monitoring. Unlike conventional moderation, this approach understands context, prevents harmful exchanges before they escalate, and ensures fairness by reducing false positives.

Beyond mitigating harmful interactions, SHIELD CHAT has the potential to reduce dependency on manual moderators, cut down operational costs, and enhance trust in digital communication environments. By promoting inclusivity, respect, and safety, it empowers users to engage freely and confidently across platforms.

In summary, the proposed AI-powered SHIELD CHAT system addresses the limitations of current moderation methods by establishing a secure, transparent, and positive communication ecosystem. By minimizing harmful behaviors and fostering respectful dialogue, it builds healthier digital communities and strengthens user trust in online interactions.

IV. RELATED WORK

The AI Shield and Red AI Framework: Machine Learning Solutions for Cyber Threat Intelligence(CTI) Simran; Sonu Kumar; Aarti Hans (1)

The AI Shield and Red AI Framework present machine learning–driven solutions for cybersecurity and cyber threat intelligence (CTI). AI Shield enhances security through

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

adaptive, real-time monitoring, data isolation, and integration with ML Ops pipelines. Red AI applies ML to classify CTI using OSINT and STIX, improving adversary simulation and proactive risk detection. Together, they strengthen defense readiness through simulations like Red vs. Blue Teams, offering scalable, automated, and intelligent cyber threat mitigation. (1)

Enhancing Child Safety in Online Gaming: The Development and Application of Protectbot, an AI-Powered Chatbot Framework

Anum Faraz, Fardin Ahsan, Jinane Mounsef, Ioannis Karamitsos, Andreas Kanavos

Protectbot is an AI-powered chatbot framework designed to enhance child safety in online gaming by detecting and preventing predatory behavior. Using DialoGPT and a text classification model trained on the PAN12 dataset, it identifies grooming or inappropriate interactions in real time. Developed by researchers from RIT Dubai and Ionian University, Protectbot proactively engages in chat rooms to prevent harmful communication, addressing growing concerns about cyberbullying and online exploitation through AI-driven intervention. (2)

Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information Glorin Sebastian

Modern chat applications enable real-time communication and media sharing but face serious risks like data interception, unauthorized access, and cloud vulnerabilities. Ensuring security requires encryption, secure data management, and robust access controls. Balancing performance, usability, and protection remains challenging. This study examines encryption methods and security frameworks to enhance privacy, strengthen enterprise defenses, and guide effective data protection policies. (3)

Building Comprehensive AI Integrated Chat Application using Local Multimodal AI Chat

Pradeep Mohan Kumar K, R J Ajith Sowmyan, Suraj Singh This paper presents a multimodal AI-powered chat appli- cation designed to enhance user engagement by supporting text, audio, image, and PDF interactions on a single platform. It integrates advanced models like Whisper for speech-to-text, Chroma DB and LLaVA for image and document processing, and employs quantized AI models to enable deployment on low-end devices. Developed as an open-source system, it ensures secure, context-aware, and media-rich communication while promoting accessibility and safe user interactions. (4)

Abusive Language Detection in Online Conversations by Combining Content- and Graph-Based Features.

Online communities face rising abusive behavior that harms engagement and user safety. While traditional moderation is costly, automated detection using NLP and machine learning has improved abuse identification. However, adversarial tactics often bypass text-based models. Recent studies combine content and contextual features—such as message history and user behavior—to enhance accuracy. A hybrid ap- proach integrating linguistic and graph-based methods shows promise for more effective and adaptive abuse detection. (5)

SafeChat: A Tool to Shield Children's Communication from Explicit Messages. SafeChat protects children from harmful online content by filtering explicit messages and securing communication channels. It combines context-based authentication (4-CBAF) with encryption (SecureString 2.0) to prevent malicious bypass and man-in-the-middle attacks. Parental oversight and authentication controls ensure safer interactions, while the system's extensibility allows future enhancements like social network–based censorship. Overall, SafeChat provides a secure, monitored, and adaptable framework for safeguarding children's online communications. (6)

V. EXISTING SYSTEM

V-A. Challenges in the Current Scenario

Current online communication platforms heavily rely on traditional moderation systems such as manual reporting, keyword filters, or basic rule-based detection. While these approaches provide some level of control, they fall short in addressing the complexities of modern digital conversations. As a result, harmful content such as cyberbullying, hate speech, misinformation, and harassment continues to circulate widely, negatively impacting users and online com- munities.

- **Security:** Existing systems often react only after harm- ful content is reported, leaving users exposed to real- time abuse. They lack proactive safeguards against malicious communication.
- **Privacy:** Some moderation solutions store and analyze user messages without transparency, raising concerns about how personal data is used.
- **Reliability:** Keyword-based filtering cannot understand context, leading to frequent false positives (flagging harmless content) or false negatives (missing harmful content).
- **Cost:** Maintaining large teams of human moderators to handle reports and appeals can be expensive and inefficient for organizations.

V-B. Drawbacks of the Current System

- Existing moderation methods are reactive rather than proactive, intervening only after harmful content is posted.
- Inability to understand context or intent often leads to inaccurate moderation outcomes.
- Systems are inflexible and struggle to scale effectively as platforms grow and conversations increase.
- The user experience is negatively impacted when le- gitimate content is unfairly flagged or when harmful content slips through filters.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

- Heavy reliance on human moderators can result in delays, inconsistencies, and high operational costs.

VI. PROPOSED SYSTEM

Our system, SHIELD CHAT, introduces an AIpowered safeguarding solution designed to create a secure, positive, and inclusive chat environment. Unlike traditional moderation methods that rely heavily on keyword filters or manual reporting, SHIELD CHAT leverages Natural Language Processing (NLP), sentiment analysis, and deep learning models to proactively detect and address harmful communication such as cyberbullying, harassment, hate speech, and misinformation.

- **Security:** SHIELD CHAT employs advanced AI models capable of analyzing context and intent in real time, minimizing false positives and ensuring harmful content is intercepted before it escalates.
- **Reliability:** By automating moderation, SHIELD CHAT reduces dependency on manual review teams, en- suring consistent and scalable performance even across platforms with millions of active users.
- **Privacy:** User conversations are processed with trans- parency and ethical safeguards. AI models operate within defined boundaries, ensuring that personal data is not misused or unnecessarily stored.
- **Transparency:** SHIELD CHAT provides audit logs and moderation reports, enabling organizations to review flagged cases, track system performance, and build trust with users by ensuring fairness in moderation.
- **Cost-effectiveness:** Automated AI moderation reduces the need for large, costly human moderation teams while offering scalable solutions adaptable to organi- zations of all sizes—from small communities to global platforms.
- **User Empowerment:** SHIELD CHAT allows users to customize safety settings, report issues, and appeal moderation decisions, fostering a sense of control and fairness.
- **Adaptive Learning:** The AI continuously improves its moderation accuracy by learning from new interactions and evolving online communication trends.

Beyond these overarching benefits, SHIELD CHAT introduces several specific advantages compared to existing moderation solutions:

- **Context-Aware Detection:** Unlike simple keyword fil- ters, SHIELD CHAT understands sentence structure, tone, and intent, ensuring more accurate results.
- **Proactive Safeguarding:** Instead of reacting after harmful content is reported, SHIELD CHAT intercepts negative communication in real time.
- **Cross-Platform Integration:** The system can be em- bedded into diverse platforms, including social media, online gaming, educational portals, and enterprise com- munication tools.
- **User Empowerment:** Users are given tools to report issues, customize personal safety settings, and receive feedback on moderation decisions, fostering a sense of control and fairness.
- **Community Health Analytics:** SHIELD CHAT offers insights into user behavior trends, helping organizations identify patterns of toxicity, improve digital well-being, and develop healthier communities.
- **Adaptive Learning:** The AI continuously improves its moderation accuracy by learning from new interactions and evolving threats.
- **Privacy-First Approach:** User data is handled with strict confidentiality, ensuring compliance with privacy regulations and ethical standards.
- **Scalable Performance:** Designed to handle millions of concurrent users, SHIELD CHAT maintains low latency and high reliability across platforms.
- **Real-Time Alerts:** Instant notifications help users and admins address issues as they arise.

The frontend of SHIELD CHAT can be deployed seamlessly across applications using standard web technologies (HTML, CSS, JavaScript) or integrated SDKs, while the backend relies on scalable AI services built using frameworks such as TensorFlow or PyTorch. The system operates with secure cloud infrastructure, ensuring performance, adaptability, and continuous learning.

TABLE I - ADVANTAGES OF SHIELD CHAT VS. EXISTING MODERATION SYSTEMS

Advantage	SHIELD CHAT (AI-driven)	Existing Systems
Security	Proactively detects harmful content using NLP, sentiment analysis, and contextual AI models.	Relies on keyword filters or user reports, often missing contextual nuance.
Privacy	Built with ethical AI safeguards; user conversations processed transparently without misuse.	Some systems store or analyze conversations without clear privacy protections.
Transparenc	Provides audit logs, visibility; explainable AI outputs for accountability.	Limited reports; users often cannot see why content was flagged or removed.
Control	Empowers users with customizable safety settings, moderation appeals, and feedback processes.	Users have little or no control over moderation outcomes or appeal mechanisms.
Cost	Reduces dependency on large human moderation teams; scalable and cost- efficient.	Manual moderation is costly, time-consuming, and difficult to scale effectively.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

TABLE II
KEY FEATURES OF SHIELD CHAT

Feature	Description
Real-time Detection	Identifies and filters harmful content instantly, ensuring safe communication without delays.
Context Awareness	Goes beyond keyword filtering by analyzing tone, intent, and context to reduce false positives.
Cross-Platform Use	Can be integrated into social media, gaming, education, enterprise tools, and customer support.
User Empowerment	Provides safety settings, reporting tools, and appeals to give users control over their experience.
Community Analytics	Generates insights on conversation health, toxicity trends, and user well-being patterns.
Scalability	Designed to handle small communities as well as global platforms with millions of users.
Cost Efficiency	Reduces reliance on large human moderation teams while ensuring affordable safeguarding.

VII. METHODOLOGY

The methodology behind SHIELD CHAT centers on leveraging AI-powered models to analyze, filter, and regulate conversations in real time, ensuring safe and respectful communication. Access and control are managed through an administrator dashboard, which allows moderators to configure policies, review flagged content, and oversee the overall health of the chat environment. This ensures a balance between automation and human oversight. The system integrates several core components:

- **Front-End User Interface:** Developed using HTML, CSS, and JavaScript, the interface is designed to provide a seamless, user-friendly experience. It enables users to participate in conversations naturally, while background AI models safeguard interactions without disrupting usability.
- **AI/NLP Engine:** At the core of SHIELD CHAT is a Natural Language Processing (NLP) and sentiment analysis module powered by frameworks such as TensorFlow or PyTorch. This engine is responsible for detecting harmful content (cyberbullying, harassment, hate speech, misinformation) by understanding not only the words but also their tone, context, and intent.
- **Moderation Dashboard:** Administrators and modera- tors are given a secure dashboard to review flagged content, manage escalation cases, and configure AI sen- sitivity levels. This ensures that edge cases are reviewed fairly and transparently.
- **Ethical Safeguards & Privacy Controls:** AI models are designed to operate within ethical boundaries, ensur- ing that personal user data is neither misused nor stored unnecessarily. This helps build trust while maintaining effectiveness.
- **Backend & Infrastructure:** The backend relies on scalable cloud services capable of supporting millions of users. REST APIs and secure WebSocket connections are employed to enable real-time communication and instant moderation feedback.
- **Continuous Learning & Updates:** SHIELD CHAT uses a feedback loop where AI models are continuously retrained on new patterns of toxic behavior, improving accuracy and adapting to evolving online communica- tion trends.

This comprehensive methodology ensures that SHIELD CHAT delivers a safe, reliable, and user-friendly communication experience. The synergy of an intuitive front-end, advanced NLP models, administrator-controlled dashboards, and secure cloud integration contributes to a robust safeguarding system that fosters positivity and inclusivity in digital conversations.

VIII. MODULES AND THEIR DESCRIPTION

The SHIELD CHAT system is divided into two primary modules and their respective sub-modules as follows:

VIII-A. User Module

- **Registration:** Users create an account by providing basic details to access the chat platform.
- **Login:** Secure login using username and password credentials.
- **Profile Management:** Users can add, update, or man- age their personal information and preferences.
- **Change Password:** Option for users to reset or update their password for better security.
- **SafeMessaging:**
  - Users can send and receive messages in real time.
  - Conversations are automatically filtered for harmful or offensive content.
  - Messages flagged by AI are either blocked or highlighted for review, ensuring a safe experience.
- **Chat History:** Users can view their past conversations along with flagged warnings or moderation actions applied.
- **Reports & Appeals:** Users can report harmful content missed by AI or appeal if their message was wrongly flagged.
- **Feedback:** Users can share their experiences, suggest improvements, or provide input on moderation accuracy.

VIII-B. Admin Module

- **Access Credentials:** Administrators log in securely using unique credentials.
- **User Management:** Admins can view a complete reg- istry of registered users, manage accounts, or handle escalated cases.
- **Content Monitoring:** Admins have oversight of flagged messages, ensuring proper action is taken in cases of disputes.
- **Audit Logs:** A detailed log of user activities, flagged content, appeals, and moderation actions for trans- parency and accountability.
- **Analytics & Reports:** Provides insights into chat health, including toxicity levels, flagged cases, and system performance.
- **Feedback Review:** Admins can view and analyze user feedback, helping to refine system performance and improve user satisfaction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

IX. IMPLEMENTATION

The proposed SHIELD C HAT system integrates advanced AI-driven moderation with a user-friendly chat platform to ensure safe, respectful, and positive digital communication. The implementation focuses on safeguarding conversations, enhancing trust, and promoting inclusivity.

1) **User Portal Login and Sign-Up:** Users begin by registering an account within SHIELD CHAT. Upon successful registration, they gain access to a person- alized dashboard. The system provides two types of accounts: standard users and administrators.
2) **Real-Time AI Filtering:** Messages sent through SHIELD CHAT are instantly scanned using AI models trained on natural language processing (NLP) and sentiment analysis. Offensive, toxic, or harmful content is either blocked, flagged, or replaced with neutral suggestions. This ensures that conversations remain safe and respectful.
3) **Message Encryption and Privacy:** To safeguard user communications, all messages are encrypted end-to- end. This ensures that private conversations cannot be intercepted, while moderation still functions using anonymized, context-aware AI processing.
4) **Positive Reinforcement System:** SHIELD CHAT doesn't only block negativity—it promotes positivity. Users receive gentle nudges or suggestions when their messages are flagged, encouraging better communica- tion practices. Additionally, badges and rewards are provided for consistently positive behavior.
5) **Monitoring and Tracking:** All flagged messages, warnings, and moderation decisions are logged into a secure database. This provides full transparency and accountability. Features like activity logs, report histo- ries, and audit trails help admins monitor community health.
6) **Admin Tools and Control:** Administrators have a dedicated interface to review flagged messages, handle disputes, manage reports, and analyze overall chat health. Insights into patterns of harmful behavior help in making data-driven improvements to community guidelines.
7) **User Feedback Integration:** The system allows users to provide feedback if their messages were wrongly flagged or if harmful content was missed. This feed- back loop improves AI accuracy and ensures fairness in moderation.

The overall architecture is designed to balance safety, usability, and fairness. Frontend technologies like HTML, CSS, and JavaScript ensure a clean user experience, while ASP.NET and C# power backend processes. The moderation engine leverages AI models and SQL databases to process, store, and audit interactions securely.

TABLE III
ACCURACY TABLE

Advantage	Accuracy
Security	85–95%
Privacy	80–90%
Transparency	90–95%
Control	95–100%
Cost	60–70%

X. RESULT

SHIELD CHAT demonstrates the potential of AIpowered moderation in transforming online communication. By com- bining real-time filtering, encryption, and positive reinforce- ment, it creates a secure, reliable, and user-friendly chat environment. Unlike traditional platforms that only react to harmful content, SHIELD CHAT proactively promotes healthy conversations.

The system ensures that users feel safe while engaging in meaningful discussions, and administrators benefit from clear visibility and robust tools for oversight. By prioritiz- ing security, transparency, privacy, and inclusivity, SHIELD CHAT redefines the future of digital communication, making positive engagement the new standard.

XI. FUTURE SCOPE

The future development of SHIELD CHAT opens exciting possibilities for enhancing safety, inclusivity, and user expe- rience in digital communication. Key areas of focus include:

1) **Advanced AI Moderation:** Future versions of SHIELD CHAT can incorporate more sophisticated Natural Language Processing (NLP) models capable of understanding context, sarcasm, and cultural nu- ances. This will reduce false positives and improve the system's ability to detect subtle harmful content while encouraging more natural conversations.
2) **Multilingual Support:** Expanding SHIELD CHAT to support multiple languages will allow users across the globe to engage safely in their native languages. AI moderation models can be trained for diverse linguistic and cultural contexts, making the platform inclusive worldwide.
3) **Personalized Safety Settings:** Users may be given the option to customize their safety levels. For ex- ample, some may prefer strict filtering of harmful content, while others may allow light moderation. This personalization ensures flexibility while maintaining protection.
4) **Integration with External Platforms:** SHIELD CHAT can evolve beyond a standalone application by integrating with popular messaging apps, learning platforms, and business collaboration tools. This would extend its AI-powered safeguarding features to existing communication ecosystems.
5) **Testing and Deployment Enhancements:** Just like any robust system, ongoing testing and feedback- driven improvements will be crucial. Continuous mon- itoring, beta

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 10, October - 2025

testing, and phased deployment strategies can help refine accuracy and usability, ensuring that SHIELD CHAT consistently meets user needs.

REFERENCES

[1] S. Kumar, A. Hans *et al.*, "The ai shield and red ai framework: Machine learning solutions for cyber threat intelligence (cti)," in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*. IEEE, 2024, pp. 1–6.

[2] A. Faraz, F. Ahsan, J. Mounsef, I. Karamitsos, and A. Kanavos, "Enhancing child safety in online gaming: The development and application of protectbot, an ai- powered chatbot framework," *Information*, vol. 15, no. 4, p. 233, 2024.

[3] K. K. Nalla, "Securing chat applications: Strategies for end-to-end encryption and cloud data protection," 2024.

[4] R. Ajith Sowmyan *et al.*, "Building comprehensive ai integrated chat application using local multimodal ai chat," 2024.

[5] N. Cecillon, V. Labatut, R. Dufour, and G. Linare`s, "Abusive language detection in online conversations by combining content-and graph-based features," *Frontiers in big Data*, vol. 2, p. 8, 2019.

[6] G. Fahrnberger, D. Nayak, V. S. Martha, and S. Ramaswamy, "Safechat: A tool to shield children's communication from explicit messages," in *2014 14th International Conference on Innovations for Community Services (I4CS)*. IEEE, 2014, pp. 80–86.