Literature Review - LLM based chatbot

Paper 1: An LLM-Driven Chatbot in Higher Education for Databases and Information Systems

Citation (IEEE Format):

A. T. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke, “An LLM-Driven Chatbot in Higher Education for Databases and Information Systems,” *IEEE Transactions on Education*, vol. 68, no. 1, pp. 103–115, Feb. 2025, doi:10.1109/TE.2024.3467912.

Problem Addressed

Large class sizes, information overload inside LMSs (Moodle), and limited TA/teacher availability reduce students’ ability to self-regulate learning and timely seek help. Classical script-based bots lack contextual depth; LLMs introduce opportunities (tutoring, exercise generation) but also risks (hallucinations, bias, misinformation). The paper targets how an LLM-driven chatbot can be integrated into Moodle to support self-regulated learning (SRL) and help-seeking while measuring acceptance and response congruency with course materials.

Method / Approach

1. **Architecture:** Retrieval-Augmented Generation (RAG) pipeline with LangChain, LlamaIndex, and Weaviate for embeddings.
2. **Agent Tools:** Separate Answer Generator and Question Generator modules coordinated via BaseMultiActionAgent.
3. **Prompting:** Defined role-specific system prompts, restricted scope, and deterministic output (temperature = 0).
4. **Evaluation:** Conducted with TAM questionnaires (acceptance) and correctness checks (manual TA + LLM fact-checker).

Key Results / Findings

1. **High Acceptance:** Students rated the chatbot as useful, easy to use, and supportive, though not a tutor replacement.
2. **Accuracy:** Achieved 81–88% correct responses across different datasets and evaluation methods.
3. **Automated Check:** LLM-based fact-checker showed ~82% accuracy but struggled with false negatives.
4. **Cost Analysis:** Chat interaction costs were manageable (~$1.65 per participant), though GPT-4 was expensive.

Limitations / Drawbacks

1. **Sample Size:** Limited and homogenous participant pool led to possible selection bias.
2. **Verification:** Single-TA manual checks may introduce evaluator subjectivity.
3. **Fact-Checking Weakness:** Automated LLM checks lacked reliability in catching incorrect answers.
4. **Scalability:** Per-course tailoring and GPT-4 reliance increased maintenance effort and cost.

Future Enhancements

1. **Improved Verification:** Combine fact-checking with retrieval-based and rule-based validation.
2. **Alternative Models:** Explore cheaper or self-hosted LLMs to reduce long-term costs.
3. **Expanded Testing:** Conduct larger-scale studies with diverse learners and multiple evaluators.
4. **Adaptive Pedagogy:** Encourage self-regulated learning by providing hints instead of direct answers.

Relevance to Final-Year Project: LLM-Based Chatbot

1. **Adopt RAG Design:** Use course-specific document ingestion, embeddings, and retrieval to reduce hallucinations.
2. **Agent Modularity:** Implement separate tools (Answer, Question, Fact-Checker) for flexibility and testing.
3. **Evaluation Strategy:** Follow TAM for usability and combine manual + automated correctness checks.

4. **Scalable Approach:** Prototype with GPT-4 but plan migration to GPT-3.5 or open-source models for deployment.

Paper 2: A Complete Survey on LLM-based AI Chatbots

Citation (IEEE Format):

S. K. Dam, C. S. Hong, Y. Qiao, and C. Zhang, “A Complete Survey on LLM-based AI Chatbots,” *IEEE Access*, vol. 12, pp. 158923–158947, Nov. 2024, doi:10.1109/ACCESS.2024.3526789.

Problem Addressed

Conversational AI has evolved rapidly, but early chatbots lacked contextual understanding, scalability, and human-like interaction. The emergence of LLMs (e.g., GPT, BARD, Claude) has enabled more sophisticated chatbots, but challenges such as hallucinations, bias, privacy risks, and academic misuse persist. This survey addresses the need for a structured, comprehensive overview of LLM-based chatbots—their evolution, applications, challenges, and future directions.

Method / Approach

1. **Historical Review:** Traces chatbot evolution from ELIZA and SmarterChild to LLM-powered systems like GPT-4 and Gemini.
2. **Taxonomy Development:** Organizes applications into education, research, healthcare, software engineering, and finance.
3. **Comparative Analysis:** Reviews existing surveys and identifies their gaps (limited scope, lack of categorization, missing technical/ethical issues).
4. **Structured Framework:** Frames challenges around three perspectives—technical, ethical, and misuse—providing sub-categories for clarity.

Key Results / Findings

1. **Applications Across Domains:** LLM chatbots enhance learning, research efficiency, healthcare diagnosis, coding support, and financial analysis.

2. **Technical Advances:** Highlights key LLM innovations such as transformers, in-context learning, and chain-of-thought prompting.
3. **Identified Challenges:** Categorizes issues into recency, reasoning, hallucination, transparency, bias, privacy, and misuse.
4. **Comparative Superiority:** Unlike earlier surveys, this work covers multiple LLMs (ChatGPT, BARD, Bing Chat, Claude, Ernie) with a broader, deeper taxonomy.

Limitations / Drawbacks

1. **Rapid Evolution:** Fast-changing LLM landscape risks survey becoming outdated quickly.
2. **Surface-Level Evaluation:** Provides broad coverage but limited experimental validation of chatbot performance.
3. **Bias in Sources:** Relies heavily on published studies, which may carry methodological or cultural biases.
4. **Generalization Limits:** Taxonomy is comprehensive but may not capture niche or domain-specific applications.

Future Enhancements

1. **Dynamic Updates:** Maintain continuously updated surveys to match rapid LLM advancements.
2. **Cross-Disciplinary Studies:** Expand into underexplored domains like law, public policy, and creative arts.
3. **Robust Benchmarks:** Establish standardized evaluation metrics for chatbot performance, ethics, and safety.
4. **Ethical Frameworks:** Develop global guidelines for responsible use, covering fairness, transparency, and misuse prevention.

Relevance to Final-Year Project: LLM-Based Chatbot

1. **Taxonomy as Blueprint:** Provides structured inspiration for categorizing chatbot functions in your project.
2. **Challenge Awareness:** Identifies pitfalls (hallucinations, bias, misuse) you must address with safeguards.
3. **Evaluation Insights:** Suggests using both technical benchmarks and ethical criteria in assessing your chatbot
4. **Innovation Pathways:** Encourages extending beyond education into multidisciplinary features for broader impact.

Paper 3: Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions

Citation (IEEE Format):

C. Kooli, “Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions,” *Sustainability*, vol. 15, no. 7, p. 5614, Mar. 2023, doi:10.3390/su15075614.

Problem Addressed

AI chatbots are increasingly used in education and research, yet they raise significant ethical issues such as cheating, bias, privacy risks, and misuse. Traditional assessment methods are vulnerable to AI-driven dishonesty, threatening academic integrity. Current research has not adequately examined the ethical implications of chatbots in academia. This study explores how AI-based chatbots impact learning, research, and ethics, while proposing solutions for sustainable adoption.

Method / Approach

1. **Qualitative Exploratory Study:** Relied on expert analysis and interpretation under an interpretivist philosophy.
2. **Thematic Analysis:** Organized findings around challenges, ethical dilemmas, and potential solutions.
3. **Research Questions:** Focused on chatbots’ impact on education (RQ1), research transformation (RQ2), and ethical challenges (RQ3).
4. **Comparative Review:** Examined prior studies on AI in education, highlighting gaps in ethical-focused research.

Key Results / Findings

1. **Education Impact:** Chatbots improve accessibility, engagement, and personalization but also risk enabling cheating in assessments.

2. **Research Role:** Chatbots act as research assistants by automating data tasks and supporting collaboration, though they lack critical thinking and empathy.
3. **Ethical Concerns:** Issues include bias, privacy violations, misinformation, and replacement of human expertise.
4. **Need for Adaptation:** Calls for innovative assessment methods, awareness campaigns, and ethical regulations to ensure sustainable integration.

Limitations / Drawbacks

1. **Exploratory Nature:** Findings are based on qualitative expert opinion, lacking large-scale empirical validation.
2. **Context Dependence:** Insights may not generalize across different educational systems and cultural settings.
3. **Rapid AI Evolution:** Risk of the study becoming outdated as LLMs and chatbots advance quickly.
4. **Subjectivity:** Reliance on interpretivist methods introduces researcher bias in interpretation.

Future Enhancements

1. **Innovative Assessments:** Develop creative, hands-on, and authentic tasks to discourage AI misuse in exams.
2. **Ethical Safeguards:** Establish strong regulations, transparency measures, and anti-bias training datasets.
3. **Awareness & Training:** Educate students, teachers, and researchers on ethical use of AI tools.
4. **Mixed-Methods Research:** Complement qualitative insights with quantitative studies for stronger validation.

Relevance to Final-Year Project: LLM-Based Chatbot

1. **Ethical Design:** Reinforces the need to embed safeguards against cheating, bias, and misuse in your chatbot.
2. **Assessment Strategy:** Suggests innovative evaluation methods (hints, problem-solving, oral components) instead of direct answers.
3. **Sustainability Focus:** Encourages treating chatbots as aids to human learning, not replacements.
4. **Research Alignment:** Validates the importance of addressing ethics explicitly in your project's scope.

Paper 4: A Survey on ChatGPT: Technology, Applications, and Challenges

Citation (IEEE Format):

M. S. R. Chowdhury, M. A. Rahman, and T. Rahman, “A Survey on ChatGPT: Technology, Applications, and Challenges,” *Computers*, vol. 12, no. 3, p. 60, Mar. 2023, doi:10.3390/computers12030060.

Problem Addressed

ChatGPT, based on large language models, has become one of the most influential AI systems, transforming human-computer interaction. Despite its popularity, there is limited structured research on its architecture, real-world applications, and challenges. The paper addresses the need for a comprehensive survey that explains ChatGPT’s underlying technology, evaluates its applications, and examines ethical, technical, and social issues.

Method / Approach

1. **Architectural Analysis:** Explains GPT’s foundation on the Transformer model, self-attention, and fine-tuning.
2. **Application Review:** Surveys ChatGPT’s use cases in education, healthcare, business, creative writing, and programming.
3. **Challenge Categorization:** Groups issues into accuracy, hallucination, ethics, bias, transparency, and misuse.
4. **Comparative Insight:** Contrasts ChatGPT with earlier AI approaches, highlighting advancements and limitations.

Key Results / Findings

1. **Wide Adoption:** ChatGPT has shown strong utility in domains like tutoring, medical advice, content creation, and software support.
2. **Strengths:** Provides natural, context-aware, and coherent responses compared to traditional chatbots.

3. **Weaknesses:** Prone to hallucinations, lacks explainability, and sometimes generates biased or unsafe content.
4. **Research Gap:** Demonstrates that while ChatGPT excels practically, academic studies on its reliability and ethics are still scarce.

Limitations / Drawbacks

1. **Lack of Empirical Data:** Survey is descriptive with minimal quantitative testing of ChatGPT.
2. **Dynamic Nature:** Rapid evolution of GPT models risks survey findings becoming outdated.
3. **Scope Restriction:** Focuses mainly on ChatGPT, not broader LLMs like Claude or Bard.
4. **Ethical Ambiguity:** Provides limited solutions to issues like bias, misinformation, and misuse.

Future Enhancements

1. **Robust Evaluation:** Incorporate benchmark datasets to assess ChatGPT's performance systematically.
2. **Ethical Safeguards:** Build fairness, bias-reduction, and transparency into chatbot systems.
3. **Model Comparisons:** Extend surveys to include other LLM-based chatbots for balanced insights.
4. **Domain-Specific Research:** Explore ChatGPT's impact in specialized areas such as law, healthcare, and engineering.

Relevance to Final-Year Project: LLM-Based Chatbot

1. **Technical Foundation:** Explains GPT and Transformer architecture, useful for understanding your project's core.
2. **Application Examples:** Provides reference use cases (education, coding support) relevant to your chatbot.
3. **Challenge Awareness:** Highlights key risks (hallucinations, bias) you must mitigate in implementation.
4. **Research Motivation:** Supports your project by showing gaps in empirical testing and ethical solutions.

Paper 5: Unsupervised Occupation Extraction and Standardization Leveraging Large Language Models

Citation (IEEE Format):

N. Li, B. Kang, and T. De Bie, “LLM4Jobs: Unsupervised occupation extraction and standardization leveraging Large Language Models,” *arXiv preprint* arXiv:2309.09708v2, 2023.

Problem Addressed

Job postings contain unstructured and inconsistent occupation titles, making it difficult for employers, recruiters, and policy makers to analyze labor market data effectively. Traditional NLP approaches struggle with domain variation, ambiguity, and the need for manual labeled datasets. The paper addresses how Large Language Models can be leveraged to automatically extract and standardize occupations from job descriptions without supervised training.

Method / Approach

1. **LLM Prompting:** Uses GPT-based models to extract occupations directly from raw job descriptions through carefully engineered prompts.
2. **Unsupervised Pipeline:** Employs zero-shot and few-shot prompting strategies, avoiding the need for labeled training data.
3. **Standardization Step:** Aligns extracted occupations with hierarchical taxonomies like ESCO (European Skills, Competences, Qualifications, and Occupations).
4. **Evaluation:** Compares LLM4Jobs with baseline methods across datasets, focusing on accuracy, robustness, and adaptability.

Key Results / Findings

1. **High Accuracy:** LLM4Jobs outperformed rule-based and traditional ML methods in occupation extraction tasks.
2. **Adaptability:** Showed strong generalization across different job posting datasets and domains.

3. **Scalability:** Unsupervised approach reduced the need for costly annotation, making it efficient for large-scale labor data.
4. **Standardization Success:** Effectively mapped extracted roles to ESCO taxonomy, enabling structured labor market analysis.

Limitations / Drawbacks

1. **LLM Dependency:** Relies heavily on GPT models, raising issues of cost, availability, and API dependency.
2. **Ambiguity Handling:** Struggled with vague or multi-role job postings where occupation was unclear.
3. **Bias Concerns:** Inherited biases from pretrained LLMs could affect fairness in occupation classification.
4. **Lack of Domain Customization:** Performance may degrade in highly specialized or niche industries.

Future Enhancements

1. **Hybrid Models:** Combine LLM-based extraction with domain-specific rules or ontologies for improved precision.
2. **Bias Mitigation:** Introduce fairness-aware prompts and datasets to reduce occupational stereotypes.
3. **Cost Optimization:** Explore lightweight or open-source LLMs for large-scale deployment.
4. **Broader Validation:** Test across more languages and global labor markets beyond ESCO framework.

Relevance to Final-Year Project: LLM-Based Chatbot

1. **Unsupervised Learning Insight:** Demonstrates how LLMs can perform domain-specific tasks without labeled training data.
2. **Pipeline Inspiration:** Extraction + standardization workflow can inspire chatbot's fact-checking and structured response generation.
3. **Scalability Lessons:** Shows efficiency gains from prompt engineering and unsupervised methods, useful for chatbot deployment.
4. **Risk Awareness:** Highlights LLM dependency, bias, and domain adaptation issues—important to consider in your project design.

Paper 6: Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAGAS)

Citation (IEEE Format):

M. L. Husain, Y. Wibisono, and A. Anisyah, “Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAGAS),” *Brilliance: Research of Artificial Intelligence*, vol. 7, no. 2, pp. 111–121, 2024.

Problem Addressed

1. Traditional chatbots in academic services lack contextual awareness and often generate irrelevant or incorrect answers.
2. LLMs improve fluency but suffer from hallucinations and domain inaccuracy.
3. Students require fast, reliable responses for academic information like course structures and policies.
4. The paper addresses how RAG can anchor chatbot answers to trusted knowledge sources.

Method / Approach

1. Implemented a RAG pipeline combining OpenAI GPT with domain-specific document retrieval.
2. Vector database used to embed and store academic resources for efficient retrieval.
3. Prompt engineering tailored queries to reduce hallucinations and improve precision.
4. System deployed as an academic services chatbot prototype accessible to students.

Key Results / Findings

1. Improved accuracy compared to baseline LLM responses.
2. Enhanced trustworthiness due to reliance on institutional documents.
3. Positive feedback from students in pilot testing.
4. RAG demonstrated scalability for multi-department integration.

Limitations / Drawbacks

1. Relies on availability of well-structured academic documents.
2. Initial system dependent on proprietary APIs, increasing costs.
3. Limited evaluation sample—restricted to small student group.
4. Focus only on academic FAQs, not broader tutoring.

Future Enhancements

1. Expand chatbot to cover multiple faculties and academic processes.
2. Introduce multilingual support for diverse student populations.
3. Add feedback loops for continuous improvement via user input.
4. Explore cost-effective alternatives to proprietary LLMs.

Relevance to Final-Year Project: LLM-Based Chatbot

1. Demonstrates how RAG improves factual accuracy in chatbots.
2. Encourages linking chatbot responses directly to course data.
3. Highlights scalability for university-wide adoption.
4. Validates RAG as a strong architectural choice for your project.

Paper 7: Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education

Citation (IEEE Format):

M. Abedi, et al., “Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education,” *arXiv preprint* arXiv:2309.13059, 2023.

Problem Addressed

1. Engineering courses often face limited instructor availability and high workloads.
2. Students require personalized tutoring beyond lectures.
3. Existing digital tools lack adaptability to complex problem solving.
4. The paper explores integrating LLM chatbots to support graduate education.

Method / Approach

1. Integrated ChatGPT into a graduate fluid mechanics course.
2. Designed prompts to generate worked examples and tailored hints.
3. Linked chatbot to external computational tools (e.g., Wolfram Alpha).
4. Collected student feedback on chatbot usefulness and acceptance.

Key Results / Findings

1. Students reported improved self-paced learning and deeper understanding.
2. Chatbot reduced reliance on instructors for routine questions.
3. Prompts enabled generation of customized exercises and explanations.
4. Demonstrated feasibility of LLMs in advanced STEM teaching.

Limitations / Drawbacks

1. Chatbot prone to hallucinations in technical problem solving.
2. Dependence on external APIs raised cost and sustainability issues.
3. Study limited to one course and a small student cohort.
4. Lacked systematic benchmarking of accuracy against verified solutions.

Future Enhancements

1. Incorporate structured fact-checking for technical correctness.
2. Expand deployment across multiple engineering disciplines.
3. Develop hybrid tutoring combining chatbot and instructor oversight.
4. Introduce adaptive personalization based on learner profiles.

Relevance to Final-Year Project: LLM-Based Chatbot

1. Highlights the value of LLMs for technical education support.
2. Reinforces the need for fact-checking to ensure chatbot reliability.
3. Suggests hybrid chatbot + teacher approach for greater trust.
4. Provides evidence of student acceptance of AI-based tutoring.

Paper 8: Memory-Augmented Large Language Model for Enhanced Chatbot Services in University LMS

Citation (IEEE Format):

J. Kim, et al., “Memory-Augmented Large Language Model for Enhanced Chatbot Services in University Learning Management Systems,” *Applied Sciences*, vol. 15, no. 17, p. 9775, 2025.

Problem Addressed

1. Conventional LLM chatbots lack memory, causing them to lose context in extended conversations.
2. Students need continuity and personalization across multiple interactions.
3. LMS-integrated chatbots often fail to adapt to evolving student requirements.
4. The study addresses how memory modules can enhance chatbot responsiveness and intelligence.

Method / Approach

1. Developed a memory-augmented framework with short-term, long-term, and event memory layers.
2. Integrated chatbot with a university LMS for course-aware tutoring.
3. Implemented filtering mechanisms to retrieve only relevant contextual data.
4. Evaluated using automated NLP metrics and student user studies.

Key Results / Findings

1. Improved coherence and continuity across multiple chat sessions.

2. Enabled personalization through long-term student tracking.
3. Users rated responses as more contextually relevant and accurate.
4. Automated evaluation confirmed higher semantic similarity to reference answers.

Limitations / Drawbacks

1. Complex memory management significantly increased computational overhead.
2. Storing student data raised privacy and security concerns.
3. Evaluation restricted to one institution's LMS, limiting generalizability.
4. Does not directly mitigate hallucinations or factual errors.

Future Enhancements

1. Introduce privacy-preserving methods for memory storage and retrieval.
2. Optimize memory retrieval to minimize processing costs.
3. Expand deployment across multiple universities and languages.
4. Combine memory modules with fact-checking systems for better accuracy.

Relevance to Final-Year Project: LLM-Based Chatbot

1. Demonstrates importance of memory retention for meaningful student interactions.
2. Encourages personalization features in educational chatbots.
3. Highlights privacy and cost trade-offs in system design.
4. Provides an evaluation framework blending metrics with user feedback.

Paper 9: RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots

Citation (IEEE Format):

P. Feldman, J. R. Foulds, and S. Pan, "RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots," *arXiv preprint* arXiv:2403.01193, 2024.

Problem Addressed

1. Retrieval-Augmented Generation (RAG) chatbots improve factual grounding but are not foolproof.
2. Contradictory or misleading retrieved data can reduce chatbot accuracy.
3. Over-reliance on retrieval risks propagating biased or incorrect knowledge.
4. The study investigates both the strengths and vulnerabilities of RAG systems.

Method / Approach

1. Conducted empirical evaluations of RAG-based chatbot performance.
2. Compared chatbot outputs under correct vs. misleading retrieval conditions.
3. Measured accuracy, robustness, and reliability against baseline LLMs.
4. Analyzed the effects of retrieval quality on overall response trustworthiness.

Key Results / Findings

1. RAG improved factual accuracy in most tested scenarios.
2. Performance dropped significantly when retrieval contained contradictions.
3. Biased or poor-quality sources negatively impacted chatbot outputs.
4. Demonstrated that retrieval quality is as critical as model capability.

Limitations / Drawbacks

1. Heavy reliance on external knowledge bases with varying reliability.
2. Did not propose strong mechanisms for error detection or correction.
3. Evaluation limited to a narrow set of domains and datasets.
4. Lacked user perception or trust-based evaluation studies.

Future Enhancements

1. Develop filters to detect and exclude poor-quality retrieval data.
2. Incorporate cross-source verification for consistency checks.
3. Expand evaluations across diverse domains and use cases.
4. Add user-centered trust studies to evaluate real-world reliability.

Relevance to Final-Year Project: LLM-Based Chatbot

1. Warns of the risks of blindly trusting retrieval results in RAG systems.
2. Suggests integrating source quality control in chatbot pipelines.
3. Supports combining retrieval with fact-checking for higher accuracy.
4. Highlights the importance of building user trust in educational chatbots.

Paper 10: Towards Optimizing and Evaluating a Retrieval-Augmented QA Chatbot Using LLMs with Human-in-the-Loop

Citation (IEEE Format):

A. Afzal, A. Kowsik, R. Fani, and F. Matthes, “Towards Optimizing and Evaluating a Retrieval-Augmented QA Chatbot Using LLMs with Human-in-the-Loop,” in *Proc. DASH Workshop at EMNLP*, 2024.

Problem Addressed

1. Fully automated QA chatbots often produce inaccurate or misleading answers.
2. Existing chatbot evaluation lacks consistent benchmarks and reliability.
3. Current optimization methods do not effectively integrate user feedback.
4. The paper addresses how human-in-the-loop strategies can enhance QA chatbot performance.

Method / Approach

1. Built a QA chatbot using GPT-4 within a Retrieval-Augmented Generation (RAG) framework.
2. Collected datasets iteratively with human reviewers correcting model outputs.
3. Optimized prompts and retrieval strategies using user-provided feedback.
4. Evaluated with both automatic NLP metrics and human judgment comparisons.

Key Results / Findings

1. Human-in-the-loop significantly improved accuracy and relevance of responses.
2. Automatic metrics showed strong correlation with human evaluations.
3. GPT-4 with optimized RAG pipeline outperformed baseline QA systems.
4. Demonstrated scalability of feedback-driven optimization workflows.

Limitations / Drawbacks

1. Dependent on consistent availability of human evaluators.
2. Focused only on QA chatbots, not general-purpose conversational agents.
3. Optimization requires ongoing resource and time investment.
4. Risk of human bias in evaluation affecting training outcomes.

Future Enhancements

1. Automate parts of the feedback collection process to reduce reliance on humans.
2. Expand evaluation across multiple domains and user types.
3. Explore cost-efficient smaller LLMs for scalable deployment.
4. Integrate multi-agent systems for collaborative verification of answers.

Relevance to Final-Year Project: LLM-Based Chatbot

1. Shows the value of human feedback in refining chatbot accuracy.
2. Provides a combined framework of automatic + human evaluation.
3. Encourages iterative improvement of prompts and retrieval in your chatbot.
4. Supports building a scalable QA chatbot with reliability and trust.
