CAPSTONE PROJECT

INTELLIGENT RESEARCH AGENT USING IBM GRANITE

Presented By:

1. Akhil T V – Rajalakshmi Institute of Technology – Computer Science Engineering

edunet foundation

OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Algorithm & Deployment
- Result
- Conclusion
- Future Scope
- References

PROBLEM STATEMENT

In today's fast-paced research landscape, students and professionals often spend significant time manually performing repetitive research tasks — such as finding relevant academic papers, reading and summarizing them, formatting citations, and drafting structured reports. These tasks are not only time-consuming but also prone to human error and inefficiencies.

As the volume of research material grows rapidly, there is a pressing need for a system that can intelligently automate these processes. Researchers require a tool that can understand natural language queries, search for academic literature, summarize key findings, generate formatted citations, suggest new research hypotheses, and compile structured reports — all with minimal manual effort.

This project aims to develop a Research Agent, an AI-powered assistant that leverages large language models and IBM's Granite Foundation Models to streamline and automate the end-to-end academic research workflow. The goal is to enhance research productivity, accuracy, and accessibility for both students and professionals in academic and industrial domains.

PROPOSED SOLUTION

- To address the inefficiencies in academic research workflows, we propose the development of an AI-powered Research Agent a system that can autonomously process natural language queries and assist in end-to-end research tasks.
- The solution involves the following key components:
  - **Query Understanding and Planning:** The system breaks down the user's research question into logical, actionable steps using IBM Granite Foundation Models.
  - **Automated Literature Search:** It searches academic databases like arXiv to retrieve relevant research papers using a refined query.
  - **Summarization of Papers:** Each paper is summarized using NLP techniques to extract key insights and methods.
  - **Hypothesis Generation:** Based on the literature, the system proposes new research hypotheses and areas for exploration.
  - **Citation Management:** It automatically generates formatted citations in APA and IEEE styles.
  - **Research Report Drafting:** A structured report is generated including Introduction, Related Work, Hypothesis, and Conclusion sections.
  - **Reflection and Self-Correction:** The agent evaluates each step's success and autonomously retries or improves failed steps enabling a true agentic AI behavior.
- The entire pipeline is integrated using IBM Cloud Lite services and leverages IBM Granite LLMs for intelligent decision-making.

SYSTEM APPROACH

The Research Agent system was developed with a modular, agentic architecture using IBM's foundation models and a modern full-stack web framework. The approach integrates frontend, backend, and AI model components seamlessly.

1. System Requirements:

- **Cloud Platform:** IBM Cloud Lite
- **LLM Infrastructure:** IBM Watsonx.ai
- **Foundation Model Used:** ibm/granite-3-8b-instruct
- **Backend Runtime:** Python 3.11+
- **Backend Framework:** FastAPI
- **Frontend Framework:** React.js
- **Package Managers:** pip (Python), npm (React)
- **Development Tools:** VS Code, Postman (API Testing)

SYSTEM APPROACH

2. Libraries and Tools Used

Library	Tools
NLP & LLM	IBM Watsonx.ai, ibm/granite-3-8b-instruct
Backend Development	FastAPI, Uvicorn, Jinja2
Frontend Development	React.js, npm
Data Parsing & Search	feedparser, requests, json, re
Summarization Logic	Prompt-driven chunked summarization
Reflection Mechanism	Self-evaluating LLM feedback with retry logic
Deployment	IBM Cloud Lite (LLM + API backend hosting)

ALGORITHM & DEPLOYMENT

Agentic Reasoning Workflow (Algorithmic Logic)

Instead of a fixed ML algorithm, the Research Agent uses a modular agent loop powered by IBM Granite, enabling autonomous decision-making via natural language prompts.

I. Core Logic Flow:

1. **Plan Generation:**
   - LLM (Granite-3-8b-instruct) breaks the user query into step-wise tasks using planning prompts.
2. **Tool Selection & Execution:**
   - Each step is tied to a specific tool: search, summarize, citation, hypothesis, or report.
3. **Reflection & Retry:**
   - After each step, the agent reflects using another prompt to determine if the step succeeded.
   - If failed, it automatically retries the step up to a safe limit.

This approach mimics cognitive reasoning, making the system adaptable to various research domains.

II. Input Features (User Query as Input)

- Natural Language Research Query (e.g., "AI for climate change mitigation")
- Dynamically rewritten for search optimization (e.g., "Applications of AI in climate change adaptation and mitigation strategies")

ALGORITHM & DEPLOYMENT

III. Execution & Response Generation

- LLM fetches papers via ArXiv API
- Summarizes relevant papers chunk-wise
- Generates structured hypotheses and reports
- Reflects and retries failed steps, improving output quality without human feedback

IV. Deployment Setup

- **Frontend:** React.js app (served via IBM Cloud Object Storage or simple static host)
- **Backend:** FastAPI server hosted on IBM Cloud Lite
- **LLM API:** Watsonx.ai endpoint with Granite-3-8b-instruct model
- **Interfacing:** RESTful API between frontend and backend
- **Hosting Constraints:** All services optimized to run within IBM Cloud Lite's resource limits

RESULT

I. Output Quality & Effectiveness

The Research Agent successfully performs the following tasks with high consistency:

- Identifies and retrieves relevant academic papers for complex research questions using optimized queries.
- Generates accurate, readable summaries of research abstracts and full papers using IBM Granite LLM.
- Creates well-formatted citations in APA and IEEE style with minimal human input.
- Proposes insightful hypotheses based on synthesized information.
- Builds structured research reports combining multiple sources and logical reasoning.

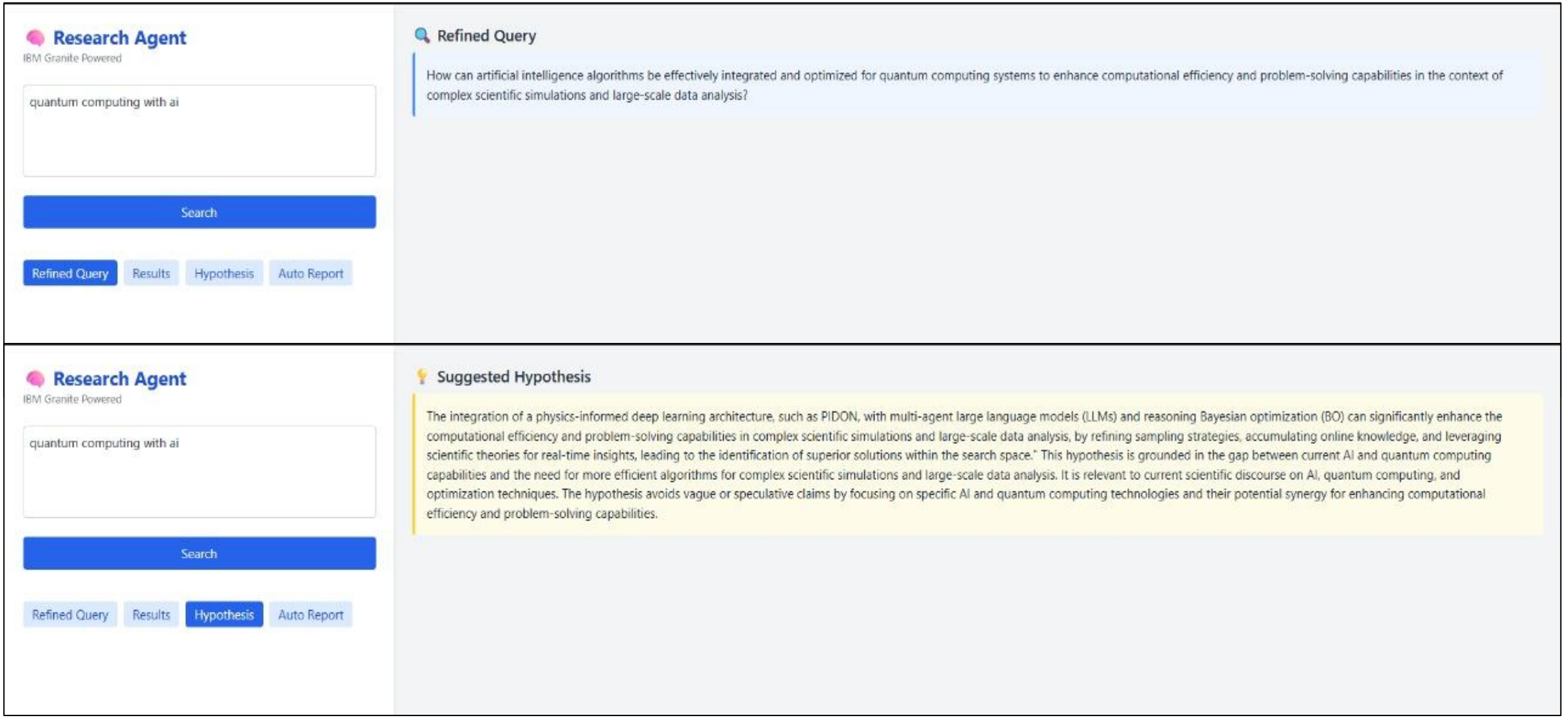
II. Autonomous Reflection & Improvement

- The system includes a self-evaluation mechanism where each step is reflected upon by the LLM.
- Unsuccessful steps are retried automatically, leading to improved reliability over time.

III. Performance Highlights

- Handles full research workflows without human-in-the-loop intervention
- Performs successfully on a wide range of topics, including:
    i. AI in healthcare
    ii. Climate change adaptation
    iii. Educational technologies
- Delivers final research output in structured, publication-ready format

RESULT


Research Agent
IBM Granite Powered
quantum computing with ai
Search
Refined Query
Results
Hypothesis
Auto Report
Refined Query
How can artificial intelligence algorithms be effectively integrated and optimized for quantum computing systems to enhance computational efficiency and problem-solving capabilities in the context of complex scientific simulations and large-scale data analysis?
Research Agent
IBM Granite Powered
quantum computing with ai
Search
Refined Query
Results
Hypothesis
Auto Report
Suggested Hypothesis
The integration of a physics-informed deep learning architecture, such as PIDON, with multi-agent large language models (LLMs) and reasoning Bayesian optimization (BO) can significantly enhance the computational efficiency and problem-solving capabilities in complex scientific simulations and large-scale data analysis, by refining sampling strategies, accumulating online knowledge, and leveraging scientific theories for real-time insights, leading to the identification of superior solutions within the search space." This hypothesis is grounded in the gap between current AI and quantum computing capabilities and the need for more efficient algorithms for complex scientific simulations and large-scale data analysis. It is relevant to current scientific discourse on AI, quantum computing, and optimization techniques. The hypothesis avoids vague or speculative claims by focusing on specific AI and quantum computing technologies and their potential synergy for enhancing computational efficiency and problem-solving capabilities.

RESULT

Research Agent
IBM Granite Powered

quantum computing with ai

Search

Refined Query | Results | Hypothesis | Auto Report

Search Results

Iterative Markov Chain Monte Carlo Computation of Reference Priors and Minimax Risk

Authors: John Lafferty, Larry A. Wasserman

1. The authors propose an iterative Markov chain Monte Carlo (MCMC) algorithm for calculating reference priors and minimax risk in general parametric families. 2. The algorithm is based on the Blahut-Arimoto algorithm from information theory, which is used to compute channel capacity. 3. The authors provide a statistical analysis of the algorithm, bounding the number of samples needed for the stochastic algorithm to closely approximate the deterministic algorithm in each iteration, with simulations presented for several examples from exponential families. The methods and analysis are applicable to a broader class of optimization problems and iterative algorithms.

▼ Citations

```
APA: John Lafferty, Larry A. Wasserman (2013). Iterative Markov Chain Monte Carlo Computation of Reference Priors and
  Minimax Risk. Retrieved from http://arxiv.org/abs/1301.2286v1IEEE: [1] John Lafferty, Larry A. Wasserman, "Iterative Markov Chain Monte Carlo Computation of Reference Priors and
  Minimax Risk," 2013. [Online]. Available: http://arxiv.org/abs/1301.2286v1BibTeX:@article{lafferty2013,
  title={ Iterative Markov Chain Monte Carlo Computation of Reference Priors and
  Minimax Risk },
  author={ John Lafferty and Larry A. Wasserman },
  year={ 2013 },
  url={ http://arxiv.org/abs/1301.2286v1 }
}
```

View on arXiv →

Harnessing Multi-Agent LLMs for Complex Engineering Problem-Solving: A Framework for Senior Design Projects

Authors: Abdullah Mushtaq, Muhammad Rafay Naeem, Ibrahim Ghaznavi, Muhammad Imran Taj, Imran Hashmi, Junaid Qadir

1. The paper introduces the concept of using Multi-Agent Large Language Models (LLMs) in senior design projects, which are crucial in engineering education for integrating theoretical knowledge with practical application. 2. The proposed framework involves distinct LLM agents representing various expert perspectives, such as problem formulation, system complexity, societal and ethical considerations, or project management, to facilitate a holistic problem-solving approach. 3. The framework leverages standard multi-agent system concepts, including coordination, cooperation, and negotiation, and adapts techniques from swarm AI to balance individual contributions towards a unified solution, encouraging interdisciplinary reasoning and negotiation similar to real-world senior design projects.

▶ Citations

View on arXiv →

Reasoning BO: Enhancing Bayesian Optimization with Long-Context Reasoning Power of LLMs

Authors: Zhuo Yang, Lingli Ge, Dong Han, Tianfan Fu, Yuqiang Li

1. The paper introduces Reasoning BO, a novel framework that combines Bayesian Optimization (BO) with reasoning models, multi-agent systems, and knowledge graphs to guide the sampling process

RESULT

Research Agent
IBM Granite Powered

quantum computing with ai

Search

Refined Query | Results | Hypothesis | Auto Report

Full Report

Introduction

The advent of quantum computing has ushered in a new era of computational possibilities, promising exponential speed-ups in solving complex scientific simulations and analyzing large-scale data sets. However, the integration and optimization of artificial intelligence (AI) algorithms within quantum computing systems to fully harness these potential benefits remain largely unexplored. This paper aims to address this gap by examining how AI algorithms can be effectively integrated and optimized for quantum computing systems, thereby enhancing computational efficiency and problem-solving capabilities.

To begin, we draw inspiration from the iterative Markov chain Monte Carlo (MCMC) algorithm proposed by the authors, which leverages the Blahut-Arimoto algorithm from information theory to compute channel capacity. This algorithm, rooted in statistical analysis, offers valuable insights into bounding the number of samples needed for the stochastic algorithm to closely approximate the deterministic algorithm in each iteration. The methods and analysis presented are applicable to a broader class of optimization problems and iterative algorithms, suggesting potential avenues for integration within quantum computing systems.

In parallel, the concept of using Multi-Agent Large Language Models (LLMs) in senior design projects, as introduced by another research group, presents a compelling framework for integrating theoretical knowledge with practical application. The proposed framework involves distinct LLM agents representing various expert perspectives, facilitating a holistic problem-solving approach. This approach could be adapted for quantum computing systems, where multiple AI agents could collaborate to address complex scientific simulations and data analysis tasks.

Furthermore, the introduction of Reasoning BO, a novel framework that combines Bayesian Optimization (BO) with reasoning models, multi-agent systems, and knowledge graphs, offers another promising avenue for AI-quantum computing integration. Reasoning BO leverages LLMs for real-time sampling recommendations and critical insights grounded in scientific theories, enhancing the BO process and aiding in the discovery of superior solutions within the search space. The framework's evaluation across diverse tasks demonstrates its ability to refine sampling strategies through real-time insights and hypothesis evolution, leading to the identification of higher-performing regions of the search space.

Lastly, the Physics-Informed DeepONet (PIDON) architecture, which extends the capabilities of conventional neural operators, provides a robust model for complex engineering systems across high-dimensional design spaces and various dynamic design configurations. PIDON's end-to-end gradient-based optimization framework accelerates the design process, demonstrating a 3x speedup in obtaining optimal design variables for aerospace-grade composites curing processes compared to gradient-free methods. This architecture could be further optimized for quantum computing systems, potentially enabling more efficient automated scientific experimentation while maintaining computational feasibility.

In conclusion, this paper seeks to explore the potential of integrating and optimizing AI algorithms within quantum computing systems to enhance computational efficiency and problem-solving capabilities. By drawing on the insights from the MCMC algorithm, multi-agent LLM framework, Reasoning BO, and PIDON architecture, we aim to pave the way for a new era of AI-enhanced quantum computing, where the power of both technologies can be harnessed to tackle complex scientific simulations and large-scale data analysis tasks.

Conclusively, the integration and optimization of AI algorithms for quantum computing systems present a significant opportunity to revolutionize scientific simulations and data analysis. By leveraging the strengths of each technology and addressing the unique challenges posed by quantum computing, we can unlock unprecedented computational power and problem-solving capabilities.

Related Work

The integration of artificial intelligence (AI) algorithms with quantum computing systems has garnered significant attention in recent years, as both technologies continue to evolve. The proposed iterative Markov chain Monte Carlo (MCMC) algorithm for calculating reference priors and minimax risk in general parametric families (Summaries 1 and 2) provides a foundation for understanding the statistical underpinnings of such integration. This algorithm, based on the Blahut-Arimoto algorithm from information theory, offers a method for optimizing computational efficiency in quantum computing systems. The statistical analysis presented by the authors, bounding the number of samples needed for the stochastic algorithm to closely approximate the deterministic algorithm in each iteration, is particularly relevant for ensuring the reliability and accuracy of quantum computations (Summary 3).

CONCLUSION

The proposed Research Agent system effectively automates the end-to-end academic research pipeline using IBM's Granite-3B-Instruct LLM on Watsonx.ai. It demonstrates strong capability in decomposing research queries, retrieving relevant literature, summarizing insights, generating citations, and compiling structured reports.

Key Strengths:

- Fully autonomous execution of complex research tasks.
- Integrated reflection and retry mechanism for improving result quality.
- Modular architecture with independent frontend and backend APIs.

Challenges Faced:

- IBM Cloud Lite's token and compute limitations occasionally restricted long-running sessions.
- Citation formatting and search relevance required fine-tuning to avoid inconsistencies.
- Reflection parsing had to be carefully controlled to avoid unstructured LLM responses.

FUTURE SCOPE

The Research Agent system lays the foundation for scalable, intelligent academic automation. Future development can significantly enhance its capabilities and impact.

System Enhancements

- **Smarter Reasoning:** Incorporate multi-agent collaboration and deeper reasoning chains for more advanced research workflows.
- **Improved Reflection:** Make step validation more nuanced using confidence scoring and output grading.
- **Multi-step Query Understanding:** Enable dynamic context tracking for ongoing, evolving research questions.

Expanded Coverage

- **Cross-domain Integration:** Extend support beyond academic literature to include datasets, reports, and patents.
- **Multi-lingual Support:** Support research queries and results in multiple languages for broader global access.

Advanced AI & Tech Stack

- **Model Upgrade:** Transition to larger foundation models (like Granite-13B or multimodal variants) as IBM expands access.
- **Emerging Tech:** Leverage edge AI or federated learning for on-device research assistants.

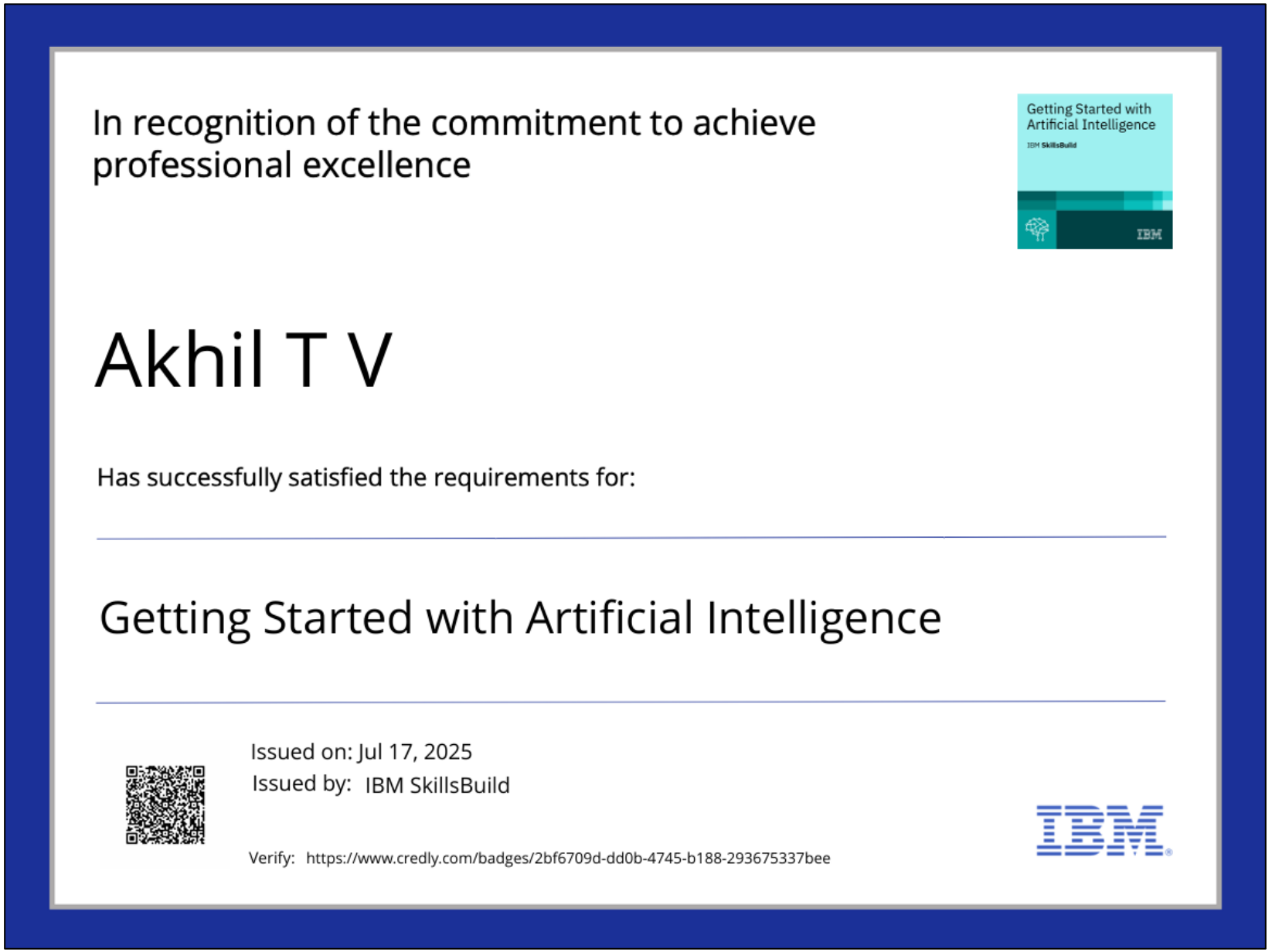
Practical Deployments

- Use cases in academic research, industrial R&D, policy research, and technical literature review in corporations.

REFERENCES

- Gu, N., & Hahnloser, R. H. R. (2023). SciLit: A Platform for Joint Scientific Literature Discovery, Summarization and Citation Generation. arXiv preprint. https://arxiv.org/abs/2306.03535
- Rouzrokh, P., & Shariatnia, M. (2025). LatteReview: A Multi-Agent Framework for Systematic Review Automation Using Large Language Models. arXiv preprint. https://arxiv.org/abs/2501.05468
- Li, Y., Chen, L., Liu, A., Yu, K., & Wen, L. (2024). ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary. arXiv preprint. https://arxiv.org/abs/2403.02574
- AgentX (2025). How to Build an AI Agent Research Team: From Concept to Automation. https://www.agentx.so/mcp/blog/how-to-build-an-ai-agent-research-team-from-concept-to-automation
- Semantic Scholar. AI-powered academic search engine providing automated summaries and citation tracking. https://en.wikipedia.org/wiki/Semantic_Scholar

IBM CERTIFICATIONS



In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence
IBM SkillsBuild

IBM

Akhil T V

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

Issued on: Jul 17, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/2bf6709d-dd0b-4745-b188-293675337bee

IBM

edunet foundation

IBM CERTIFICATIONS

In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution
IBM SkillsBuild

Akhil T V

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

Issued on: Jul 19, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/d024fee0-9c8d-416e-bf50-df26c718ba89

IBM

edunet foundation

IBM CERTIFICATIONS

IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Akhil T V

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 28 Jul 2025 (GMT)

Learning hours: 20 mins

edunet foundation

THANK YOU

edunet foundation