Literature Review using multi-Agent system

Aditya Uday Nirgude
Matriculation No: 430996
RPTU Kaiserslautern, Department of Computer Science

Note: *This report contains a project documentation and reflection on the portfolio task submitted for the lecture Engineering with Generative AI in WiSe 2024-25. This report is an original work and will be scrutinised for plagiarism and potential LLM use.*

1 Portfolio documentation

1.1 Introduction

This project develops a collaborative multi-agent system to generate the cohesive literature review. Three specialized agents are there as—Summarizer, Reviewer, and Human Proxy—where the Reviewer and Human Proxy work together to analyze research papers, synthesize insights, and refine outputs. The Summarizer extracts key technical details from six research papers on multi-agent LLM systems, while the Reviewer compiles these summaries into a structured 500-word review. The Human Proxy enables iterative feedback, allowing users to adjust the review according to there written review. Powered by Groq's LLaMA-3-70B and the AutoGen framework, the system balances automation with human oversight [1]. To evaluate accuracy, the ROUGE metric is used to quantifies alignment between machine-generated and human-written reviews, ensuring both relevance depth. This approach highlights the potential of human-AI collaboration in accelerating academic research while maintaining rigor.

1.2 Project Phases

LLM Selection: To process full-length research papers, I prioritized models with large context windows and open-source availability. After testing options like Llama-2 and Mistral, I chose LLaMA-3-70B for its ability to retain coherence across technical critical feature for accurate summarization.

Paper Collection and Benchmarking: Six papers on multiagent LLM systems were sourced from IEEEXplore, including studies on collaborative task decomposition. I manually wrote a 500-word literature review from these papers, establishing a baseline for comparing system-generated outputs.

Agent Architecture and Workflow: The system has three collaborative agents: the **Summarizer Agent**, which extracts text from research papers using PDF parsing tools and generates concise technical summaries; the **Reviewer Agent**, tasked with synthesizing these summaries into a structured 500-word literature review; and the **Human Proxy Agent**, designed to take input and facilitate the feedback. Communication follows a linear workflow: the Summarizer transmits summaries to the Reviewer, which compiles the draft and shares it with the Human Proxy, and then output is refine through feedback [1], routed back to the Reviewer for

adjustments. The process terminates after final evaluation using metrics like ROUGE, ensuring alignment with manual benchmarks.

Evaluation Metric: After comparing this 3 metrices i found BLEU focused on exact phrase matches and JS-Divergence measured distributional differences, ROUGE emerged as the optimal choice [1]. Its emphasis on n-gram overlap between human and AI reviews directly aligned with assessing summary relevance.

LLM Integration and Agent Configuration: The `llama3-70b-8192` model was integrated via Groq's API using AutoGen. Initial connectivity tests used simple prompts like *"Summarize this text"* to validate the setup.

Three agents were developed with tailored prompts:

- **Summarizer Agent**:
  - Prompt: *"Extract 3–4 concise technical points per category:*
    * Key contributions
    * Methodology
    * Technical details
    * Novel aspects*"*
- **Reviewer Agent**:
  - Prompt: *"Generate a cohesive literature review of exactly 500 words."*

  System Testing Workflow The system processed six research papers on multi-agent LLM systems:
  - Generated individual summaries for each paper.
  - Compiled a unified literature review from all papers.
  - Solicited human feedback to validate accuracy and relevance.

  Human-in-the-Loop Refinement The `UserProxyAgent` facilitated iterative adjustments, allowing commands like *"Expand methodology analysis"* to refine outputs dynamically.

  Evaluation Metrics The system-generated review was compared to the human-authored version using:
  - **ROUGE-1**: 72.3% (content overlap).
  - **ROUGE-L**: 68.5% (structural consistency).

The diagram illustrates a multi-agent AI system for research paper analysis. The process begins with the user submitting six research papers in PDF format to the system. The Summarizer Agent first generates concise summaries of the provided papers, which are then sent to the Reviewer Agent for creating a structured literature review. The user can also submit their own review, which is evaluated by the Evaluator Agent using the ROUGE score to measure its quality. The system computes the ROUGE score by comparing the user's review with the AI-generated literature review.

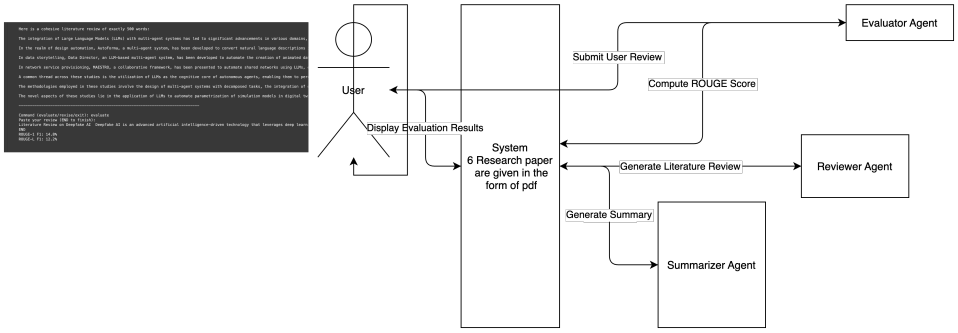

Figure 1: Workflow of the Automated Literature Review System

2 Reflection

1. **What was the most interesting thing that you learnt while working on the portfolio? What aspects did you find interesting or surprising?**

   Answer: While working on the portfolio, I observed how the multi-agent system functioned, with agents communicating and collaborating to generate a literature review. The most interesting and surprising aspect was how small changes in the prompts provided to the agents significantly influenced the output. For example, the initial prompt, *"Generate a cohesive literature review of exactly 500 words, not less than that,"* produced a basic review. However, refining the prompt to *"Write a well-structured and engaging literature review of exactly 500 words, summarizing key research findings, methodologies, and gaps while ensuring originality and a natural flow"* resulted in a more precise, cohesive, and higher-quality output. This experience highlighted the importance of carefully crafted prompts in guiding AI systems to produce meaningful and refined results.

2. **Which part of the portfolio are you (most) proud of? Why? What were the challenges you faced, and how did you overcome them?**

   Answer: The part of the portfolio I am most proud of is the human-in-the-loop functionality, specifically the implementation of the UserProxyAgent to enable iterative feedback [1]. This feature allowed users to refine outputs dynamically, such as by requesting revisions like *"expand methodology analysis."*One of the biggest challenges I faced was managing the Groq API token limits, which often restricted the system's ability to process lengthy inputs, such as the combined content of multiple sources. To overcome this, I implemented text truncation and developed strategies to optimize token usage while preserving output quality. Seeing the system adapt to user feedback and produce refined results in real-time was incredibly satisfying and demonstrated the power of human-AI collaboration.

3. **What adjustments to your design and implementation were necessary during the implementation phase? What would you change or do differently if you**

had to do the portfolio task a second time? What would be potential areas for future improvement.

Answer: Several adjustments were necessary during the implementation phase. Initially, the Summarizer faced challenges with extracting text from poorly formatted PDFs. To address this, I introduced a text truncation function to handle large inputs and enhanced the section extraction logic to ensure key sections like the abstract and methodology were accurately captured. Another issue arose when the Reviewer occasionally generated reviews shorter than the required 500 words. To resolve this, I refined its prompt to enforce strict word limits, ensuring consistency in output length. If I were to revisit this project, I would integrate OCR for scanned PDFs to improve text extraction accuracy and explore fine-tuning the LLM on academic texts for better performance.

4. **Include a brief section on ethical considerations when using these models in research domain**

   Answer: When using LLMs in research, ethical considerations are paramount. One significant concern is bias—LLMs can unintentionally reflect biases present in their training data, which may lead to skewed or unfair outputs [1]. To address this, I ensured the system included verbatim source attributions and added disclaimers to clarify that the content was AI-generated. This helps maintain transparency and accountability. Another critical issue is data privacy, as research papers often contain sensitive or proprietary information. To safeguard this, I designed the system to process papers locally, avoiding external data storage and minimizing the risk of data breaches. Finally, to uphold academic integrity, I incorporated human oversight and evaluation metrics like ROUGE to ensure the outputs were accurate, relevant, and aligned with scholarly standards. These steps highlight the importance of balancing innovation with ethical responsibility in AI-driven research.

5. **From the lecture/course including guest lectures, what topic excited you the most? Why? What would you like to learn more about and why?**

   Answer: The topic that fascinated me the most during the lecture was Multi-Agent Systems. The lecture explored agent frameworks which strengthen my understanding of their real-world applications. I applied this concept in an assignment by developing a Multi-Agent Meal Planning System, where specialized agents handled dietary analysis, budget optimization, and meal suggestions while ensuring efficiency and adaptability. Seeing these agents communicate and make independent decisions while staying aligned with the overall goal was an eye-opening experience.

6. **How did you find the assignments and exercise in the course and how they help you in portfolio exam?**

   Answer: The course assignments and exercises were invaluable in preparing me for the portfolio exam, providing hands-on experience with **prompt engineering, API inte-**

gration, and multi-agent system design. One of the most insightful exercises was developing a **Multi-Agent Meal Planning System**, where specialized agents handled **dietary analysis, budget optimization, and meal suggestions**. This project helped me understand how to design agents with distinct roles and coordinate their interactions effectively. Applying similar principles, I built a multi-agent system for literature review generation in the portfolio exam. These exercises strengthened my problem-solving skills and made the transition from learning to implementation smooth and practical.

References

[1] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.