Mixture of Experts for Interactive Literature Analysis

Stanley Joel Gona
University of Potsdam
gona@uni-potsdam.de

Abstract

As the volume of research literature in natural language processing (NLP) continues to expand, extracting actionable insights from large collections of papers has become increasingly challenging. In this work, I introduce a modular, interactive system that leverages the Mixture of Experts (MoE) paradigm in combination with large language models (LLMs) to provide users with on-demand analytical capabilities for research papers. My system offers three LLM-powered experts such as summarization, contribution extraction and question answering, allowing users to flexibly obtain targeted insights from any selected paper. I describe the design and implementation of my system and discuss the impact of MoE-inspired modularity on literature analysis workflows. **The code is available at**[1].

1 Introduction

Large language models (LLMs) such as GPT-4(OpenAI, 2023), Gemini(Team et al., 2023), BART(Lewis et al., 2019) and FLAN-T5(Chung et al., 2022) have transformed the landscape of natural language processing, enabling a wide array of applications from document summarization to interactive question answering. As academic literature in fields like NLP continues to grow at an unprecedented pace, researchers face increasing difficulty in rapidly extracting key insights from the vast body of published work.

While automated literature review tools exist, they often lack flexibility, modularity or the ability to offer users a range of tailored analytical functions. Meanwhile, the Mixture of Experts (MoE) paradigm(Shazeer et al., 2017; Fedus et al., 2022) has shown promise for improving both efficiency and task specialization in complex machine learning systems. Traditionally, MoE involves orchestrating a collection of specialized models or modules, each expert in a specific subtask, to collaboratively solve more complex problems.

In this work, I present a modular, interactive research paper assistant based on the Mixture of Experts approach. My system enables users to select any research paper and consult a suite of LLM-powered experts including a summarizer, a contribution extractor and a question answering module to receive focused, high-quality insights on demand. By allowing users to flexibly route their queries to different experts for the same paper, my system blends the strengths of MoE architectures with an intuitive, human-in-the-loop interface.

Unlike traditional MoE systems that employ learned gating networks to automatically route inputs to experts, my approach adopts a human-in-the-loop design where users themselves act as the routing mechanism. This design choice offers greater transparency and interpretability, allowing researchers to iteratively consult different experts based on their evolving information needs. Each expert is optimized through task-specific prompting strategies, mirroring the specialization principle central to MoE while maintaining the flexibility required for exploratory literature analysis.

I describe the design and implementation of my system, analyze its outputs qualitatively and discuss the opportunities and challenges that arise when applying MoE-inspired modularity to real-world literature analysis workflows.

The system leverages Gradio[2] to provide an accessible web-based interface, enabling researchers to interact with the expert modules without requiring technical expertise in API integration or command-line tools.

[1] https://github.com/stanley7/Interactive-Literature-Analysis

[2] https://gradio.app

2 Related Work

2.1 Automated Literature Review Tools

The challenge of efficiently analyzing research literature has led to the development of various automated tools. Semantic Scholar(Lo et al., 2020) provides AI-powered search and recommendation features for academic papers, while systems like TLDR(Cachola et al., 2020) focus on generating concise summaries of scientific documents. More recently, tools such as Elicit and Consensus have emerged, leveraging large language models to answer research questions by synthesizing information across multiple papers. However, these systems typically offer fixed functionality and limited customization, restricting users to predefined analytical workflows rather than allowing flexible, task-specific exploration.

2.2 Document Summarization and Analysis

Neural approaches to document summarization have evolved significantly with the advent of pre-trained language models. BART(Lewis et al., 2019) and PEGASUS(Zhang et al., 2020) demonstrated strong performance on abstractive summarization tasks, while more recent work has explored instruction-tuned models like FLAN-T5(Chung et al., 2022) for improved controllability. In the scientific domain, systems like SciSummPip(Goldsack et al., 2022) and LongT5(Guo et al., 2022) have addressed the challenge of summarizing lengthy technical documents. Despite these advances, most approaches focus on single-task optimization rather than providing modular, multi-functional analysis capabilities.

2.3 Mixture of Experts Architectures

The Mixture of Experts paradigm was introduced by Shazeer et al.(Shazeer et al., 2017) as a method to scale neural networks by conditionally activating specialized sub-networks. Switch Transformers(Fedus et al., 2022) further demonstrated that sparse MoE models could achieve better performance with fewer computational resources by routing inputs to task-specific experts. Recent work has extended MoE to multimodal domains(Riquelme et al., 2021) and even to billion-parameter language models like Mixtral(Jiang et al., 2024). These systems typically employ learned gating mechanisms that automatically determine expert selection based on input characteristics.

2.4 Interactive Systems for Research

Interactive question-answering systems for scientific literature have gained attention with the rise of conversational AI. Systems like QASPER(Dasigi et al., 2021) provide datasets and benchmarks for answering questions about research papers, while tools like Iris.ai and Scholarcy offer commercial solutions for literature exploration. However, these systems rarely combine multiple analytical functions in a unified, modular framework that allows users to iteratively refine their understanding through different types of queries.

2.5 Positioning of This Work

Unlike existing literature analysis tools that provide fixed analytical pipelines, my system adopts an MoE-inspired modular architecture where users act as the intelligent routing mechanism. This human-in-the-loop design combines the specialization benefits of MoE with the flexibility required for exploratory research, allowing researchers to consult different experts based on their evolving information needs. By integrating summarization, contribution extraction, and question answering in a single interactive interface, my system bridges the gap between automated analysis and user-driven exploration.

3 Methodology

3.1 System Overview

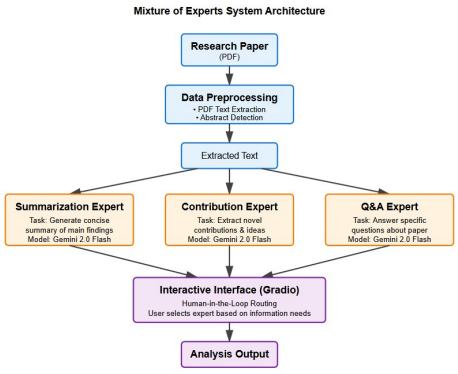

Figure 1: System architecture showing the flow from PDF input through preprocessing to three specialized LLM experts, with human-in-the-loop routing via the Gradio interface.

My system follows a modular pipeline architecture consisting of three main components: data preprocessing, expert modules, and an interactive user interface. Figure 1 illustrates the overall sys-

tem design.[3] When a user selects a research paper, the system first extracts and preprocesses the text content, then makes it available to three specialized LLM-powered experts: a summarization expert, a contribution extraction expert, and a question-answering expert. Unlike traditional MoE systems with learned routing mechanisms, users directly select which expert to consult based on their analytical needs, enabling flexible and transparent exploration of the paper's content. All expert modules are powered by Google's Gemini 2.0 Flash model, accessed via API, with each expert employing task-specific prompting strategies to optimize performance for its designated function.

3.2 Data Preprocessing

The data preprocessing pipeline handles the extraction and preparation of textual content from PDF research papers. The system uses PyMuPDF (fitz)[4] to parse PDF files and extract raw text from each page. This library was chosen for its efficiency in handling academic papers with complex layouts, mathematical notation, and multi-column formats.

Once the full text is extracted, the system employs a pattern-matching approach to identify and isolate the abstract section, which typically contains the most concentrated information about the paper's contributions. The extraction uses a regular expression that searches for common abstract delimiters, including section headers (e.g., "Abstract", "ABSTRACT") and subsequent section markers (e.g., "Introduction", "1.", "I."). If the abstract cannot be reliably identified, the system defaults to using the first 1,500 characters of the document, which typically encompasses the abstract in most academic paper formats.

This preprocessing step serves two critical functions: first, it reduces the input length sent to the LLM, thereby improving response time and reducing API costs; second, it focuses the expert modules on the most information-dense portion of the paper, leading to more accurate and relevant outputs. The extracted text is then passed unchanged to all three expert modules, ensuring consistency across different analytical perspectives.

[3] System architecture diagram showing the flow from PDF input through preprocessing to the three expert modules and user interface.

[4] https://pymupdf.readthedocs.io

3.3 Expert Design

Each expert module is designed to excel at a specific analytical task through carefully crafted prompting strategies. All experts utilize the Gemini 2.0 Flash model, which offers a strong balance between performance, cost-efficiency, and response latency. The specialization occurs not through different model architectures, but through task-specific prompt engineering that guides the model's behavior toward the desired analytical function.

3.3.1 Summarization Expert

The summarization expert is designed to provide users with a concise overview of the paper's main findings and contributions. The prompt instructs the model to generate a summary in 3-5 sentences, focusing specifically on the core results while ignoring metadata such as author names and institutional affiliations. This constraint ensures that the summary remains focused on substantive content rather than peripheral information. The prompt is structured as follows:

> *"Summarize the following research paper abstract in 3-5 sentences, focusing on the main findings and contributions. Ignore author names and affiliations."*

By limiting the output length, the summarization expert provides users with a rapid overview suitable for initial screening of papers or quick reference during literature review workflows.

3.3.2 Contribution Extraction Expert

The contribution extraction expert identifies and articulates the novel contributions and key ideas presented in the research paper. Unlike the summarization expert, this module is prompted to produce structured output in bullet-point format, making it easier for users to quickly scan and identify specific innovations. The prompt instructs the model to focus on novelty and significance:

> *"List the main contributions and novel ideas of the following research paper as bullet points. Ignore author names and affiliations."*

This expert is particularly valuable when users need to understand what distinguishes a paper from prior work or when comparing multiple papers' contributions side-by-side.

3.3.3 Question Answering Expert

The question answering expert provides flexible, interactive analysis by responding to user-specified questions about the paper's content. This expert receives both the paper's abstract and the user's question as input, allowing for targeted information retrieval. This open-ended design enables users to pursue diverse analytical goals, from clarifying methodological details to understanding implications or relating the work to other research. The Q&A expert represents the most flexible component of the system, adapting to the user's evolving information needs throughout the exploration process.

3.4 Human-in-the-Loop Routing

A key distinction between my system and traditional MoE architectures lies in the routing mechanism. While conventional MoE systems employ learned gating networks that automatically route inputs to appropriate experts based on learned representations, my system places the human user at the center of the routing decision. This design choice reflects the inherently exploratory and iterative nature of literature analysis, where researchers' information needs evolve as they develop understanding of a paper's content.

In practice, users can consult multiple experts for the same paper in any order, allowing them to triangulate understanding from different analytical perspectives. For example, a researcher might first request a summary to determine the paper's relevance, then extract contributions to identify novel ideas, and finally pose specific questions to clarify technical details or methodology. This flexibility is difficult to capture in automated routing systems, which must commit to a single expert or set of experts based on static input features.

The human-as-router approach offers several advantages: **(1) Transparency** - users explicitly understand which analytical function is being applied. **(2) Control** - researchers can pursue their specific analytical goals without being constrained by system predictions. **(3) Iterative refinement** - users can sequentially consult different experts as their understanding deepens. **(4) Interpretability** - the relationship between user intent and system output remains clear throughout the interaction.

This design acknowledges that for complex analytical tasks like literature review, the "best" expert to consult depends not only on document characteristics but also on user context, prior knowledge, and current analytical objectives, factors that are difficult to capture in automated gating mechanisms.

3.5 Interface Design

The system's user interface is implemented using Gradio, a Python library designed for rapid prototyping of machine learning applications. The interface employs a tab-based design that naturally maps to the three expert modules, providing clear visual separation between different analytical functions while maintaining a unified workflow.

The main interface consists of two primary components: a paper selection panel and an expert consultation area. In the selection panel, users choose from available PDF papers via a dropdown menu, and a preview window displays the first 600 characters of the selected paper, allowing users to verify their selection and gain initial context. This preview serves as a lightweight confirmation mechanism before invoking any LLM-powered analysis.

The expert consultation area is organized into three tabs that is "Summarize", "Extract Contributions" and "Q&A" each corresponding to one of the specialized expert modules. Each tab contains a simple interface with a button to trigger the analysis and a text area to display results. The Q&A tab additionally includes an input field where users can type their specific questions. This design minimizes cognitive load by presenting only the controls relevant to each analytical task.

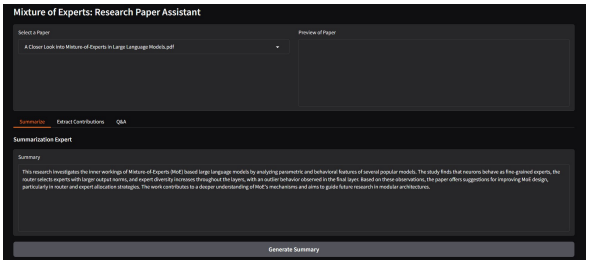

Figure 2: Screenshot of the Gradio-based user interface showing the paper selection dropdown, preview panel, tab-based expert modules, and sample output from the summarization expert.

All LLM interactions are handled asynchronously to maintain interface responsiveness, with error messages displayed directly in the output area if API calls fail. The stateless design means that each expert consultation is independent, allowing users to freely switch between tabs and papers without losing functionality. This architecture prioritizes simplicity and accessibility, ensuring that

researchers can focus on analysis rather than navigating complex interface mechanics.

4 Ethical Considerations

The deployment of LLM-powered systems for academic literature analysis raises important ethical concerns. First, the risk of **hallucinations** means that the system may occasionally generate plausible but factually incorrect information, requiring users to verify critical details against original papers. Second, **oversimplification** through automated summarization may omit important nuances, caveats, or contextual details present in full papers. Third, over-reliance on the system may lead to **reduced engagement with primary sources**, potentially causing researchers to miss subtle insights that emerge only through careful reading. Finally, **bias propagation** from the underlying LLM may reflect biases related to author demographics, institution prestige, or research topics. To mitigate these concerns, users should treat system outputs as preliminary insights rather than authoritative interpretations, always supplementing automated analysis with thorough reading of original papers.

5 Conclusion

In this work, I presented a modular, interactive research paper assistant that adapts the Mixture of Experts paradigm for literature analysis. My system provides three specialized LLM-powered experts like summarization, contribution extraction, and question answering accessible through an intuitive web-based interface. By adopting a human-in-the-loop routing mechanism, the system offers researchers greater transparency, control, and flexibility compared to automated analysis tools.

The key contributions of this work include: **(1)** a practical application of MoE principles to literature analysis with human-guided routing, **(2)** task-specific prompt engineering strategies for three complementary analytical functions, and **(3)** an accessible implementation using Gradio that lowers the barrier to entry for researchers seeking LLM-assisted paper analysis.

5.1 Limitations and Future Work

Several directions for future development merit consideration. First, extending the system to analyze full paper text rather than only abstracts would enable deeper insights, though this requires addressing the computational and cost challenges of processing lengthy documents. Second, implementing a learned routing component that suggests relevant experts based on user queries could combine the benefits of automation with human oversight. Third, expanding language support and document format compatibility would increase the system's applicability across diverse research communities.

Additional enhancements could include citation network analysis, comparative analysis across multiple papers, and integration with reference management tools. Evaluation through user studies would provide valuable insights into how the system impacts real-world literature review workflows and which expert configurations best serve different research tasks.

The code and documentation for this system are publicly available to support reproducibility and community extensions.[5]

References

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Findings of EMNLP*, pages 4766–4777.

Hyung Won Chung, Le Hou, and Shayne Longpre. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL-HLT*, pages 4599–4610.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of EMNLP*, pages 10589–10604.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. *Findings of NAACL*, pages 724–736.

Albert Q Jiang, Alexandre Sablayrolles, and Arthur Mensch. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mike Lewis, Yinhan Liu, and Naman Goyal. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[5] https://github.com/stanley7/Interactive-Literature-Analysis

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of ACL*, pages 4969–4983.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Carlos Riquelme, Joan Puigcerver, and Basil Mustafa. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:1539–1551.

Noam Shazeer, Azalia Mirhoseini, and Krzysztof Maziarz. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Gemini Team, Rohan Anil, and Sebastian Borgeaud. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of ICML*, pages 11328–11339.