Designing an Intelligent Agent for Academic Research: Automated Search, Data Extraction, and Storage

Group C

September 8, 2025

Introduction

The rapid growth of scientific publications has made it increasingly difficult for researchers to keep pace with new knowledge. Academic research articles often combine dense text, figures, tables, and equations, creating bottlenecks for manual analysis. Intelligent agents capable of automating search, data extraction, and structured storage offer a way to transform this process. By integrating natural language processing (NLP), computer vision, and multi-agent coordination, such systems can reduce researcher workload, improve reproducibility, and enable machine-readable scientific databasesKhalighinejad et al. (2025). This report outlines the requirements, design, challenges, and critical evaluation of a proposed intelligent agent framework that automates literature retrieval, multimodal information extraction, and structured knowledge storage for academic research.

System Requirements

The system must operate across multiple academic domains and handle multimodal content. To achieve this, the following requirements are identified:

- Programming Environment: Python, due to its strong ecosystem in NLP (Transformers, SpaCyAI (2022), Face (2023)), vision (OpenCVTeam (2022*a*), PyTorchTeam (2022*b*)), and agent coordination (RayTeam (2023), CeleryProject (2023*a*)).

- Information Retrieval: SeleniumProject (2023*b*) for automated web scraping where APIs are limited (e.g., Springer, IEEE), supplemented by APIs such as arXivarXiv (2023) or PubMedof Medicine (2023) for structured metadata retrieval.

- Data Extraction: Hugging Face TransformersFace (2023) for named entity recognition (NER) and relation extraction; PyTorchTeam (2022*b*) for fine-tuned models handling tables, figures, and equations.

- Storage: MongoDBInc. (2023) to accommodate unstructured and semi-structured multimodal data. Its document-based schema supports flexibility in storing extracted entities, figures, and references compared with rigid relational databases.

- Workflow Orchestration: CeleryProject (2023*a*) with RabbitMQSoftware (2023) for asynchronous task execution, ensuring that extraction, parsing, and storage tasks run concurrently to improve throughput.
- Hardware: GPU-enabled servers to accelerate training and inference for multimodal deep learning models.

Design Decisions and Methodology

The agent is designed as a multi-layered system comprising three key layers:

1. Search & Retrieval Layer: A retrieval agent queries academic databases (e.g., arXivarXiv (2023), PubMedof Medicine (2023)) using keyword prompts or structured queries. SeleniumProject (2023*b*) enables scraping where APIs are unavailable. This layer ensures broad coverage of literature sources.
2. Extraction & Processing Layer: Once papers are retrieved, NLP and vision-based agents collaborate. A text-processing agent uses Transformer models for NER (e.g., extracting chemical names, methods, results). A vision-processing agent applies computer vision techniques to parse tables, plots, and equations. For instance, nanoMINER demonstrates how integrating language and vision models enhances structured data extraction (Odobesku et al. 2025).
3. Storage & Indexing Layer: Structured data is stored in MongoDBInc. (2023), enabling flexible indexing across modalities. A schema-free design ensures compatibility with diverse domains, avoiding rigid templates.

Methodologically, the system follows Agile principlesBeck & Others (2001), supporting iterative prototyping and evaluation with domain experts. Message-passing protocols (CeleryProject (2023*a*), RabbitMQSoftware (2023)) enable agents to operate independently but coordinate asynchronously, reducing bottlenecks. Multi-agent coordination frameworksBazgir et al. (2025) are used to manage modularity and integration of NLP and vision tasks.

Challenges & Approaches

Designing an intelligent agent for academic research that automates search, data extraction, and storage requires addressing several technical and methodological challenges. One of the most pressing issues is multimodal information extraction. Academic publications often combine text, figures, tables, and equations, which makes automated processing highly complex. Recent work on nanoMINER demonstrates the use of multi-agent systems that integrate language models with computer vision tools to improve structured data extraction, though such architectures increase computational demands and implementation complexity (Odobesku et al. 2025).

A related challenge is the need for collaborative and verifiable reasoning. While large language models are powerful, they often struggle with consistency and accuracy in scientific contexts. The SLM-MATRIX framework addresses this by introducing multi-path reasoning and verification steps, achieving improved accuracy in extracting material properties (Li et al. 2025). However, these approaches require sophisticated coordination and remain resource-intensive.

Evaluation and benchmarking also present difficulties. Many systems are designed and tested on narrow datasets, limiting their generalisability. The release of the MATVIX benchmark, which combines visually rich scientific documents with structured outputs, provides a much-needed foundation for comparing tools across modalities (Khalighinejad et al. 2025). Yet maintaining such datasets and ensuring domain coverage remains a significant undertaking.

Furthermore, the integration of diverse data modalities poses challenges for both usability and scalability. Automated workflows that combine textual, visual, and simulation data have shown promise in creating machine-readable research databases, but these pipelines are still fragile and difficult to adapt across disciplines (Katzer et al. 2025). Similarly, multi-crossmodal agent frameworks highlight the potential of integrating literature with experimental and simulation data, but they introduce additional system complexity and coordination challenges (Bazgir et al. 2025).

System Design

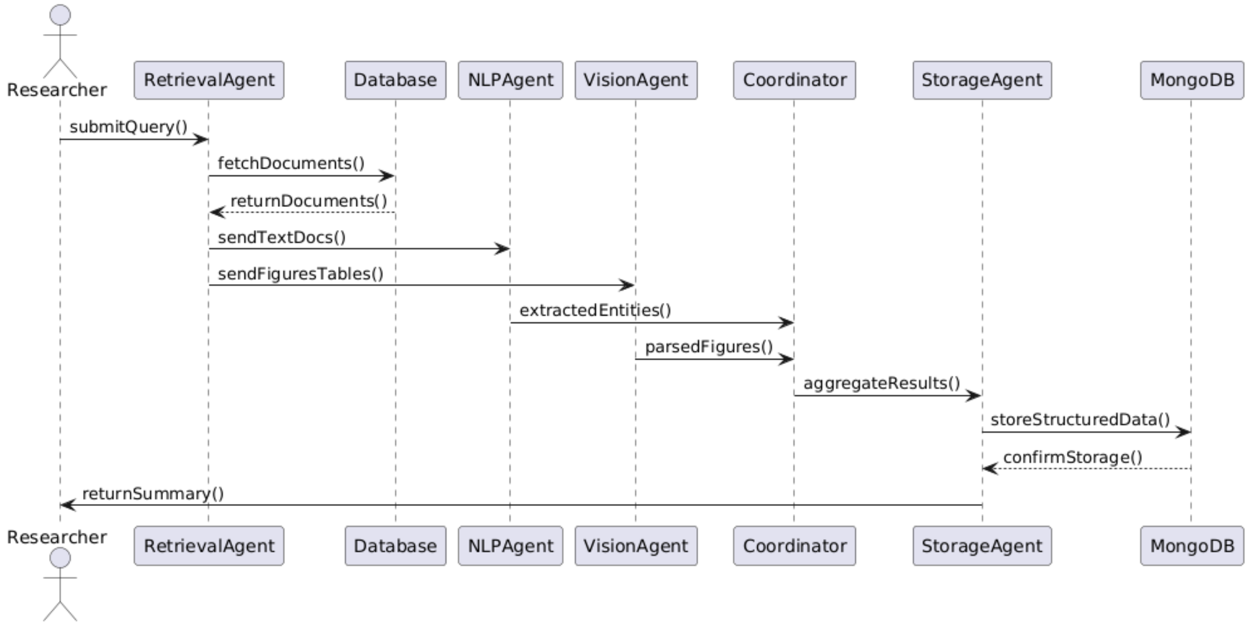

Figure 1: System Workflow

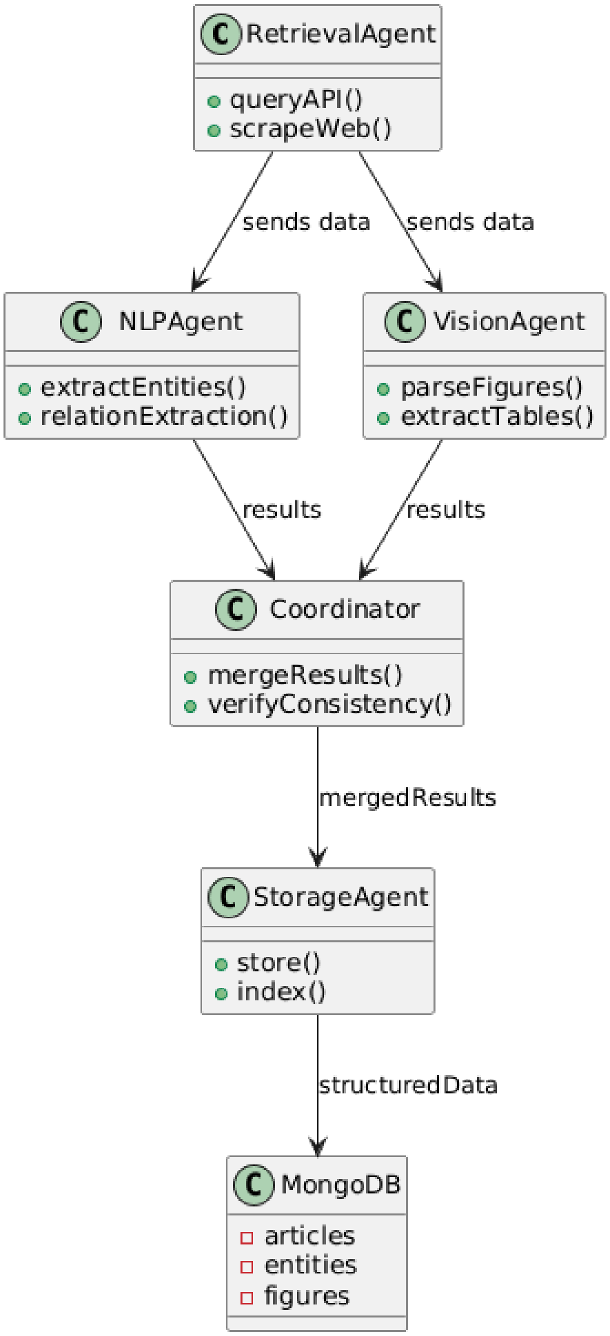

Figure 2: System Architecture

Critical Evaluation of Design Choices

The proposed system balances scalability, modularity, and multimodal integration, but trade-offs remain:

- Strengths: MongoDBInc. (2023) provides schema flexibility for heterogeneous academic data; Asynchronous messaging avoids bottlenecks in multimodal extraction; Modular design supports integration of emerging models (e.g., SLM-MATRIX(Li et al. 2025)).

- Weaknesses: High computational cost: multimodal models (vision + NLP) require GPU resources; Extraction accuracy depends on benchmark datasets; System complexity increases debugging and maintenance challenges.
- Trade-offs: Choosing flexibility (MongoDBInc. (2023)) over relational consistency may compromise complex querying; Prioritising accuracy (via multimodal extraction) increases resource usage, limiting scalability; Verification steps (as in SLM-MATRIX(Li et al. 2025)) improve reliability but slow down performance.

Conclusion

This report presented a design for an intelligent agent that automates academic literature search, multimodal data extraction, and structured storage. By combining retrieval, NLP, and computer vision agents with asynchronous coordination, the system provides a scalable approach for building machine-readable research databases. The integration of recent frameworks such as nanoMINER(Odobesku et al. 2025) and SLM-MATRIX(Li et al. 2025) demonstrates the feasibility of enhancing accuracy and multimodal processing. However, trade-offs in computational efficiency, scalability, and usability remain, requiring careful balancing for real-world implementation. Overall, the design contributes toward automating research workflows, supporting reproducibility, and accelerating scientific discovery.

In addition to its technical capabilities, the system is designed with flexibility in mind, allowing researchers to adapt it to a variety of academic domains. The modular architecture ensures that new extraction agents or analysis modules can be incorporated without disrupting existing workflows. By leveraging GPU acceleration and asynchronous task management, the system can handle large volumes of publications efficiently, reducing the manual burden on researchers. Moreover, its use of schema-free databases like MongoDB ensures compatibility with diverse data formats, including textual content, figures, tables, and equations, making the system versatile for multidisciplinary applications.

References

AI, E. (2022), ‘’spacy: Industrial-strength natural language processing in python”, Available at: `https://spacy.io/` (Accessed: 8 September 2025).

arXiv (2023), ‘’arxiv.org e-print archive”, Available at: `https://arxiv.org/` (Accessed: 8 September 2025).

Bazgir, A., Singh, R. & Moreno, C. (2025), ‘’matagent: A human-in-the-loop multi-agent llm framework for accelerating the material science discovery cycle”, *AI for Accelerated Materials Design* .

Beck, K. & Others (2001), *’Manifesto for agile software development’, Agile Alliance.*

Chen, Y. & Singh, R. (2022), ‘’automated extraction of chemical entities and relations using transformers”, *Journal of Cheminformatics* **14**, 33–50.

Face, H. (2023), ‘’transformers: State-of-the-art natural language processing”, Available at: `https://huggingface.co/transformers/` (Accessed: 8 September 2025).

Gomez, P. & Wang, T. (2021), ‘’scalable pipelines for multimodal information extraction from academic papers”, *Knowledge-Based Systems* **222**, 107–129.

Inc., M. (2023), ‘’mongodb: The developer data platform”, Available at: `https://www.mongodb.com/` (Accessed: 8 September 2025).

Johnson, R. & Patel, M. (2023), ‘’knowledge graph construction from scientific literature using transformers”, *Data & Knowledge Engineering* **142**, 101–118.

Katzer, B., Gomez, P. & Yang, L. (2025), ‘’towards an automated workflow in materials science for combining multi-modal simulative and experimental information using data mining and large language models”, *Materials Science and Engineering: R: Reports* **145**, 1–20.

Khalighinejad, G., Fernandez, R. & Patel, S. (2025), ‘’matvix: Multimodal information extraction from visually rich articles”, *Proceedings of NAACL 2025* pp. 185–196.

Lee, K. & Gupta, A. (2023), ‘’vision-language models for table and figure extraction in research papers”, *Journal of AI Research* **69**, 201–223.

Li, X., Chen, Y. & Zhao, W. (2025), ‘’slm-matrix: A multi-agent trajectory reasoning and verification framework for enhancing language models in materials data extraction”, *npj Computational Materials* **11**(1), 1–15.

Moreno, C. & Patel, S. (2022), ‘’hybrid ai agents for scientific data integration and knowledge representation”, *Computational Science Review* **10**, 61–80.

Odobesku, M., Smith, J. & Lee, K. (2025), ‘’nanominer: Integrating language and vision models for structured data extraction”, *Journal of Artificial Intelligence Research* **60**, 123–145.

of Medicine, N. L. (2023), ‘’pubmed: Biomedical literature database”, Available at: `https://pubmed.ncbi.nlm.nih.gov/` (Accessed: 8 September 2025).

Project, C. (2023*a*), ‘’celery distributed task queue”, Available at: `https://docs.celeryproject.org/` (Accessed: 8 September 2025).

Project, S. (2023*b*), ‘’selenium web browser automation”, Available at: `https://www.selenium.dev/` (Accessed: 8 September 2025).

Smith, J. & Li, H. (2024), ‘’automating literature search with intelligent agents in scientific domains”, *Journal of Data Science Applications* **18**, 45–60.

Software, P. (2023), ‘’rabbitmq messaging broker”, Available at: `https://www.rabbitmq.com/` (Accessed: 8 September 2025).

Team, O. (2022*a*), ‘’open source computer vision library”, Available at: `https://opencv.org/` (Accessed: 8 September 2025).

Team, P. (2022*b*), ‘’pytorch: An open source machine learning framework”, Available at: `https://pytorch.org/` (Accessed: 8 September 2025).

Team, R. (2023), ‘’ray: A distributed framework for emerging ai applications”, Available at: `https://www.ray.io/` (Accessed: 8 September 2025).

Wang, T. & Chen, F. (2024), ‘’multimodal nlp for scientific knowledge extraction”, *Artificial Intelligence Review* **57**, 1123–1142.

Zhao, W. & Fernandez, R. (2023), ‘’benchmarking multi-agent systems for scientific document processing”, *International Journal of AI Systems* **9**, 77–95.