Project number: CS55

Project Source: School of Computer Science

Project Title: Agentic Multimodal RAG: An Intelligent Framework for Scientific Concept Discovery from Text and Visuals

Project Description and Scope: Motivation
Despite advances in large language models and RAG systems, current approaches struggle to reason across text and visual data in scientific domains. Most systems passively retrieve documents and treat images and diagrams as secondary. Meanwhile, human researchers rely heavily on visual cues (figures, captions, diagrams) to interpret scientific results.
There is a need for an agentic multimodal RAG system that not only retrieves relevant information across modalities but actively plans, reasons, and generates structured scientific insights, such as abstracts or hypotheses — grounded in open-access research content.

Project Description
We propose a modular framework that combines agentic planning, multimodal retrieval, and structured generation to explore and summarize scientific concepts. The system acts as an autonomous research assistant, capable of:
Planning multi-step reasoning tasks (e.g., gather papers, extract figures, compare methods)
Retrieving text, figure captions, and images using contrastive models (e.g., CLIP, BLIP, SciBERT)
Generating mini-publications (title, abstract, intro) grounded in the retrieved context
The project leverages open-source datasets of scientific papers, figures, and image-caption pairs for training and evaluation.
Real uses Cases:
Autonomous Research Assistant (Scientific Discovery & Review)
Use Case: Help scientists explore new topics or write literature reviews
How: Given a query like "Latest methods in cancer image segmentation using transformers", the agent retrieves papers, figures, and captions — then synthesizes a summary or hypothesis.
Medical Literature Analysis & Clinical Decision Support
Use Case: Summarize radiology papers with images for hospital staff
How: Query "MRI-based tumor detection CNNs," and the agent finds papers, reads images, and returns grounded insights

Goals
Develop an agentic RAG pipeline that performs iterative goal planning and multimodal retrieval.
Integrate multimodal retrievers (CLIP/BLIP + SciBERT) for figure + text understanding.
Generate structured scientific outputs (e.g., abstract, hypothesis) from retrieved

multimodal context.
Evaluate grounding and novelty using automatic and human-in-the-loop methods.
Publish an open benchmark for multimodal agentic scientific reasoning.

Datasets: any open source datasets
https://github.com/ibm-aur-nlp/PubLayNet
https://github.com/allenai/s2orc
https://github.com/lupantech/ScienceQA
https://pmc.ncbi.nlm.nih.gov/tools/openftlist/

Related Papers & Prior Work
Lewis et al. (2020) – Retrieval-Augmented Generation (RAG)
https://arxiv.org/abs/2005.11401
Baseline for RAG in text domains; passive, not agentic or multimodal.
Liu et al. (2021) – DocPrompting: Generating Scientific Documents with Retrieved Context
https://arxiv.org/abs/2112.09334
Structured generation but text-only.
Lu et al. (2022) – BLIP: Bootstrapped Language–Image Pretraining
https://arxiv.org/abs/2201.12086
Multimodal pretraining useful for figure retrieval.
Mialon et al. (2023) – Augmented Language Models: a Survey
https://arxiv.org/abs/2302.07842
Overview of agent tools and planning, but not applied to science.
Chalkidis et al. (2023) – Prompt2Paper: Drafting Scientific Papers via Large Language Models
https://arxiv.org/abs/2305.14780
Text-only generation of papers; does not use visual input or retrieval.
KAT (2024) – Knowledge-Augmented Toolformer for Scientific QA
https://arxiv.org/abs/2404.06815

Expected outcomes/deliverables: Report and code

Specific required knowledge, skills, and/or technology: Machine Learning, Natural Language Processing, Algorithm

Fields that this project may involve: Data Science/Analytics;Artificial Intelligence;NLP;

Dataset provided by the client: Yes, the supervisor will provide dataset.

Resources provided by the client: Datasets: any open source datasets
https://github.com/ibm-aur-nlp/PubLayNet
https://github.com/allenai/s2orc
https://github.com/lupantech/ScienceQA
https://pmc.ncbi.nlm.nih.gov/tools/openftlist/

Related Papers & Prior Work
Lewis et al. (2020) – Retrieval-Augmented Generation (RAG)
https://arxiv.org/abs/2005.11401
Baseline for RAG in text domains; passive, not agentic or multimodal.
Liu et al. (2021) – DocPrompting: Generating Scientific Documents with Retrieved Context
https://arxiv.org/abs/2112.09334
Structured generation but text-only.
Lu et al. (2022) – BLIP: Bootstrapped Language–Image Pretraining
https://arxiv.org/abs/2201.12086
Multimodal pretraining useful for figure retrieval.
Mialon et al. (2023) – Augmented Language Models: a Survey
https://arxiv.org/abs/2302.07842
Overview of agent tools and planning, but not applied to science.
Chalkidis et al. (2023) – Prompt2Paper: Drafting Scientific Papers via Large Language Models
https://arxiv.org/abs/2305.14780
Text-only generation of papers; does not use visual input or retrieval.
KAT (2024) – Knowledge-Augmented Toolformer for Scientific QA
https://arxiv.org/abs/2404.06815