1st CAPSTONE PROJECT STATUS CHECKING FORM

Project Title:	CS55-Agentic Multimodal RAG_An Intelligent Framework for Scientific Concept Discovery from Text and Visuals
Project Client:	Ali Braytee, Ali Anaise
Group Number:	CS55-1
Date of client meeting:	

Project Information	**Summary of report & feedback**
Overall Information	
Project Description and Scope based on Group's understanding	• Build an agentic multimodal RAG system that can answer questions about scientific documents by retrieving and fusing text, figures and tables (no video). • Use an open-source stack only: e.g. SciBERT for text embeddings, CLIP for image embeddings, FAISS for indexes; LLMs like Llama/Mistral/Qwen (Llama-Vision if needed); LightRAG ideas for lightweight, stepwise retrieval. • Core pipeline: document & figure indexing → multimodal retrieval → evidence fusion → LLM reasoning → grounded answer with citations. • Datasets: shortlist public, labelled sets for DocQA (e.g. ScienceQA, DocVQA; plus PubLayNet/S2ORC for pages/figures). • Evaluation-first scope: choose 1–2 datasets with clear ground truth so we can measure accuracy and trace evidence. • Constraints: run locally / university GPU, no paid APIs, keep models lightweight. • Out of scope: custom model pretraining at scale; video understanding; closed/commercial APIs.
Client's feedback	You must choose at least 3 datasets for evaluation
Project Expected Outcomes based on Group's understanding	• A working prototype that performs multimodal document QA with text+image+table evidence and produces quote-backed answers. • Reproducible environment (scripts, configs, small demo corpus) and an open repo with code + README. • Architecture & research proposal documenting the agent plan, retrieval modules, fusion strategy, and design choices. • Baseline vs. our method comparison (standard RAG vs. agentic multimodal RAG), with ablations on retrieval/fusion. • Quantitative results on chosen datasets using agreed metrics: retrieval Hit@5 / NDCG@5 / Evidence Completeness, answer Quote-F1 / Acc@1 (CleanEval normalisation). • Qualitative demos: screenshots/notebook showing evidence traces and step-by-step reasoning. • Next steps pack for the client: finalized dataset choice, evaluation protocol, and a plan for scaling/optimisation on uni GPUs.

Client's feedback	Must check if there are any SOTA to comapre with

Status Highlight	Summary of report & feedback
Overall Project Status	The project is progressing as planned. The team has completed the initial investigation and framework design. Our focus is on improving the overall architecture of LightRAG. Instead of fine-tuning base models, we aim to enhance performance through knowledge graph construction, multi-modal retrieval, and agent-based strategy optimization.
Progress and Achievements	• Completed review of the LightRAG framework and identified optimization entry points. • Proposed a **multi-agent collaboration approach** (Text / Visual / Table Agents) with a planner-based tool routing design. • Designed the initial **Knowledge Graph Schema v0.1** (Paper → Section → Figure/Caption → Method → Dataset). • Confirmed the strategy of improving pipeline and retrieval methods rather than fine-tuning large models.
	Client's feedback: You need to think of the novelty of your work
Key Issues	• Knowledge graph construction still needs to be implemented, including efficient extraction of nodes and relations. • Cross-modal alignment between figures and text remains technically challenging (e.g., caption binding, semantic consistency). • Time pressure to deliver a **quick proof-of-concept (PoC)** demonstration.
Obstacles & Risks	• **Technical:** Retrieval accuracy may depend heavily on embedding model selection. • **Engineering:** If knowledge graph construction or indexing is inefficient, system responsiveness could be affected. • **Management:** The team has not fully finalized the focus (Graph-first vs Agent-first vs Retrieval-first approaches).

	Client's feedback:
Next steps	• Finalize and implement **KG Schema v0.1** and ingest a small-scale paper dataset. • Build dual indexes (text + image) and test retrieval fusion (RRF). • Implement minimal planner tool APIs (search_text, search_figure, graph_query). • Prepare 10–20 scientific Q&A cases for initial PoC testing.

Plan & Milestones	**Phase 1: Agentic Multimodal RAG Prototype (PoC) (W5–6)**
	• Index text and image embeddings separately using FAISS
	  ○ **Text:** SciBERT
	  ○ **Image:** CLIP
	• Implement retrieval fusion (e.g., Reciprocal Rank Fusion)
	• Develop minimal planner-based agent execution:
	  ○ search_text(), search_figure(), search_table(), graph_query()
	• Run initial PoC with 10–20 scientific questions using small document corpus
	  ○ Answers must be grounded and quote-backed
	
	Phase 2: Knowledge Graph Schema Implementation (W7)
	• Extract key nodes and relations from scientific documents
	  ○ *(Paper → Section → Figure → Caption → Method → Dataset)*
	• Build initial knowledge graph using lightweight extraction methods
	  ○ Rule-based extraction or small LLM tools only
	• Integrate graph_query() as a callable tool for the agent planner
	• Focus on small-scale data only
	  ○ (In line with constraint: no large-scale pretraining or commercial APIs)

Phase 3: Retrieval Alignment & Evaluation (W8–9)

- Improve cross-modal alignment between figures and text
    - Caption binding, semantic consistency
- Evaluate retrieval module performance:
    - **Metrics:** Hit@5, NDCG@5, Evidence Completeness
- Evaluate answer generation quality:
    - **Metrics:** Quote-F1, Acc@1 (CleanEval normalization)
- Perform ablation studies to validate pipeline components:
    - Without Knowledge Graph
    - Without Agent Planner
    - Without Multimodal Fusion (text-only baseline)

Phase 4: Final Deliverables & Client Pack (W10–12)

- Finalize open-source code repository with:
    - Scripts, configs, README, and demo corpus
    - Clean agent routing interface and modular components
- Complete system architecture and design documentation
    - Agent roles, retrieval strategies, fusion logic, KG schema
- Prepare qualitative demo (screenshots, reasoning traces)
    - Use notebooks or rendered outputs to highlight agent steps
- Deliver final client pack including:
    - Evaluation results
    - Dataset & scope summary
    - Plan for scaling/optimization on university GPUs

	Client's feedback: Adding KG into the framework is great. Will you automatically create KG using LLM? How you will represent KG (neo4j?)

All group member signatures (either handwritten or digital signatures):

Xiaoran Wang

Hanyu Wang Junbo Liu Kunming Lyu.

zhencheng Huang

Client signatures (either handwritten or digital signatures):

A.B.

Time: 13/08/2025, 5pm – 5:20pm(W2)

Venue: Google Meet, https://meet.google.com/zkj-rgqm-xhc

Meeting Minute Taker: Xiaoran Wang, Hanyu Wang

Attendances: Hanyu Wang, Xiaoran Wang, Zhencheng Huang, Jinlin Zhong, Kunming Lyu, Junbo Liu

Apologies: N/A

Main Contents

- Scope: build an **agentic multimodal RAG** to discover/summarize scientific concepts from **text + figures/captions**; produce structured outputs (abstracts, mini-reviews, hypotheses).
- Approach: **agentic planning** + **multimodal retrieval** (eg. PubMed/arXiv, figures) using **recent models** (eg. CLIP/BLIP, SciBERT) and **open datasets**.
- Evaluation/Outcome: compare generated abstracts to originals and show **use-case demos**; aim for **novelty** and potential **publication**; comms via **Slack**; meet next week then **bi-weekly**.

Key Takeaways

- Project scope is clear: build an **autonomous, multimodal RAG** research assistant for scientific literature.
- **Novelty and grounding** are crucial; evaluation will include **abstract-level similarity** and **use-case demonstrations**.
- Use **open/public datasets** and **state-of-the-art** multimodal/text models; model choices are flexible.
- Deliverables will emphasize **working code** plus **well-structured, evidence-grounded write-ups** (mini-papers/abstracts).
- Meeting cadence: **next week** then **every two weeks**; **Slack** is the primary channel.

What's Next

- As a team, read all references provided by the client and survey additional relevant resources (multimodal RAG, CLIP/BLIP, SciBERT, scientific figure understanding) to build strong foundational knowledge**.**
- In-group meeting on Saturday (16/08/2025) to discuss findings, agree on initial use cases, pick baseline models/datasets, and outline the first prototype plan.

We acknowledge Mr. Hunter Xu in assistance of preparing this template.

Time: 20/08/2025, 5pm – 5:20pm(W3)

Venue: Google Meet, https://meet.google.com/fsn-apcx-htc

Meeting Minute Taker: Xiaoran Wang, Hanyu Wang

Attendances: Hanyu Wang, Xiaoran Wang, Zhencheng Huang, Jinlin Zhong, Kunming Lyu, Junbo Liu

Apologies: N/A

Main Contents

- **Scope**: Focus the project on **multimodal Document QA** (text + figures/images + tables). Aim for a new framework, not a reproduction.
- **Datasets**: Use **public QA datasets with ground truth** (e.g. ScienceQA) so we can measure accuracy.
- **Models/Tools**: **Open-source only** (e.g. LLaMA/Mistral/Qwen). **LangChain is OK**. No budget for closed APIs.
- **Method**: Image embeddings via **CLIP**; **fuse** text/table/image evidence before generation; evaluate by answer accuracy.

Key Takeaways

- Keep the problem to **DocQA across modalities**; domain is flexible but evaluation must be objective.
- **GPU access** is available via the university—email the client to be connected.
- **Slack** remains the primary channel; bring **progress slides** next check-in (papers, datasets, initial plan).
- Baseline to start: CLIP + text embeddings → multimodal retriever → open-source LLM → accuracy metrics.

What's Next

- **Literature scan (this week):** review recent multimodal DocQA/RAG papers; extract 3–5 design ideas for our framework.
- **Dataset shortlist:** choose 1–2 public QA datasets and note licenses + evaluation protocol.
- **Baseline prototype:** implement the simple pipeline above; document setup; request **GPU access** via email.
- **Slides for next meeting:** problem statement, related work, chosen dataset(s), proposed architecture, evaluation plan, risks.

We acknowledge Mr. Hunter Xu in assistance of preparing this template.

Time: *W4*

Venue: *N/A*

Meeting Minute Tak*er: N/A*

Attendances: N/A

Apologies: N/A

Main Contents

No client meeting on week4 agreed by both client and group members.

We acknowledge Mr. Hunter Xu in assistance of preparing this template.