Short Story Presentation

MA-RAG: Multi-Agent Retrieval-Augmented Generation

Rishi Patel

CMPE 297 Special Topics

MA-RAG: Multi-Agent Retrieval-Augmented Generation

A breakthrough framework that uses specialized AI agents working together to answer complex questions more accurately than traditional methods.

The Challenge

Why Current RAG Systems Fall Short

Ambiguous Queries

Users often ask vague or underspecified questions that confuse retrieval systems.

Multi-Hop Reasoning

Complex questions requiring information from multiple sources are difficult to answer.

Noisy Results

Retrieved documents contain irrelevant information that dilutes answer quality.

The MA-RAG Solution

MA-RAG orchestrates four specialized AI agents that collaborate through chain-of-thought reasoning to break down complex questions and retrieve precise answers.

01

Planner Agent

Analyzes the question and breaks it into clear, manageable sub-tasks.

02

Extractor Agent

Filters retrieved documents to keep only relevant information for each step.

03

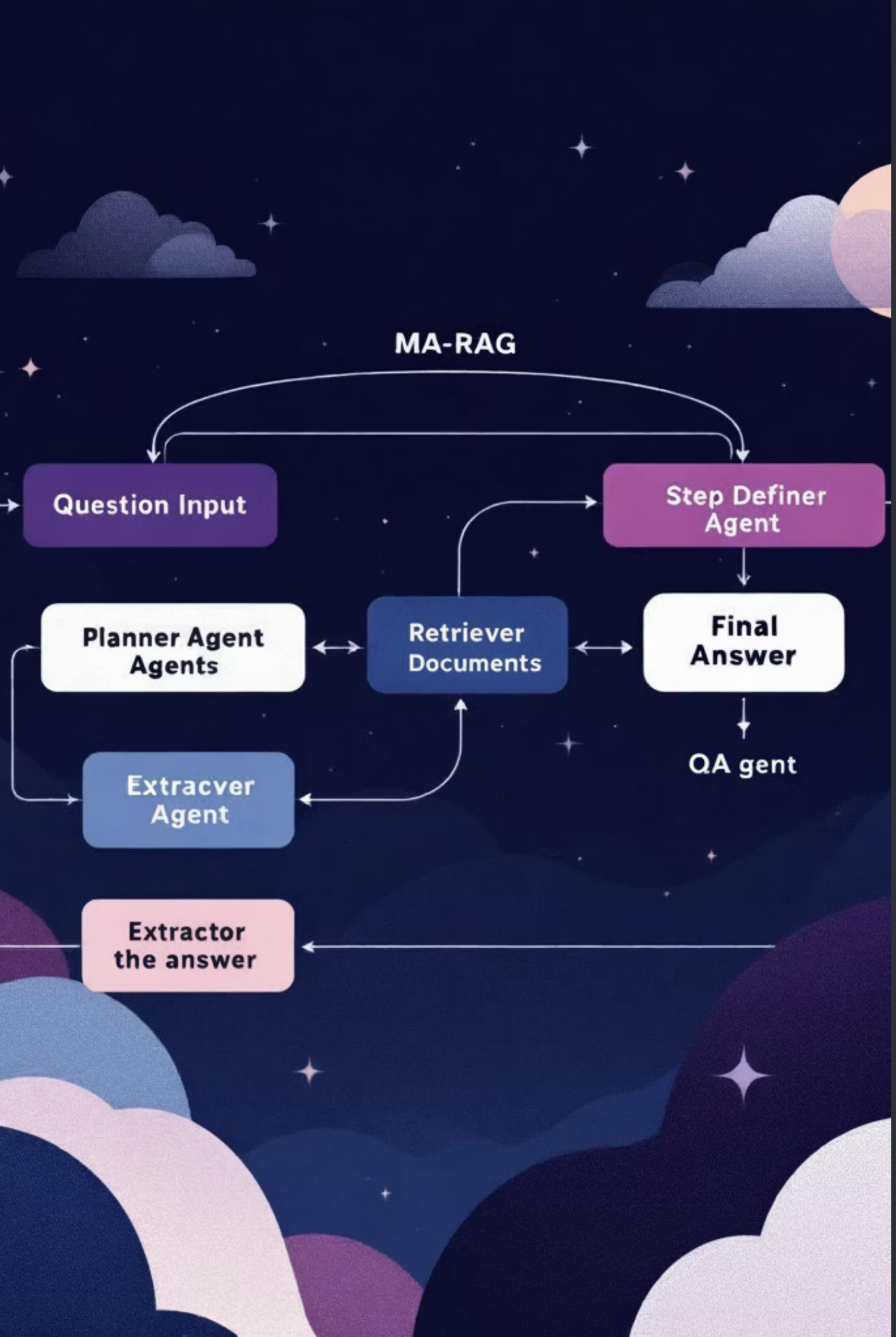
Step Definer Agent

Creates detailed search queries for each sub-task based on context and history.

04

QA Agent

Synthesizes the final answer from extracted evidence and reasoning steps.



How It Works: Step-by-Step

Example Question

> "Where was the only European Cup Final in which Jupp Heynkes played held?"

Planner Breaks It Down

1. Which year did Jupp Heynkes play in the European Cup Final?
2. Where was that year's final held?

Agents Collaborate

The Step Definer creates precise queries, the Retrieval Tool fetches documents, the Extractor filters noise, and the QA Agent builds the answer step-by-step.

Result

Accurate answer: 1977 in Rome

Impressive Performance Results

59.5

Natural Questions

Exact match score on NQ benchmark, outperforming GPT-4 (40.3) without retrieval.

52.1

HotpotQA

State-of-the-art on multi-hop reasoning questions, beating previous best of 42.7.

47.5

2WikimQA

Leading performance on complex multi-hop dataset requiring multiple sources.

MA-RAG with smaller models (8B parameters) outperforms much larger standalone LLMs, proving the power of collaborative reasoning.

Key Advantages

No Training Required

Completely training-free framework that works with any LLM backend without fine-tuning.

Interpretable Reasoning

Each agent's chain-of-thought provides transparent intermediate steps you can follow.

Precision Filtering

Extractor agent removes irrelevant content, keeping only what matters for each step.

Domain Flexibility

Generalizes to specialized domains like medical QA without domain-specific training.

What Makes Each Agent Critical

Without Planner

Performance drops dramatically on multi-hop questions (50.7 → 36.2 on HotpotQA).

System can't decompose complex queries into manageable steps.

Without Extractor

Noisy documents overwhelm the system (50.7 → 43.4 on HotpotQA).

Irrelevant information dilutes answer quality.

Both Together

Full MA-RAG achieves best results across all benchmarks.

Modular design enables fine-grained control over reasoning.

Model Size Matters for Different Agents

83%

QA Agent Impact

Largest performance drop when using smaller model—answer synthesis needs capacity.

78%

Extractor Agent Impact

Significant drop with smaller model—evidence filtering requires strong reasoning.

97%

Step Definer Impact

Minimal drop with smaller model—structured role less dependent on size.

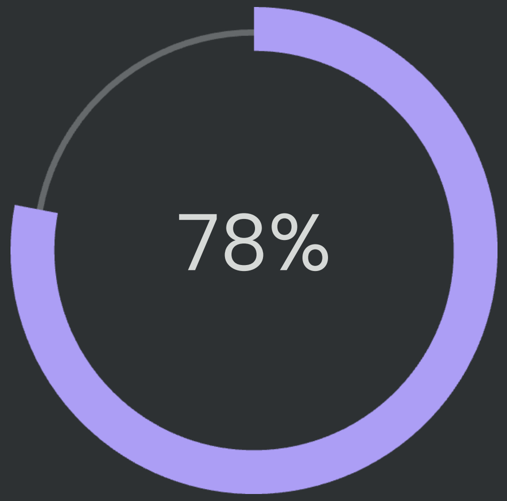
This insight enables efficient resource allocation: use larger models for QA and extraction, smaller models for step definition.

Beyond General Knowledge: Medical Domain Success

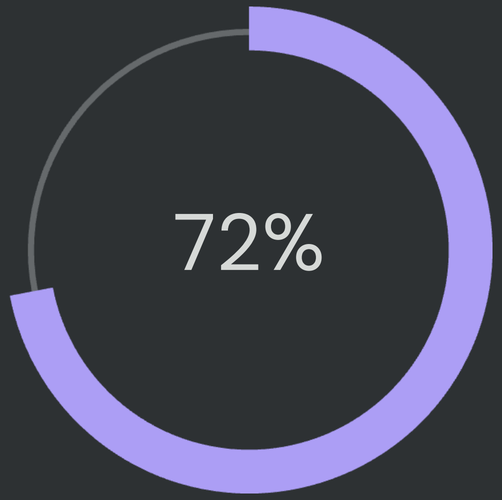
MA-RAG achieves competitive performance on medical benchmarks like PubMedQA and MedMCQA **without any medical fine-tuning**.

It outperforms domain-specific models like Meditron-70B and PMC-LLaMA-13B, demonstrating true generalization through modular reasoning.

When equipped with GPT-4o-mini, MA-RAG surpasses even GPT-4-0613 on biomedical questions.



PubMedQA



MedMCQA

The Future of Retrieval-Augmented AI

1

Collaborative Intelligence

Specialized agents working together outperform monolithic systems.

2

Transparent Reasoning

Chain-of-thought provides interpretable steps for trust and debugging.

3

Efficient Scaling

Strategic model allocation enables powerful results with fewer resources.

MA-RAG establishes a new paradigm: complex reasoning through modular, training-free agent collaboration that's both powerful and practical.