AI Agent 과정

Agent 기초

목차 Agent 기초

3. 다양한 문서 처리 Tool 구현

RAG + Agent = Agentic RAG

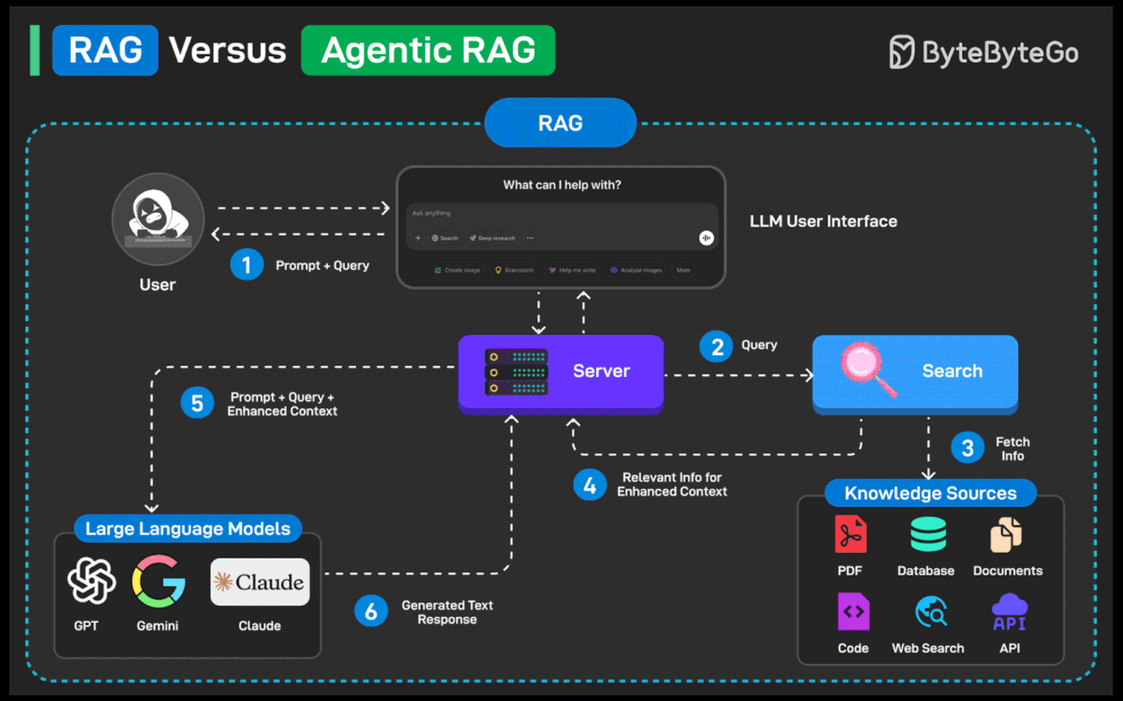
- RAG
  - 정적으로 미리 처리된 데이터베이스로부터 정보를 검색하고, 그 결과를 답변에 통합
  - 이는 쿼리가 복잡하거나 모호하거나, 다단계 추론이 필요한 경우 어려움이 있음
- AI Agent
  - 주변 환경을 인식하고, 정보를 처리하고, 결정을 내리고, 목표를 달성하기 위해 행동하는 소프트웨어 기반 시스템
  - 이용자의 지시를 이해하여, 적절한 도구를 선택하고 절차를 수행
    - 외부 도구(Tool) 사용, API 호출, 복잡한 워크플로우 실행 등

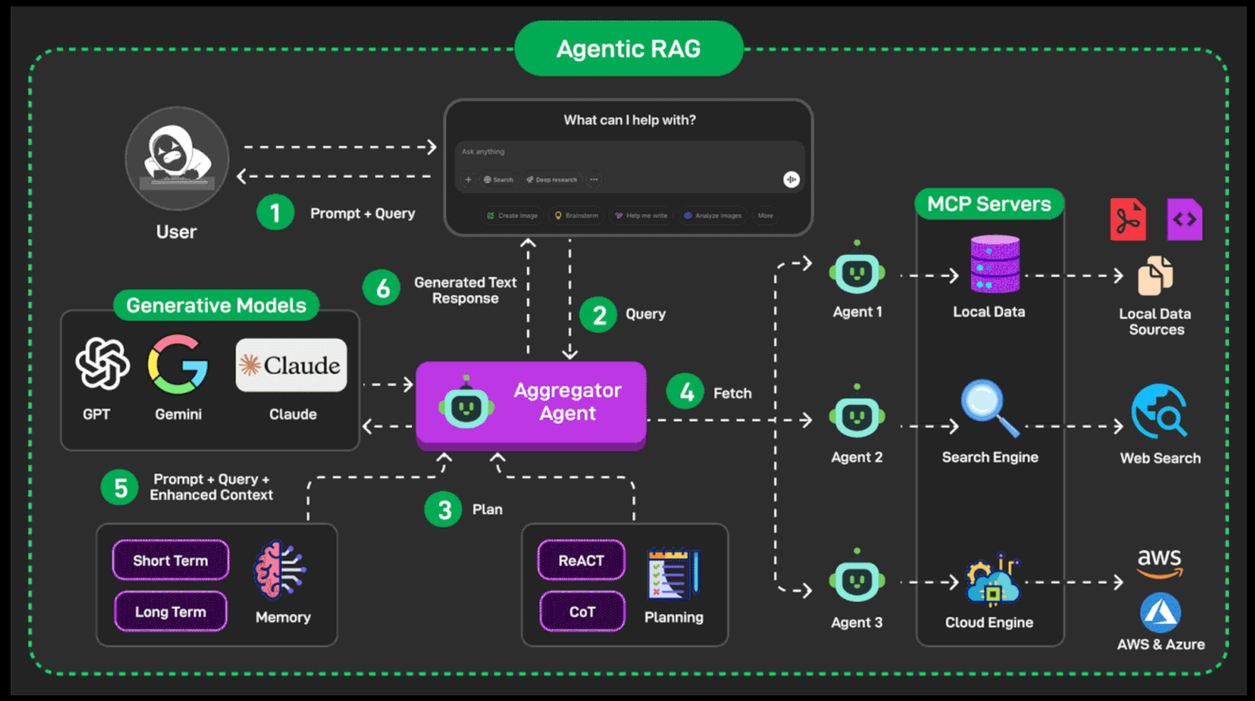
RAG + Agent = Agentic RAG

- Agentic RAG
  - 기존 RAG를 기반으로, 동적으로 작업을 실행할 수 있는 자율적인 의사 결정 주체인 Agent를 도입
  - Agent는 질문을 분해하고, 어떤 도구를 호출할지 결정한 뒤, 검색과 추론 단계를 조율
  - 검색, 추론, 평가 등 하위 작업을 전문으로 하는 여러 Agent가 서로 상호작용
  - 작업을 Agent들에게 분산시키고, 오케스트레이터가 조율하여 검색과 추론을 가속하면서 정확도 유지
  - Evaluator Agent는 생성된 답변의 품질을 검사하고, Optimizer Agent는 필요한 경우 계획 수정

https://levelup.gitconnected.com/rag-is-smart-but-agentic-rag-with-langgraph-is-smarter-a-practical-guide-ac2a3b0bc3bc

RAG + Agent = Agentic RAG





https://medium.com/@vipra_singh/ai-agents-agentic-rag-part-10-f1f7d6d5c8a9

RAG + Agent = Agentic RAG

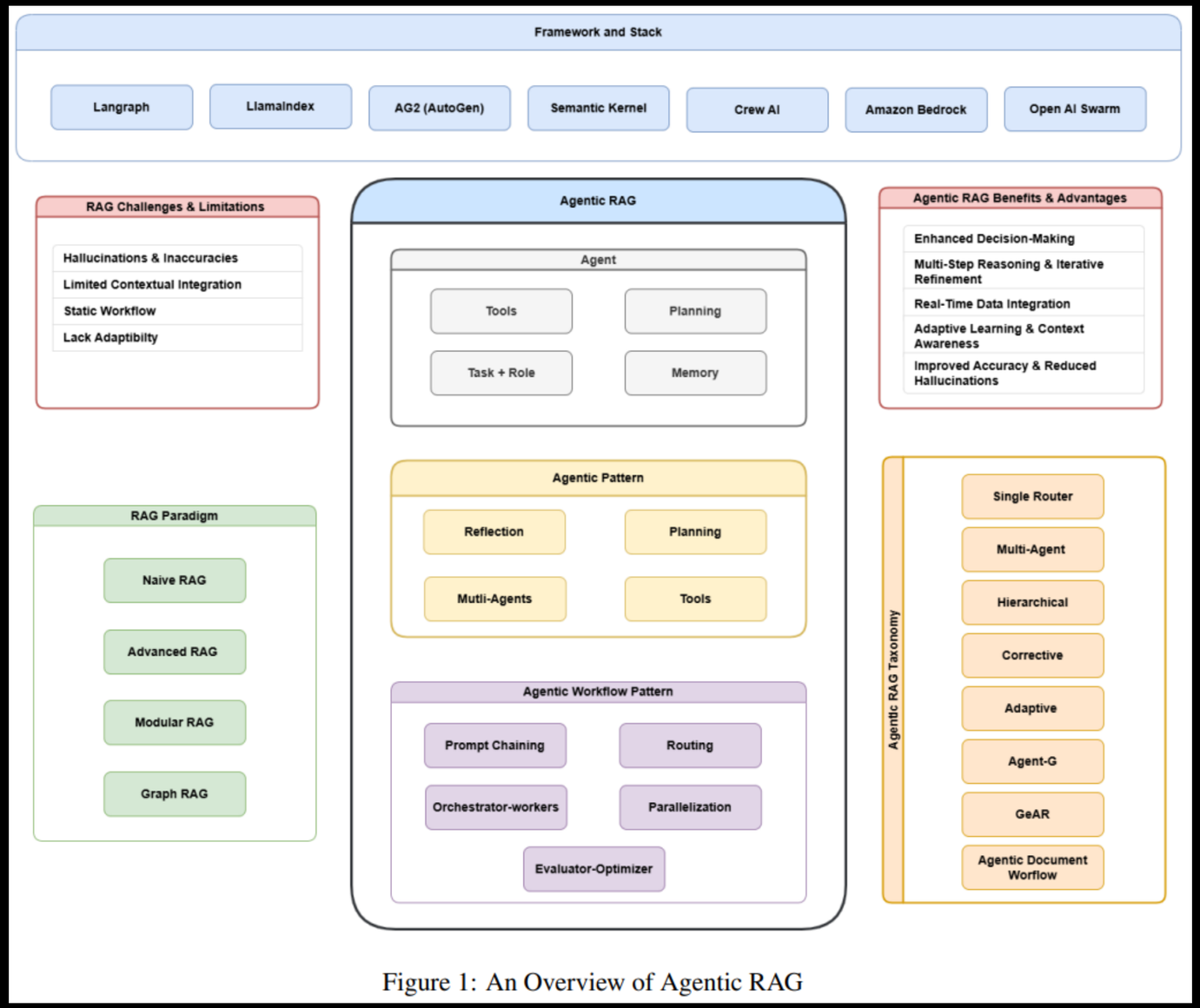


Figure 1: An Overview of Agentic RAG

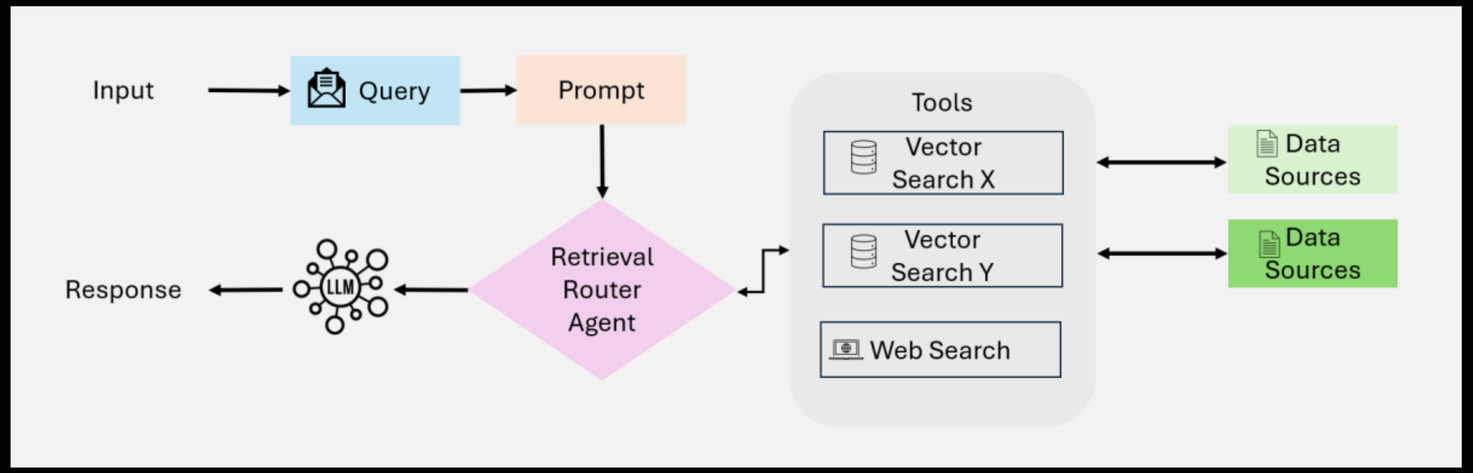
https://arxiv.org/pdf/2501.09136

Agentic RAG 과정

- 능동적 쿼리 요청 분석
  - 쿼리가 AI Agent에게 전달되면, 해당 쿼리의 의도와 맥락 해석
- 메모리 및 전략 수립
  - 단기(세션) 및 장기(과거 기록) 메모리를 활용해 맥락 추적 후 동적인 검색 및 추론 전략 수립
- 도구 선택 및 데이터 수집
  - 벡터 검색, API 커넥터, 다른 Agent 같은 도구들을 선택하고, 관련된 지식 베이스로부터 데이터 검색
- 프롬프트 구성
  - 검색 데이터와 원래의 쿼리, 시스템 프롬프트를 합쳐 최종 프롬프트를 구성하여 LLM에게 전달
- LLM 응답 생성
  - LLM은 최적화되고 맥락이 부여된 프롬프트를 처리하여, 신뢰도 높은 응답을 생성

Agentic RAG 종류

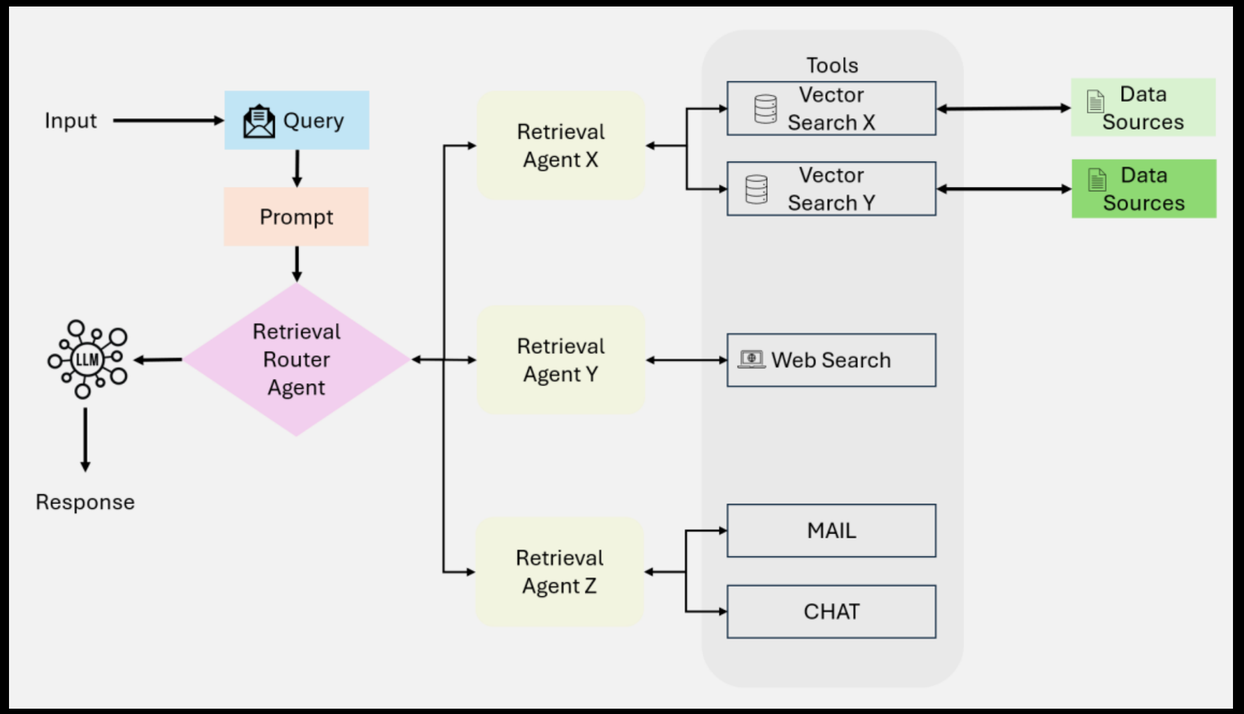
- Single-Agent Agentic RAG : Router
  - 단일 Agent가 전체 검색과 생성 과정 제어하므로 아키텍처 설계, 구현 및 유지 관리가 간편
  - 간소화된 시스템으로 도구나 데이터 소스가 제한된 환경에 특히 효과적
  - 순차적 추론이 필요한 작업에 적합



https://arxiv.org/pdf/2501.09136

Agentic RAG 종류

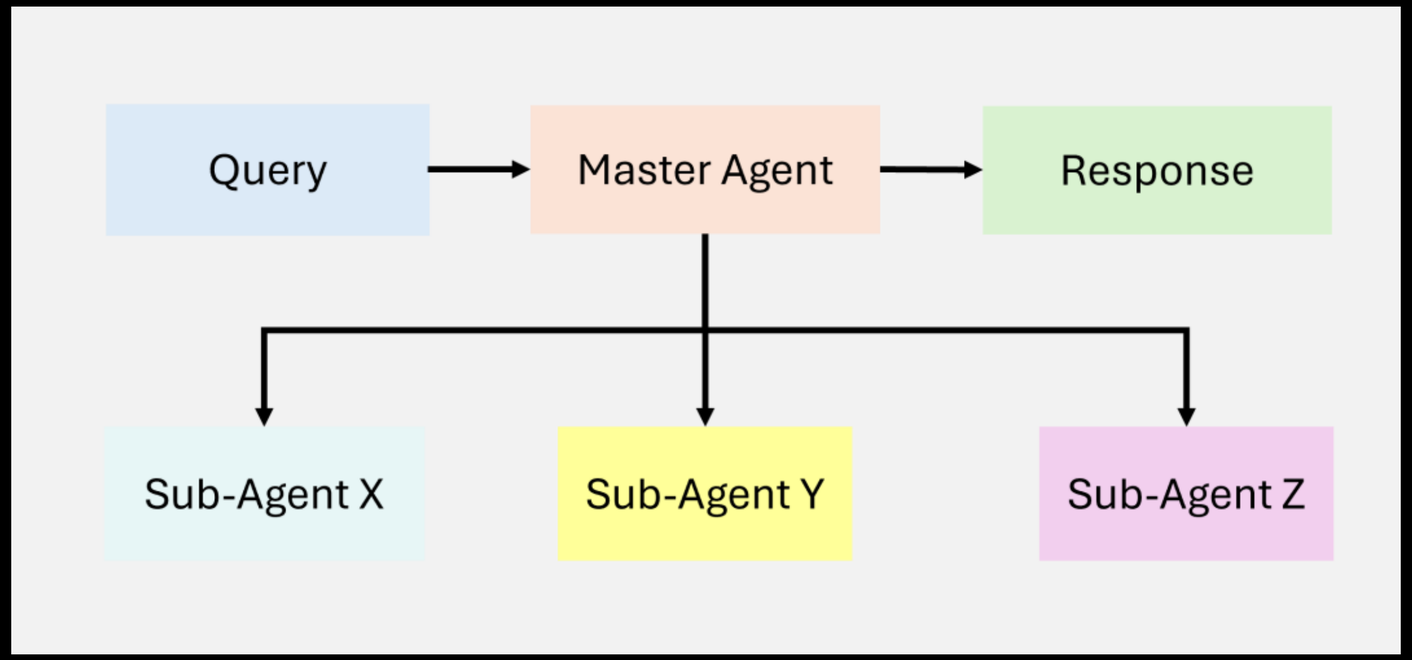
- Multi-Agent Agentic RAG Systems
  - 여러 에이전트가 검색, 계획, 작성, 평가 등 전문 역할을 나누어 협업
  - 협력과 경쟁을 통해 답변의 품질을 높이고 환각을 줄일 수 있으나,
  - 조정 및 통신 오버헤드로 효율이 떨어질 수 있음



https://arxiv.org/pdf/2501.09136

Agentic RAG 종류

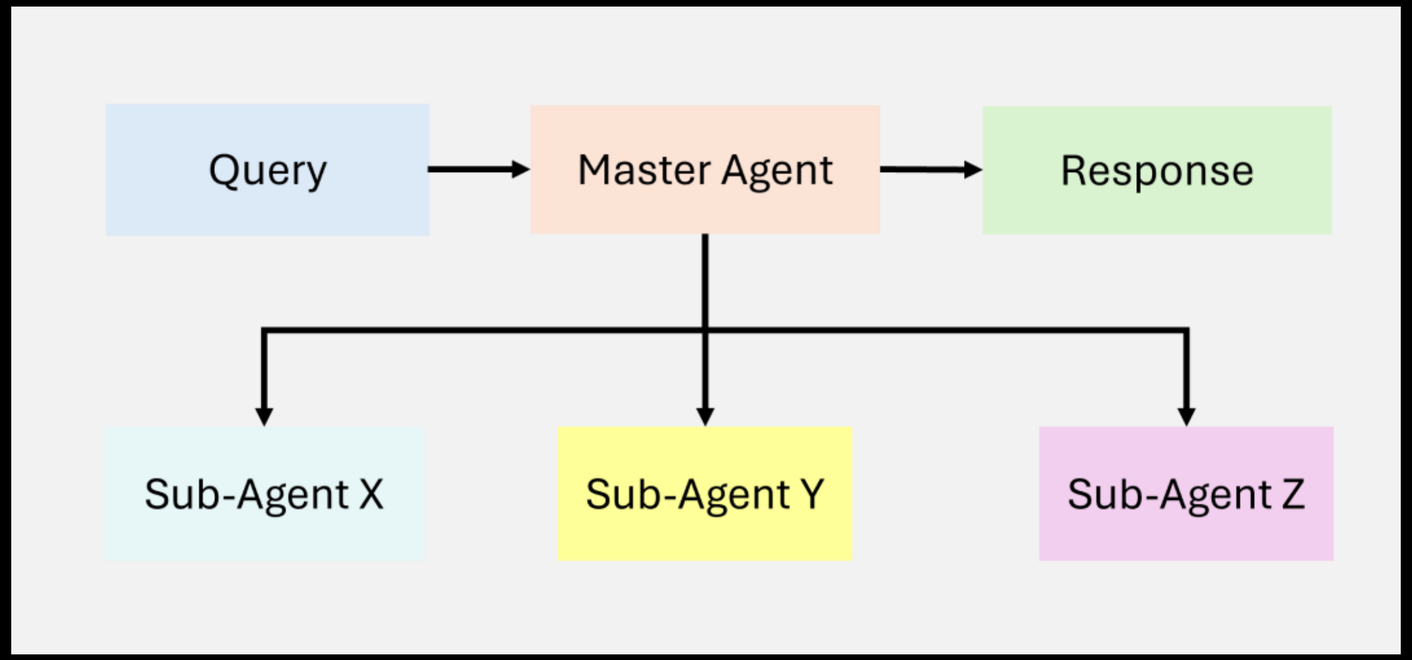
- Hierarchical Agentic RAG Systems
  - 계층형 구조는 Agent를 여러 층으로 조직
  - 상위 Agent가 작업을 계획하고 하위 Agent에게 하위 작업을 위임
  - 긴 문서 요약이나 지식 베이스 구축 같은 복잡한 워크플로우 지원



https://arxiv.org/pdf/2501.09136

Agentic RAG 종류

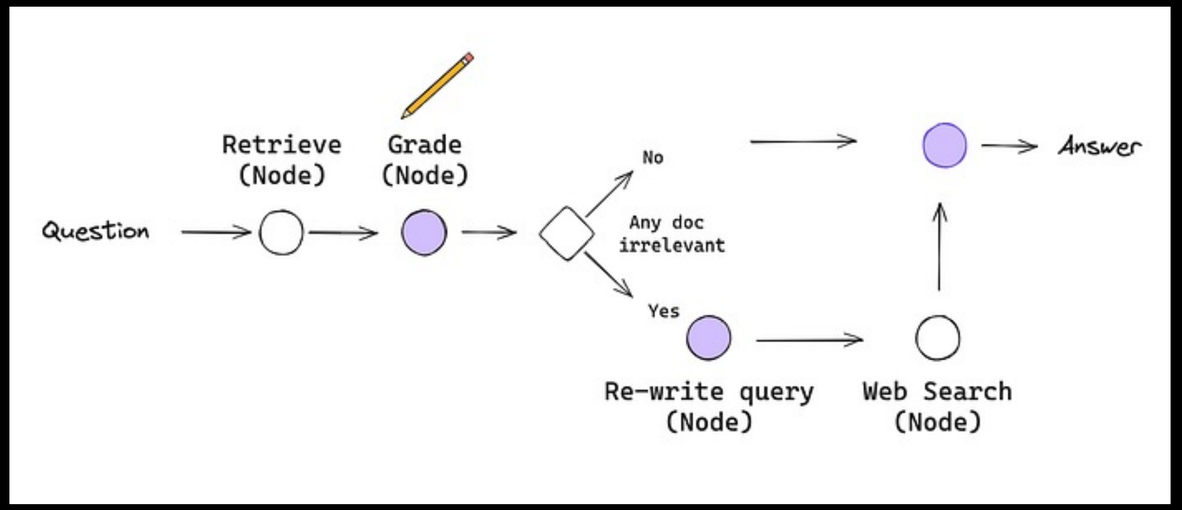
- Hierarchical Agentic RAG Systems
  - 계층형 구조는 Agent를 여러 층으로 조직
  - 상위 Agent가 작업을 계획하고 하위 Agent에게 하위 작업을 위임
  - 긴 문서 요약이나 지식 베이스 구축 같은 복잡한 워크플로우 지원



https://arxiv.org/pdf/2501.09136

Agentic RAG 종류

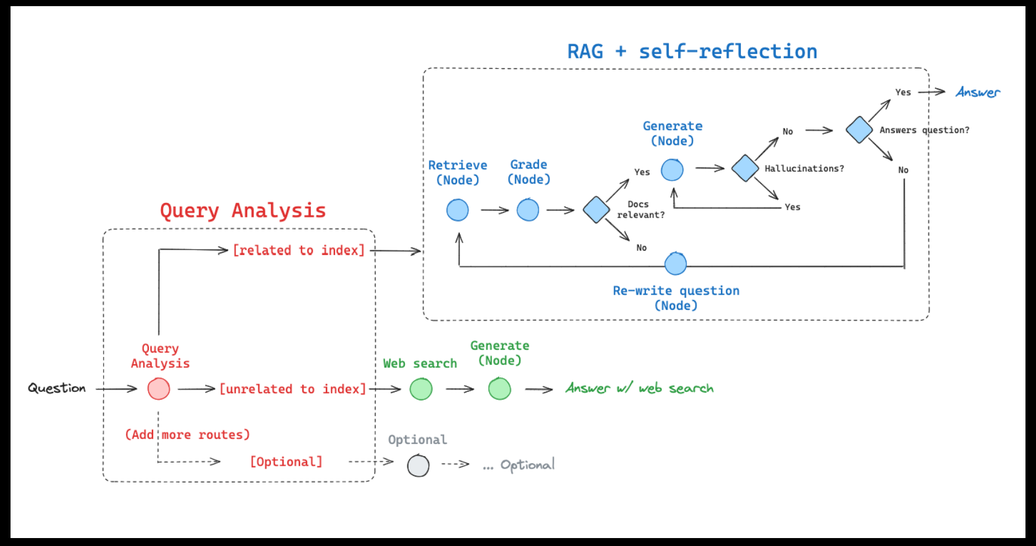
- Agentic Corrective RAG
  - 검색 문서에 대한 self-reflection 과 self-grading을 포함하는 RAG
  - 검색 결과를 반복적으로 개선하고, 피드백 루프를 도입하여 문서 활용도와 응답 품질 향상
  - 검색된 문서를 평가하고, 관련성이 낮은 문서는 시정 조치
  - 쿼리 정제 에이전트가 검색 성능을 개선하고 더 나은 응답 생성을 위해 쿼리 재작성



https://medium.com/@mohitbasantani1987/agentic-rag-vs-corrective-rag-the-battle-for-smarter-ai-retrieval-systems-7d9076309fb0

Agentic RAG 종류

- Adaptive Agentic RAG
  - (1) 쿼리 분석과 (2) active/self-corrective RAG를 결합한 RAG
  - 쿼리를 복잡도에 따라 분류하는 분류기를 도입하여 적절한 검색 전략 선택하고 필요할 때 검색 반복
  - 각 쿼리에 대한 검색 프로세스를 맞춤 설정하여 효율성과 정확성을 향상시키는 것이 목표
  - 단순한 질의에는 불필요한 검색을 피하고, 복잡한 질의에는 다단계 검색을 수행하도록 함



https://langchain-ai.github.io/langgraph/tutorials/rag/langgraph_adaptive_rag/