ChatGPT

Executive Summary: Autonomous Research Assistant (Auto-Analyst)

The **Autonomous Research Assistant (Auto-Analyst)** is a free, open-source system designed to automate the end-to-end research process. It leverages Retrieval-Augmented Generation (RAG) and open-source large language models (LLMs) to plan research, search trusted sources, extract information, store relevant context and generate fact-checked answers.

Project Purpose

Manual research is slow, inconsistent and prone to error. Auto-Analyst addresses this pain point by combining a *retriever* that pulls up relevant documents with a *generator* that synthesises a coherent answer. The RAG architecture reduces hallucinations by grounding responses in external data [1] and enables AI systems to return verifiable, real-time facts instead of outdated guesses [2]. This makes Auto-Analyst suitable for professionals, students and domain specialists who need trustworthy information quickly.

Key Features

- **Free and open-source:** All components—including models, vector stores and search APIs—are free to use. There are no paid APIs or proprietary dependencies.
- **Agentic workflow:** A planner decomposes the user's question into search tasks, a retriever gathers information from public sources, and a generator composes answers. Each step is independent, enabling flexibility and fault isolation.
- **Stateful orchestration:** Built on LangGraph, the system maintains memory across steps, supports loops and retries, and allows human-in-the-loop approval. LangGraph's graph-based architecture gives explicit control over flows and enables persistence and streaming [3] [4].
- **Citation tracking and verification:** The answer generator annotates each claim with an inline citation and a corresponding entry in a sources section. A verification agent reviews the draft answer and removes unsupported statements.
- **Streamlit interface:** Users interact with a simple web UI to submit questions, adjust the number of sources and view results. Past queries are stored in session state for convenient history.

Why It Matters

Modern knowledge workers need tools that are both **accurate** and **cost-effective**. RAG reduces hallucinations and improves reliability by grounding each response in real documents [1]. Auto-Analyst delivers this reliability without vendor lock-in or expensive usage fees. It showcases mastery of LLM orchestration, vector search, semantic embeddings and evaluation metrics. The resulting application is a showcase of best practices in building production-grade GenAI systems with a focus on transparency, correctness and user trust.

---

1 2 The Science Behind RAG: How It Reduces AI Hallucinations
https://zerogravitymarketing.com/blog/the-science-behind-rag/

3 4 A Developer's Guide to LangGraph for LLM Applications | MetaCTO
https://www.metacto.com/blogs/a-developer-s-guide-to-langgraph-building-stateful-controllable-llm-applications