DEEP RESEARCH MULTI-AGENT SYSTEM (DR-MAS)

Project Kickoff: System 2 Reasoning Applications on GCP

Target Audience:

Data Scientists, Research Analysts, DevOps

Goal:

Autonomous, verifiable deep research and insight generation.

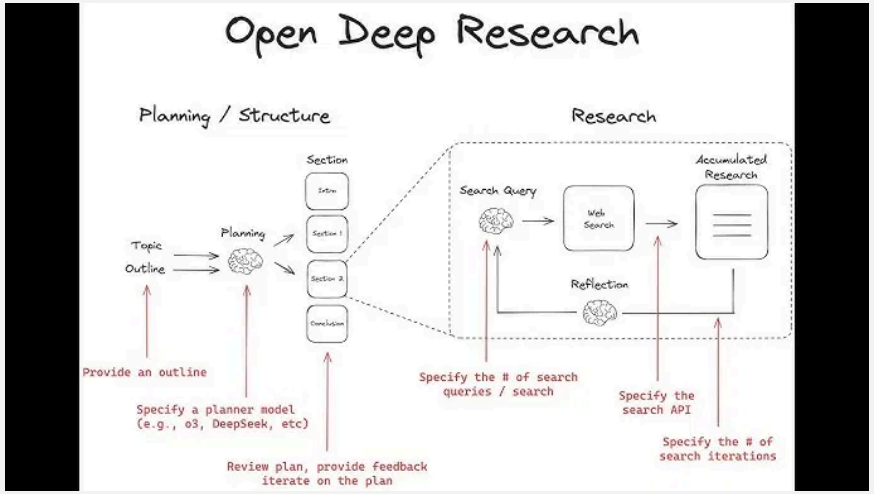
THE EPIC: ELEVATING AI RESEARCH TO SYSTEM 2 REASONING

Problem: Current AI research lacks the rigor, self-correction, and verifiable accuracy required for high-stakes, open-ended problems.

Solution: Implement a sophisticated, GCP-native multi-agent system for autonomous deep research.

The "Why"

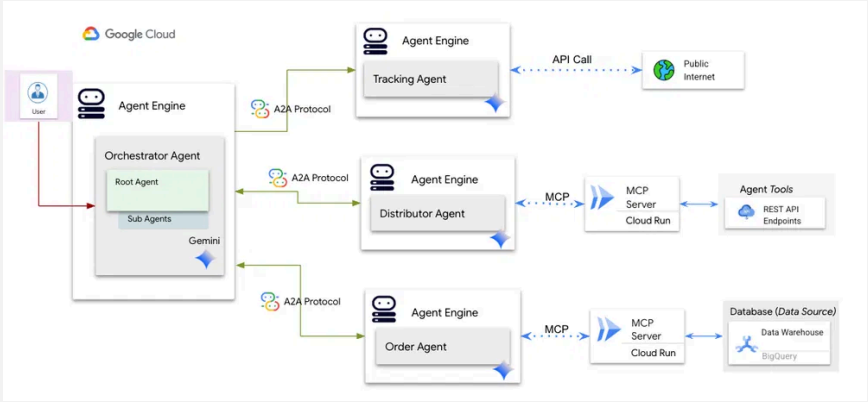
- **Elevate Quality:** Move beyond simple retrieval to true analytical reasoning and synthesis.
- **Reduce Time:** Significantly reduce manual research time and improve decision quality.
- **Establish Leadership:** Position our platform as the leader in enterprise-grade, verifiable AI research.



CORE ARCHITECTURE: COLLABORATIVE AGENTS ON VERTEX AI

Architecture: Hierarchical Multi-Agent System using the **Agent-to-Agent (A2A) Protocol on Vertex AI Agent Engine.**

AGENT ROLE	CORE FUNCTION	MODEL STRATEGY
Researcher	Data gathering, BigQuery/Vertex Search tool use.	**Gemini 1.5 Flash (Speed)**
Critic	Factual validation, self-correction loop trigger.	**Gemini 1.5 Pro (Reasoning)**
Synthesizer	Aggregation, long-context data processing.	**Gemini 1.5 Pro (Synthesis)**
Reviewer	Final quality gate, objective alignment check.	**Gemini 1.5 Pro (Alignment)**



USER STORY 1: AUTONOMOUS DEEP RESEARCH EXECUTION

USER & GOAL

Research Analyst

Submit a complex, open-ended research question and receive a comprehensive, synthesized report without manual intervention.

TECHNICAL REQUIREMENTS

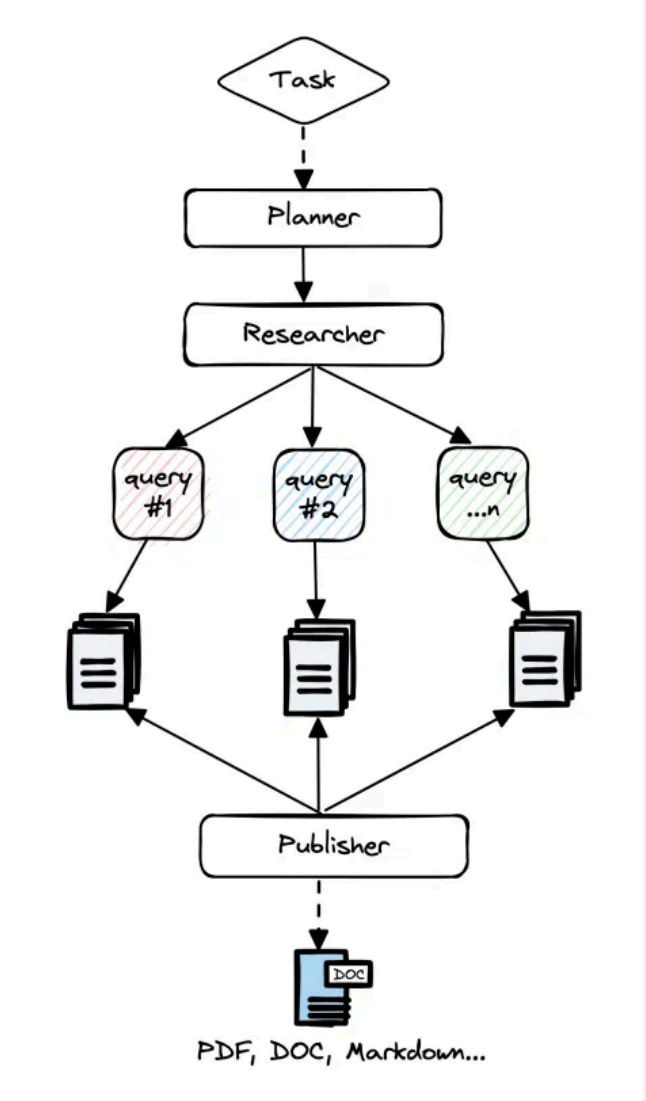
Orchestration

Full Research → Critique → Synthesis → Review flow using ADK/LangGraph.

State Management

Robust state persistence using Cloud SQL/AlloyDB for long-running tasks.

< 15 MIN TARGET RESEARCH LATENCY



USER STORY 2: ACCURACY & SELF-CORRECTION

USER: QUALITY ASSURANCE MANAGER

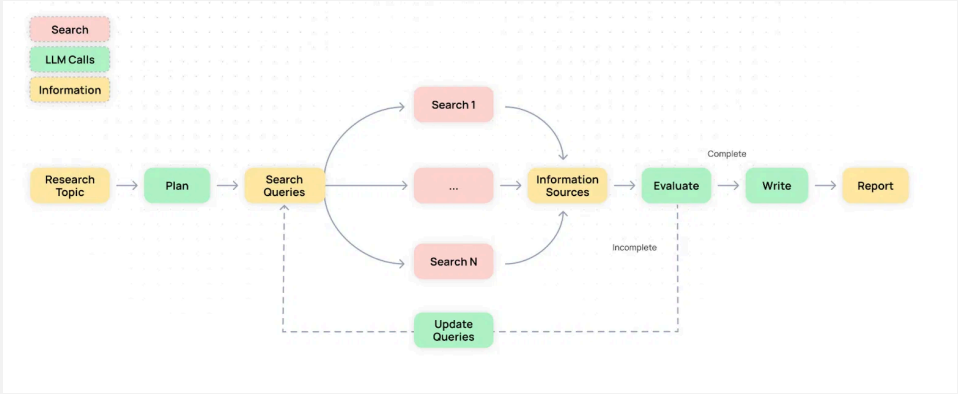
Goal: Critic agent automatically validates findings and triggers a re-research loop if errors are found.

Key Technical Requirements

→ **Critique Pipeline:** A2A Protocol communication for targeted correction requests between Critic and Researcher.

→ **Model Balancing:** Dynamic routing of high-stakes validation tasks to **Gemini 1.5 Pro** for superior reasoning.

→ **Reflection Loop:** Automated re-triggering of search tools based on Critic feedback to resolve inaccuracies.



Pass/Fail Metric: **Factual Accuracy > 95%**

USER STORY 3: LONG-CONTEXT DATA PROCESSING

USER: DATA SCIENTIST

"I want to process very large datasets (500k+ tokens) efficiently, so that I can generate comprehensive summaries without context limits or excessive latency."

TECHNICAL REQUIREMENTS

- **Vertex AI Context Caching:** Reduce TCO and latency for repeated data access.
- **Gemini 1.5 Pro Integration:** Leverage 1M+ token context window for deep synthesis.
- **Vertex AI Search:** Seamless retrieval from enterprise knowledge sources.

PASS/FAIL METRIC

> 50%

Latency reduction for subsequent accesses via Context Caching.

USER STORY 4: GOVERNANCE & OBSERVABILITY

Role: DevOps Engineer | Goal: Full visibility into execution path and security posture.

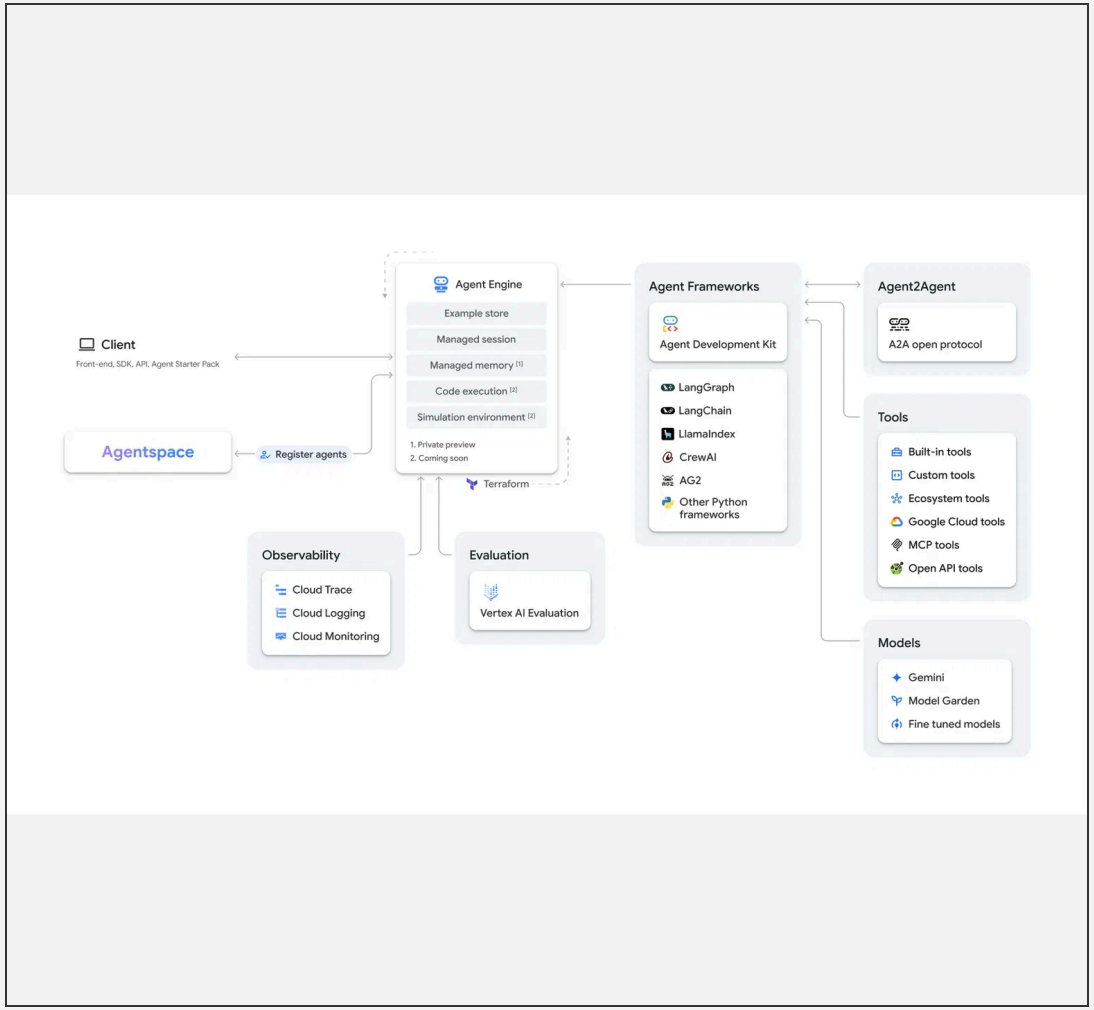
KEY TECHNICAL REQUIREMENTS

- **Observability:** End-to-end tracing via **OpenTelemetry** and Cloud Trace to capture all agent steps and tool calls.
- **Security:** Strong guardrails using **Vertex AI Model Armor** to block PII and prohibited content.
- **LLM Ops:** Evaluation pipeline using **AutoSxS** for systematic model benchmarking against golden datasets.
- **Auditability:** Comprehensive Cloud Logging for all agent trajectories and model invocations.

PASS/FAIL METRIC

Security Compliance: 100%

Zero tolerance for PII or prohibited content processing.



SUMMARY: PRODUCTION-READY AGENTIC AI

PROJECT SUCCESS CRITERIA

- **Verifiable Accuracy:** >95% Factual Accuracy via Critic/Reviewer agents.
- **GCP-Native Scale:** Leveraging Vertex AI Agent Builder and Gemini 1.5.
- **Measurable Performance:** Meeting all story-specific Pass/Fail thresholds.

IMMEDIATE NEXT STEPS

- Finalize technical design documents and A2A protocol schemas.
- Begin Phase 1: Foundational Infrastructure Setup (AlloyDB, Redis).
- Establish **AutoSxS** golden dataset for initial benchmarking.