Agentic AI Research: Architecture of Gemini Deep Research and Comparative Patterns

Figure: Example multi-stage data analysis workflow (knowledge discovery process). Gemini's Deep Research similarly decomposes a complex question into a staged pipelinegemini.googleblog.google. First, the system **plans** by translating the user's query into a detailed multi-point research plangemini.googlebleepingcomputer.com. The user can review and refine this outline before execution. Next, in the **searching/browsing** stage, Gemini autonomously issues web searches and deeply browses relevant sites to gather up-to-date informationgemini.googlebleepingcomputer.com. Throughout the session Gemini **reasons** iteratively, showing "thoughts" and keeping a shared internal state so it can decide on the next stepsgemini.googlegemini.google. Finally, once enough information has been collected, Gemini **synthesizes** the findings into a comprehensive, well-organized reportgemini.googlegemini.google. This report is formatted like an academic paper, with structured sections and source citations. The pipeline is designed to be agentic: Gemini uses a very large context window (≈1 million tokens) supplemented by retrieval so it can "remember" everything learned during the sessiongemini.google.

- **Planning:** Transforms the prompt into a multi-point research plangemini.googlebleepingcomputer.com. Users can approve or adjust this plan before the agent proceeds.
- **Searching/Browsing:** Uses Google Search and web tools to autonomously find relevant, up-to-date sourcesgemini.googleblog.google. Gemini may open dozens or hundreds of pages, extracting and summarizing data.
- **Reasoning:** Iteratively analyzes retrieved content, shows intermediate reasoning steps (via a "thinking panel"), and decides which sub-tasks to tackle nextgemini.googlegemini.google. Sub-tasks can be done in parallel or sequence as determined by the agent.
- **Reporting:** Synthesizes the gathered information into a structured report (including tables, figures, and citations), performing self-critiques and edits to improve clarity and completenessgemini.googleblog.google.

Comparison to Other AI-Assisted Research Methods

AI "deep research" tools vary in strategy. OpenAI's *Deep Research* (an agentic ChatGPT mode) emphasizes a dynamic, interactive workflowopenai.comanalyticsvidhya.com. It may first ask clarifying questions of the user, then embark on a multi-step search trajectory, **backtracking and pivoting** as needed when new information emergesopenai.comanalyticsvidhya.com. The model continuously adjusts its research path in real time, integrating web browsing, code execution (via a Python tool), and multimodal analysis (text, images, PDFs) into its loopsectionai.comopenai.com. In contrast, Gemini's workflow is more **plan-driven**: it lays out a fixed series of tasks that the user approves, then executes them with little deviationsectionai.comanalyticsvidhya.com. Reviewers note that Gemini's structured planning gives users control and organizes output neatly, but can be less flexible than ChatGPT's free-form agentic approachsectionai.comanalyticsvidhya.com.

Beyond these individual agents, there are multi-agent frameworks. For example, Anthropic's Claude Research uses a **lead agent with parallel sub-agents**anthropic.com. The lead agent decomposes the problem and spawns specialized agents (tools and workers) to tackle different facets simultaneouslyanthropic.com. Each sub-agent autonomously searches or analyzes a subset of the question, then returns its findings. This **parallel breadth-first** strategy enables very thorough coverage; in tests a multi-agent system found correct answers by delegating sub-tasks (e.g. different companies in a list) more effectively than a single-agent systemanthropic.com. The drawback is cost and complexity: multi-agent systems can use roughly **10–15× more computation (tokens)** than simpler approachesanthropic.com, and require coordination logic to avoid redundant work.

A more traditional AI approach is **Retrieval-Augmented Generation (RAG)**. In a RAG pipeline, the system retrieves relevant document snippets from a fixed database (e.g. cached Wikipedia or a scientific corpus) and then uses an LLM to generate an answeranthropic.com. This is essentially a one-shot process: static retrieval followed by single-step synthesis. RAG is fast and stable for well-covered topics, but it cannot "pivot" or learn new information beyond its database. Agentic methods like Deep Research differ fundamentally: they perform a *multi-step search loop* (possibly updating queries based on intermediate results) rather than fetching a predetermined set of chunksanthropic.com. In practice, RAG-based tools often handle straightforward factual queries well, while agentic LLMs handle complex, evolving questions with more nuance.

Comparison with Traditional Research

Traditional (non-AI) research is manual and iterative. A human researcher typically **defines a research question**, then conducts literature and web searches using keywords and academic databases. Methods sections of papers explicitly list search terms, databases (e.g. PubMed, Google Scholar), and inclusion/exclusion criteriaunr.edu. The researcher then reads, annotates, and synthesizes sources. As guides note, an effective literature review organizes findings into themes or "moves" (introductory context, identified gaps, proposed contributions)unr.edu. This process is highly flexible and nuanced: humans evaluate source credibility, infer context, and write with style. However, it is **time-consuming and labor-intensive**. No human can truly skim hundreds of websites in minutes; doing so might yield superficial results. The key differences are speed and scale: AI agents automate the search and basic synthesis, while humans apply deeper critical judgment. In summary:

- **Human research:** Involves deliberate keyword search, reading full texts, note-taking, and writing. For example, literature reviews document exactly which keywords and databases were used[unr.edu](unr.edu) and then organize insights by common themes[unr.edu](unr.edu). This yields high-quality analysis but takes hours or days per topic.
- **Agentic AI research:** Automates search and assembly of information at scale. Tools like Gemini can browse broadly in minutes and generate an initial draft, but they may miss subtleties.
- **RAG/one-shot search:** Quicker for factual queries (like a scholar using Google), but prone to hallucination if sources are incomplete.

Evaluation of Research Patterns

These patterns offer trade-offs between depth, control, and cost:

- **Structured planning (Gemini's approach):** *Benefits* – clear workflow, user can steer the focus, outputs are well-organized. *Drawbacks* – inflexible to new leads; if the approved plan was incomplete, the agent won't explore unplanned angles[sectionai.com](sectionai.com).
- **Dynamic/iterative planning (ChatGPT's approach):** *Benefits* – adaptive exploration, can dive deeper or change directions mid-task[openai.com](openai.com). *Drawbacks* – less predictable for users, and can be slower because it may repeatedly rewrite its plan.
- **Multi-agent parallel search:** *Benefits* – very broad coverage and speed via parallelism[anthropic.com](anthropic.com). *Drawbacks* – high resource use (tokens, time) and complexity of coordination, plus potential for context fragmentation (agents acting independently).
- **Single-agent search (non-parallel):** *Benefits* – simpler architecture and lower cost. *Drawbacks* – may miss some parallel lines of inquiry and can only track one line of reasoning at a time.
- **One-shot RAG:** *Benefits* – efficient for known queries and easy to implement. *Drawbacks* – static and superficial; cannot verify new information or browse live content.
- **Human researcher:** *Benefits* – critical thinking, domain expertise, handling ambiguity. *Drawbacks* – slow, unscalable for very large searches.

In terms of control flow (see Figure 2), one can categorize: **branching** occurs when a query splits into independent sub-questions (as in multi-agent or parallel subtasks). **Looping/iteration** appears when the agent refines its plan or searches (e.g. Gemini repeating searches after learning new facts, or ChatGPT backtracking). Non-iterative (straight) modes include single-pass RAG or an initial plan being executed without revision. Agentic frameworks often combine loops and branches: e.g. Gemini may spawn parallel searches (branch) and then loop back to synthesize results, while Claude's orchestrator loops through multiple agentic search cycles[anthropic.com](anthropic.com).

Potential Improvements and Unified Architecture

A unified deep-research architecture could adaptively combine these approaches. For example, a **meta-controller** could first categorize the query's complexity: for straightforward factual queries, use a lightweight RAG or direct answer model; for complex open-ended questions, trigger the full agentic workflow. Within the agentic flow, improvements might include:

- **Adaptive planning:** Allow the plan itself to evolve. Instead of a fixed "7-point plan," incorporate dynamic re-planning loops (reflection steps) where the agent reviews what it has found and proposes additional sub-questions if needed. This is akin to the "reflection" pattern of having the model check its own work[medium.com](medium.com).
- **Flexible branching:** Enable optional parallelism: small queries might be solved sequentially, while large queries launch parallel sub-agents (like Claude) to cover multiple angles. The system could estimate which approach minimizes time vs cost.
- **Human-in-the-loop toggles:** Let users easily dial the level of structure. Novices could skip the plan stage (letting the agent assume a plan), while experts might want detailed outlines. Users could also flag insufficient sub-answers to trigger further search loops.
- **Integrated memory and retrieval:** Combine RAG retrieval with live search. The system might begin by retrieving known facts via a knowledge base, then use web browsing to update or verify information. This hybrid could reduce web queries for well-documented topics while still enabling freshness.
- **Specialized tools:** Incorporate additional tools (e.g. PDF readers, databases, calculators) in the agent's toolkit so that each sub-agent is tailored (e.g. an academic-paper agent, a data-analysis agent, etc.).
- **Simplification when possible:** For simpler tasks, the system could skip some stages. For example, if the question is a summary request, it might jump directly to synthesizing high-level answers rather than a full browse. This reduces latency and cost.

Each addition has trade-offs. **Adding complexity** (e.g. multi-agent parallelism, richer toolsets) can improve depth and accuracy but raises computational cost and potential for coordination failures. **Simplification** (e.g. fewer steps or hidden plans) makes the system faster and easier to use but risks missing nuance or requiring more user prompting. A modular design would allow toggling these modes based on task value: low-stakes queries use fast simple paths, while high-stakes research invokes full agentic machinery. Ultimately, a versatile architecture might look like a decision tree: the query passes through classifiers that route it to the appropriate sub-module (static RAG vs. dynamic planning vs. multi-agent), and results are merged at the end. This could be represented as a comprehensive diagram with branching (parallel paths) and loops (iterative refinement) in a Mermaid graph.

Conclusion

Gemini Deep Research exemplifies an **agentic multi-step pipeline**: it explicitly plans, searches, reasons, and writes, giving users visibility at each stage[gemini,googleblog,google](). This stands in contrast to simpler RAG models and more dynamic agents like ChatGPT's research mode[sectionai.comopenai.com](). Each pattern—structured vs. dynamic, single-agent vs. multi-agent, iterative vs. one-shot—brings its own strengths and weaknesses. By analyzing these patterns, one can envision a future unified system that flexibly routes a question to the most suitable mode of analysis. Such a system would balance **efficiency** (avoiding unnecessary complexity) against **comprehensiveness** (employing advanced search and reasoning when needed). The design of these deep-research AI agents is still evolving; studying current approaches reveals that adding or removing layers of planning, parallelism, and iteration can be tuned to match different kinds of research tasks, providing a spectrum of modes from quick answers to exhaustive reports.

Sources: Descriptions of Gemini Deep Research and its workflow are drawn from Google's documentation[gemini,googleblog,google]() and reviews[bleepingcomputer.com](). Comparative details about OpenAI's and Anthropic's systems are from OpenAI's research announcement[openai.com]() and Anthropic's engineering blog[anthropic.comanthropic.com](). Traditional research steps are summarized from academic writing guides[unr.eduunr.edu](). Analytical comparisons rely on published blog posts and evaluations[sectionai.comanalyticsvidhya.comanthropic.com]().

Citations

[

gemini.google

Gemini Deep Research — your personal research assistant

Planning

](https://gemini.google/overview/deep-research/#:~:text=Planning)[

blog.google

Gemini: Try Deep Research and Gemini 2.0 Flash Experimental

Under your supervision, Deep Research does the hard work for you. After you enter your question, it creates a multi-step research plan for you to either revise or approve. Once you approve, it begins deeply analyzing relevant information from across the web on your behalf.

](https://blog.google/products/gemini/google-gemini-deep-research/#:~:text=Under%20your%20supervision%2C%20Deep%20Research,the%20web%20on%20your%20behalf)[

bleepingcomputer.com

Google Gemini's Deep Research is finally coming to API

Deep Research is not just about writing long papers, but also about transforming a prompt into a personalised multi-point research plan.

](https://www.bleepingcomputer.com/news/artificial-intelligence/google-geminis-deep-research-is-finally-coming-to-api/#:~:text=Deep%20Research%20is%20not%20just,point%20research%20plan)[

gemini.google

Gemini Deep Research — your personal research assistant

Searching

](https://gemini.google/overview/deep-research/#:~:text=Searching)[

gemini.google

Gemini Deep Research — your personal research assistant

Reasoning

](https://gemini.google/overview/deep-research/#:~:text=Reasoning)[

gemini.google

Gemini Deep Research — your personal research assistant

Context management

](https://gemini.google/overview/deep-research/#:~:text=)|

gemini.google

Gemini Deep Research — your personal research assistant

Reporting

](https://gemini.google/overview/deep-research/#:~:text=Reporting)|

gemini.google

Gemini Deep Research — your personal research assistant

- Synthesis: Once the model determines enough information has been gathered, it synthesizes the findings into a comprehensive report. In building the report, Gemini critically evaluates the information, identifies key themes and inconsistencies, and structures the report in a logical and informative way,

](https://gemini.google/overview/deep-research/#:~:text=,a%20logical%20and%20informative%20way)|

blog.google

Gemini: Try Deep Research and Gemini 2.0 Flash Experimental

Over the course of a few minutes, Gemini continuously refines its analysis, browsing the web the way you do: searching, finding interesting pieces of information and then starting a new search based on what it's learned. It repeats this process multiple times and, once complete, generates a comprehensive report of the key findings, which you can export into a Google Doc. It's neatly organized with links to the original sources, connecting you to relevant websites and businesses or organizations you might not have found otherwise so you can easily dive deeper to learn more. If you have follow up questions for Gemini or want to refine the report, just ask! That's hours of

](https://blog.google/products/gemini/google-gemini-deep-research/#:~:text=Over%20the%20course%20of%20a,just%20ask%21%20That%E2%80%99s%20hours%20of)|

openai.com

Introducing deep research | OpenAI

it learned to plan and execute a multi-step trajectory to find the data it needs, backtracking and reacting to real-time information where necessary. The model is also able to browse over user uploaded files, plot and iterate on graphs using the python tool, embed both generated graphs and images from websites in its responses, and cite specific sentences or passages from its sources. As a result of this training, it reaches new highs on a number of

](https://openai.com/index/introducing-deep-research/#:~:text=it%20learned%20to%20plan%20and,highs%20on%20a%20number%20of)|

analyticsvidhya.com

OpenAI Deep Research vs Gemini Deep Research

relevant sources and generate the report.

](https://www.analyticsvidhya.com/blog/2025/02/openai-vs-google-who-does-deep-research-better/#:~:text=relevant%20sources%20and%20generate%20the,report)[

sectionai.com

We tested two Deep Research tools. One was unusable.

They both conduct multi-step research to generate detailed reports – but ChatGPT's Deep Research feature does multimodal analysis (including text, images, and PDFs) while Gemini's Deep Research feature only does text-based research and synthesis.

](https://www.sectionai.com/blog/chatgpt-vs-gemini-deep-research#:~:text=They%20both%20conduct%20multi,based%20research%20and%20synthesis)[

sectionai.com

We tested two Deep Research tools. One was unusable.

The other big difference in their performance is that OpenAI's Deep Research feature adjusts its research path in real time, while Google's Gemini Deep Research feature follows a structured research plan that users can review and modify before execution. This means Gemini offers more control, but allows for less nuance.

](https://www.sectionai.com/blog/chatgpt-vs-gemini-deep-research#:~:text=The%20other%20big%20difference%20in,but%20allows%20for%20less%20nuance)[

analyticsvidhya.com

OpenAI Deep Research vs Gemini Deep Research

Q1. How is OpenAI Deep Research different from Google Gemini Deep Research?

](https://www.analyticsvidhya.com/blog/2025/02/openai-vs-google-who-does-deep-research-better/#:~:text=Q1,from%20Google%20Gemini%20Deep%20Research)[

analyticsvidhya.com

OpenAI Deep Research vs Gemini Deep Research

Both OpenAI Deep Research and Google Gemini Deep Research bring powerful AI- driven research capabilities to the table. OpenAI’s Deep Research focuses on real-time, interactive analysis with transparency. Meanwhile, Google Gemini Deep Research offers a more affordable yet structured research methodology, presenting a well-formatted, document-friendly report. While OpenAI Deep Research provides deeper insights with a more iterative approach, Gemini Deep Research remains a strong contender for users who prefer a straightforward research output at a lower price.

](https://www.analyticsvidhya.com/blog/2025/02/openai-vs-google-who-does-deep-research-better/#:~:text=Both%20OpenAI%20Deep%20Research%20and,output%20at%20a%20lower%20price)[

anthropic.com

How we built our multi-agent research system \ Anthropic

critical lessons about system architecture, tool design, and prompt engineering. A multi-agent system consists of multiple agents (LLMs autonomously using tools in a loop) working together. Our Research feature involves an agent that plans a research process based on user queries, and then uses tools to create parallel agents that search for information simultaneously. Systems with multiple agents introduce new challenges in agent coordination, evaluation, and reliability.

](https://www.anthropic.com/engineering/multi-agent-research-system#:~:text=critical%20lessons%20about%20system%20architecture%2C,agent%20coordination%2C%20evaluation%2C%20and%20reliability)[

anthropic.com

How we built our multi-agent research system \ Anthropic

ImageThe multi-agent architecture in action: user queries flow through a lead agent that creates specialized subagents to search for different aspects in parallel.

](https://www.anthropic.com/engineering/multi-agent-research-system#:~:text=ImageThe%20multi,for%20different%20aspects%20in%20parallel)[

anthropic.com

How we built our multi-agent research system \ Anthropic

for breadth-first queries that involve pursuing multiple independent directions simultaneously. We found that a multi-agent system with Claude Opus 4 as the lead agent and Claude Sonnet 4 subagents outperformed single-agent Claude Opus 4 by 90.2% on our internal research eval. For example, when asked to identify all the board members of the companies in the Information Technology S&P 500, the multi-agent system found the correct answers by decomposing this into tasks for subagents, while the single agent system failed to find the answer with slow, sequential searches.

](https://www.anthropic.com/engineering/multi-agent-research-system#:~:text=for%20breadth,answer%20with%20slow%2C%20sequential%20searches)[

anthropic.com

How we built our multi-agent research system \ Anthropic

There is a downside: in practice, these architectures burn through tokens fast. In our data, agents typically use about 4× more tokens than chat interactions, and multi-agent systems use about 15× more tokens than chats. For economic viability, multi-agent systems require tasks where the value of the task is high enough to pay for the increased performance. Further, some domains that require all agents to share the same context or involve many dependencies between agents are not a good fit for multi-agent systems today. For instance, most coding tasks involve fewer truly parallelizable tasks than research, and LLM agents are not yet great at coordinating and delegating to other agents in real

](https://www.anthropic.com/engineering/multi-agent-research-system#:~:text=There%20is%20a%20downside%3A%20in,to%20other%20agents%20in%20real)[

anthropic.com

How we built our multi-agent research system \ Anthropic

Traditional approaches using Retrieval Augmented Generation (RAG) use static retrieval. That is, they fetch some set of chunks that are most similar to an input query and use these chunks to generate a response. In contrast, our architecture uses a multi-step search that dynamically finds relevant information, adapts to new findings, and analyzes results to formulate high- quality answers.

](https://www.anthropic.com/engineering/multi-agent-research-system#:~:text=Traditional%20approaches%20using%20Retrieval%20Augmented,quality%20answers)[

unr.edu

CHS 211 Literature Review | Writing & Speaking Center | University of Nevada, Reno

Where you conducted your research

](https://www.unr.edu/writing-speaking-center/writing-speaking-resources/chs-211-literature-review#:~:text=Where%20you%20conducted%20your%20research)[

unr.edu

CHS 211 Literature Review | Writing & Speaking Center | University of Nevada, Reno

An important aspect of breaking down and organizing your findings results is to identify themes; these will be headings and subheadings when you format your paper. Themes are developed from the main findings of the results; they are concepts or ideas that reoccur throughout your study of the literature.

](https://www.unr.edu/writing-speaking-center/writing-speaking-resources/chs-211-literature-review#:~:text=An%20important%20aspect%20of%20breaking,your%20study%20of%20the%20literature)[

anthropic.com

How we built our multi-agent research system \ Anthropic

ImageProcess diagram showing the complete workflow of our multi-agent Research system. When a user submits a query, the system creates a LeadResearcher agent that enters an iterative research process. The LeadResearcher begins by thinking through the approach and saving its plan to Memory to persist the context, since if the context window exceeds 200,000 tokens it will be truncated and it is important to retain the plan. It then creates specialized Subagents (two are shown here, but it can be any number) with specific research tasks. Each Subagent independently performs web searches, evaluates tool results using interleaved thinking, and

](https://www.anthropic.com/engineering/multi-agent-research-system#:~:text=ImageProcess%20diagram%20showing%20the%20complete,results%20using%20interleaved%20thinking%2C%20and)[

medium.com

5 Agentic AI Patterns to Build Smarter LLM Systems | Data Science Collective

Reflection: Teach Your Agent to Check Its Own Work

](https://medium.com/data-science-collective/stop-prompting-start-designing-5-agentic-ai-patterns-that-actually-work-a59c4a409ebb#:~:text=1,to%20Check%20Its%20Own%20Work)

All Sources

[

gemini

2

](https://gemini.google/)[

sectionai

](https://www.sectionai.com/)[

anthropic

](https://www.anthropic.com/)[

openai

2

](https://openai.com/)[

bleepingcomputer

](https://www.bleepingcomputer.com/)[

blog

2

](https://blog.google/)[

reddit

](https://www.reddit.com/)[

analyticsvidhya

2

](https://www.analyticsvidhya.com/)[

cdn.prod...ite-files

](https://cdn.prod.website-files.com/)[

researchgate

2

](https://www.researchgate.net/)[

medium

](https://medium.com/)[

commons.wikimedia

2

](https://commons.wikimedia.org/)[

upload.wikimedia

3

](https://upload.wikimedia.org/)[

unr

](https://www.unr.edu/)