ChatGPT

Advanced AI Agentic Research Platforms (2025)

Several new AI products now rival ChatGPT's Deep Research capability by using multi-step, agentic workflows over large models. For example, **Anthropic's Claude API** now includes a *Web Search* tool that lets the model autonomously query the internet, retrieve results, and synthesize answers with citations [1] [2]. When Claude detects a need for up-to-date or specialized information, it generates targeted searches, fetches relevant pages, and reasons over them in sequence to produce thorough answers [1] [2]. This enables tasks like developer documentation lookup or GitHub issue analysis within a single conversation. Similarly, **OpenRouter** provides a unified API layer for *any* LLM, routing calls to 300+ models from 50+ providers (OpenAI, Anthropic, Qwen, etc.) with pay-as-you-go credits [3] [4]. OpenRouter's distributed infrastructure boasts high availability (auto-failover) and minimal added latency (~25ms) while enforcing fine-grained data policies, and it accepts one API key for all supported models [3] [4].

Other platforms focus on AI-powered search and summarization tailored to developers: **Tavily Search API** is a cloud service built specifically for LLMs and agents. It "reviews multiple sources to find the most relevant content" and returns concise, citation-backed information optimized for retrieval-augmented generation (RAG) workflows [5]. Unlike generic search APIs, Tavily lets you control search depth and domains, and it integrates readily with LangChain or LlamaIndex [5] [6]. Its pricing is usage-based (e.g. a ~$30/month plan for 4,000 queries) with a free tier for students [5] [6]. Tavily's strength is real-time, up-to-date web search tuned for AI, but it may require custom ingestion if you need internal sources (see Limitations below).

Felo Search Agent is a commercial platform that combines AI search with reporting automation. It runs multi-step queries and "automatically breaks down tasks and conducts multi-step searches," then compiles the findings into reports with one click [7]. Felo emphasizes enterprise features: it can ingest uploaded documents and output formatted deliverables (PPT slides, mind maps, PDFs) based on the gathered information [8]. In practice, Felo will gather answers from web, documentation, and other inputs, filter the best snippets ("industry-leading answer quality"), and stitch them together. Pricing is tiered (enterprise quotes on request), but it offers a free or trial search interface. Strengths include end-to-end automation of research reports; limitations include relative opacity (no public API details) and possible reliance on web sources unless internal data is uploaded.

Khoj AI positions itself as an "AI research copilot" that you can customize with different LLMs. It supports a wide range of large models – from Anthropic's Claude and Google's Gemini to OpenAI's GPT and even open-weight models on Hugging Face [9]. Khoj provides a chat interface where you can upload knowledge sources (Discord chat logs, PDFs, code, etc.) and query them with the selected model. It offers a free "Humanist" plan (unlimited chat with free models, 10 MB of personal data) and a $30/mo "Futurist" tier (any model, 500 MB of context, longer windows) [10] [11]. By allowing you to plug in custom data and pick the model, Khoj is flexible: e.g. you could use Llama 3.1 or DeepSeek R1 via Hugging Face for unconstrained reasoning, or go with Claude/GPT-4o for more reliable answers. In sum, Khoj's strength is its flexibility and data privacy (user data isn't used to train models). Its limitations include context-size caps and the need to manage data upload manually; very niche sources (like private Discord servers) may require extra steps to ingest.

Phind (formerly Sides) is an AI search engine optimized for programmers. It is “built for developers, by developers,” using a stack of large-code models. Its core answer model is a fine-tuned CodeLlama (Phind-70B and larger Phind-405B models) designed for code understanding [12] [13]. Phind provides instant, code-aware answers (typically within ~15 seconds) to technical questions, integrates with your codebase for context-aware results [14], and even supports code testing and image analysis via GPT-4. Pricing is subscription-based: **Phind Pro** is around $20/month (annual $17/mo) for unlimited use of its own 70B/405B models plus daily access to GPT-4o and Claude [13]. There is also a business tier ($40/mo) with enterprise features. Strengths of Phind include its multi-query search, support for 32K token contexts, and focus on code diagnostics. However, it’s primarily a developer Q&A engine and not a general agent; it may not easily ingest non-public data (e.g. private issue trackers) without custom connectors.

Besides these hosted services, some open-source and enterprise tools merit note. **Perplexica** is a self-hosted, open-source AI search engine (a Perplexity AI clone) that you can deploy yourself [15]. It provides chat-style answers with citations, using any open LLM backend you choose. As an open MIT-licensed project, it’s free to use (assuming you supply the model). Another example is **CrewAI**, an agent orchestration platform: it’s open-source for developers and offers enterprise deployment. CrewAI lets you build multi-agent workflows across any cloud/LLM, but pricing for enterprise support is custom [16]. Finally, Hugging Face’s ecosystem supports similar capabilities: their **smolagents** library and *Agents* framework let you compose LLMs with tools (search, code execution, etc.) [17], and their Inference API hosts models like NVIDIA’s Llama-3.1 Nemotron Ultra 253B (open-sourced on HF in 2025 [18] [19]) or Google’s DeepSeek R1. Using Hugging Face or OpenRouter, one can swap in these state-of-the-art models behind any agentic pipeline.

Each option has trade-offs. For up-to-date web research, Claude’s API and Tavily excel with streaming citations and broad coverage [1] [5]. For developer-focused troubleshooting, Phind and Felo offer tailored interfaces. For maximal flexibility or privacy, self-hosted solutions like Perplexica, or frameworks like LlamaIndex (for custom corpora) [20], let you target niche sources (Discord archives, mailing lists, etc.). OpenRouter and Hugging Face give you the plumbing to use the biggest models (LLama-3.1/Nemotron Ultra, Gemini, Claude-4o, etc.) with pay-as-you-go billing [3] [18]. In practice, a combination is often best: for example, one might use Claude’s web search tool or Tavily to gather raw data, then feed it through a reasoning-capable LLM (like Llama-3.1 Nemotron Ultra) via OpenRouter or a custom Hugging Face agent.

Comparison of Notable Tools:

Tool / Platform	Key Features	Models Supported	Pricing	Strengths	Limitations
Tavily Search API [5] [6]	Web search API optimized for AI/RAG; multi-source aggregation; easy LangChain integration	Any (via API input; uses OpenAI, Anthropic, etc. behind the scenes)	Usage-based (~$30/mo for 4k queries; free student plan) [5]	Fast, up-to-date info; built for LLM workflows; citations included	Focuses on public web – private/closed sources not indexed by default

Tool / Platform	Key Features	Models Supported	Pricing	Strengths	Limitations
Felo Search Agent [7] [8]	Automated research agent; multi-step query planning; report/PPT generation; collaborative enterprise features	Proprietary AI models + GPT/others for integration	Enterprise pricing (contact sales); offers limited free/demo tier	End-to-end report automation; format outputs (PPT, mind map) [8]	Less transparency (closed tech); costly for heavy use; not open-source
Khoj AI Copilot [9] [10]	AI chat copilot with custom agents; upload documents/logs; multi-model support (Claude, Gemini, OpenAI, open models)	Anthropic Claude, Google Gemini, OpenAI GPT, Grok, HF models, etc. [9]	Free tier ($0, 10 MB data) and $30/mo premium (500 MB, any model) [10] ; enterprise custom	Highly flexible; can plug in own data and models; on-prem options	Limited by upload size & context length; needs manual data ingestion; some source types (Discord) need extra handling
Phind (developer search) [21] [13]	Developer-focused AI search; code-aware Q&A; codebase integration; 32K token context	Phind's own 70B/405B models (fine-tuned on code) plus GPT-4o/Claude daily [13]	Phind Pro ~$20/mo (yearly $17) [13] ; Business ~$40/user/mo	Very fast, precise coding answers; supports many languages [12] ; PASS1 ~75%; live code testing	Narrower focus (tech Q&A); not easily extensible to arbitrary corpora; requires web/internet (no self-host)
Anthropic Claude (API w/ web search) [1] [2]	General AI assistant with integrated web search tool; multi-turn reasoning with citations; agentic query refinement	Claude Large/Opus (4o) and Claude+ with browsing	Pay-as-you-go tokens (Anthropic API rates); plus free/pro Chat plans	Up-to-date info access; proven reasoning on tech topics [2] [22] ; built-in citation	API cost can be high for long jobs; web search cannot index private data; model cap in tokens

Tool / Platform	Key Features	Models Supported	Pricing	Strengths	Limitations
OpenRouter [3] [4]	Unified LLM API for 300+ models (OpenAI, Mistral, Qwen, etc.); distributed for reliability; single-key billing	All major LLMs (GPT-4, Claude-4o, Gemini, Llama-3.1, etc.) via partners [3]	Credit-based pay-as-you-go (no subscription) [3]	Flexibly switch models mid-workflow; global failover and low overhead [4] ; fine-grained control	Dependent on external providers' uptime/prices; abstracting many models can obscure performance differences
Perplexica (OSS) [15]	Open-source AI search/chat engine (Perplexity clone); you deploy it yourself	Any open LLM backend (customizable)	Free (self-host)	Fully customizable; no vendor lock-in [15] ; immediate citations	Must host and scale infrastructure; setup and maintenance required; out-of-box performance depends on chosen model
CrewAI (OSS/ Enterprise) [16]	Multi-agent orchestration framework; build automated workflows with any LLMs	Any LLM (via integration); cloud-agnostic	Free (community); Enterprise version (custom pricing) [16]	Supports complex, multi-agent pipelines [16] ; open source core	Enterprise features (UI, support) cost extra; code-level configuration needed (not plug-n-play)

Each platform's strengths align with different research needs. For rapid web-based answers and analysis, Claude's browsing API [1] and Tavily [5] are compelling. Developer-specific issues are well served by Phind's specialized knowledge base [21] . Projects requiring maximum customization or internal data use (GitHub issues, company docs, Discord threads) may lean on open solutions (Perplexica, LlamaIndex/RAG frameworks [20]) or API platforms where you feed in your own data (Khoj, CrewAI, etc.). All these tools leverage very large models – for instance Nvidia's 253B Nemotron Ultra or Google's DeepSeek R1 are available via Hugging Face or OpenRouter [3] [18] – and most offer flexible billing (credit/token usage) rather than flat fees. In summary, practitioners now have a diverse toolkit of agentic AI platforms that go beyond simple web search, each with trade-offs in ease-of-use, openness, and specialty domains.

Sources: OpenAI and product docs [1] [2] [5] [6] [7] [8] [9] [10] [21] [12] [13] [3] [4] [15] [16] [18] . (See table for specific citations per tool.)

[1] [2] [22] Introducing web search on the Anthropic API \ Anthropic
https://www.anthropic.com/news/web-search-api

[3] [4] OpenRouter
https://openrouter.ai/

[5] [6] Tavily
https://tavily.com/

[7] [8] Enterprise Pro - Felo - Your Free AI Search Engine
https://felo.ai/enterprise

[9] [10] [11] Khoj AI
https://khoj.dev/

[12] [13] [14] [21] Phind: AI-Powered Search Engine for Developers | Solve Coding Problems Faster
https://www.phindai.com/

[15] GitHub - ItzCrazyKns/Perplexica: Perplexica is an AI-powered search engine. It is an Open source alternative to Perplexity AI
https://github.com/ItzCrazyKns/Perplexica

[16] [20] 35+ Agentic AI Tools to Watch in 2025
https://akka.io/blog/agentic-ai-tools

[17] Agents
https://huggingface.co/docs/transformers/en/agents

[18] [19] Nvidia's new Llama-3.1 Nemotron Ultra outperforms DeepSeek R1 at half the size | VentureBeat
https://venturebeat.com/ai/nvidias-new-llama-3-1-nemotron-ultra-outperforms-deepseek-r1-at-half-the-size/