

In-Product Search: How the UX Is Changing

Deep Research report · 150 sources

Search is being pulled in two directions at once: consumer and enterprise products are collapsing keyword boxes into conversational, multimodal interfaces, while the infrastructure underneath (retrieval, ranking, reranking) is quietly being rebuilt around embeddings, agents, and LLM-driven relevance. Below is what actually shipped, organized by the part of the search experience it touches.

1. Keyword boxes are becoming conversational, intent-first interfaces

The clearest UX trend is the replacement of literal keyword matching with natural-language, intent-driven query understanding across very different product categories:

- **Amazon** rebuilt Alexa for Shopping's search bar to accept natural-language questions ("where are my AA batteries?") and return personalized answers instead of keyword results, including order-tracking, product comparisons, and preference inference from shopping history¹⁸.
- **Salesforce** launched agentic commerce search that interprets shopper intent and context rather than matching keywords, explicitly framing this as the next step after 20 years of keyword-based commerce search³⁵.
- **DoorDash** rolled out Ask DoorDash, letting users describe what they want ("quick dinner under \$30 near me") instead of browsing categories; about 70% of early messages are discovery-type queries, and most users iterate across multi-turn sessions rather than issuing one-shot searches^{20 86}.
- **Gemini** and **Chrome** are absorbing search-adjacent tasks directly into the assistant: Gemini Live can now analyze photos or screenshots a user shows it mid-conversation¹⁰⁸, Chrome added a Gemini-powered side-by-side product comparison tool for shopping¹¹⁷, and Chrome integrated Google Lens natively for real-time visual search without leaving the browser¹⁵⁰.
- **Apple TV** built a system that ranks search results after every keystroke, blending semantic text embeddings with collaborative ID-based embeddings so 1-3 character prefixes still produce usable results, a direct response to how little text impatient searchers actually type¹⁴⁹.

The throughline: search UX is shifting from "type keywords, scan results" to "state intent, get an answer," with the system doing the work of disambiguating what the user actually wants.

2. Visual and multimodal search is expanding beyond text entirely

- **Yelp** built real-time dish recognition that matches restaurant photos to menu items, prices, and reviews, eliminating manual dish-by-dish searching, particularly useful in-restaurant⁷. It also

replaced its rigid menu-tree support chatbot with conversational AI that handles queries outside predefined phrase patterns⁵⁴.

- **Amazon's** Alexa for Shopping can analyze a photo of a fridge or bookshelf and recommend products based on what it sees, combined with contextual signals like apartment size¹⁸.
- **Pinterest's** homepage still foregrounds visual search mechanics like "search by skin tone" and object-tagging based discovery as headline product features⁵⁸.
- **Google Deepmind** is prototyping an AI-enhanced mouse pointer that infers not just what a user is pointing at but why it matters in context, treating a 50-year-old UI primitive as ripe for an intent layer¹⁰⁹.

3. Search results are becoming richer objects, not lists of links

Several products shipped features that change what a "result" even looks like:

- **Amplitude's** Resource Center now highlights matched search terms directly in article text and autoscrolls to the relevant passage when an article opens, a small but table-stakes UX pattern for documentation search¹.
- **Lattice's** AI Agent now answers data questions with charts and tables plus direct links to full reports, replacing text-only summaries⁹.
- **Google Ads and Analytics** added AI Overviews summarizing performance shifts at the top of dashboards, text-prompt-driven visual report generation, and benchmarking against anonymized competitor data, all built on Gemini [15][32] .
- **Contentful** added an Optimization Agent that answers natural-language questions about content performance with text and visual responses, plus follow-up querying without manual report-building¹⁰⁴.
- **Cloudflare's** Radar Researcher converts plain-language queries into interactive charts of internet traffic data⁸⁷.
- **OpenAI** added natural-language reservation search to ChatGPT (describing an outdoor patio for two next Saturday) and quiz generation from any topic, moving structured-search tasks into conversational flow¹²⁹.

The pattern: search results are increasingly generated artifacts (charts, summaries, structured answers) rather than ranked lists a user has to interpret themselves.

4. Feed and homepage ranking are moving from static slots to generative, retention-aware systems

This is a step beyond query-based search: platforms are rethinking how content surfaces without an explicit query at all.

- **Netflix** replaced its slot-based homepage architecture with GenPage, which generates entire personalized layouts end-to-end rather than ranking pre-defined slots, tested across millions of users with measurable engagement gains^{47 81}.
- **Pinterest** redesigned home feed ranking around User Interest Clusters, using hierarchical clustering on embeddings of a user's last 500 actions to model interest lifecycle stages, shifting optimization from pure engagement to retention, with state-dependent scoring that treats new interests differently from mature ones¹³.
- **Netflix** also built GenRec, an LLM-backed ranking system that matches its production recommender using 10 to 40 times less training data, suggesting the underlying ranking logic behind "search-adjacent" surfaces like homepage and recommendations is consolidating around foundation models rather than hand-built features³⁰.

5. Search inside AI agents and assistants is being exposed as infrastructure

A distinct trend: search is being unbundled from any single product and offered as a composable capability for agents to call.

- **Perplexity** shipped a Search SDK letting agents fan out multiple queries and programmatically filter, dedupe, and rank results in code¹⁹, plus a remote MCP server giving Claude, Cursor, and VS Code access to Perplexity search with just an API key¹³⁵.
- **Ahrefs** launched an MCP server so ChatGPT and Claude can query live SEO and brand-visibility data remotely [45][11] .
- **Mistral** released Agentic Search, a five-tool retrieval loop (search, open, navigate, read, grep) for iterative document inspection, claiming 3x accuracy gains on financial-filing benchmarks and up to 39.6% latency reduction versus single-pass retrieval¹¹⁸.
- **Pinecone** launched Nexus, a knowledge layer with its own query language (KnowQL) meant to replace repeated agent re-derivation of context with a shared, traceable knowledge layer, reporting 12% higher accuracy and up to 80% lower cost per task on a third-party benchmark^{26 103}.
- **AWS** added native vector search to DynamoDB (single-digit millisecond latency, 99%+ recall, trillions of vectors, no separate infrastructure)²³, and launched a managed Bedrock Knowledge Base with an "agentic retriever" for multi-step queries beyond keyword matching⁶⁰.
- **Cloudflare** launched AI Search, a no-code product letting users point at their own files and websites to stand up search without stitching together separate primitives⁶⁹.

This signals that "search" is increasingly a backend service agents call programmatically, not just a front-end box users type into.

6. Relevance and ranking engineering is getting more specialized under the hood

Even where the front-end box looks unchanged, the relevance logic behind it is being rebuilt:

- **DoorDash** deployed small language models for ad relevance prediction, using an offline LLM to generate training data at scale and a low-latency SLM in production, explicitly separating relevance (precision) from ranking/personalization (recall)³. It also rebuilt ads retrieval to handle spelling correction inline, combining dense embeddings with lexical matching¹⁰², and built a three-layer memory system (long-term, session, conversational) for Ask DoorDash that lifted checkout conversion 15-24% and cut intent-misunderstanding by about a third⁹⁸.
- **Instacart** built Semantic IDs to unify fragmented product records across retailer SKUs, feeding cleaner data into downstream search and ranking¹³⁸, and is hiring specifically for Search & Recommendations Ranking roles⁶⁴.
- **Airbnb** is hiring for Relevance and Personalization / Query Intelligence¹¹³, and is testing an AI-search toggle with natural-language input, AI-generated listing titles, and real-time personalized highlights, letting users opt in without disrupting the existing filter-based search¹⁴⁸.
- **Meta** is building a hierarchical, unified embedding system spanning users, advertisers, and products for deep-funnel ad matching⁴⁰, alongside a sequence-modeling ranking architecture that replaces static features with temporal user-action sequences at billions-of-interactions scale⁹¹.
- **Cohere** released Rerank 4, targeting both relevance accuracy and lower latency for RAG and search pipelines³¹, and **HubSpot** built a custom vector database handling over 20 billion embeddings in production to support search, recommendations, and other AI features at scale⁶.
- **Anthropic's** Contextual Retrieval technique (Contextual Embeddings + Contextual BM25) reduces failed RAG retrievals by 49% alone and 67% combined with reranking, addressing the problem of chunks losing surrounding context¹⁴.
- **Hugging Face's** Sentence Transformers v6.0 added a MultiVectorEncoder for ColBERT-style late-interaction retrieval, keeping one vector per token instead of compressing to a single vector, including support for visual document retrieval without OCR⁵⁷.
- **Pinecone** added fuzzy matching and n-gram substring search to its full-text search, enabling typo tolerance and autocomplete without reindexing⁵².

7. Enterprise knowledge search is consolidating into cross-platform "context layers"

Internal/enterprise search is converging on a pattern: pull everything (docs, tickets, CRM records, meeting notes) into one queryable layer rather than searching tool by tool.

- **Atlassian's** Rovo now indexes Google Drive, SharePoint, and OneDrive alongside native content, making external documents queryable through Rovo Search, Chat, and Agents¹⁶. It also added a Salesforce connector to the Teamwork Graph so agents can pull account and opportunity data alongside Jira and Confluence content⁸⁴, and launched Code Context for multi-repo codebase understanding feeding into the same graph¹²⁵.

- **Glean** added native Snowflake integration letting users query approved tables in natural language, combined with context pulled from email, calendars, and task tools⁷¹. Glean's CEO Arvind Jain argues that real context requires know-how, know-why, and know-who, not just document retrieval, and frames this as the seven-year bet behind the product [130][89] .
- **Granola** added MCP integrations across six specialized tools plus Claude, ChatGPT, and Cursor, treating itself as a context source rather than a closed note-taking app⁵¹, and integrated with Gemini for natural-language querying of meeting notes⁸⁸.
- **Spotify** built a context layer encoding domain expertise into its internal data assistant specifically to reduce the friction of finding the right dashboard and interpreting its data⁴³.
- **Dropbox** used the DSPy framework to automate optimization of its Dash relevance judge, replacing manual prompt engineering with a measurable, repeatable tuning loop⁶⁵.

What this adds up to

The UX signal across these sources is consistent: the search box itself is disappearing as a distinct interaction pattern in favor of ambient, conversational, and often proactive retrieval, while the mechanics powering relevance (embeddings, reranking, agentic multi-step retrieval, memory) are getting measurably more sophisticated and are increasingly exposed as callable infrastructure for agents rather than locked inside a single product's UI. Google's leadership has pointed to this shift showing up in the numbers already, with Pichai attributing 17% search revenue growth to AI integration into the core product², while Shopify's CTO has pushed back on the assumption that this is zero-sum, noting Shopify's AI-referred traffic grew 3x year over year while organic traffic still grew 12% on a larger base¹⁰⁵.

Sources

Launch

- 19 [Released Perplexity Search SDK for agent-first Python applications](#)
- 20 [Rolled out AI conversational shopping assistant to U.S. iOS, powered by internal LLM agent](#)
- 26 [Launched Pinecone Nexus, a knowledge layer for enterprise AI agents](#)
- 31 [Released Rerank 4, claiming improved relevance and latency over prior versions](#)
- 60 [Managed RAG service with native connectors cuts enterprise data pipeline setup](#)
- 69 [Launched AI Search, a no-code search engine for enterprise data](#)
- 70 [Launches Agent B AI tool; adds AI indexes to Brand Radar](#)
- 87 [Launched Radar Researcher, an AI tool for querying Internet traffic data in plain language](#)
- 94 [Launched Perplexity Computer integration for Microsoft Teams](#)
- 103 [Launches Pinecone Nexus, a knowledge layer for AI agents with Knowledge Query Language](#)
- 118 [Launched Agentic Search with multi-step retrieval; 3x accuracy on financial filings](#)

- 122 [Launched Seller, standalone Marketplace app with AI-powered photo-to-listing auto-fill](#)
- 125 [Atlassian Code Context brings multi-repo codebase understanding to Teamwork Graph](#)
- 134 [Introduced Scouts, autonomous agents that wrangle and contextualize PostHog data](#)
- 140 [Brave Search API reaches 700,000 users via OpenClaw integration](#)

Feature

- 1 [Resource Center articles now highlight search terms and autoscroll to matches](#)
- 7 [Built real-time dish recognition from restaurant photos and reviews](#)
- 8 [Accelerated project search in version 214.4 without adopting ripgrep](#)
- 9 [Agent returns charts and tables for data queries, not just text](#)
- 12 [Opened Chrome DevTools debugging data to AI coding assistants](#)
- 15 [Google Ads and Analytics gain AI summaries, visual reporting, and performance benchmarking](#)
- 16 [Rovo gains Google Drive and SharePoint integration for unified search and chat](#)
- 17 [Gemini adds 14 app integrations for planning, creation, and task management](#)
- 18 [Alexa for Shopping adds conversational search with personalized recommendations and image](#)
- 23 [DynamoDB adds native vector search with single-digit millisecond latency](#)
- 29 [Claude Code gains browser debugging via Chrome extension integration](#)
- 32 [AI Max gains multi-campaign A/B testing and Performance Planner integration](#)
- 33 [Confluence agents now callable via @mention, work across Claude and IDEs](#)
- 34 [Gemini Spark gained Chrome web browsing integration](#)
- 35 [Released agentic commerce search that interprets shopper intent beyond keywords](#)
- 36 [Integrated Deep Research into Perplexity Computer as native skill](#)
- 45 [Launched MCP Server enabling ChatGPT and Claude to query live Ahrefs data](#)
- 47 [Replaced slot-based homepage with AI system generating fully personalized layouts](#)
- 49 [Integrated design creation into Google Gemini and AI Search](#)
- 51 [Granola now connects via Model Context Protocol to six AI apps and services](#)
- 52 [Full-text search gains fuzzy matching and n-gram substring search](#)
- 57 [Sentence Transformers v6.0 adds MultiVectorEncoder for ColBERT-style late-interaction retrieval](#)
- 59 [Shipped rainbow brackets feature after 3 years of user requests](#)
- 62 [Launched Bot Analytics tool; 24 updates across platform including AI prompts and API access](#)
- 63 [Added code review tools to Windsurf editor workspace](#)
- 67 [Rovo AI agents automate release admin tasks for enterprise customers](#)
- 68 [Integrates with Superhuman Go to surface GTM context across 1M+ apps](#)
- 71 [Glean Assistant gains native Snowflake integration for SQL-free data queries](#)
- 75 [Pomelli Photoshoot gains preset templates and video animation](#)
- 76 [Mobile app redesigned navigation bar around user identity and preview upcoming features](#)
- 79 [Shop Campaigns data now available through ShopifyQL for third-party analytics apps](#)
- 80 [Logs now cross-reference traces in search results](#)
- 83 [Added AI-powered citations feature that auto-suggests and formats sources inline](#)
- 84 [Salesforce connector added to Teamwork Graph for unified customer context](#)
- 85 [Supabase integration added to Perplexity Computer AI agent](#)

- 88 Granola now integrates with Gemini for natural-language meeting queries
- 93 Added custom chat triggers on any website element
- 97 Claude Code desktop gains built-in browser for doc and design interaction
- 100 Atlassian Rovo enables custom AI agents for release management automation
- 102 Integrated spell-correction into ad retrieval to handle millions of misspelled grocery queries
- 104 Adds AI-powered Optimization Agent to Personalization product
- 108 Gemini Live gains ability to analyze images users show it in real time
- 112 Windsurf Wave 12 integrates Devin AI capabilities into IDE
- 115 Published video showing how PMs cut status meetings by centralizing launch tracking in Jira
- 116 Gong arrives in ChatGPT Apps directory; August release adds API controls, scorecard templates,
- 117 Gemini AI in Chrome gains side-by-side comparison tool for shopping decisions
- 124 Added developer tools to help sites work with AI agents
- 126 Added phonetic controls and pronunciation glossary to fix mispronounced words
- 127 Jira Summer 2026 release adds MCP capabilities for coding agents, GA Teamwork Graph CLI
- 128 Databricks AI Search now handles high-throughput production queries
- 129 ChatGPT adds quiz generation, reservation search, Google Drive integration
- 135 Perplexity search available as remote MCP server for Claude, Cursor, VS Code
- 136 Added phonetics and voice-note controls to fix AI mispronunciations
- 141 DevTools stabilized agent debugging and integrated AI assistance across audits
- 143 Collections now support multi-source composition and variant-level targeting
- 148 Tests natural-language search with AI-generated titles and highlights; expands support bot to 50+
- 149 Dual embeddings let Apple TV rank search results from single keystrokes
- 150 Chrome integrates Google Lens for real-time visual search within browser

Internal tool

- 6 Built custom vector database handling 20 billion+ embeddings in production
- 13 Shifted home feed ranking to retention-first via User Interest Clusters, replacing engagement-only
- 24 Researchers revealed vision-language models fail at inferring visual concepts from images
- 28 Embedded hidden offers in website code to test whether AI crawlers respond to marketing
- 40 Built unified embedding system to match user interests with ads across entire network
- 43 Embedded domain expertise into data assistant to reduce dashboard search friction
- 46 Built containerized Claude Code with proprietary context mesh for internal development
- 65 Replaced manual prompt engineering with automated optimization system for search ranking
- 82 Metrics Revamped
- 96 Compresses production failures into model weights daily, beats frontier models, cuts serving costs
- 99 Performance as a Core Product Feature
- 109 AI mouse pointer infers user intent from cursor position and context
- 114 Built ML system to deduplicate billions of pins across different URLs
- 137 Built AOSC, an open-source OpenSearch plugin for zero-downtime index migration
- 138 Built unified product IDs across millions of retailer SKUs to fix fragmentation

AI adoption

- 2 AI tools drive 17% search revenue growth in latest quarter
- 3 Deployed small language models for search ad relevance prediction at scale
- 48 How we built a high quality Q&A assistant
- 54 Replaced rigid menu-tree chatbot with conversational AI for support
- 77 Built internal AI agent platform connecting to 1,000+ tools for knowledge work
- 78 Agents wrote 30% of code in a week; global tax filing shipped in 1/3 the time
- 98 Built agent memory system to personalize Ask DoorDash ordering assistant
- 147 CTO describes iterative flywheel for aligning LLM agents: optimize ground truth, then prompt,

Org

- 64 Hiring Senior Machine Learning Engineer II, Search & Recommendations Ranking
- 113 Hiring Senior Machine Learning Engineer, Relevance and Personalization (Query Intelligence)

Research

- 14 Contextual Retrieval reduces RAG retrieval failures by up to 67%
- 30 Built GenRec, an LLM-backed ranking system matching production recommender on 10–40× less
- 37 Distilled few-shot prompts cut B2B conversation classification tokens 99%, improve accuracy
- 91 Developed multi-stage ads ranking architecture using temporal user sequences and scaling laws
- 119 27B model trained on synthetic law-firm data outperforms frontier models at 10x lower cost

Perspective

- 4 Half of major software companies won't survive five years without intent-first design
- 5 CEO argues AI's power centralization is not structural, disputes scaling-law inevitability
- 10 Monorepo architecture best for agentic software factories, CEO argues
- 21 Only 24% of agent steps actually need AI, research finds
- 22 Co-founder argues measuring quality and designing systems are now engineering's hardest
- 27 Anthropic engineer argues SaaS companies leave revenue on table by not exposing APIs for agent
- 38 AI should proactively assist work, not wait for explicit requests
- 39 Video will be the final interface for all software, CEO predicts
- 44 Pay-per-use model will become standard offering for AI agents, with upgrade paths
- 55 Stacked defenses reduce indirect prompt injection to near-zero on unseen attacks
- 56 CEO argues Google sacrificed AI dominance by withholding frontier models behind APIs
- 61 GitHub is most-used connector in AI agent platforms, reaching non-technical users
- 72 Model routing and context efficiency, not model power alone, drive AI economics
- 73 Musk backs space economy as essential for AI compute and internet scaling
- 89 Intelligence layer ownership requires control over context, quality definition, and user workflows
- 90 CEO: AI tools enable building products previously impossible to ship alone
- 95 Armstrong: agents will outnumber humans; crypto infrastructure must scale to serve them
- 105 CTO challenges AI-will-kill-organic-search narrative with Shopify data
- 106 Musk backs claim that orbital datacenters will become necessary for AI scaling by 2030s

- 120 Frontier AI will scale reasoning via automated symbolic world models, not raw scale
- 121 Canva ranks second only to ChatGPT in web traffic among AI tools
- 130 Context is not more data; enterprise AI needs know-how, know-why, know-who
- 131 Head of Engineering: AI agents will reshape issue-tracking workflows
- 139 Eich argues scraping differs legally and commercially from crawling; defends search engine
- 144 CEO warns AI sycophancy erodes product ideas; proposes adversarial design patterns