Modernizing search:

An enterprise guide to AI-powered information retrieval

OpenSearch

An executive guide to enterprise search

OpenSearch

1: Executive summary

Enterprise search is a fundamental capability that enables teams to discover, access, and act on an organization's collective knowledge—across systems, formats, and silos—through a unified interface. Unlike consumer search engines that index public web content, enterprise search operates on proprietary data of your choosing: documents, emails, tickets, code, policies, chat threads, and the vast stores of unstructured information where institutional knowledge lives.

Your enterprise's code repositories, CRM systems, ticketing tools, data warehouses, and internal knowledge bases are a powerful source of competitive advantage, but only if you can access that information with accurate, trustworthy results.

Today, enterprise search is in the midst of a fundamental transformation. We are moving beyond traditional one-shot searches and introducing conversational capabilities made possible by agentic AI. Traditional keyword-based and lexical search techniques are now being combined with modern semantic approaches, creating hybrid systems that offer faster retrieval, higher quality results, and less user friction, empowering companies to generate more value from their data than ever before.

To deliver real business impact, enterprise search solutions must provide clear answers, actionable insights, and contextualized intelligence, rather than returning a list of links or partial matches. Recent advances in AI have reshaped what search can deliver: no longer limited to simple retrieval, search is evolving into an intelligent interface for enterprise knowledge, capable of interpreting user intent, integrating information across silos, and supporting real-time decision-making. And this technology is ready to be deployed to enterprises at nearly any scale.

This book is written for enterprise leaders, architects, and practitioners seeking to modernize their organization's search capabilities. It outlines the technological shifts, key architectural decisions, and operational requirements for deploying search responsibly at scale and addresses critical considerations such as security, compliance, governance, and rollout.

2: Developments and opportunities in enterprise search

Enterprise search is the capability that enables an organization's people to discover, access, and act on its collective knowledge—across systems, formats, and silos—through a unified interface.

Traditionally, enterprise search solutions used lexical or keyword search models, which are optimized for highly structured data. An employee's queries would return documents, not answers or context. For a time, this is all that was required and it functioned well.

However, today's modern companies produce massive quantities of unstructured or semi-structured data (see Diagram 1) across an expanding ecosystem of platforms. Traditional search models can struggle to retrieve relevant results from unstructured and semi-structured content, leaving users without immediate access to the insights their data contains.

Semantic and hybrid search improve relevance by retrieving information based on meaning and context, not just keywords. While a meaningful step forward, these systems still retrieve results in a single pass, leaving users to interpret and act on what they find.

The next evolution is AI-driven, agentic search: systems that retrieve, reason, and evaluate results iteratively—running modified queries across sources until the response most closely matches the user's intent. This transformation from single-pass retrieval to multi-pass reasoning is what makes modernized enterprise search a strategic capability rather than a utility feature.

The fundamental business requirement of deriving value from proprietary data remains as strong as ever—which means enterprise search must evolve into something more powerful and tailored to stay competitive.

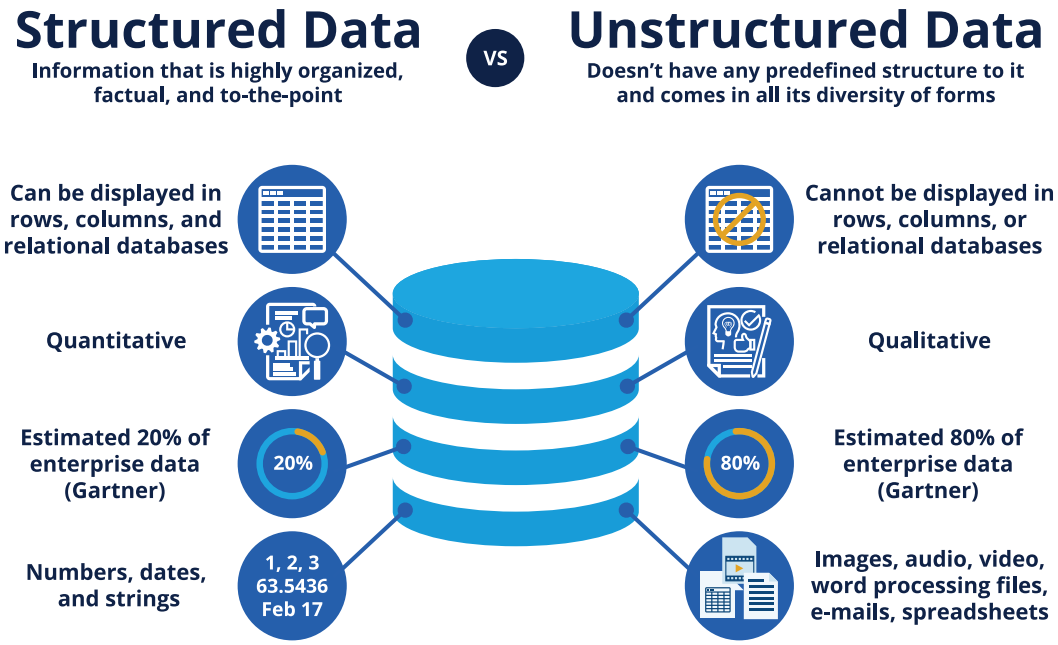

Diagram 1. An estimated 20% of enterprise data is structured, and the other 80% is unstructured in formats like documents, email, code repos, and chat logs. Statistics source: MIT Sloan (https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data)

From document retrieval to decision acceleration

Traditional search systems were built for structured data and static repositories, not for how modern enterprises work. Users must guess the right terms, sift through long result lists, and manually piece together context across systems. Even with tuning and filters, the burden remains on the user. The result is friction for employees, inefficiency for teams, and direct business consequences in the form of lost time, slower execution, and insights that arrive too late or never reach the right people. For many use cases, modern enterprises need search systems that understand intent, context, and meaning, delivering clear and actionable answers—rather than static lists of links.

AI-powered search is enabling new ways to interact with data and surface meaning, particularly from massive volumes of unstructured data which comprise the majority of today's enterprise data. The rest of this chapter explores this transformation in search technology over the last thirty years.

The hidden knowledge problem

Early enterprise search relied primarily on keyword matching which is effective when queries and content are precise, but less suited to the contextual, implicit language that fills most organizational knowledge today. According to a recent article by MIT, 80% to 90% of enterprise content lives in unstructured formats such as emails, chat threads, tickets, call transcripts, PDFs, meeting notes, internal wikis, and code repositories. This is where decisions are explained, trade-offs are debated, and lessons are learned, but it's also where knowledge is most easily lost.

Because critical insights are often buried inside long documents, scattered across multiple conversations, or expressed in informal language, traditional search tools are ill-equipped to surface what matters. For teams on the ground, this means repeatedly solving problems that have already been addressed, receiving inconsistent answers across departments, and spending weeks onboarding when days should suffice. When enterprise search cannot surface these insights, crucial knowledge remains locked in the heads of a few experts, turning them into bottlenecks and slowing progress across the organization. For leaders, the impact is compounding: reduced productivity, increased operational risk, and a slower path from insight to value.

Establishing a modern foundation for intelligent search

Because most enterprise data is unstructured, effective search has become a core organizational capability. Modern enterprise search must move beyond basic keyword matching to understand context, connect disparate pieces of information, and surface actionable insights. By doing so, organizations can fully leverage their knowledge, improve decision-making, and ensure that critical information remains accessible at scale.

To truly accomplish this goal, solutions need to be purpose-built for the requirements of modern enterprises, enabling them to unify structured and unstructured data within a single, scalable search platform. These are the core capabilities required for modern enterprise search at scale:

- **Hybrid retrieval**, combining keyword precision with semantic understanding
- **Vector storage and similarity search** for meaning-based discovery
- **Evaluation and ranking frameworks** to continuously improve relevance
- **Flexible, open architectures** that integrate across data sources and applications

TThis foundation enables organizations to move beyond syntax-based search and toward systems that operate on meaning, making it possible to surface insights, summarize content, and connect related knowledge across silos.

Core AI advances transforming enterprise search

Modern enterprise search is being transformed by representational learning through vector embeddings. By capturing semantic meaning rather than relying solely on exact word matches, this approach enables search experiences driven by intent and context. Users can surface relevant information even when terminology differs, making search more natural and effective.

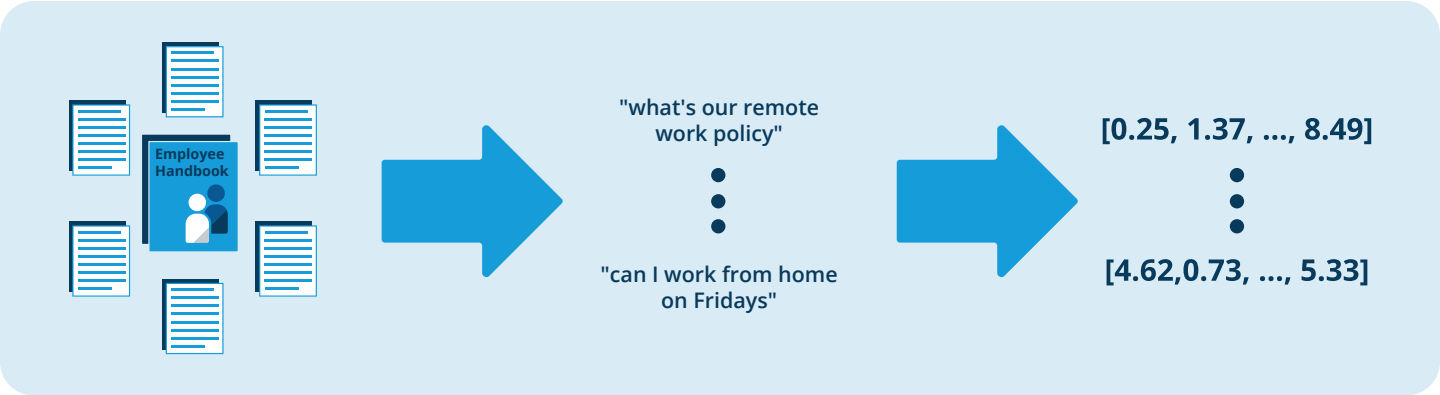

Diagram 2: Vector embeddings capture semantic meaning.

Modern search platforms extend these capabilities to the enterprise by representing content as vectors that encode meaning and relationships. This foundation supports semantic and hybrid search at scale, unifying the precision of keyword search with deeper contextual understanding in a single platform.

For organizations, this evolution moves search beyond basic document retrieval toward true insight delivery. Employees can ask questions in everyday language, uncover related knowledge across systems, and connect information that was previously siloed, resulting in faster access to insights, more confident decision-making, and a search experience that scales with the growing complexity of the business.

	Search type	How it works	Strengths	Limitations	Best use cases
1990s	**Lexical search**	Matches exact keywords and terms using inverted indexes (for example, BM25).	Fast, predictable, transparent, easy to tune.	Struggles with synonyms, ambiguity, domain-specific jargon, and natural language queries.	Compliance search, exact document retrieval, log search, known-term queries.
Late 1990s - 2010s	**Semantic search**	Uses vector embeddings to retrieve results based on meaning rather than exact terms.	High recall, understands intent, handles synonyms and paraphrasing well.	Less transparent, more compute-intensive, relevance can be harder to explain.	Knowledge discovery, natural-language queries, Q&A over unstructured content.
2015 - early 2020s	**Hybrid search**	Combines lexical and semantic techniques with configurable weighting (boosting) to balance keyword precision and semantic understanding across result sets.	Balances precision and recall, improves relevance across query types.	More complex to configure and evaluate.	Enterprise search, customer support, internal knowledge bases.
2020s - now	**AI search**	Uses LLMs across the search pipeline: understanding query intent, enriching data at ingestion, and synthesizing retrieved results to surface insights.	Can explain results, connect insights, and support decision-making.	Requires strong governance, has a higher cost and risk of hallucinations without safeguards.	Research, strategic analysis, agentic workflows, complex investigative tasks.

Table 1: The enterprise search continuum.

Modern enterprise search platforms dynamically apply the right retrieval techniques for each query, using fast keyword search for simple requests and semantic or AI-powered methods for more complex questions. The result is accurate, scalable search that improves efficiency, enables insight reuse, and makes enterprise knowledge easier to discover, understand, and apply.

3: AI-powered Search: RAG to agentic riches

Retrieval-augmented generation (RAG) has become foundational to how organizations use AI to access and apply knowledge. RAG helps AI give better answers by looking up the right information first and then responding with accuracy and context. This makes it easier to turn large volumes of scattered information into insights people can actually use.

Traditional search tools like keyword search work well when users know the exact terms to look for. Semantic and hybrid search improve upon this with added relevance by understanding meaning and context. But these approaches often fall short when users need answers that synthesize information from multiple documents or systems. RAG addresses this gap by augmenting search with AI-generated responses, allowing systems to deliver clearer, more complete, and more actionable answers.

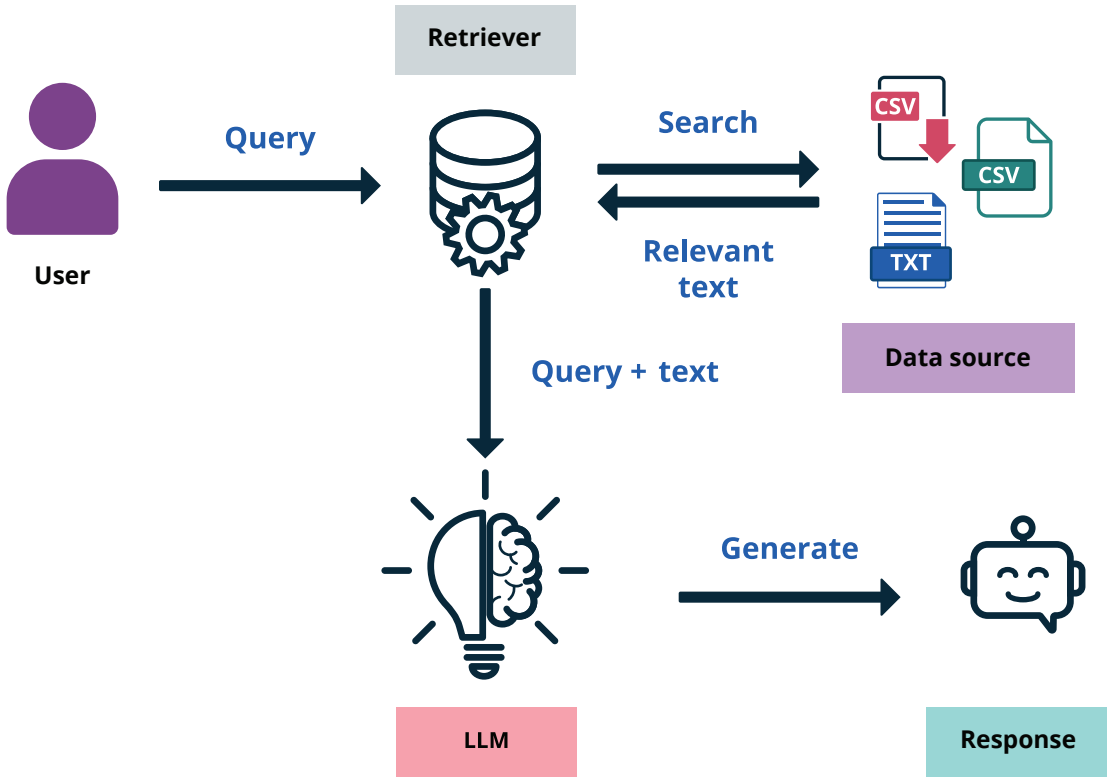

Diagram 3: A rudimentary RAG workflow.

RAG works by first finding the most relevant information across large data sets, then using AI to summarize, explain, or synthesize that information in plain language. This turns enterprise search from a basic document lookup into a practical knowledge engine that unlocks insights hidden in emails, reports, tickets, and other unstructured content. The result is faster access to information, improved productivity, and better decision-making.

Most RAG systems today follow a straightforward pattern: a user asks a question, the system interprets the semantic meaning of the request, retrieves relevant documents, and the AI generates a single response grounded in that retrieved content.

From retrieval to reasoning at scale

Most search systems, including many RAG-based solutions, assume that a single query is sufficient to answer a user's question. While this works for simple requests, it quickly breaks down when questions are ambiguous, multi-part, or require synthesizing information across different areas of the business.

These limitations point to the next stage in search evolution: agentic search. Instead of generating a single response and stopping, agentic systems can plan, reason, and iteratively take multiple steps to arrive at a complete and accurate answer. Agentic search addresses the limitations of traditional approaches by decomposing complex questions into smaller, manageable steps. Instead of attempting to generate a complete answer in a single pass, these systems plan and reason through a sequence of actions, gathering information incrementally and building toward a coherent response.

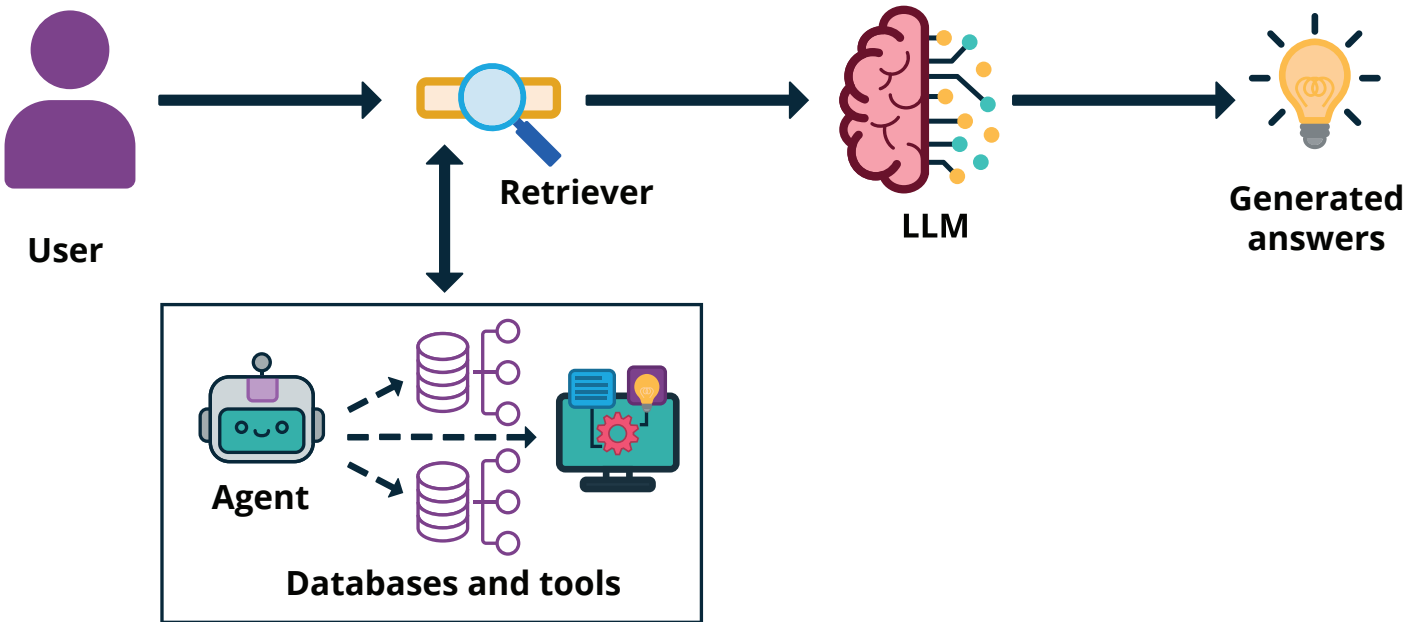

Diagram 4: A basic agentic search workflow.

In practice, agentic RAG systems determine what information is needed first, retrieve relevant context in stages, and refine their queries as new insights emerge. This iterative process allows them to adapt their approach, incorporate additional context, and handle more complex, multi-part requests similar to a knowledgeable human assistant.

For example, a seemingly simple question such as "What is our data retention policy?" might require synthesizing guidance from multiple policy documents, regional compliance requirements, and approved internal exceptions. Using an agentic approach, platforms can first identify relevant policy documents, then assess whether retention rules differ by geography or data type, and finally cross-check results for consistency. Each step builds on the previous one, enabling the system to refine its understanding rather than relying on a single retrieval pass.

The strategic opportunity: when agentic search creates business value

The shift from single-query RAG to multi-step agentic systems represents more than an incremental improvement in search quality. It fundamentally changes what kinds of business problems become solvable through enterprise search.

Organizations that move early on agentic search gain measurable advantages in three areas. First, they create competitive separation in knowledge-intensive work categories where speed and accuracy compound over time: due diligence, regulatory response, strategic planning, and cross-functional coordination. Second, they compress decision-making cycles by reducing the time senior staff spend hunting for context across systems—turning hours of research into minutes of synthesis. Third, they reduce operational risk by ensuring that complex questions receive thorough, consistent answers with complete audit trails (see Chapter 4) rather than relying on whoever happens to remember the right policy or procedure.

The decision to deploy agentic search isn't binary. Simple RAG remains the right solution for straightforward, single-domain questions: HR policy lookups, product specification queries, or FAQ-style helpdesk tickets. Agentic search justifies its complexity when questions require cross-referencing multiple knowledge domains, when thoroughness carries regulatory or competitive stakes, or when the cost of incomplete answers is high. **The question for most organizations isn't whether to adopt agentic capabilities—it's which use cases justify the complexity and investment today.**

Where agentic search delivers real impact: Use cases

Consider three scenarios where agentic search directly impacts business outcomes rather than simply improving search results.

Mergers and acquisitions due diligence requires synthesizing insights across legal agreements, financial statements, operational systems, and cultural assessments, often under compressed timelines. An agentic system can incrementally work through a diligence checklist, pulling relevant passages from hundreds of documents, cross-referencing claims against source materials, and flagging gaps or inconsistencies. What traditionally takes a team days of coordination happens in hours, with a complete record of what was examined and why.

Regulatory compliance and policy synthesis becomes exponentially more complex in global organizations where rules vary by jurisdiction, industry, and data type. When a compliance officer asks "What are our obligations for customer data collected in the EU and transferred to our US systems?", an agentic system can identify the relevant regulations (GDPR, state privacy laws, industry requirements), locate internal policies addressing each, check for approved exceptions or transfer mechanisms, and flag any conflicts or gaps. The result isn't just a faster answer but a more thorough one, with lower risk of overlooking a critical requirement.

Product incident response and post-mortem analysis relies on connecting information across monitoring systems, ticket histories, code repositories, runbooks, and prior incident reports. When production breaks at 2:00 AM, an agentic search system can trace similar past incidents, identify what resolved them, pull relevant runbooks, and surface recent changes that might be contributing factors—all before the on-call engineer finishes reading the initial alert. Speed matters, but so does comprehensiveness; missing critical context can turn a 10-minute fix into a three-hour outage.

These scenarios share a common requirement: synthesizing context across systems, formats, and organizational boundaries to compress decision cycles, reduce operational risk, and create competitive separation through superior knowledge work.

Designing agentic systems for enterprise trust

This strategic potential comes with obligations. Building agentic systems requires careful orchestration and management. Unconstrained agents can introduce risk, unpredictability, and compliance concerns, especially in highly regulated or zero-trust environments.

The foundational principle is that AI agents must function within well-defined boundaries, with clear limits on the actions they can perform and the systems they can access. Their behavior should be fully observable, enabling teams to understand how decisions are made and why specific sources are consulted. Comprehensive auditing and logging are crucial, not only for troubleshooting but also for meeting regulatory requirements.

4: Trust through observability

AI-powered search systems make autonomous decisions about what information to surface, synthesize, and present to users. As these systems incorporate semantic search, machine learning models, and generative AI, they introduce new capabilities and new risks. Without clear visibility into how and why these decisions are made, enterprises face three interconnected challenges.

Adoption risk: Users won't trust opaque systems in high-stakes scenarios. When search results lack explanation or attribution, knowledge workers default to manual research rather than relying on AI-assisted answers.

Governance risk: Organizations cannot audit or explain system decisions to regulators, compliance teams, or customers without understanding how results were produced.

Performance risk: Teams cannot diagnose failures, optimize relevance, or prevent regressions without systematic insight into system behavior.

The solution is observability: the capability to understand, audit, and improve AI search systems through systematic measurement and transparency. For enterprises, observability transforms AI search from an experimental tool into governable infrastructure that users and stakeholders can trust

What observability means for AI search

Observability in AI search extends beyond traditional infrastructure monitoring—tracking not just whether systems are running efficiently, but what decisions they're making and why. This focus on decision transparency complements operational observability and is essential for building trust in AI-powered systems.

Rather than monitoring cluster health, latency, or error rates alone, observability for AI search requires making AI decisions interpretable. The three dimensions that matter most to enterprise decision-makers are result explainability, source transparency, and confidence signaling.

Result explainability: Why was specific content surfaced or prioritized?

Users and operators need visibility into the factors that influence ranking and retrieval. This includes understanding which relevance signals were applied, how keyword and semantic scores were combined, and why certain documents were prioritized over others.

Results for query: *laptop*

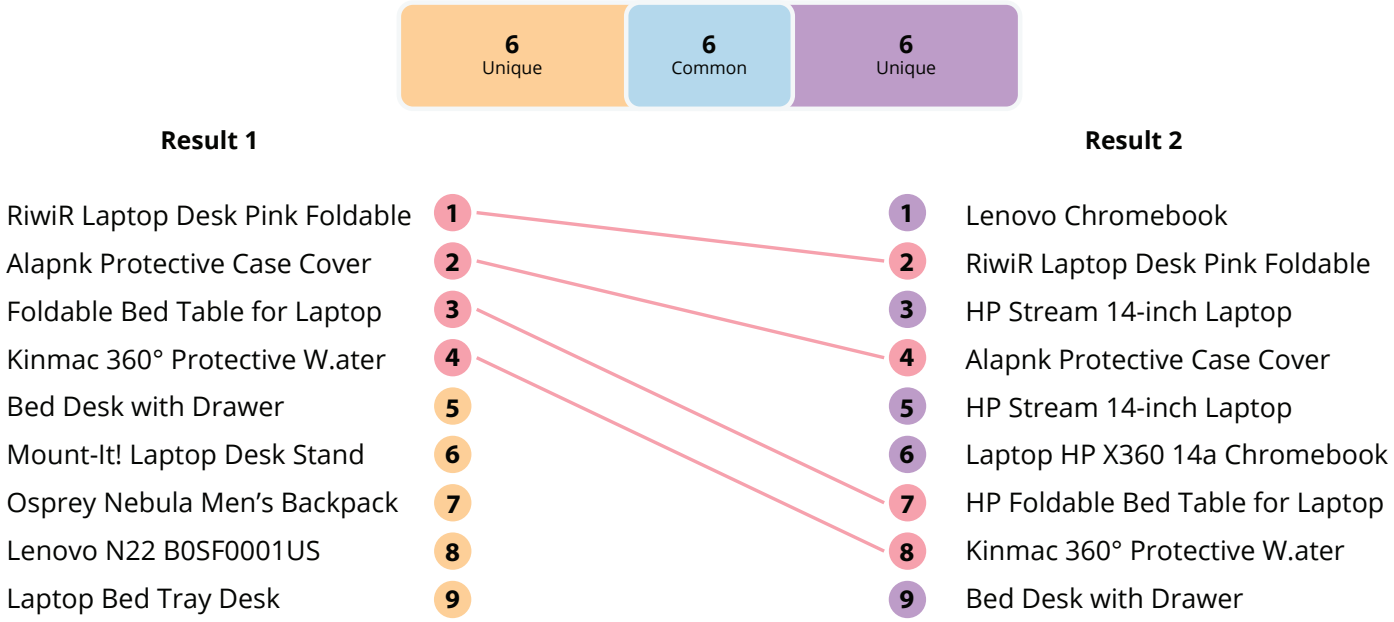

Diagram 5: Side-by-side results show what changed in search results and why.

Without these insights, it becomes difficult to distinguish between expected behavior and system error.

Source transparency: Where did this information originate?

In enterprise environments, content originates from many systems with varying levels of authority and freshness. The system must clearly indicate which data sources contributed to a given result. Source transparency is especially critical in AI-assisted search and question-answering scenarios, where users need to know whether answers are grounded in official documentation, internal policies, or informal knowledge.

Confidence signaling: How certain is the system in its outputs?

AI-powered search systems often return results that appear authoritative, even when system confidence is low. The search platform should provide signals that reflect uncertainty, such as relevance scores, confidence indicators, or coverage metrics. These signals enable users and downstream applications to make informed decisions about how much to trust a given result.

Explainability, transparency, and signaling connect directly to business outcomes: faster troubleshooting when systems underperform, regulatory compliance through auditable decision trails, and user confidence that drives adoption. This is how your system earns trust from users.

Observability in practice: Key signals and examples

Observability depends on capturing the right signals—metrics that reveal both system performance and decision quality. These signals serve dual purposes: they enable teams to diagnose and improve system behavior internally, and they provide the transparency users need to trust AI-generated outputs externally.

Result quality metrics help teams assess whether relevant content is being surfaced for given queries and track changes in retrieval quality over time, allowing teams to detect regressions and measure the impact of tuning efforts.

User behavior patterns provide valuable feedback on effectiveness. Metrics such as query abandonment rates, reformulation rates, and click-through patterns reveal whether users are finding what they need. High abandonment or repeated reformulation often signals relevance issues, unclear results, or excessive latency.

System performance and latency remain critical factors in user satisfaction. AI-powered search pipelines can introduce additional processing steps, making end-to-end latency monitoring essential to identify bottlenecks and ensure performance remains within acceptable thresholds.

These signals converge in practice. When an agentic system answers, "What is our data retention policy?", observability should reveal which policy documents were consulted first, whether the system identified regional variations, which sources were deemed authoritative, and where it was cross-checked for consistency. Each decision point should be traceable, allowing compliance teams to audit the reasoning and users to verify the completeness of the answer. Without this visibility, even correct answers lack the credibility required for high-stakes decisions.

Trust but verify

Observability only creates trust when it's surfaced to users in meaningful ways. Internal metrics and logs are necessary but insufficient: users need transparency at the moment of interaction.

User-facing features like citations, confidence scores, and reasoning traces transform your complex query into verifiable insights completely derived from your enterprise data. Observability isn't operational overhead but is what makes AI-powered search scalable, governable, and suitable for enterprise deployment.

Systems that can't explain their behavior are difficult to improve, govern, or scale safely. Without observability, AI search remains an experimental capability confined to low-stakes use cases. With robust observability—supported by meaningful metrics, transparent sourcing, and explainable outputs—enterprises can deploy AI search confidently across high-value, regulated, and mission-critical functions.

5: Platform choices and deployment models at enterprise scale

Enterprises approaching AI-powered search must make a foundational decision that shapes implementation timelines and long-term flexibility, control, and operational costs: How will this capability be deployed, operated, and evolved?

This decision involves weighing two types of investment. Financial investment includes software licensing, infrastructure costs, and managed service fees that scale with usage. Engineering investment includes the time and expertise required to build, tune, and maintain the platform. Organizations rarely optimize for both simultaneously. Instead, the question becomes which resource you can afford to spend more of: capital or engineering capacity.

Three primary deployment models dominate the landscape, progressing from turnkey simplicity to maximum control and customization. Most enterprises ultimately adopt hybrid approaches that balance these trade-offs.

Managed services: Speed and simplicity

Managed search services prioritize ease of use and rapid deployment. In this model, a service provider handles infrastructure provisioning, scaling, security, upgrades, and ongoing maintenance. Internal teams can focus on delivering features and optimizing user experiences rather than managing systems.

Services like Azure AI Search, Amazon Kendra and Amazon OpenSearch Service, and managed Elasticsearch deployments exemplify this approach. They typically offer strong defaults for security, monitoring, and backups, dramatically shortening time to value—especially for teams without deep search infrastructure expertise.

Benefits:

- Rapid provisioning with minimal operational overhead
- Provider handles scaling, availability, and infrastructure reliability
- Quick time-to-value for teams prioritizing feature delivery
- Protect against lock-in with managed services built on open-source technology and open standards

Drawbacks:

- Limited customization of search behavior and ranking logic
- Data residency options might be constrained by provider infrastructure
- Pricing can become unpredictable at scale, particularly with AI-driven workloads
- Reduced flexibility as search evolves from a feature into strategic infrastructure

Best for: Teams prioritizing speed to market and operational simplicity, especially where search supports the business rather than defining it.

Open-source platforms: Control and flexibility

Open-source platforms offer transparency, customization, and deployment flexibility at the cost of operational complexity. Organizations gain full control over architecture and infrastructure placement—deploying on-premises, in private clouds, or in hybrid configurations that meet strict data sovereignty and regulatory requirements.

OpenSearch, an Apache 2.0-licensed search and analytics platform, serves as the representative example throughout this guide. Like other open-source options (Apache Solr, Meilisearch), OpenSearch allows teams to customize behavior and integrate deeply with internal systems. This guide uses OpenSearch as an illustrative foundation, recognizing that the principles apply broadly across open source search deployments.

Benefits:

- High transparency and control over architecture and ranking logic
- Deployable wherever business and regulatory requirements dictate
- Supports strict data sovereignty, compliance, and security policies
- No vendor lock-in; organizations control upgrade timing and feature adoption

Drawbacks:

- Requires significant engineering and operational expertise
- Teams must manage scaling, upgrades, security, and performance tuning
- Operational complexity increases with AI-driven workloads
- Engineering investment compounds over time

Best for: Enterprises that view search as strategic infrastructure, require deep customization or regulatory compliance features, and can invest in skilled teams for long-term ownership.

Build from scratch: Differentiation at maximum cost

Fully custom search systems—built entirely in-house from foundational components—represent the highest-commitment deployment model. Organizations pursuing this path typically do so because search is central to their product offering or constitutes core intellectual property that drives competitive differentiation.

When executed well, greenfield development can deliver highly specialized experiences perfectly aligned with unique business requirements. However, this approach demands sustained engineering investment and continuous adaptation as search and AI technologies evolve.

Best for: Organizations where search is the product itself (search engines, discovery platforms) or a fundamental competitive advantage (recommendation systems at scale). This is the rarest deployment model and typically justified only when off-the-shelf solutions cannot meet domain-specific requirements. Examples include companies like Google (search is the product) and Netflix (recommendation and discovery are core differentiators).

Hybrid approaches: The pragmatic reality

Most enterprises adopt hybrid strategies that combine elements from multiple deployment models. Organizations might choose to maintain control over search logic and data while outsourcing infrastructure operations; combining the flexibility of open source with the operational simplicity of managed services. Many managed service providers build upon open-source projects, allowing teams to start with a managed deployment and selectively customize or migrate components as requirements evolve.

This approach balances control, speed, and operational pragmatism. It allows organizations to maintain stability in foundational systems while iterating rapidly on AI capabilities that change frequently. Hybrid deployments also provide natural migration paths—starting with managed services for speed, then selectively moving critical components to self-managed infrastructure as requirements crystallize.

The hybrid model acknowledges that different components have different strategic importance. Not everything needs to be managed, and not everything needs to be custom.

Comparing deployment options

Deployment model	Key benefits	Key drawbacks	Best use cases
Managed services	Fast deployment, minimal operations, provider handles infrastructure	Limited customization, potential cost unpredictability at scale	Teams prioritizing speed and simplicity where search is a supporting feature
Open source	Full control and transparency, flexible deployment, no vendor lock-in	Requires operational expertise, ongoing engineering investment	Search as strategic infrastructure with customization or compliance requirements
Hybrid infrastructure	Balances control with operational efficiency, natural migration paths	Requires architectural planning, managing multiple operational models	Most enterprise implementations—pragmatic balance of flexibility and efficiency
Build from scratch	Complete control and differentiation, optimized for specific needs	Highest cost, continuous adaptation required, significant opportunity cost	Search is central to the product or business model with unique requirements

Table 2: Comparing deployment models for enterprise search implementations.

Making the strategic choice

Selecting the right deployment model requires evaluating how search will evolve over time, its importance to core business workflows, the degree of customization required, internal operational readiness, and any security, compliance, or data residency considerations.

The decision is rarely binary, which is why hybrid approaches to infrastructure have become the dominant enterprise pattern. Enterprises often choose to optimize different components independently, maintaining control where it matters strategically while using managed services where operational simplicity creates more value than customization.

Enterprises that treat search as a long-term capability rather than a point solution are best positioned to take advantage of advances in semantic search, agentic AI, and intelligent knowledge systems where they deliver measurable value. By choosing a deployment model that balances flexibility, operational reality, and future readiness, organizations can build search systems that scale not just with data volume, but with organizational ambition.

The remainder of this guide uses OpenSearch as an illustrative open source foundation, recognizing that most enterprise implementations will combine open source flexibility with managed operational layers and custom components tailored to specific business requirements.

[1] This mirrors broader cloud infrastructure trends, where Gartner forecasts 90% of organizations will adopt hybrid deployments (combining on-premises and public cloud) by 2027. (Gartner, “Gartner Forecasts Worldwide Public Cloud End-User Spending to Total $723 Billion in 2025,” November 19, 2024)

6: Risk, security, compliance, and sovereignty

Search systems sit at the intersection of data access, discovery, and decision-making, often touching the most sensitive information within an organization: intellectual property, customer data, employee records, and regulated content. Introducing AI into search amplifies both the value and the risks. Without careful governance, AI-powered search can expose data in unintended ways, undermine compliance efforts, and erode trust.

For enterprises, security and governance must be foundational design principles. AI-powered search systems must operate within existing security boundaries, respect regulatory requirements, and behave predictably even in the face of adversarial inputs.

Key security and governance considerations

As organizations integrate AI into their search infrastructure, they introduce new categories of risk that traditional search didn't face. The following areas deserve particular attention during planning and implementation.

Data leakage and over-permissive access

A primary risk in AI-powered search is unintended data exposure. AI-generated responses might summarize, combine, or infer information across multiple sources, increasing the risk that sensitive data is revealed to unauthorized users. Over-permissive access at the retrieval layer is a common root cause. If search indexes aggregate content without enforcing fine-grained permissions, AI systems might retrieve data that users shouldn't see. Preventing data leakage requires strict enforcement of access controls during retrieval, ensuring that only authorized content is ever passed to downstream AI components.

Regulatory compliance

Enterprises in regulated industries must ensure AI-powered search complies with laws such as GDPR and HIPAA. AI adds complexity by creating new regulated artifacts—embeddings, logs, generated outputs—that must follow retention and access rules. Compliance requires transparent systems with consistent access controls, auditability, and governance across the entire retrieval and generation pipeline.

Regional data residency

For global enterprises, data residency is a key governance requirement. AI-powered search systems must ensure data is indexed, processed, and queried within approved regions to avoid regulatory violations. Region-aware architectures with localized indexes help maintain compliance while delivering consistent user experiences.

Enforcing security at the retrieval layer

A core security principle of AI-powered search is the enforcement of access controls at the retrieval layer, not at the user interface level. UI-level filtering occurs too late, after data might already be processed or exposed.

Enforcing permissions during retrieval ensures unauthorized content is never seen by ranking algorithms, generative models, or logs, reducing risk and simplifying compliance. AI and search systems must integrate directly with enterprise identity and permission frameworks, inheriting the same access policies as the underlying data.

7: Operating principles and governance

Modernizing search is an ongoing effort, not a one-time project. As data, user needs, and AI capabilities evolve, search systems must continuously adapt. Long-term ownership, clear accountability, and strong governance are essential to sustaining value and alignment with organizational goals.

Defining clear ownership across the search lifecycle

Platform operations

Platform operations teams ensure the reliability, performance, and scalability of search infrastructure. As AI components are added, they must also manage vector search, model inference, and integrations. Clear operational ownership enables early issue detection and rapid resolution, minimizing user impact.

Data stewardship

High-quality search depends on high-quality data. Data stewards ensure accurate, well-governed data by defining schemas, maintaining metadata, and managing data lifecycles. In AI-powered search, they must also govern embeddings, logs, and derived data to maintain search quality and compliance.

Model governance

Model governance refers to selecting, configuring, evaluating, and updating AI models over time. This includes decisions about which models are approved for use, how prompts and retrieval strategies are managed, and how model behavior is monitored in production. Governance in this area ensures that AI-powered search systems behave predictably and responsibly, reducing the risk of bias, hallucinations, or unintended behavior.

User experience ownership

Search systems ultimately succeed or fail based on user experience. UX ownership includes defining success metrics, collecting user feedback, and translating insights into improvements. It also involves coordinating changes across teams so that updates to data, models, or infrastructure do not degrade the overall experience. Without dedicated UX ownership, search improvements may become fragmented or misaligned with user needs.

Agent readiness

As AI agents become consumers of search alongside human users, organizations need to plan for how those agents interact with the retrieval layer. This includes monitoring agent query patterns and failure modes, enforcing the same access controls for agent-initiated queries as for human users, and designing retrieval pipelines that serve both conversational interfaces and automated workflows. Teams that treat agent interactions as a first-class concern now will be better positioned as agentic architectures mature.

8: Exploring your use case for enterprise search

Enterprise search has become one of the most strategically consequential investments an organization can make in its data infrastructure. Exploring the potential business case for search modernization is a prudent first step. Each use case below could be served by multiple search approaches with different cost, complexity, and capability trade-offs. The following table summarizes which approaches deliver the most value for each category, while the sections that follow explore the business case in detail.

Internal knowledge and productivity

Employee knowledge assistants help employees quickly find and understand information across collaboration tools, wikis, and document repositories. These systems must deliver fast, accurate results while protecting sensitive information and providing clear explanations to build user trust.

Engineering and code search helps development teams quickly navigate large codebases, APIs, and documentation. Successful implementations depend on semantic understanding, high accuracy, and explainable results to support developer productivity.

Use Case	Lexical search	Hybrid Search	AI-powered Search
Employee knowledge assistants	Limited. Requires users to know the right keywords across fragmented repositories.	Recommended. Retrieves relevant content regardless of terminology differences across teams and document types.	Highest value. Delivers direct answers from wikis, collaboration tools, and document repositories through conversational interfaces.
Engineering and code search	Strong for exact API names, function signatures, error strings, and log patterns.	Recommended. Adds conceptual search for “how do we handle authentication” across code and documentation.	Highest value. Connects related code, runbooks, and prior incident reports to surface context that spans repositories and systems.

Customer-facing applications

Customer support resolution surfaces relevant knowledge articles, case histories, and troubleshooting guides in real time, reducing response times and improving satisfaction. The system must deliver high accuracy, low latency, and clear citations to support agent decisions and meet compliance requirements.

Sales and account intelligence surfaces customer insights, deal history, and cross-sell opportunities from CRM, internal documents, and market data. These use cases require fast, accurate, and explainable results supported by flexible architectures that integrate multiple data sources.

Use Case	Lexical search	Hybrid Search	AI-powered Search
Customer support resolution	Sufficient for known error codes and documented product issues.	Recommended. Handles the natural-language queries that customers actually submit, improving resolution rates.	Highest value. Generates synthesized troubleshooting answers grounded in knowledge articles and case histories, reducing time-to-resolution.
Sales and account intelligence	Sufficient for retrieving specific account records and deal histories by name or ID.	Recommended. Surfaces relevant customer context across CRM, documents, and market data using both exact and intent-based retrieval.	Highest value. Synthesizes cross-sell opportunities and account insights across multiple data sources, connecting signals that manual review would miss.

Compliance, governance, and risk

Compliance and policy discovery helps enterprises stay compliant by quickly surfacing relevant policies, regulations, and audit records. These activities demand high precision, clear traceability, and explainable results, ensuring decisions can be justified and risk minimized.

M&A due diligence requires synthesizing insights across legal agreements, financial statements, and operational data under compressed timelines. Like compliance workflows, the underlying requirements are the same: precision, traceability, and the ability to surface risk across large, heterogeneous document sets.

Use Case	Lexical search	Hybrid Search	AI-powered Search
Compliance and policy discovery	Sufficient for known regulation numbers, policy titles, and exact legal terms.	Recommended. Adds intent-based retrieval for natural-language compliance questions across large policy corpora.	Highest value. Synthesizes obligations across multiple policies, jurisdictions, and document types with auditable sourcing.
M&A due diligence	Limited. Volume and variety of documents make keyword-only search impractical under compressed timelines.	Useful. Improves retrieval across legal, financial, and operational document types.	Highest value. Agentic workflows synthesize insights across document categories under time pressure, surfacing risks and connections that manual review would miss.

9: Common pitfalls to avoid

Search modernization offers significant benefits, but organizations often encounter recurring challenges that can undermine performance, reliability, and trust. Some of the most consequential mistakes happen before implementation begins.

Assuming AI search is the right starting point

Not every search problem requires AI. Many enterprise search challenges, including inconsistent results, poor recall, and user frustration, stem from weaknesses in the retrieval foundation rather than the absence of an LLM. In many cases, well-configured lexical search with relevance tuning, field boosting, and proper analyzers can resolve the majority of search quality issues at a fraction of the cost and complexity. Organizations that jump directly to semantic or AI-powered search without first evaluating what traditional techniques can deliver risk over-engineering their solution, introducing unnecessary infrastructure costs, and creating systems that are harder to maintain, debug, and explain.

Treating AI search as a UI feature

Enterprise search is evolving into a core, AI-powered system, not just an improved interface. When organizations treat it as a UI upgrade without strengthening the underlying infrastructure, they risk building fragile, hard-to-scale solutions. Without solid foundations in retrieval, indexing, and relevance, users encounter slow, inconsistent, or low-quality results, ultimately eroding trust and adoption.

Ignoring data quality and access control

Data quality and governance are critical to AI-powered search yet are often underestimated. Inconsistent metadata, stale content, and weak access controls reduce accuracy and increase security and compliance risks. Without strong permission enforcement at the retrieval layer, AI systems can expose sensitive information, making rigorous data preparation and governance essential.

Over-reliance on LLMs without strong retrieval grounding

LLMs are powerful but lack built-in awareness of enterprise knowledge and regulatory limits. Without grounding them in authoritative retrieval systems, AI search can produce inaccurate or inconsistent results. A retrieval-augmented approach is essential to ensure accuracy and maintain trust, especially in high-stakes enterprise use cases.

Deploying without feedback loops or observability

Deploying AI search without a clear rollout plan, user validation, or ongoing monitoring often results in low adoption and hidden failures. Without feedback loops, systems stagnate and degrade over time. AI search must be treated as a core enterprise system designed for adoption, observability, and continuous improvement.

10: Next steps: Making AI search a strategic reality

Enterprise AI search has evolved from a document-finding tool into a strategic capability that shapes how organizations access, interpret, and act on information. Turning this vision into reality requires clear ownership and a deliberate rollout plan.

Who should own this initiative

In most enterprises, this initiative should be sponsored by a senior technology leader—CTO, CIO, or chief data or AI officer—with execution shared across platform engineering, data governance, security, and representative business teams. Search touches data, infrastructure, and daily workflows, making it a cross-functional system rather than a feature owned by a single team.

Where to start

Begin with a focused, high-value use case where implementing next-generation enterprise search can clearly improve outcomes: customer support, sales enablement, engineering productivity, or compliance workflows. Make an informed and strategic decision on which deployment model to adopt at the start. These early deployments allow organizations to validate performance, accuracy, and trust while building operational expertise.

Start with strong retrieval foundations—accurate indexing, semantic understanding, and proper access controls—then layer in AI enhancements like contextual reasoning and agentic capabilities. This approach reduces risk while demonstrating tangible value quickly.

Building for long-term success

Invest in governance and operating discipline from the start. Data quality, access controls, model oversight, observability, and feedback loops are essential to sustaining performance and confidence as systems scale. Without these foundations, even advanced AI capabilities risk producing errors, exposing sensitive information, or eroding user trust.

Modernizing search is an investment in better decisions, faster execution, and a more intelligent organization. When thoughtfully designed and well governed, AI-powered search becomes a scalable platform for learning, collaboration, and operational excellence, delivering measurable ROI and long-term strategic value across the enterprise.

OpenSearch is the trusted vendor-neutral open source platform for AI-powered search, analytics, observability, and vector database solutions. Built for scalability, performance, and extensibility, developers and enterprises rely on OpenSearch to securely manage workloads, without costly licensing. Hosted by the Linux Foundation and backed by the Apache 2.0-license, the OpenSearch platform offers transparent governance, freedom from vendor lock-in, and community-driven innovation.

opensearch.org

OpenSearch

THE LINUX FOUNDATION