FAQs Mindbreeze InSpire July 2025

GenAI

For GenAI we support the OpenAI API also both in terms of using third party LLMs as well as exposing our own services and LLM deployments. Our architecture allows multiple deployments- physical and virtual appliances also available on hyper scalers as well as cloud native offering deployable on Kubernetes. We use OAuth/JWT as the native backend authorization method and support a variety of authentication protocols such as SAML 2.0, Keberos, as well as OAUth.

Mindbreeze InSpire uses special prompting techniques e. g. to mine for facts using semantic search or do self-querying retrieval and then to use this information as an input to the prompts for generative AI this way one can make sure that the knowledge is part of the input and presented to the LLM. This way Mindbreeze InSpire can ensure that the user really has access to the information that is then used to generate answers. With modern LLMs that have large context sizes (>= 128k tokens) this is ensures that the LLM gets all the relevant information for generating good content from the Insight Engine. Furthermore, our Evaluation Framework is used to automatically test using state of the art metrics and also security content filtering can be applied. The user can also send feedback directly on the generated content to raise knowledge if something unexpected appears, this is then automatically feedbacked for fine tuning with reinforcement learning from human feedback.

Mindbreeze InSpire uses AI throughout the processing stages. For instance, in the semantic pipeline, query and generative AI pipelines. A very important use of AI are agentic AI use cases. These are provided by the Insight Workplace where actionable Insight Apps are generated based on the users need for information. The agentic tools can be administrated via the Insight Services administration where one can fine tune and evaluate the quality using metric scores.
For indexing we provide multi modal neural networks that are trained for specific tasks such as OCR, layout detection, formatting, and then later also for segmentation and entity recognition and semantic linking.
We have an integrated vector index engine that is highly scalable and can be used to do semantic similarity search not only for text but also for images and multimedia content.
During retrieval AI is also used for reranking, relevance feedback by learning from personalization calls such as action clicks or interactions with the Insight Apps. This way Mindbreeze automatically improves and learns from the user to provide the best answers.

Mindbreeze Insight Workplace

Mindbreeze InSpire provides very strong structured exploration and context-driven formatting in its unique Insight Workplace.

Insight Workplace: Each question/search/action generates persistent Insight Touchpoints (query + results + tools + actions), grouped into Insight Journeys for topic continuity. Insight Apps offer configurable and on-the-fly creatable (360-degree) views using e.g. charts, tables, lists, maps, audio and video —adaptable by intend/user profile/context/data. All insight touchpoints and apps are integrate seamlessly into Outlook, SharePoint, Teams, Salesforce, and custom portals.

Users explore results with real-time faceted filtering, previews, secure deep links, tools and actions. Developers access all functionality via REST APIs and SDKs, enabling integration into any digital workflow.

Built-in natural language Q&A selects optimal response formats—text, table, chart, map—powered by Retrieval-Augmented Generation. Insight Services enrich content with metadata and extract facts (entities, dates, metrics) during indexing. Result presentation adapts to user roles through metadata injection and policy-aware logic.

Predefined insight apps, modules, widgets as well as on-the-fly ""created-on-purpose"" ones support domain needs: sales dashboards, HR org charts, legal clause tracking, manufacturing SPC, support trends. Mindbreeze InSpire turns enterprise search into a personalized, intelligent, and extensible business-decision-insights platform."

Mindbreeze InSpire offers a comprehensive set of prebuilt applications and solution accelerators for rapid deployment across functions like IT, HR, Legal, Finance, and industries such as Healthcare, Insurance, Manufacturing, and Utilities.

Key solutions include:

360-degree Views (Customer, Product, Corporate, Project, ...) for role-based, consolidated insights.

Experience Economy Use Cases like Expert Finding, Lessons Learned, Skill Management, and Recommendation Engines.

Audit & Compliance tools including GDPR Analytics, Contract Summarization, Preventive Compliance, and Fraud Detection.

Enterprise Search & Workflow Integration supporting contextual insights and knowledge discovery.

To develop and embed search, Mindbreeze provides:

Insight Workplace with Insight Touchpoints that use Insight Apps (no-code/low-code UI builder),

Insight Services for semantic enrichment and relevance tuning,

APIs & SDKs for seamless integration with ERP, CRM, or custom apps,

Flexible cloud-native, cloud, on-prem, and hybrid deployment options.

Besides many other use cases, Mindbreeze leverages GenAI to:

Interpret natural language queries and enrich metadata,

Generate accurate summaries and answers from content,

Enable conversational interfaces with LLMs,

Automate compliance workflows and risk detection,

Integrate securely with OpenAI, Azure OpenAI, or private LLMs.

Together, these capabilities transform search into intelligent, action-oriented insights.

Mindbreeze multi-layered enrichment pipeline

Mindbreeze InSpire applies a multi-layered enrichment pipeline optimized for enterprise-scale unstructured and semi-structured data. At ingestion, content is parsed and semantically enriched using advanced NLP techniques such as NER, entity disambiguation, topic modeling, sentiment analysis, rule-based classification, and speech/image-to-text conversion—augmented by domain-specific ontologies and vocabularies.

Transformer-based language models (pretrained or fine-tuned) embed documents, sections, or sentences into high-dimensional vectors. These are indexed both in a term vector inverted index for locality-aware search and a similarity index for semantic retrieval, enabling hybrid symbolic-neural search.

Prompt-based LLM tasks are supported at ingestion and can be distilled into efficient, domain-specific models. Input is dynamically chunked into semantically meaningful units based on task context (e.g., Q&A vs. search).

Weak and semi-supervised learning pipelines use structured enterprise data to bootstrap annotations for tasks like NER and classification, with few-shot learning minimizing manual labeling and focusing SME effort on edge cases.

AI models interpret document layouts and extract semantic structure through layout analysis, object detection, table understanding, and multimodal LLMs. This semantical structure is retained in the index and used for precise chunking and structured prompting in reasoning over complex formats.

Knowledge Graphs and Ontologies

Mindbreeze InSpire offers robust capabilities for creating and managing knowledge graphs and ontologies, tightly integrated with its core AI functionality.

Internal Graph Representation: At the heart of the platform is a powerful graph-based data model that structures and links entities, concepts, and relationships across disparate information sources.

Entity Discovery Framework: Mindbreeze automatically discovers and extracts entities from unstructured and structured data, enriching the knowledge graph with contextual relationships.

Ontology Integration: The platform supports the use of domain-specific ontologies, allowing symbolic information to be embedded directly into the graph. These ontologies power advanced features like semantic search, classification, and reasoning.

360-degree Views: Ontologies and graph structures are leveraged to build comprehensive 360-degree views of entities (e.g., customers, products, assets), enabling deep insight across connected data.

Symbolic Reasoning and Generative AI: Knowledge graphs and ontologies enhance AI capabilities by supporting symbolic reasoning, which complements generative AI and agentic workflows for more complex and explainable outcomes.

Tool Calling and Agentic AI: The graph structure enables agentic AI to reason over entities and relationships, dynamically invoking tools or data sources to enrich the graph or perform tasks.

Mindbreeze Search Methods

Mindbreeze combines traditional high-end search techniques with advanced AI, including vector search, large language models (LLMs), and Retrieval-Augmented Generation (RAG), to ensure deep semantic understanding and contextual relevance.

Search methods include:

Keyword Search with support for Boolean, proximity, and optional terms

Vector + LLM-Based Search, merging transformer-generated embeddings with LLMs for semantic, language-agnostic search, including multimedia similarity (e.g., images).

LLM-Based Search using models like LMSys, LLAMA, and GEMMA, providing Generative QA that returns direct, context-rich answers instead of simple document lists.

RAG (Retrieval-Augmented Generation) enhances LLM responses by grounding them in real-time, trusted enterprise content for accuracy and traceability.

Graph-Based Search uses a semantic graph to uncover relationships and support subqueries, such as identifying people connected to specific topics or documents.

Faceted and Guided Navigation enables real-time filtering and drill-down using an in-memory engine and dynamic expression language.

Clustering and Grouping automatically organizes content into semantically meaningful categories, supporting better discovery and analytics.

All methods share robust access control and operate seamlessly across structured, unstructured, and multimedia data types, enabling efficient, AI-driven insight discovery.

Mindbreeze Search Queries

Mindbreeze InSpire allows a multimodal, intelligent, and embedded discovery experience by supporting diverse input types and dynamic, context-aware interaction methods:

Conversational & Generative Input: Users engage in natural, back-and-forth dialogues using GenAI-powered dialogs (Insight Journey and Touchpoints). Queries evolve contextually with each interaction, enabling clarification, refinement, and proactive follow-up suggestions.

Multimodal Querying: Mindbreeze InSpire processes inputs across text, voice (via speech-to-text), images, video, documents, and even combinations—like uploading a visual and asking questions about it—creating richer, personalized search experiences.

Agentic Use Cases & Task Triggers: Beyond returning results, InSpire acts as an intelligent agent—executing workflows, surfacing next best actions, or auto-generating summaries and answers based on intent.

Embedded & Contextual Execution: Search can be initiated directly within apps, chatbots, dashboards, or workflow stages. Context (user role, screen, state) is automatically injected to tailor relevance.

Continuous Input via Feedback: The platform adapts in real time using behavioral signals (clicks, inactivity, refinements) and conversational cues, enabling dynamic reranking and better outcomes.

User intent and results

Mindbreeze InSpire understands users through a multi-layered personalization framework that models user role/context, intent detection/behavior and customization.

It uses graph-based user modeling with knowledge graphs, permission frameworks, and collaboration history to build a comprehensive profile. Custom context tags embedded in queries and machine learning pipelines enable role-based relevance tuning, while client-side telemetry and HTML context injection (Insight Apps) enrich user understanding in real time. Intent detection is achieved via a hybrid approach combining semantic and lexical search, machine learning/NLP-based classification, and retrieval-augmented generation (RAG) using internal LLMs. This allows Mindbreeze to leverage both current and past interactions, maintaining conversational context for better intent inference.

Customization is extensive—organizations can fine-tune relevance using usage analytics, configure RAG pipelines, or define query pipelines with custom metadata, facets, and entity extractors. Rule-based logic, synonym transformers, and NLP models further personalize results. Pretrained and fine-tuned transformer models semantically interpret queries by encoding them into tensor representations, which are compared for similarity. LLMs then cross-encode the most relevant results for enhanced semantic interpretation. This layered, adaptive approach ensures that user intent is deeply understood and effectively served."

Intent Recognition at Depth: Mindbreeze models intent in context of organizational structure, cross-app interactions, and evolving behavior. By combining ML, telemetry, and user signals (across systems), it understands not just what users search for, but why. This leads to business-aligned results.

Self-Tuning Personalization Engine: Unlike static search relevance, Mindbreeze adapts live, reshaping rankings for each user, refine, and explore. Predefined relevance parameters combine with auto-learned ones, meaning precision at scale.

LLM-Augmented Understanding: MB uses LLMs to semantically re-rank, extract smart snippets, and contextualize answers. We try to do deeper document comprehension and pinpointing insights fast.

Entity & Signal Fusion Across Modalities: With built-in entity recognition and multimodal indexing and surfaces relevant content even when buried in videos, PDFs, or transcripts.

Telemetry-Driven Optimization Loop: Mindbreeze anonymizes and analyzes behavior across systems, enabling a closed feedback loop that continuously enhances relevance without manual intervention.

Embedded Intelligence (IPX): Mindbreeze embeds insight directly —in apps, portals, CRMs—using injected context (role, workflow stage, etc.) to shape each result in real time. It works inside the workflow.

Insight Apps: Allow users to not only retrieve information but actively engage with it.

Mindbreeze User roles

Mindbreeze InSpire supports a spectrum of expert roles throughout the AI-powered insight lifecycle. The Solution Architect designs the semantic pipeline, increasingly accelerated by LLM-assisted configuration and generative automation. The Data Engineer implements data connectors, governs ingestion, and maintains index freshness. The SME creates user interface touchpoints using zero-code. A Relevance Engineer fine-tunes ranking logic, filters, and boosting—guided by behavioral

signals and AI-driven analytics. The Knowledge Manager curates taxonomies and semantic layers, now enriched through GenAI-powered entity linking and classification. SMEs validate relevance and insight precision, while UX Designers craft intuitive, role-aware touchpoints. The Administrator ensures governance, compliance, and operational resilience.

Key pipeline stages include: Indexing with ACL-aware connectors and delta updates; Semantic Enrichment using GenAI for entity recognition and metadata tagging; Query Processing with LLM-backed expansion and intent detection; and Ranking, optimized via selected models. The Presentation Layer delivers role-aware Insight Apps with contextual filtering.

Mindbreeze provides advanced tools such as the AI-augmented Management Center, Search Preview, Usage Analytics, and the Evaluation Workbench for controlled testing and tuning. Features like Explainability, Query Simulation, and Versioning support continuous iteration and GenAI-enhanced enterprise search.

Mindbreeze InSpire a cloud-native solution

Mindbreeze InSpire is a fully cloud-native platform, running on Kubernetes-based environments like OpenShift. It uses a multi-stage, microservice-oriented architecture—comprising crawler, filter, index, query, and client services—all communicating via HTTP/S. This modular design enables independent scaling and efficient resource allocation.

Key design maxims support growth in file sizes, document volumes, and data sources:

Scalability: Scale-Out and Scale-Up supported through microservices and flexible licensing based on indexed documents—not technical limits.

High Availability: Redundant units and index mirroring ensure resilience and minimal downtime.

Federation & Distribution: Distributed indexing across installations merges into unified search results, ideal for geographically dispersed data.

Producer & Consumer Model: Separates indexing and search, allowing each to scale without performance interference.

Hybrid Deployment: Operates on-premise or in the cloud (SaaS, AWS, Azure, Google, Oracle) with seamless interoperability.

GPU Optimization: Built on custom Dell-Nvidia hardware, leveraging heavy GPU usage for accelerated indexing and search, maximizing throughput with minimal latency.

Resource Efficiency: Kubernetes ensures auto-scaling, load balancing, and fine-grained control over compute and memory resources.

These design principles ensure high performance, availability, and adaptability for enterprise-scale search.

Mindbreeze InSpire is engineered as a cloud-native platform that adapts flexibly across deployment models—from full on-premises control to hybrid architectures and fully managed cloud operations. Its modern microservices-based design, running in containers and orchestrated via Kubernetes or OpenShift, ensures scalability, resilience, and seamless integration of GenAI capabilities.

Each component operates as an independently deployable service, enabling high elasticity, rapid updates, and continuous delivery. The platform is built for enterprise-grade reliability, offering 99.9% availability, role-based access control, and compliance with certifications such as ISO 27001, SOC 2, and EU Cloud CoC Level 3.

Mindbreeze supports multiple deployment options: on-premises for full data sovereignty and infrastructure control; cloud (SaaS) hosted in secure Mindbreeze-operated data centers; and hybrid deployments, which combine on-prem data sensitivity with cloud-based scalability—ensuring unified indexing and semantic context across environments.

In addition, Mindbreeze InSpire is available via a Bring-Your-Own-License (BYOL) model on leading cloud marketplaces including AWS, Microsoft Azure, Google Cloud, and Oracle Cloud. This gives organizations the flexibility to deploy in their preferred cloud environment using existing licenses, while taking full advantage of Mindbreeze's cloud-native architecture, enterprise security, and GenAI-powered insights.