

## deepset Cloud – An LLM Platform for Enterprise AI Teams

In today's fast-paced technological landscape, modern AI teams often struggle to prioritize when building real-world applications. The need for a reusable standard toolkit drives many teams into the undifferentiated heavy lifting of building and maintaining an LLM platform. It's also easy to overlook that model training (or fine-tuning) is just one of many organizational and technical hurdles of implementing LLM-driven applications.

## What is deepset Cloud?

**deepset Cloud** is a SaaS platform designed for enterprise AI teams focusing on delivering LLM-enabled applications within a limited timeframe. With a set of proven and trusted tools, along with practical workflows, deepset Cloud stands out for its model-agnostic and provider-neutral approach, ensuring future-proof LLM-driven solutions.

## Why use deepset Cloud?

**deepset Cloud** is the go-to platform for the teams looking to reduce the complexities associated with building LLM-powered applications. Ideal for those who wish to bypass the heavy lifting of developing own LLM platform from scratch, deepset Cloud provides AI and software engineers with:

Ability to deliver working, future-proof LLM-enabled prototypes in a matter of days

Observability and monitoring tools to streamline LLMOps

SOC2-certified, secure LLM backend

A standardized environment to optimize software development lifecycle (SDLC) for all key stakeholders

Built with trusted technology based on deepset's renowned open-source Haystack project

Scalable, low-latency inference, production-grade infrastructure

Swappable models and BYOM for model- and provider-agnostic applications

Verified tools to detect and mitigate hallucinations and inconsistencies to build AI that the end-users can trust

Easy-to-use model evaluation and end-user feedback tools

Pre-built, reusable components, templates and workflows for all major use cases: RAG, information extraction, vector-based semantic search, document similarity, and more

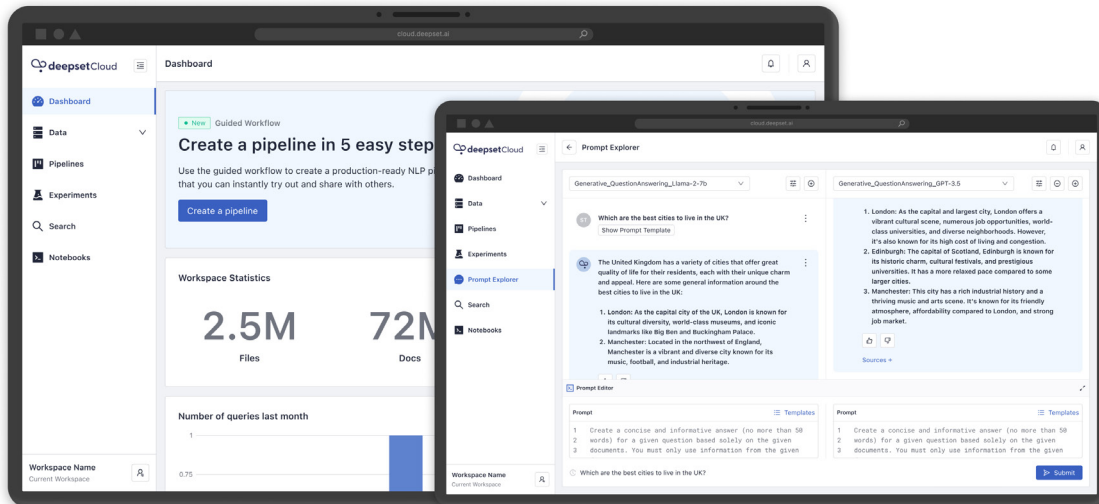
Integrations with all major LLMs: OpenAI, Cohere, Anthropic, open-source and other models

Whether you're embedding an LLM in an existing or a new application, deepset Cloud provides a comprehensive set of pre-built, standardized LLM backend components to expedite the development lifecycle, allowing AI teams to focus on what matters most – delivering tangible and measurable value to the end-users as soon as possible.

# Features and Capabilities



**deepset Cloud** supports the full software development lifecycle for LLM applications—from prototyping to testing to production.



## Start

- Guided model discovery covering all major models and model providers.
- Pre-built templates for common LLM pipelines covering key use cases.
- Ready-to-use pipeline components.

## Tune

- Prompt engineering explorer.
- Flexible configuration tools for custom pipelines and model composition.
- Model fine-tuning with GPU notebooks.

## Evaluate

- Experiment runtime and tracking.
- Easy-to-use tools for in-depth qualitative error analysis.
- Evaluation from human feedback.

## Deploy

- One-click and zero-downtime deployments.
- Expose any custom pipeline via REST API.
- Automatic pipeline scaling, including scale-to-zero.

## Observe

- Log monitoring and log export.
- Accuracy, latency and traffic monitoring.
- Hallucination auto-detection.

## Secure

- SOC 2 Type 2 certified.
- Private data plane in a VPC.
- Authentication via MFA & SSO.

## About deepset

**deepset offers enterprise developer tools to build LLM-powered applications.**

Our products and services have helped many customers across Europe and the USA to optimize information processing and discovery, and our open-source technology is used by many thousands of organizations worldwide. Founded in 2018, deepset is backed by the leading venture capital firms, such as Balderton Capital and Google Ventures.