data.world

WHITEPAPER

Building the Foundation for Scalable AI

LLMs + Knowledge Graph + Data Catalog

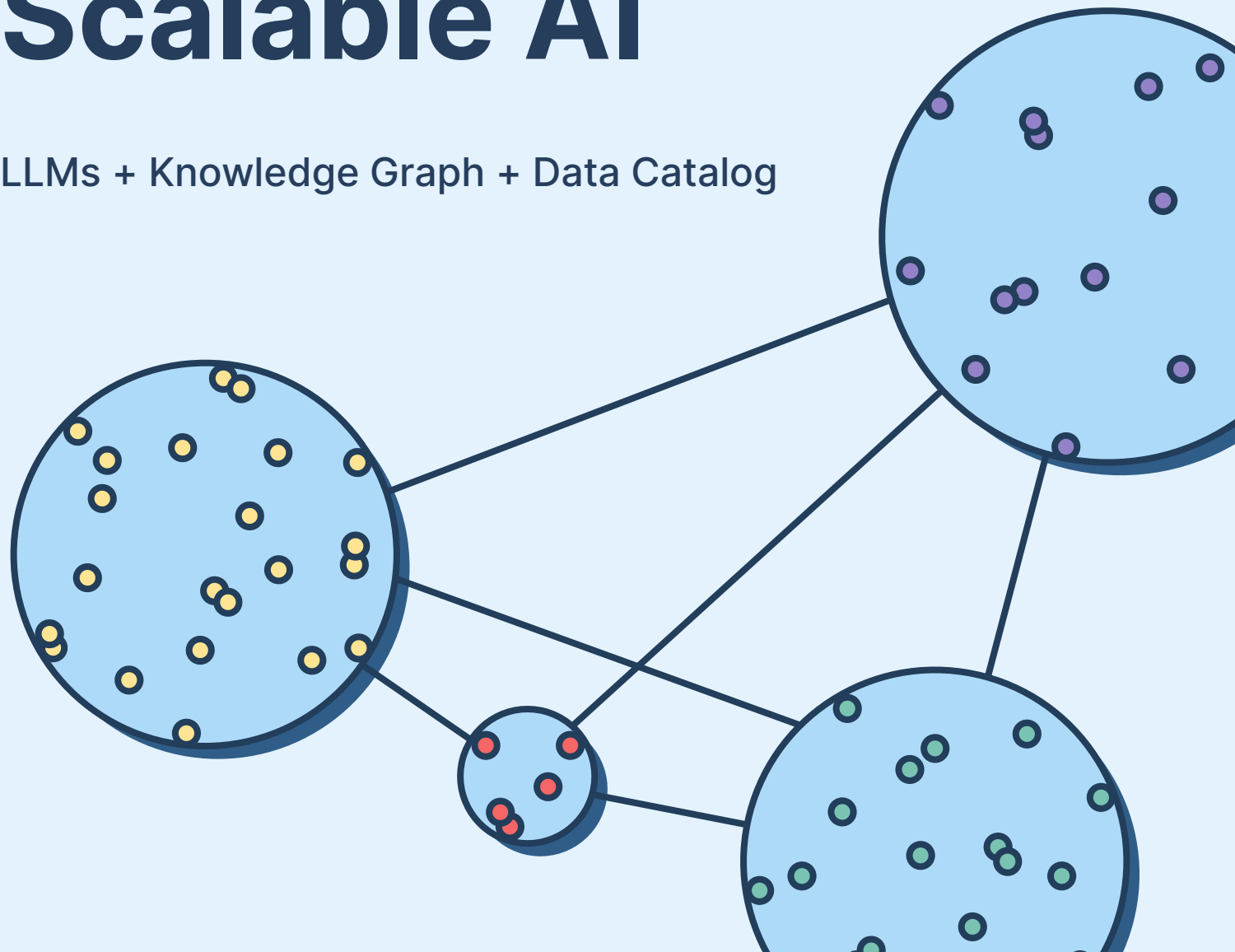
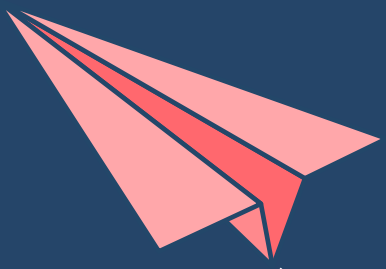


Table of contents



Introduction

Generative AI is changing everything. Things that were considered impossible twelve months ago can now be done for pennies worth of API calls. The prohibitive communications gap between humans and machines has been bridged, enabling us to speak the same language – and the resulting opportunity is immense. Goldman Sachs predicts that generative AI could raise the global GDP by 7% (or almost $7 trillion) over the next 10 years.

Despite the potential, there are critical limitations that organizations must overcome to trust the results of applications powered by Large Language Models (LLMs). Without trust, it's nearly impossible to create meaningful business value.

There are three critical limitations that erode trust:

01 **LLMs are prone to surfacing inaccurate responses, often called "hallucinations."**

02 **The black-box nature of LLMs makes it impossible to understand how a response was generated.**

03 **LLMs have the potential to expose confidential and proprietary information.**

Because of these limitations, generative AI is only half the equation needed to drive value with AI-powered applications. The other half is a data catalog built on a knowledge graph.

The data catalog brings all of the organization's data and knowledge into a governed platform. The knowledge graph can then model data and metadata in an interlinked, flexible, and open graph format. With the organization's data and knowledge governed and in this flexible format, AI-powered applications have the rich context needed to generate accurate, explainable, governed responses.

> **Knowledge graphs are the appropriate target for exploiting LLMs for business value, since they are machine-readable data structures, representing semantic knowledge of the physical and digital worlds. These worlds include entities (people, companies, digital assets) and their relationships, which adhere to a network of nodes (vertices) and links (edges/arcs), the graph data model.**
>
> **Gartner**
> "Adopt a Data Semantics Approach to Drive Business Value," Guido De Simoni, Robert Thanaraj, Henry Cook, July 28, 2023

Limitations of Generative AI

Despite their capabilities, LLMs present significant challenges when it comes to storing and retrieving facts. LLMs function as statistical pattern-matching systems. They analyze vast quantities of data to generate statistically likely responses to prompts. “Likely” being the operative word.

These systems don’t store facts in the traditional sense. They produce the most probable response based on patterns detected during training. These may be appropriate for some low stakes inquiries about general knowledge – Who is the 44th President of the United States? – but less appropriate when the inquiry is business critical and nuanced – What is the amount of Total Loss claims this quarter?

There are three key shortcomings with this approach that make it difficult for organizations to trust the responses.

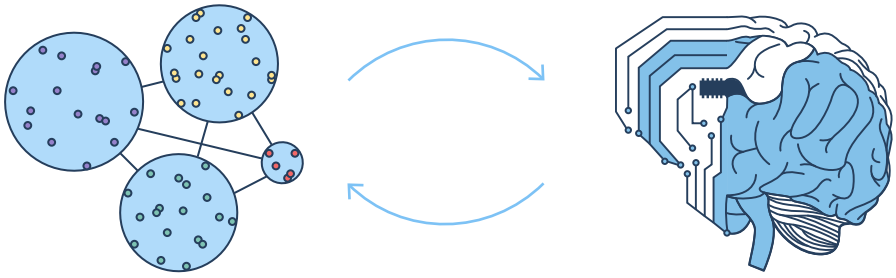
01 **A lack of accuracy:**
For LLMs, the correctness of the response is merely a statistical likelihood, not a guaranteed fact. And the smaller the datasets – say your organization’s internal data rather than the open internet – the more the model accuracy decreases. With fewer examples, LLMs can more easily make wrong assumptions, resulting in flawed responses – “hallucinations.”

02 **A lack of explainability:**
LLMs are a black box. The model exudes the same opaque confidence irrespective of the correctness of its response. The phenomenon of “hallucinations” perfectly exemplifies this issue. The LLM cannot tell you how it arrived at a specific response. If you ask it to explain how it arrived at its conclusion, it will simply produce another statistically likely response.

03 **A lack of governance:**
Without some way to control the LLM’s information source and how it is pulling information, there is always the potential for exposing confidential and proprietary information. While LLMs can produce impressive and successful results under the right conditions, the shortcoming of the models can lead to mistrust. Without trust, organizations will find it difficult to derive meaningful value.

A data catalog built on a knowledge graph

Rich organizational content



Powerful language proficiency

According to the McKinsey Global Survey, The state of AI in 2023: Generative AI's Breakout Year, "Inaccuracy," "Explainability," and "Regulatory compliance" are three of the greatest risks facing organizations when it comes to generative AI. A data catalog built on a knowledge graph is key to overcoming all three.

The data catalog creates a map of what data exists in the organization. By being built on a knowledge graph, that data is tied to meaning – semantics and context. The transformation from rigid tables to a flexible graph structure allows for more complex relationships between data points, providing a richer and more comprehensive understanding of the data, people, processes, and decisions that drive the business.

Once that data is mapped and connected, it can then be shared with the LLM. The result is the best of both worlds. The language reasoning power of the LLM and more accurate, explainable, governed responses. The data catalog built on a knowledge graph addresses three of the limitations that come with LLMs, resulting in:

01

Increased accuracy:
By sharing enterprise context (rather than relying on statistical methods), a data catalog built on a knowledge graph boosts the relevancy and correctness of LLM responses.

02

Clear explainability:
With that interlinked, flexible, and open graph format in place, now it's possible to directly trace the LLM responses to enterprise knowledge. Where LLMs were a black box, now they can literally show their work.

03

Governed responses:
With proper governance in place through the data catalog, now organizations can limit what LLMs can access — keeping confidential and proprietary information from being exposed.

data.world

About data.world

data.world is a cloud-native SaaS (software-as-a-service) Data Catalog Platform that combines an intuitive user experience with a powerful knowledge graph to deliver enhanced data discovery, agile data governance, and actionable insights.

The knowledge graph architecture leverages R2RML, a language for expressing customized mappings from relational databases to RDF datasets. This makes it possible to transform traditional relational data into a flexible and interconnected format. data.world's distinctive model-and-catalog approach makes it a foundational Platform for any organization seeking drive value with AI.

If you don't choose a data catalog platform on a knowledge-graph architecture, and bring in all data and knowledge, governed in one platform, then you are setting yourself up for failure in an AI future.

Vip Parmar
Global Head of Data, WPP

Schedule a demo →

Learn more

Want to learn more about data.world's Data Catalog application and AI capabilities? Sign up for our bi-weekly live demo or a personalized data catalog application demo with our team — bring your questions and find out more about how to get started with data.world.

Gartner