.addepto

Generative AI Europe Conference 2023:

Key Takeaways

The Addepto team participated in the Generative AI Europe Conference 2023 in Amsterdam on 5-6 December.

The event served as an in-depth dive into the world of Generative AI, covering everything from practical applications and success stories to the challenges and constraints associated with its adoption.

Here are our key takeaways:

LangChain: Strengths and Limitations

LangChain is an emerging technology that is gaining attention for its ability to handle language-based tasks efficiently. It is particularly adept at facilitating quick development and testing of ideas, making it a valuable tool for researchers and developers in the prototyping phase.

However, while LangChain excels in rapid concept validation, **it falls short in production environments** due to stability issues, scalability challenges, poor performance, and limited customizability.

LANGCHAIN LIMITATIONS IN PRODUCTION ENVIRONMENTS

Stability issues

Frequent crashes or unexpected behavior can lead to significant downtime and unreliable service, making **it less suitable for applications that require consistent performance.**

Scalability challenges

As projects grow in size and complexity, **LangChain does not scale efficiently.**

Poor performance

In comparison to more robust frameworks, **LangChain often lags in performance metrics.**

This can manifest as slower response times, reduced accuracy, or inefficiencies in handling large volumes of data, which are detrimental in a production setting.

Limited customizability

While LangChain offers some level of flexibility, **it falls short in terms of deep customization.** This restricts developers from fine-tuning the framework to meet the specific requirements of a production-level project, limiting its applicability in more specialized or demanding scenarios.

OpenAI API vs. open-source LLM

The debate between using OpenAI's API and open-source large language models (LLMs) centers around key factors such as control, privacy, and customization. OpenAI's API offers ease of use and high-quality results without the overhead of managing infrastructure, while open-source LLMs provide greater control and privacy, crucial for sensitive data handling.

RISK OF API-DEPENDENT SOLUTIONS

Using APIs such as OpenAI's **introduces potential risks, primarily in the form of stability and service accessibility.** Unpredictable service interruptions can significantly impact applications that rely heavily on these APIs.

COSTS OF API-DEPENDENT SOLUTIONS

The cost structure, typically based on predicted usage tiers, can **become prohibitively expensive for high-volume applications**, making it a less viable option for long-term, scalable solutions.

INTEGRATING OPEN-SOURCE LLMS INTO TECH STACK

A strategic approach to leveraging LLMs involves integrating open-source models into the existing technology infrastructure. By doing so, **businesses can establish their own API-like interfaces** for operations that require the capabilities of LLMs.

QUALITY AND PERFORMANCE OF OPEN-SOURCE LLMS

The output quality of open source LLMs is comparable to that of closed APIs like OpenAI's.

However, harnessing the full potential of open-source LLMs necessitates a highly optimized data pipeline. Quick deployment, effortless installations, and efficient configurations are crucial for maximizing the benefits of open-source LLMs.

COST OF FINE-TUNING

Open-source LLMs offer enhanced capabilities for customization, allowing businesses to tailor the model to their specific needs and use cases. This level of customization is often more challenging or limited with API-based solutions.

Fine-tuning open-source models can lead to better performance, more relevant outputs, and a higher degree of alignment with the specific objectives of the application.

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation significantly enhances the capabilities of Large Language Models, **making them more accurate and contextually aware.** This advancement opens new possibilities for various real-world Gen AI applications, particularly in areas where up-to-date knowledge and specialized information.

Implementing RAG in LLMs involves integrating a retrieval mechanism that can access and process information from external sources in real-time. This requires an infrastructure capable of handling large data volumes and complex queries efficiently.

What is RAG?

In traditional language models, the generation of text is based solely on the model's training data. However, RAG extends this by first retrieving relevant information from a vast external database or corpus and then using this retrieved data to inform and enrich the generation process.

Moreover, the integration must be seamless, ensuring that the retrieval process does not hinder the speed or fluency of the language model's outputs.

The benefits of using RAG with LLMs:

- Enhanced contextual understanding
- Adaptability and learning
- Increased accuracy and relevance
- Customization and flexibility

Adapting Gen AI solutions is challenging

Generative AI implies challenges that go beyond those set in classic Machine Learning. They come from several key factors, including a lack of awareness, **a lack of specialized talent, and various technological limitations.**

Technological Challenges:

Versioning and storage

Efficient management of different versions of AI models and the need for fast storage solutions to handle large datasets are crucial.

Cost management

The financial implications of developing and maintaining Gen AI systems are significant. These include **costs related to processing power, storage**, and **the development of custom solutions.**

Hallucinations

Gen AI systems sometimes generate plausible but entirely fictional information, known as hallucinations. Addressing this issue requires advanced algorithms and training techniques to ensure the reliability of the generated content.

Inference and scalability

Efficiently scaling Gen AI solutions while maintaining fast and accurate inference capabilities is a complex problem.

It involves optimizing algorithms and hardware to handle large-scale operations without compromising performance.

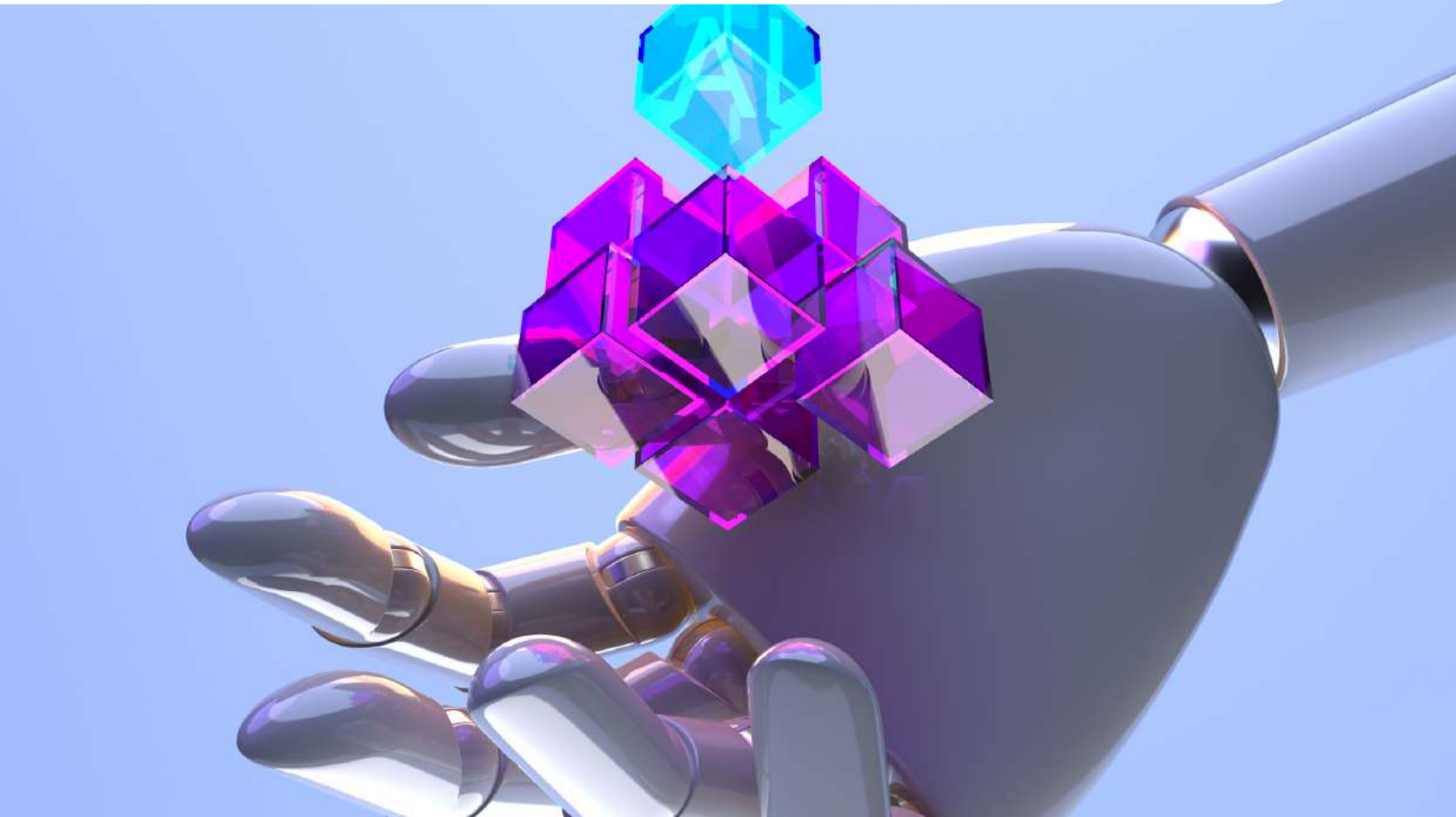
INDUSTRY-SPECIFIC DEMANDS AND SCALABILITY ISSUES

Even the largest tech companies face difficulties in building scalable Gen AI solutions. Prominent cloud services offer a foundation, **but they often fall short in terms of scalability and can be prohibitively expensive.**

Moreover, specific industries, such as legal tech, have unique requirements that go beyond generic AI capabilities. For instance, LegalTech solutions need to navigate specific workflows, legal databases, and document analysis, requiring tailored Gen AI approaches.

Conclusion: The current state of Gen AI

At present, the landscape of Gen AI is predominantly experimental. Businesses and researchers are actively testing and developing proofs of concept (PoCs), but there is a notable absence of fully operational production solutions.

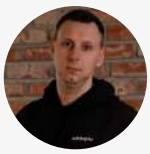


About Addepto

Addepto is an AI consulting company that delivers cutting-edge data-driven solutions supporting Digital Transformation.

We are not a newbie in AI and know how to separate the overhyped buzzwords from the market-proven solutions that lead to cost savings, performance improvements, and competitive advantages.

Book a meeting with our team and get the most out of AI:



Edwin Lisowski
CGO & Co-founder
edwin.lisowski@addepto.com

Kate Czupik
International Partnership Manager
kate.czupik@addepto.com