

Maximizing Efficiency with AI Tools as Catalysts



Richard Tweed

Senior Site Reliability
Engineer

Thought Machine

[https://infosec.exchange/
@RichardoC](https://infosec.exchange/@RichardoC)

<https://tales.fromprod.com>

DALL-E and Stable Diffusion were harmed for these image but no
AIs were harmed in the production of this content

Copyright © 2024 Richard Finlay Tweed. All rights reserved.
All views expressed are my own.

Key takeaways

- Don't overcommit - things are improving rapidly
- Build out a small suite of working examples
- Empower teams to self-service
- Ensure it's easy for teams to do the right thing
- There are limitations

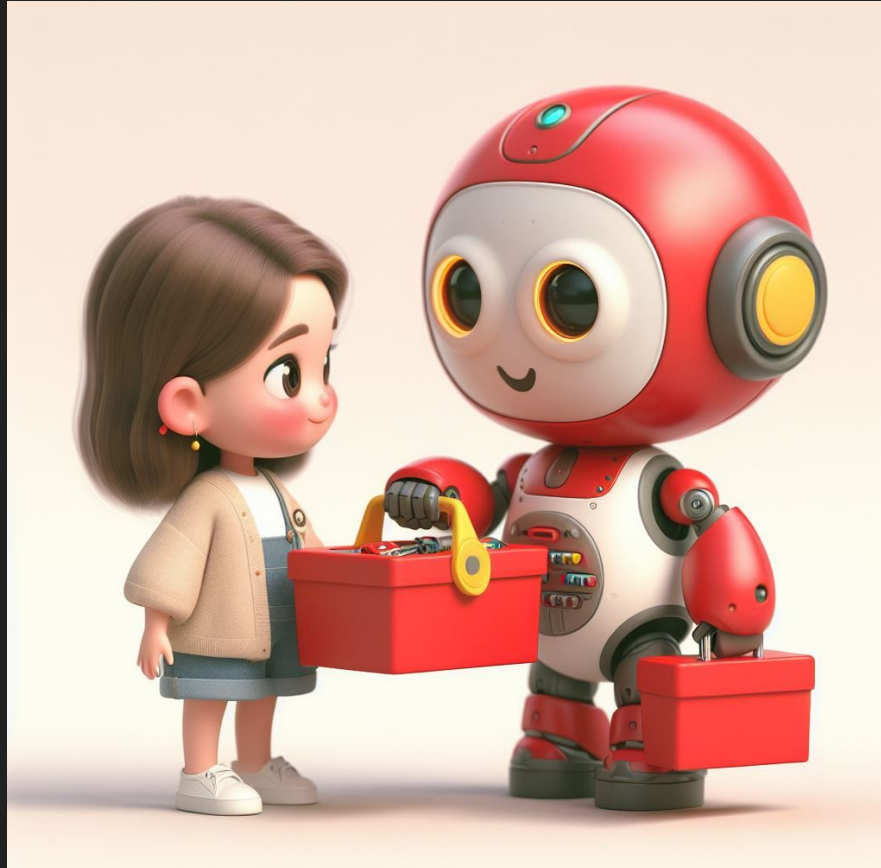
Don't overcommit - things are improving rapidly



Build out a small suite of working examples



Empower teams to self-service



Example architectures




Local processing

Preparations for conf42 ⌵

User 22:14


Generate some text which makes it look very impressive when shown in small font in a presentation




 Jan 22:16 Token Speed: 42.08t/s

Here's some text that might impress when shown in small font in a presentation:

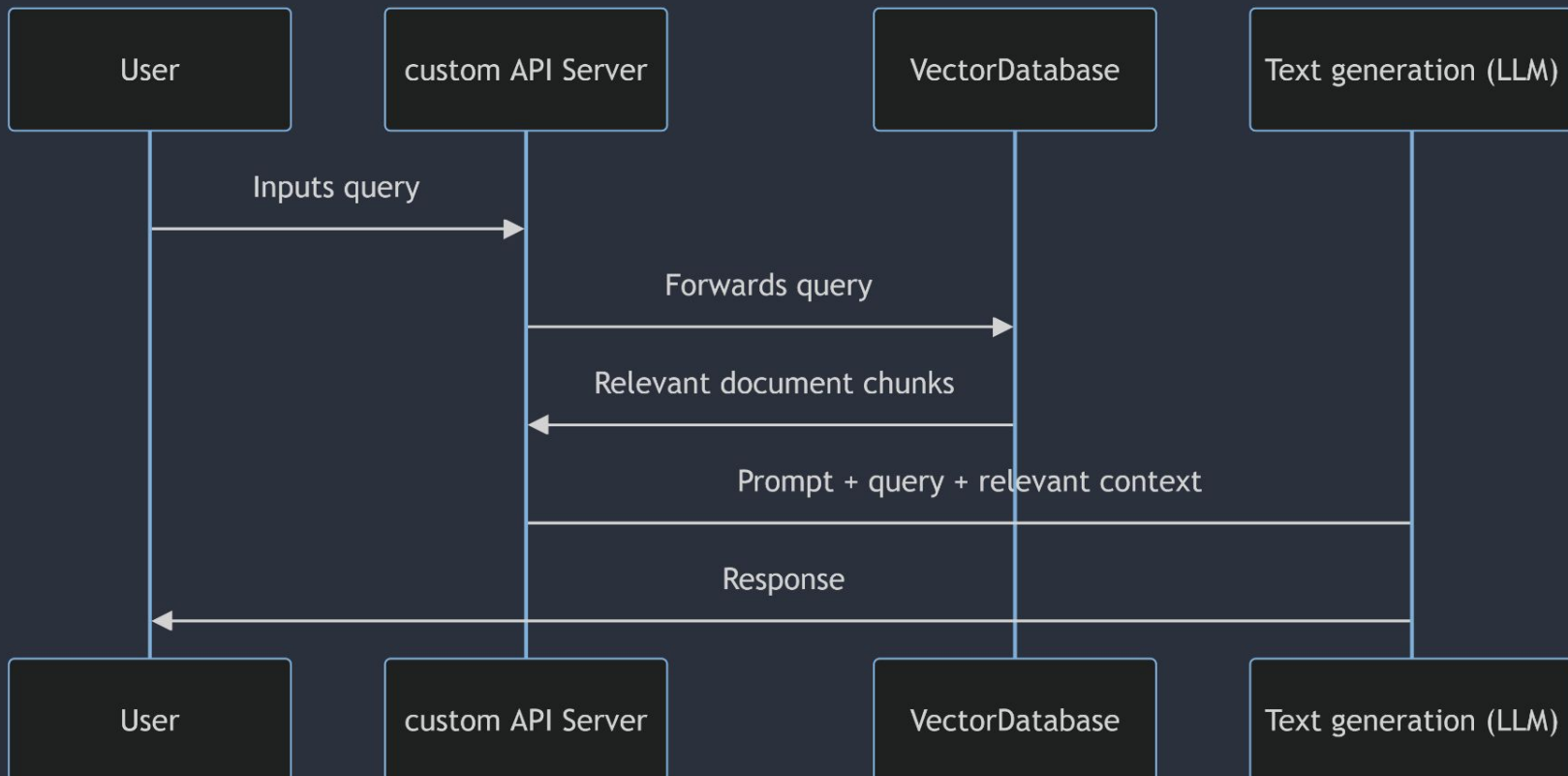
- "Innovative Solutions for the Modern Era**
- Unlocking the Full Potential of Data Analytics**
- Revolutionizing Business Operations with AI-Driven Insights**
- Transforming the Future of Industry with Cutting-Edge Technology**
- Empowering Decision-Makers with Actionable Intelligence**
- Fostering Collaboration and Innovation through Data-Driven Storytelling**
- Elevating the Standard of Excellence through Continuous Improvement**
- Pioneering the Next Generation of Business Intelligence**
- Unleashing the Power of Data to Drive Results**
- Building a Brighter Tomorrow through Data-Driven Excellence**
- Transforming the Landscape of Business with Data-Driven Innovation**

This text uses a mix of buzzwords and action verbs to create a sense of excitement and innovation. The short

Enter your message...  Send

 System Monitor Jan v0.4.13  

Retrieval Augmented Generation (RAG)



Limitations

- LLMs can make stuff up (hallucination)
- Outdated “knowledge”
- Better for templates or skeletons rather than fine detail
- Can be expensive
- Fast moving

Key takeaways

- Don't overcommit - things are improving rapidly
- Build out a small suite of working examples
- Empower teams to self-service
- Ensure it's easy for teams to do the right thing
- There are limitations
- Example architectures

Useful resources

- <https://jan.ai/>
- <https://ollama.com/>
- <https://github.com/langchain-ai/langchain>
- <https://docs.databricks.com/en/generative-ai/retrieval-augmented-generation.html>