Harnessing RAG Technology

A Practical Guide to Smarter Information Access

~~Stefan Damm~~

Jakob Reiter

Jakob Reiter

Co-Founder Mimirio

Serial AI Entrepreneur

VentureBeat: Top 100 people worldwide to watch in the AI space

Former CEO / CTO TheVentury

Former CEO BotBase

Stefan Damm (sick)

Co-Founder Mimirio

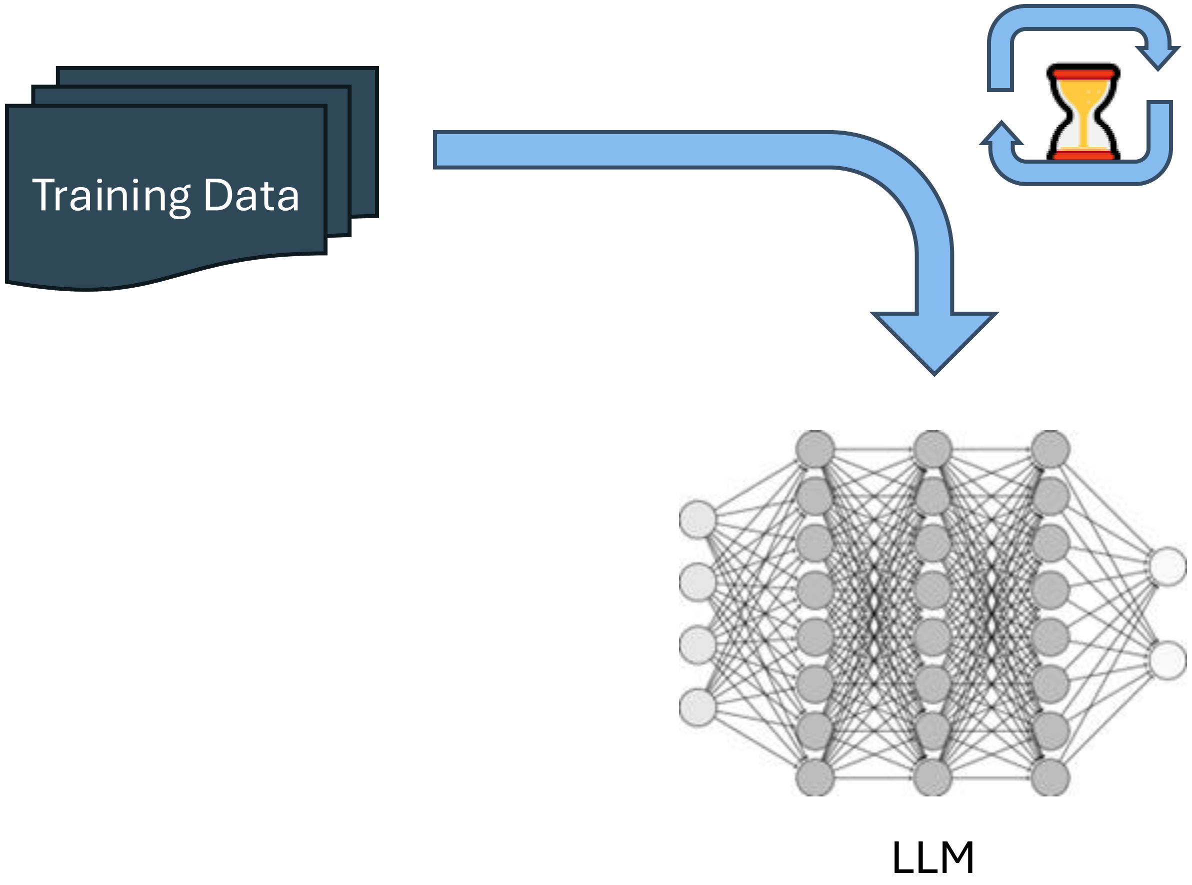
Former CTO Runtastic / Adidas

CTO Coach

Large Language Models

A quick and naïve Introduction

Large Language Models - Training



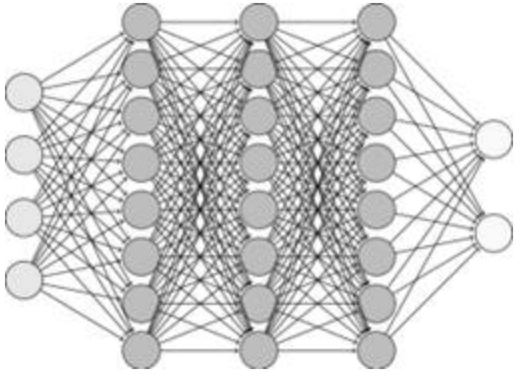
Training Data

LLM

Large Language Models - Inference

[System Message]
You are a helpful chat bot.
You answer questions.

[User Question]



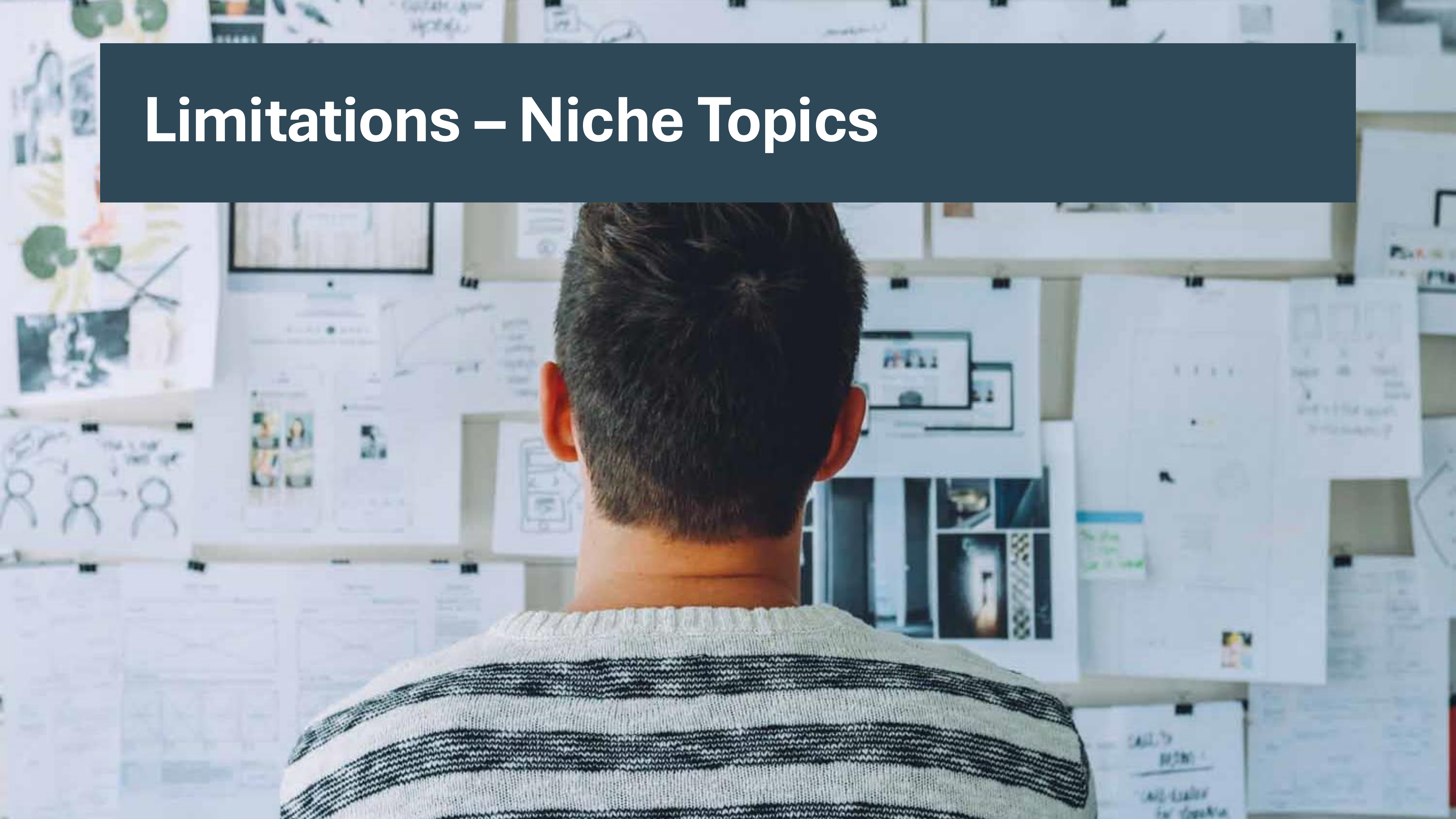
LLM

Output

Limitations – Cut off

Limitations – Hallucinations

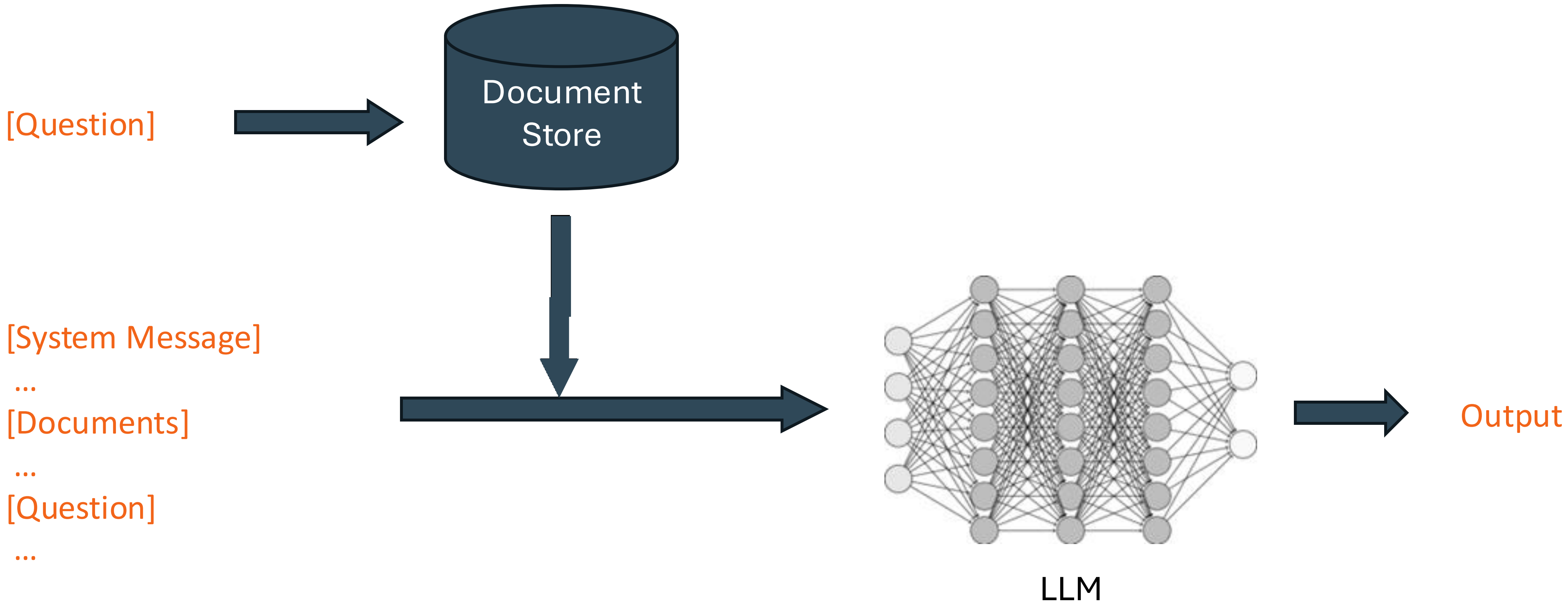
Limitations – Niche Topics



It’s RAGtime

Retrieval Augemented Generation as solution

RAG + LLM



[Question]

Document Store

[System Message]
…
[Documents]
…
[Question]
…

LLM

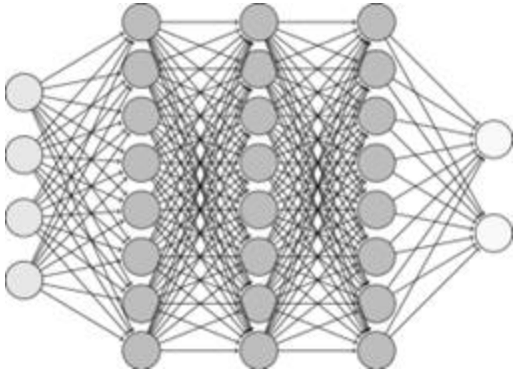
Output

RAG + LLM

[System Message]
Given these documents,
answer the question.

[Documents]
Document 1 content
Document 2 content
Document 3 content

Question: [Question]



LLM

Output

Why Retrival Augemented Generation?

Adding “live context” to LLMs

- Augments LLMs with external data sources
- Access to specific documents during inference
- Improves accuracy & relevance
- ‘Grounding”: Less Hallucination, Citation and Attribution

SHOW & TELL

Demo Time

Current Applications of RAG

Chat Agents

AI Productivity Tools

perplexity

AI features of Software / SaaS solutions

Part of most purpose-built AI agents

Data Privacy & Compliance

“**If you are logging in with your consumer google account and choose to provide feedback**, human reviewers may review your queries, uploads, and the model's responses to troubleshoot, address abuse or make improvements.”

Google (NotebookLM)

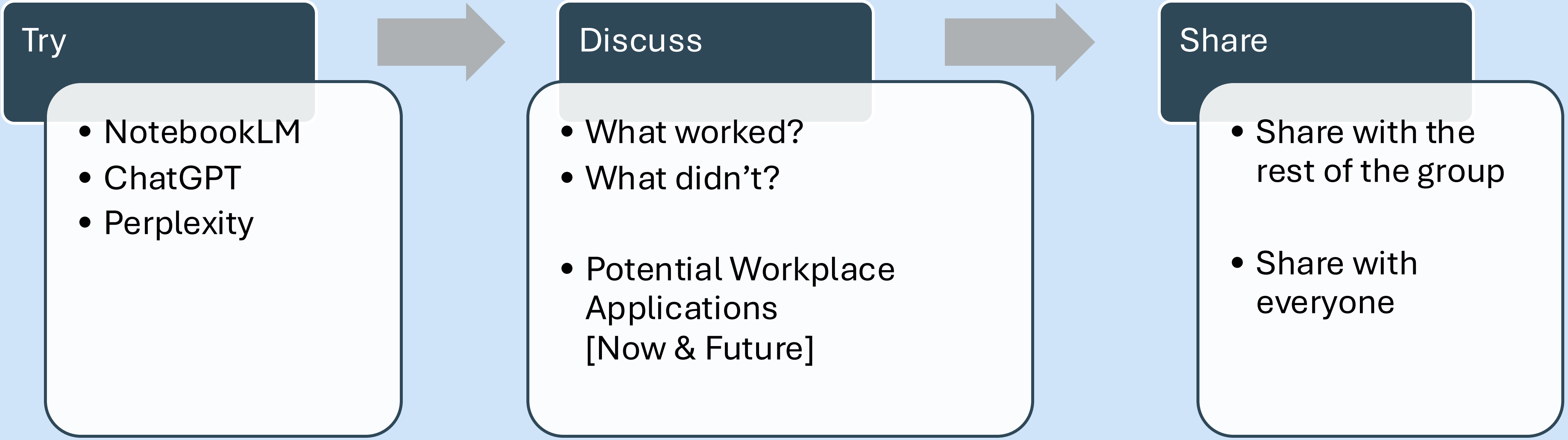
“When you use our services for individuals such as ChatGPT or DALL•E, we may use **your content to train our models**.”

OpenAI (ChatGPT)

“we may use any of the above information[User Input] to provide you with and improve the Services (including our AI models) and … monitoring and analyzing trends, conducting internal research and development, …”

Perplexity

Group Assignment



Try

- NotebookLM
- ChatGPT
- Perplexity

Discuss

- What worked?
- What didn’t?
- Potential Workplace Applications
  [Now & Future]

Share

- Share with the rest of the group
- Share with everyone

https://mimirio.com/testdata.zip

Looking Ahead …

Consider what you can do now and envision future possibilities

Future of RAG in business

- **Scaling**: just left research state (TRL 6) - big scaling challenges!
- **ARAG**: dynamically changing strategy based on context and need
- **Short term memory**
- **Feedback loop**: Improve retrieval based on user feedback
- **Cross-Modal Capabilities**:
  multiple data modalities (text, image, video, etc.) for even more nuanced responses
- **Ethical and Bias Considerations**
  fairness, transparency, and accountability

Mimirio

Trusted AI for smarter Businesses

- Autonomous RAG agent – no more upload, universal context
- Private cloud / on-premise - fully GDPR compliant, 100% loyal
- Business autoamtion – amplify your answers by taking actions

https://www.mimirio.com/tedai-slides/

jakob@mimirio.com