

Unlocking RAG's Potential

MCP and Multi-Agent Reinforcement Learning in Action

David vonThenen

Senior AI/ML Engineer

     [@davidvonthenen](https://twitter.com/davidvonthenen)



David vonThenen

- Are you Human or an AI?
- I want 5 Kubernetes
- Virtual Machines are Real
- Cloudy, cloudy, cloudy...
- There is storage for that!

     [@davidvonthenen](https://twitter.com/davidvonthenen)



Agenda

- **Let's Review Agent2Agent and MCP**
- **The Untold Stories of MCP**
 - **Live: Demo, Demo**
- **Using Agent2Agent Effectively**
 - **Live: Demo, Demo**
- **Resources**
- **Q&A**

Let's Review Agent2Agent and MCP

Model Context Protocol

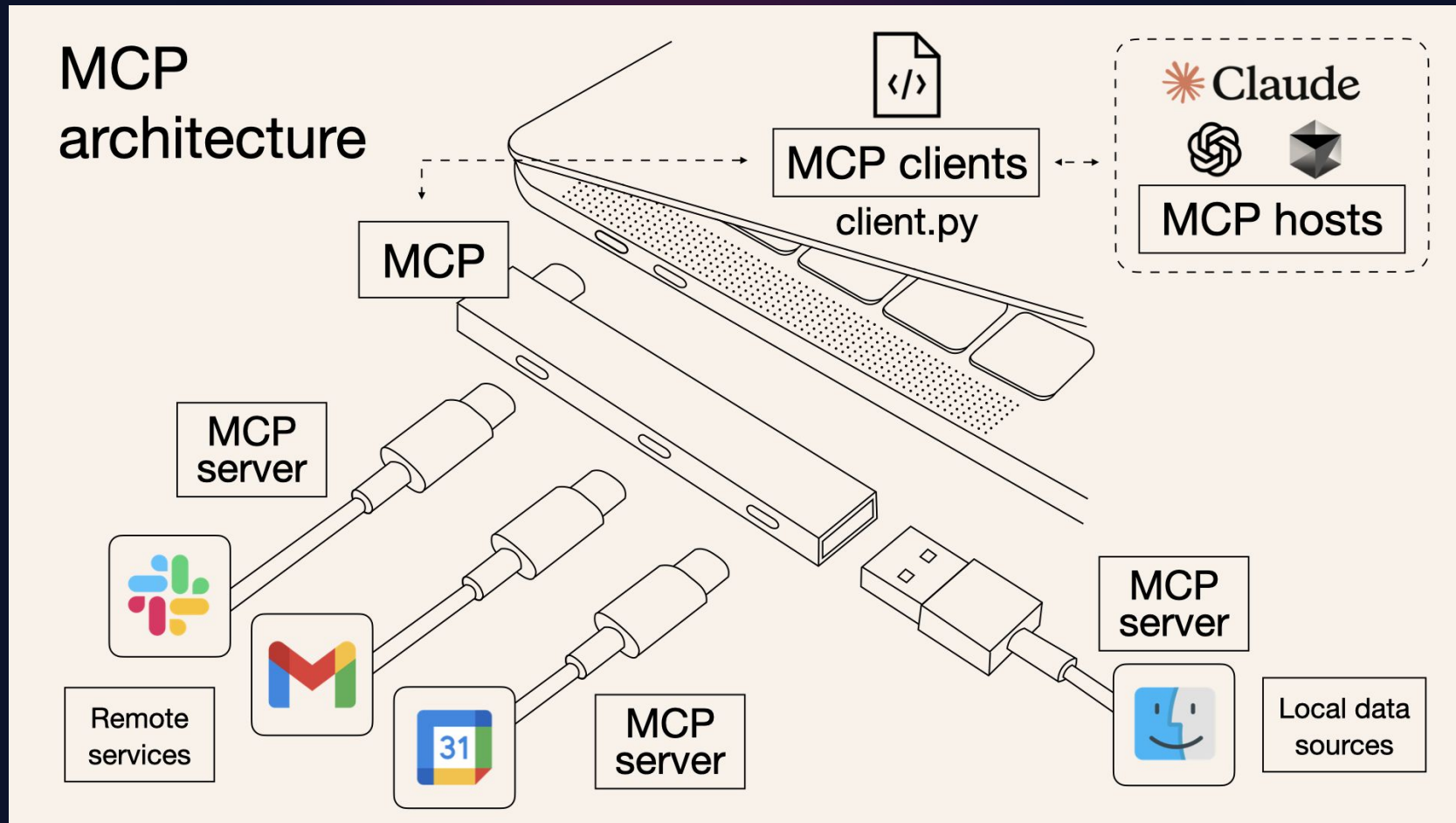
- MCP Standardizes How To Provide Content to LLMs
 - How Tools Are Connected
 - Accessing Data Sources
 - Calling APIs, Services, etc
- Access via MCP Client
- Implement an MCP Server
 - Advertise Capabilities
 - Handshake for Communication
- UNIFIED PROTOCOL!



Figurative Architecture

Image Credit:

[Norah Klintberg Sakal](#) - [LinkedIn Post](#)



What Is Agent2Agent Protocol?

- Protocol for Multi-Agent Interoperability
 - Communication via SSE, HTTPS, JSON-RPC
 - Agent Can Delegate Sub-Tasks
- github.com/a2aproject/A2A
 - [Donated to Linux Foundation](#)
- Local or Remote/Cloud Agents
- Decoupled vs Single Environment
 - Multiple Agent Services
 - Agents Talking to Agents



Combining A2A Protocol & MCP

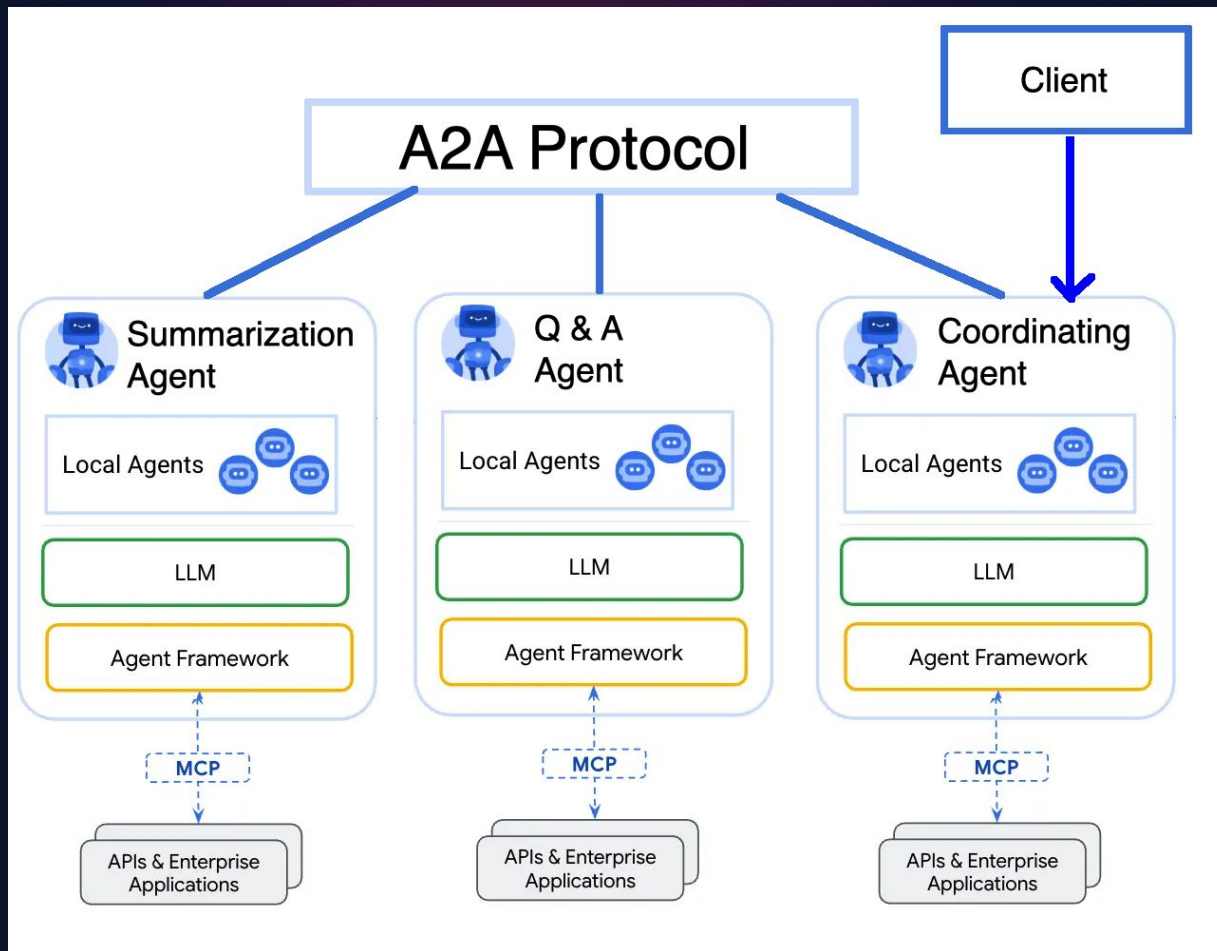


Image Attribution
Xin Cheng - [Hello World to Agent2Agent Protocol](#)


Building Blocks Of Agentic AI

- AI Agents (Typically, an LLM):
 - Singular Task, Access To Tools, etc
- Agentic AI (Workflows):
 - Make Autonomous Decisions
 - "Non-Deterministic" Path to Goals
- Plan, Iterate, and Coordinate
 - Minimal Human Intervention
 - Control Loops:
 - Analyze, Adjust, Done?
 - They Have Memory
 - Call APIs, Search the Web, etc



The Untold Stories of MCP

Swiss Cheese of Security Holes



Follow

Ad

Thousands of MCP servers are already live, but most security teams don't have a clear strategy yet. Get this guide and learn:

- Key risks with local and remote MCP servers
- Real-world threats like prompt injection and supply chain compromise
- Steps for safely using MCP tools

MCP Shortfalls

1. Limited Observability / Auditability

- Black-box Interactions
- No Transaction Logs

2. Security And Privacy

- Prompt Injection
- Token Theft and Identity Spoofing
- Privacy Leakage And Data Aggregation

3. Performance / Cost Inefficiencies

- Context Window Bloat
- High Latency For Large Datasets
- Token/Infrastructure Costs



MCP Shortfalls (Continued)

4. Architectural / Design Limitations

- Stateful Connection & Scalability Challenges (SSE vs REST)
- Lack Of State & Memory In Single Requests

5. Developer And Organizational Barriers

- Technical Complexity & Immature Tooling
- Authorization & Identity Management Gaps
- Rapidly Evolving Spec

Want To See A Big List Of Shortfalls?

<https://bit.ly/47tCL9e>



Image Attribution

https://www.reddit.com/r/windows/comments/weebhf/feel_old_yet_i_recreated_the_task_failed/

Rushing To "WRITE" Operations

- With All Of These Shortcomings...
 - "Non-Production" MCP Servers Popping Up Everywhere
 - Accessing "Destructive" Functionality
 - Replit's LLM-based "Vibe Coding" Agent
 - Leaking Sensitive Data

Example Failures:

- Anthropic MCP And Oat++
- mcp-remote Code Execution
- GitHub MCP Prompt Injection



Vector Embedding Limitations

- Amazing! Semantic Search Over Unstructured Text
- But... Semantic Similarity Limitations
 - All Knowledge is Flat
 - Difficult to Reason On Multiple Hops
 - No Holistic View
 - No Data Continuity
 - Miss Complex Entity Connections
- "Game Of Chance" On Purpose



Image Attribution
Creator: Alex Livesey | Credit: Getty Images
Copyright: 2014 Getty Images

Enhanced Data Search Using MCP

- MCP Is Incredibly Powerful For READ Operations
 - Access New Sources Of Data
 - Opportunity For ACL On Data
 - Trained Data Is General Data
 - Partition Data for Access
 - Opportunity For RBAC For Tools
 - APIs Behind Permissions
 - Fine-Grained Controls
- We Should Be Thinking About...
 - Data Access Tiers + API Access

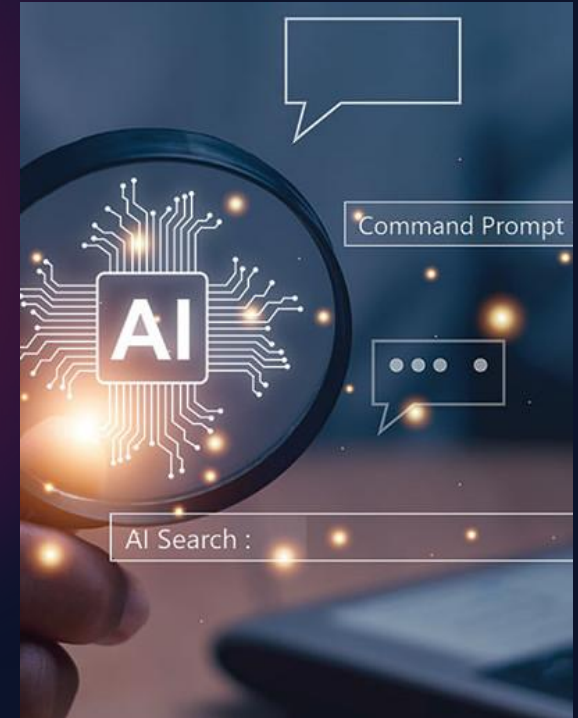


Image Attribution

<https://www.technewsworld.com/story/ai-enhanced-searches-may-pose-threat-to-creators-publishers-179193.html>

Reinforcement Learning

- Using MCP For New Sources Data
 - Reinforcement Learning Opp.
 - Store This Data / Expand Corpus
 - Tier 1: Trusted Source of Data
 - Tier 2: Everything Else
 - Validated The Data:
 - Algorithmic / Automated
 - Human-in-the-Loop
- As AI Generated Data Proliferates
 - Must Expand "Real" / Human Data



Image Attribution

<https://www.akc.org/expert-advice/training/operant-conditioning-positive-reinforcement-dog-training/>

Demo: Semantic Search Fail

<https://youtu.be/DR4zM2RF7A0>

Demo: MCP and RL

<https://youtu.be/xV9vaH9i25A>

Using A2A Effectively

All Eyes On Large Language Models

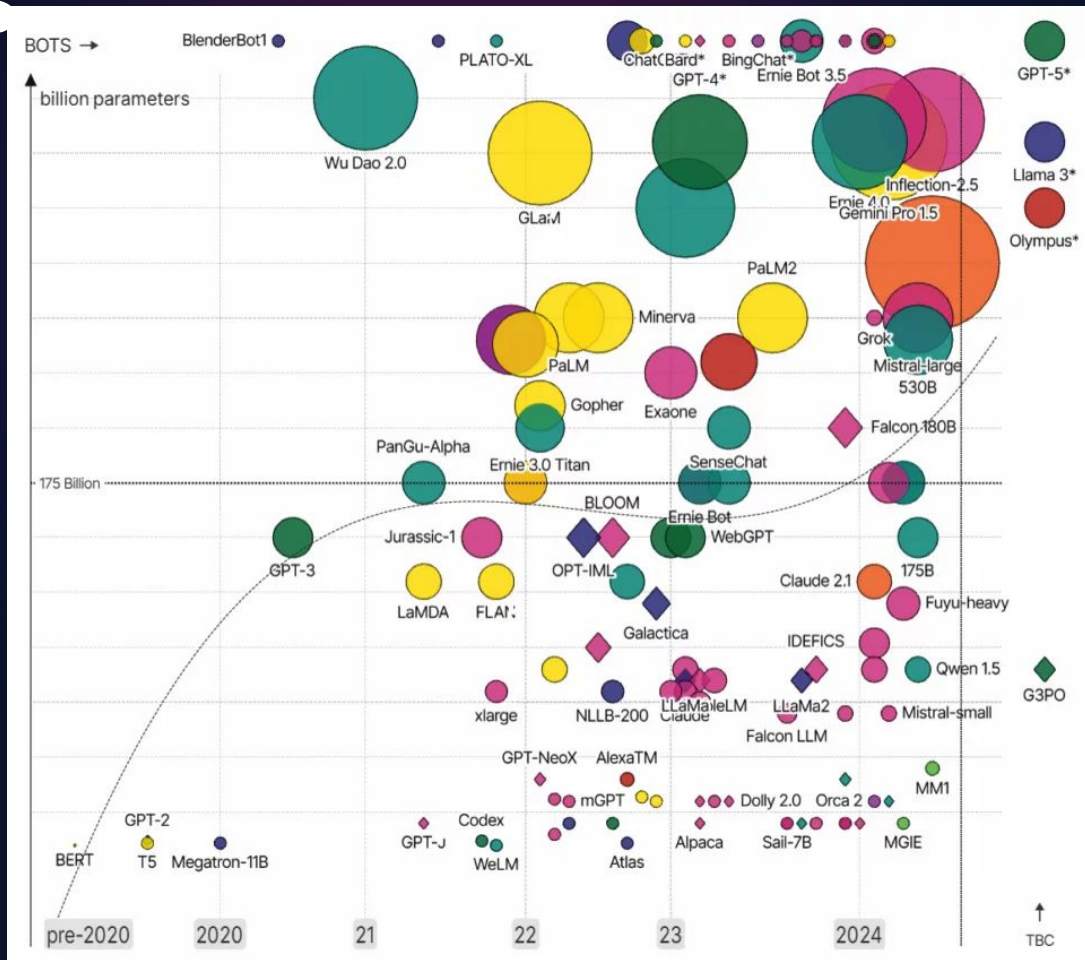


Image Attribution
How Large Language
Models are Shaping the
Future: Teaching
Machines to Understand
in 2024

<https://nestify.io/blog/top-large-language-models/>

Large Language Models Issues

- Hallucination Increase With Larger Dataset
 - Not Necessarily The Size Of The LLM
- Reward Mechanism To Provide An Answer
 - Confident Wrong Answers
 - No Reward For "I Don't Know"
- Again... All Data Is Flat
 - All Things Open AI 2025
Leveraging Knowledge
Graphs For RAG
 - No Association With OpenAI, Google, and Windsurf Articles



Consider Small Language Models

- Small Language Models (SLM) w/ RAG
 - Hallucination Rate On Par Or Better
 - Smaller Memory Footprint
 - Inference Faster / Cheaper
 - Smaller Corpus + Specialized Data
- Pick The LLM or SLM For The Task
- Easier To Scale Out SLMs
 - More Compute Targets
 - Less Dependence On \$\$\$ GPUs



Image Attribution
<https://www.dreamstime.com/illustration/david-vs-goliath.html>

A2A Decomposing Problems

- Use Agent2Agent Breakdown Problems
 - Smaller Subject Matter Expert (SME)
 - All The Benefits Of SLMs
 - Connect SMEs Via Policy & Logic
 - Bad: LLM w/ 20+ MCPs Endpoints
 - Billing Agent → Access Billing API
- Use Software Engineering Values
 - Code Reuse or Libraries → SMEs
 - Modular & Separation Of Concerns
 - Easier To Secure And Lockdown



Image Attribution

<https://www.amazon.com/YAKELUS-Russian-Nesting-Matryoshka-handmade1070/dp/B01LYU541Z>

Warnings: Policy & Business Logic



LangChain

VS



LangGraph

Blog:

[LangChain and LangGraph Agent Frameworks Reach v1.0 Milestones](#)

Session Recording:

[WeAreDevelopers World Congress 2025: Using Adversarial Techniques for Better AI and True Anonymity](#)



Combining A2A Protocol & MCP

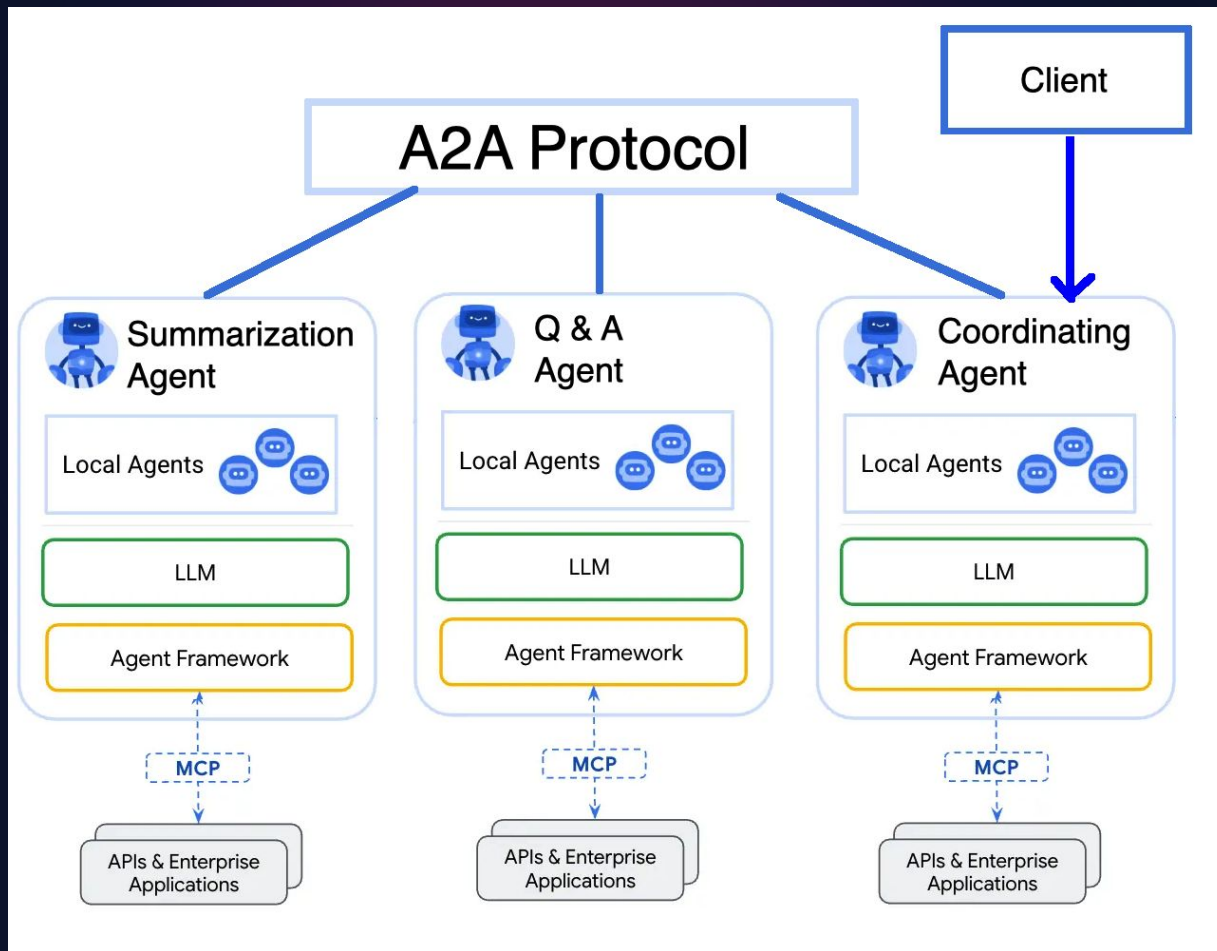


Image Attribution
Xin Cheng - [Hello World to Agent2Agent Protocol](#)

Demo: Simple A2A Example

<https://youtu.be/qiTGGDLXzG8>

Demo: Full A2A + MCP Example

<https://youtu.be/GUkTM8aXX4M>



Resources

Resources

All Materials (Slides, Code, etc):

github.com/davidvonthenen/2025-tbilisi-data-science-meetup

Want To Participate In A Podcast:

[Need AI/ML Help? Work A Problem Together](#)

Reduced Hallucinations, Better Answers:

Graph-based RAG

- github.com/davidvonthenen/graph-rag-guide

Document-based RAG

- github.com/davidvonthenen/document-rag-guide



Thank You!



David vonThenen
Senior AI/ML Engineer

     [@davidvonthenen](https://twitter.com/davidvonthenen)

