# RAG FUSION WITH COHERE & WEAVIATE

Csaba Toth
ThruThink Support Chat Agent

# CSABA TOTH

- Full stack engineer
  - Director of Product Engineering at SportsBoard (startup)
  - CTO at ThruThink (startup)
- Interested in AR/XR, AI/ML, lately Gen AI, LLM, RAG
- GDG (Google Developer Group) Fresno lead, WTM (Women Techmakers) Fresno ambassador, tech meetup junkie

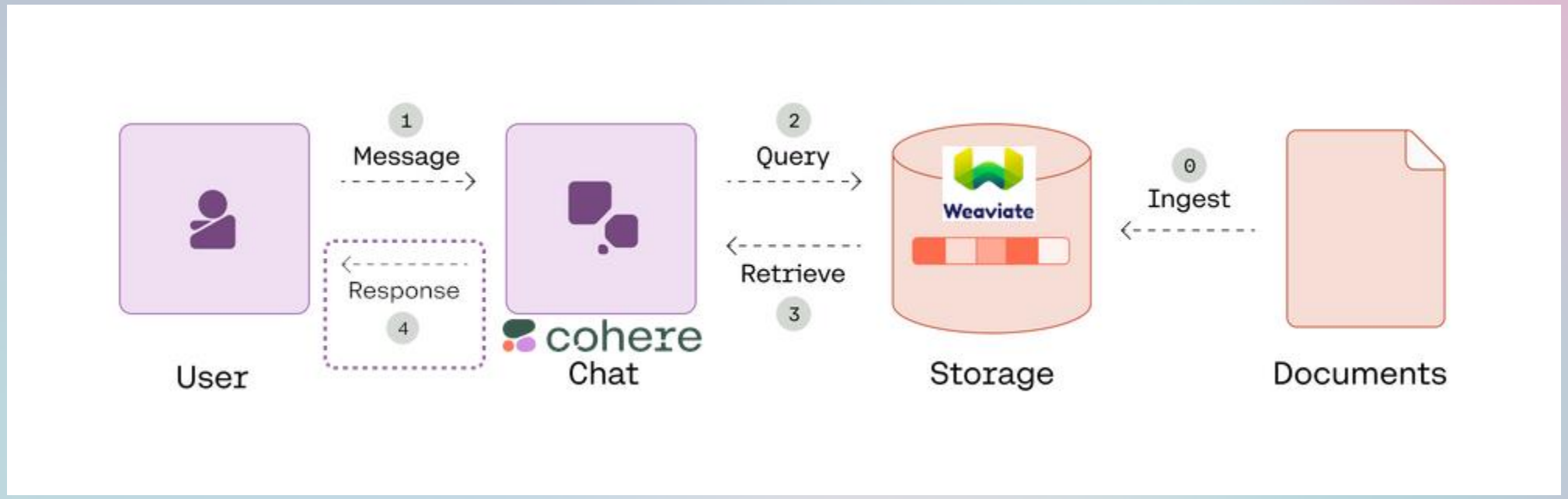
# THRUTHINK

- ThruThink® is a business budgeting on-line app to create professional budgets and forecasts
- product of literally decades of experience and careful thought, and thousands of calculations
- Thru-hiking, or through-hiking, is the act of hiking an established long-distance trail end-to-end continuously
- There are no dedicated personnel for support chat agent roles, had a “classic” chat agent integration

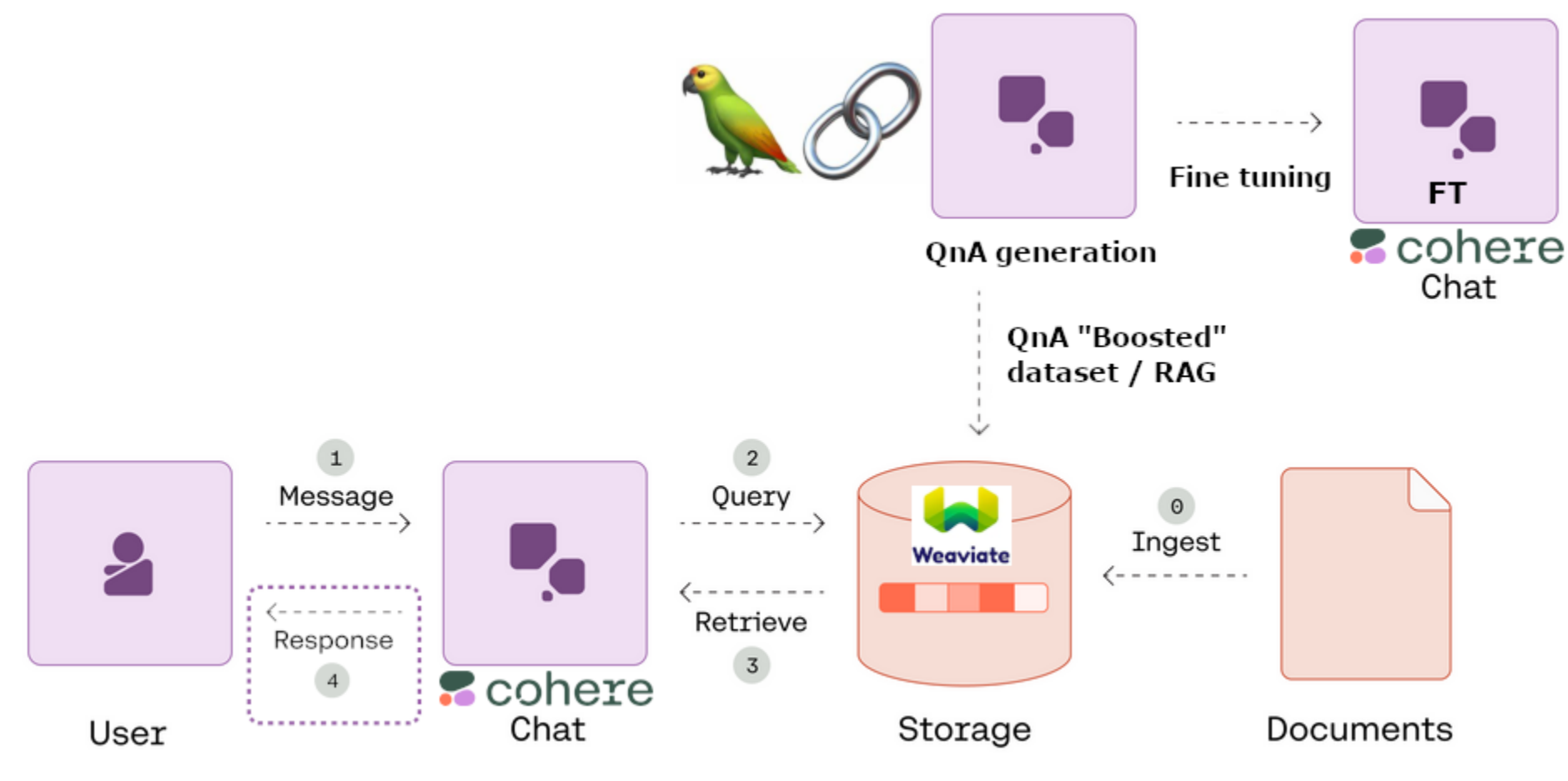
# SUPPORT CHAT AGENT

- Invaluable help, given that
  - It stays relatively grounded
  - Won’t hallucinate* wildly
- Desired abilities:
  - Main goal: answer ThruThink software specific questions: “In ThruThink can I make adjustments on the Cash Flow Control page?”
  - Nice to have: answer more generic questions such as: “How much inventory should I have?”

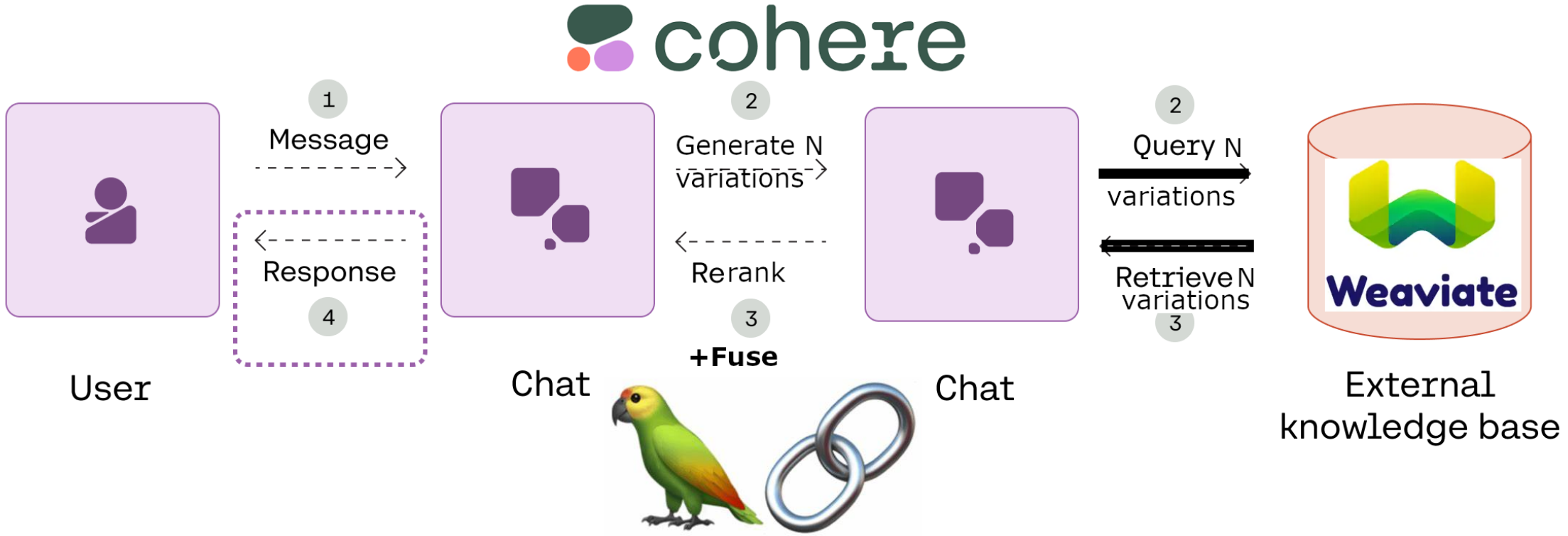
# TRADITIONAL RAG



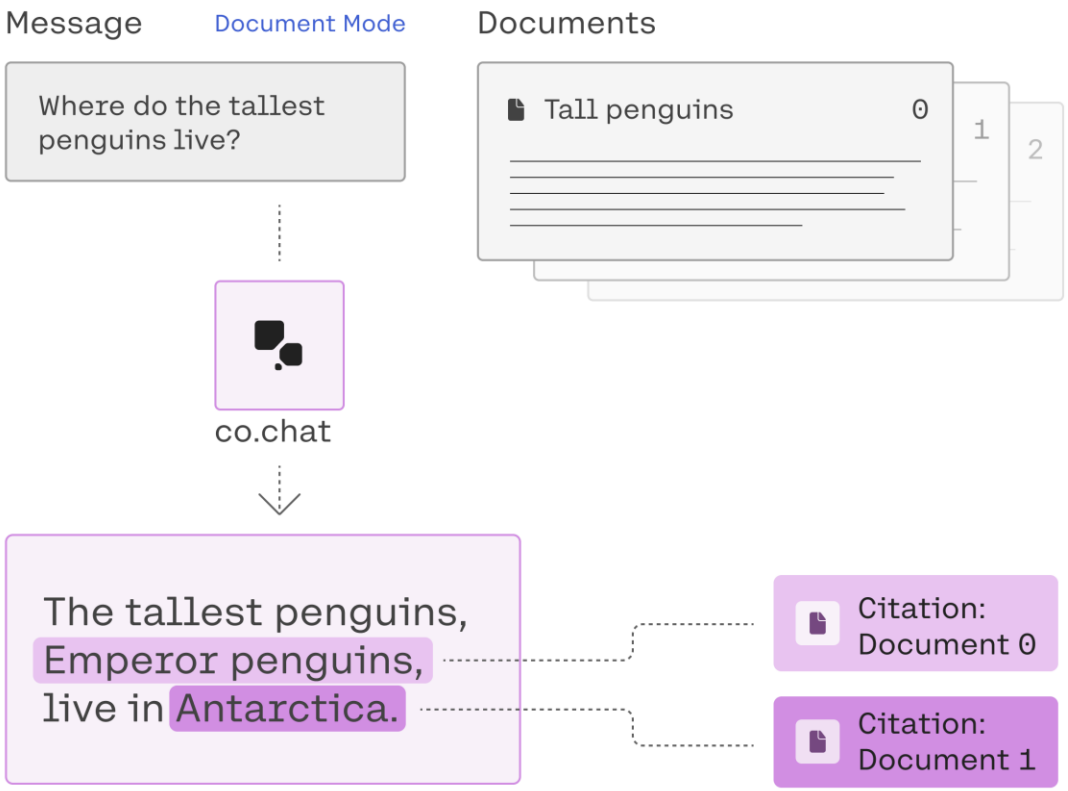
# QNA BOOSTED RAG



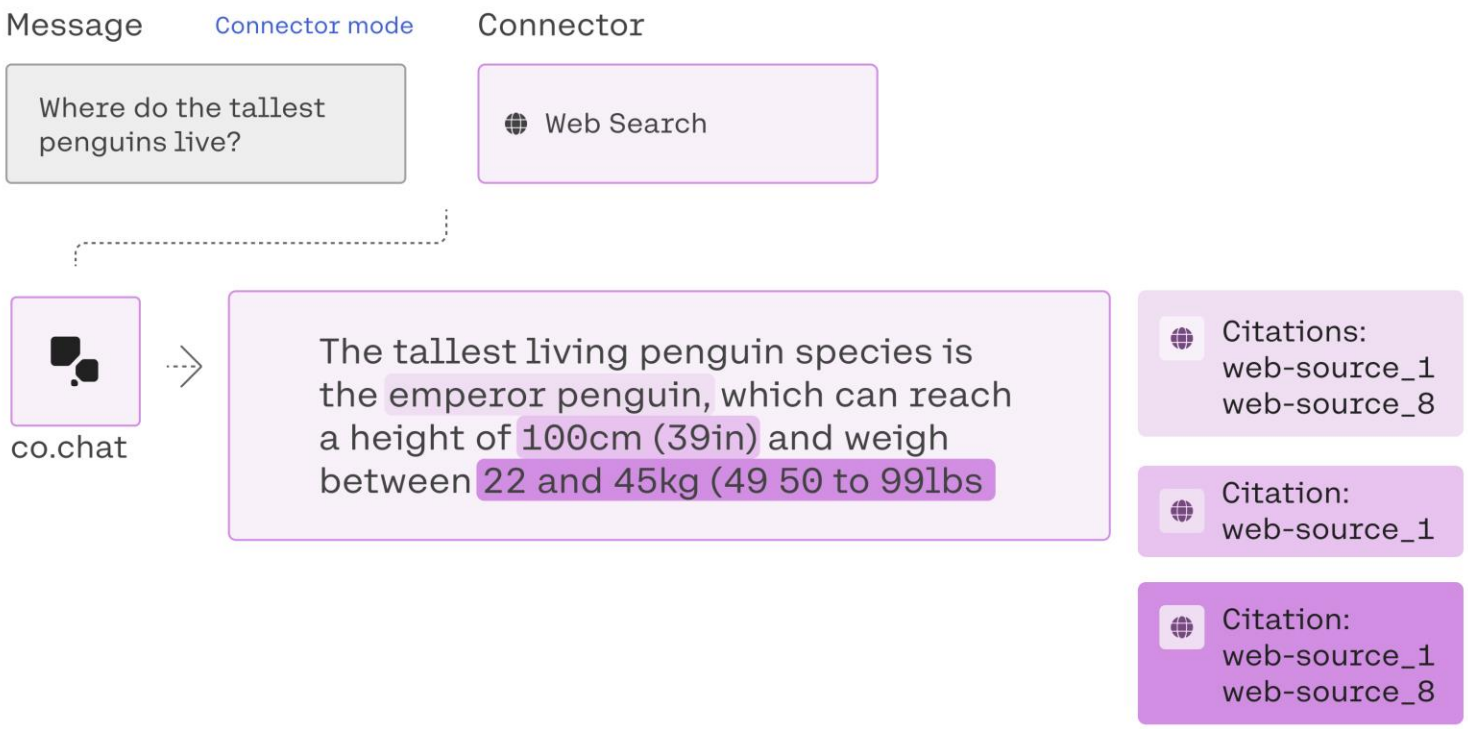
# RAG FUSION


cohere
1
Message
Response
4
User
2
Generate N variations
Rerank
3
+Fuse
Chat
Chat
2
Query N variations
Retrieve N variations
3
Weaviate
External knowledge base

# STEP 4 PT1: DOCUMENT MODE



# STEP 4 PT2: CONNECTOR MODE



# ACCOMPLISHMENTS

- Contributed to question_extractor for the QnA generation, and QnA reprocessing for ingestion
  - Check-pointing to resume generation
  - Cohere fine tuning format
  - Rate limiting

https://github.com/CsabaConsulting/question_extractor

- QBRAG with the synthetic / augmented data

# ACCOMPLISMENTS

## RAG FUSION

Using LangChain, Weaviate, Cohere

## STREAMLIT

Serving platform is streamlit, highly customized and advanced UI with references https://github.com/CsabaConsulting/ThruThinkCohereWeaviateChat

## NOTEBOOKS

https://github.com/CsabaConsulting/Cohere

# CO.CHAT

1. Document mode
2. Web connector mode

# FUTURE / TO-DO

## UX/UI

Decrease runtime by running the variation document retrievals in parallel, this is a Streamlit specific tech challenge with asyncio / await.

Decrease runtime by running the final two co.chat RAG calls in parallel, this is a Streamlit specific tech challenge with asyncio / await.

Make the citation linking nicer and other UI enhancements.

## PERFORMANCE / QUALITY

Measure how much the RAG Fusion improves answer quality.

Measure the trade-off factoring in extra latency and potential token usage increase which also means cost increase.

Integrate the agent into ThruThink which uses ASP.NET MVC / C# / Azure technology stack. I'll be able to open up referred help topics using the meta-data.

Add filter against harmful content, for example using Google PaLM2's safety attributes.

# THANK YOU

Csaba Toth

https://csaba.page/

https://www.linkedin.com/in/csabatothdev