Birthing Perjury-free AI

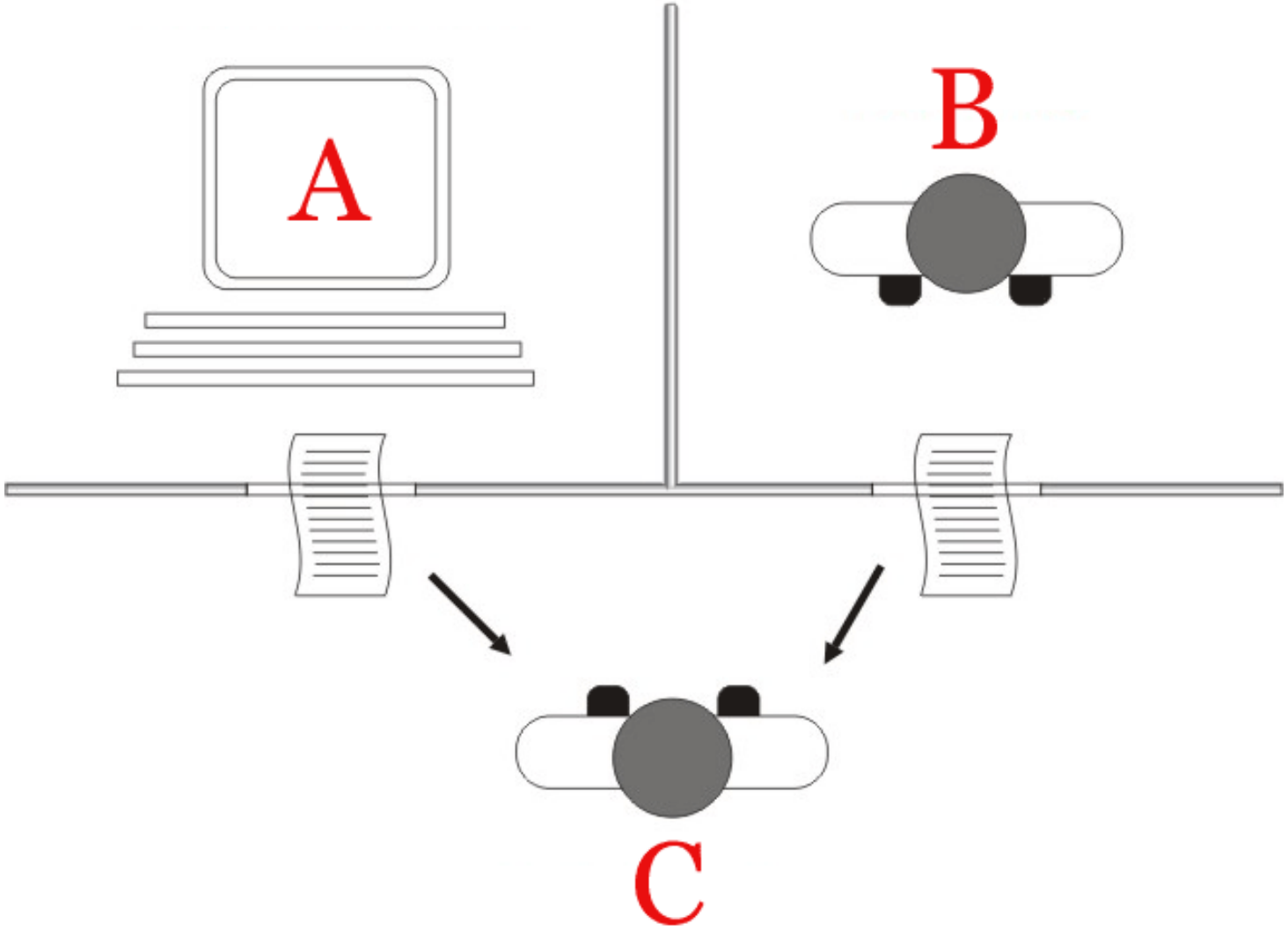
Charles D. Herring, WitFoo co-Founder & CTO
Charles@WitFoo.com
CharlesHerring.com
@charlesherring

About Charles

- WitFoo co-Founder and Project Lead (2016-)
- Cisco & Lancope Security Architect (2012-16)
- DoD Security & Data Consultant (2005-12)
- InfoWorld Test Center (2003-2008)
- US Navy Cyber Security (2002-2005)
- US Navy F/A 18 Hornet Avionics (1995-2002)
- Arkansas Drug Care Director of IT (1993-1995)

Turing Test (Imitation Game) Bias in GenAI

- “a machine's ability to exhibit intelligent behavior equivalent to, or ***indistinguishable*** from, that of a human”
- Believability > Accuracy
- Training data lacks “I don’t know” statements



GenAI Hallucination

- IBM: “AI hallucination is a phenomenon wherein a large language model (LLM)—often a generative AI chatbot or computer vision tool—perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are ***nonsensical or altogether inaccurate***.”

Perjury in Law Enforcement

- Digital records are ***evidence***
- Analysis must be verifiable, accurate and explainable
- Consequences for error are extreme

Training for Truth

- The datasets chosen must be truthful/accurate
- Removing error is hard/impossible
- Good data is expensive

Fine-tuning

- Testing of model’s accuracy
- Providing correction
- “Therapy” of the model

Safety Guardrails

- Guardrails are the “conscience” of the model
- Instructions post-training
- Protect private data

RAG – Using Reference

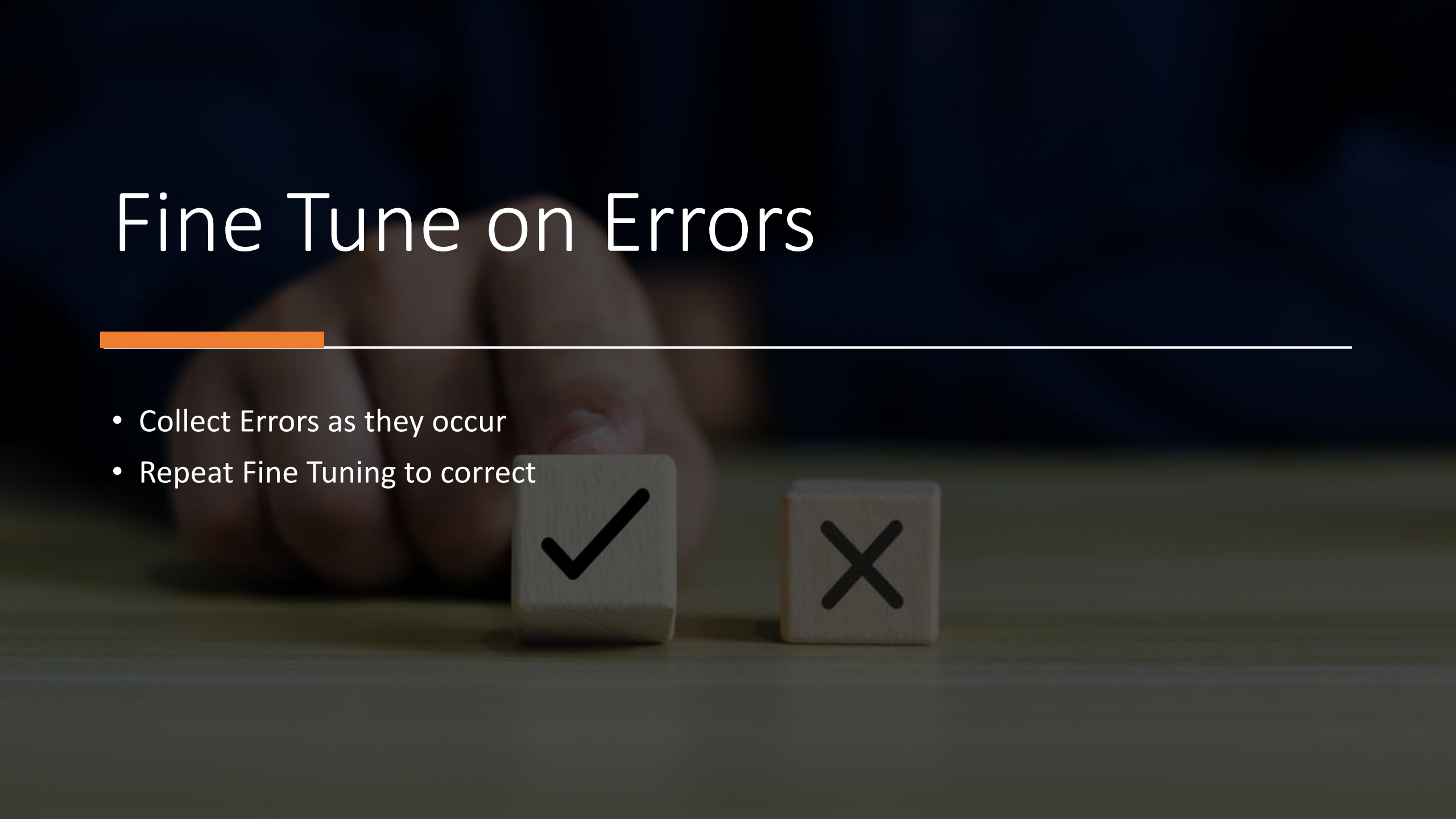
- RAG: Retrieval-Augmented Generation
- Use authoritative vector data
- “Show your work”

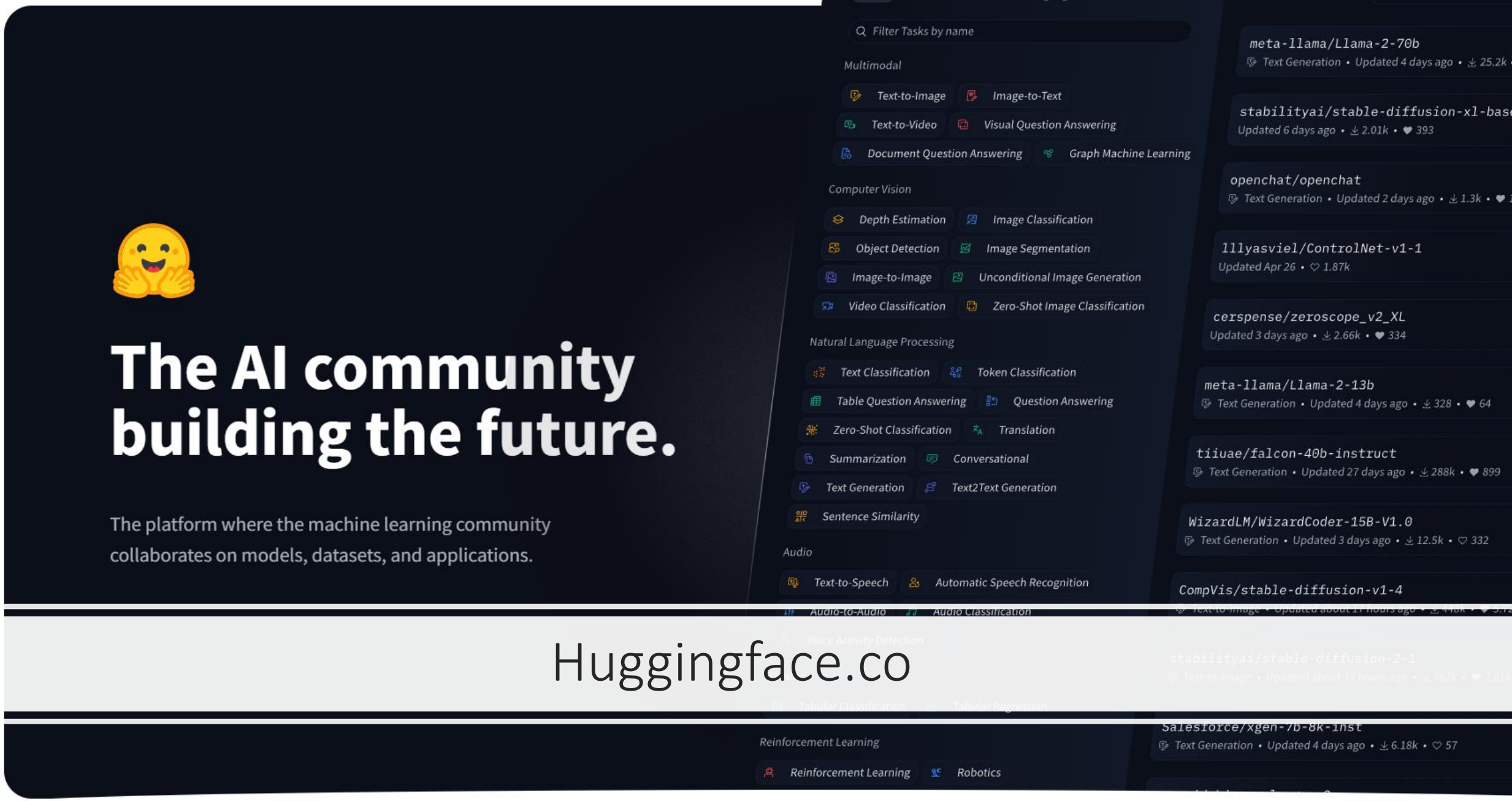
Sibling Checks

- Ask different models to verify the response
- “Is this statement accurate? {AI response}”

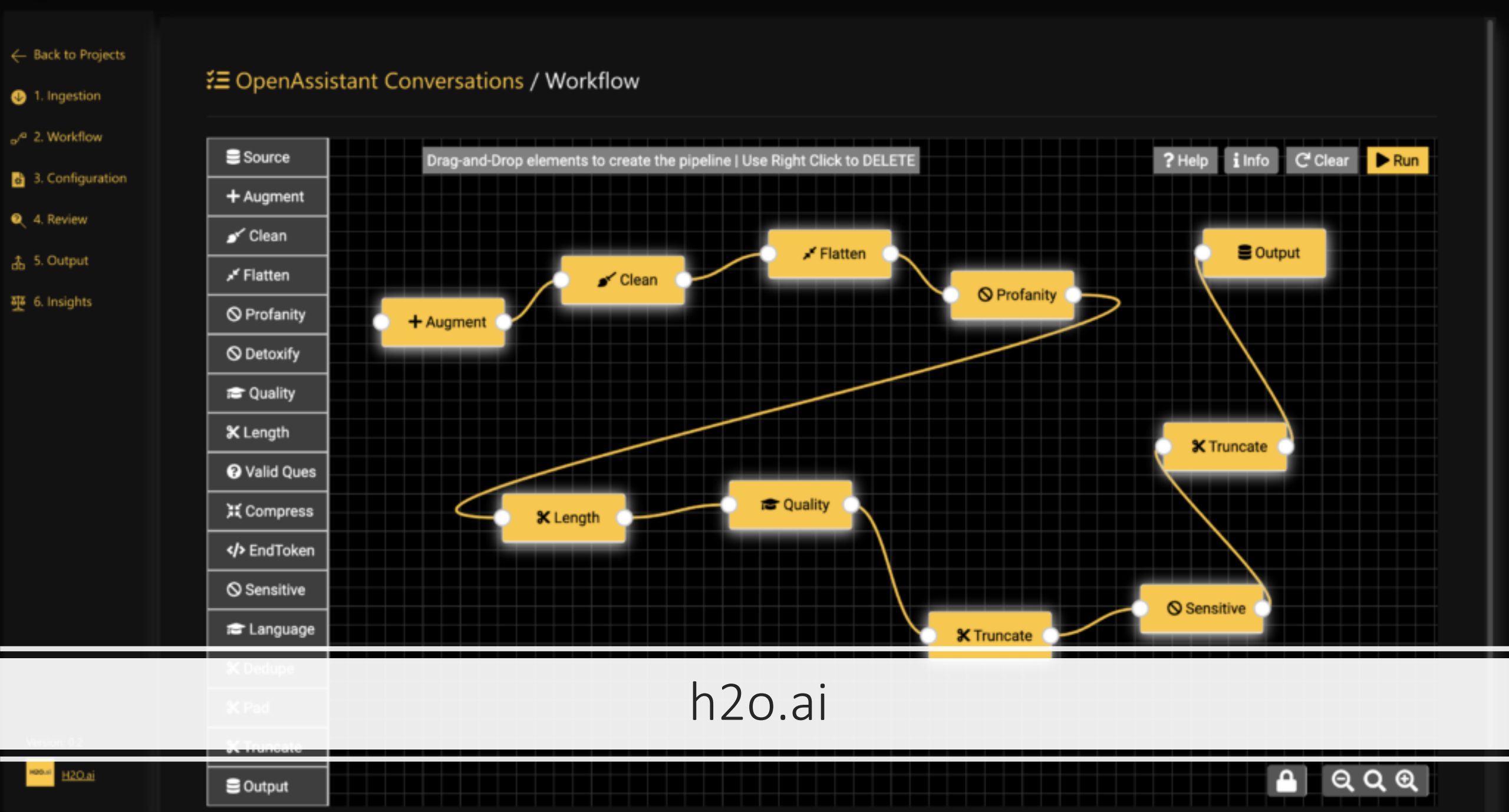
Fine Tune on Errors

- Collect Errors as they occur
- Repeat Fine Tuning to correct



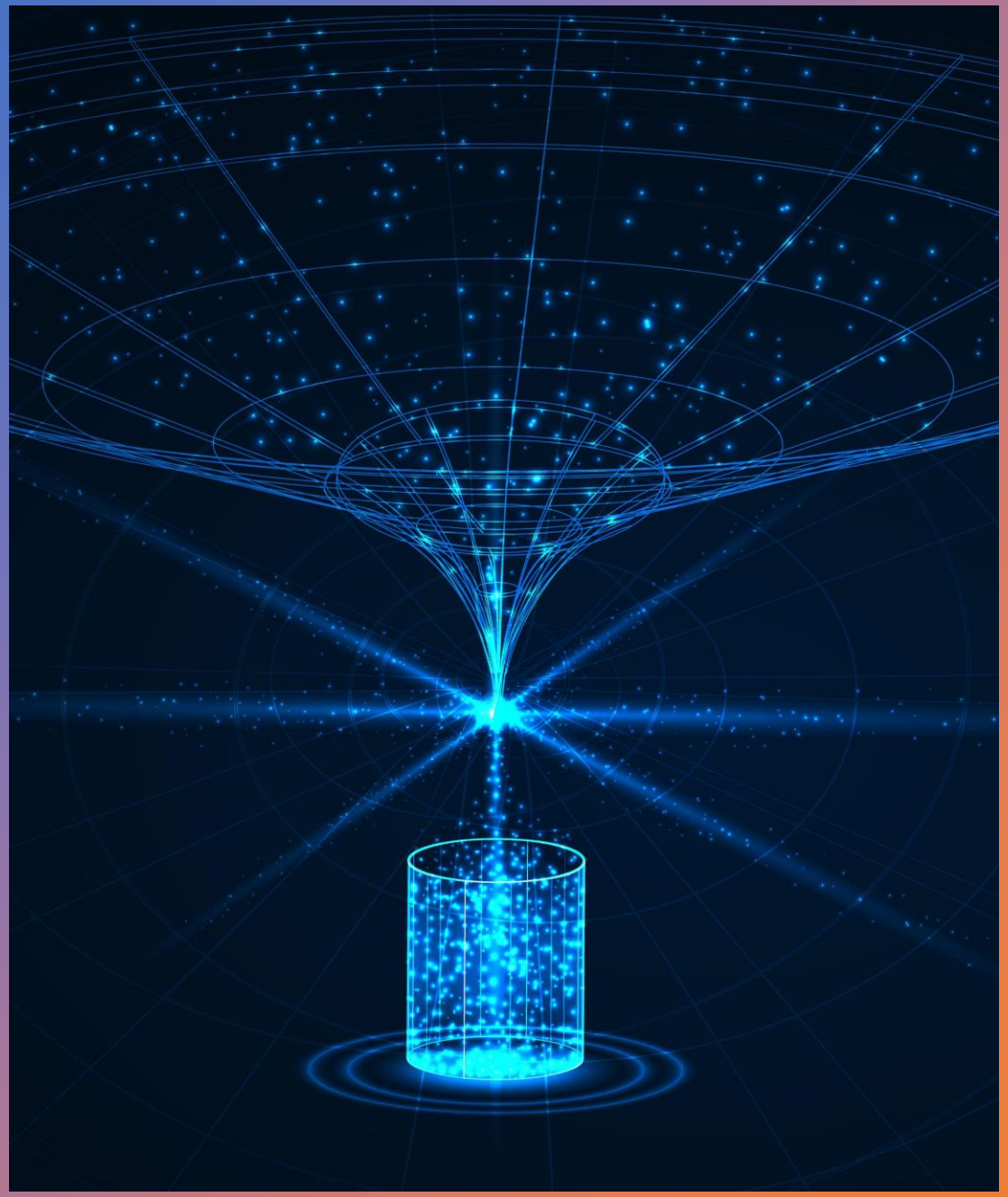


Huggingface.co


Back to Projects
1. Ingestion
2. Workflow
3. Configuration
4. Review
5. Output
6. Insights
OpenAssistant Conversations / Workflow
Drag-and-Drop elements to create the pipeline | Use Right Click to DELETE
Help
Info
Clear
Run
Source
Augment
Clean
Flatten
Profanity
Detoxify
Quality
Length
Valid Ques
Compress
EndToken
Sensitive
Language
Dedupe
Pad
Output
Augment
Clean
Flatten
Profanity
Output
Truncate
Length
Quality
Truncate
Sensitive
H2O.ai

h2o.ai

Big Data Vector NoSQL Database



Vector Data

- Storing Data by geo-spatial distance
- Referenced in RAG
- Langchain for extraction (Cassio)
- Also store original data (forensics)

GenAI Maturation Pipeline

Artificial Narrow Intelligence (ANI)

- Coded (not trained) for specific tasks
- Faster, Cheaper, Predictable
- Defendable in Court

Resources

- CharlesHerring.com
- www.witfoo.com/blog
- Huggingface.co
- H2o.ai
- Cassandra.Apache.org

Birthing Perjury-free AI

Charles D. Herring, WitFoo co-Founder & CTO
Charles@WitFoo.com
CharlesHerring.com
@charlesherring