

# Birthing Perjury-free AI

Charles D. Herring, WitFoo co-Founder & CTO

[Charles@WitFoo.com](mailto:Charles@WitFoo.com)

CharlesHerring.com

@charlesherring

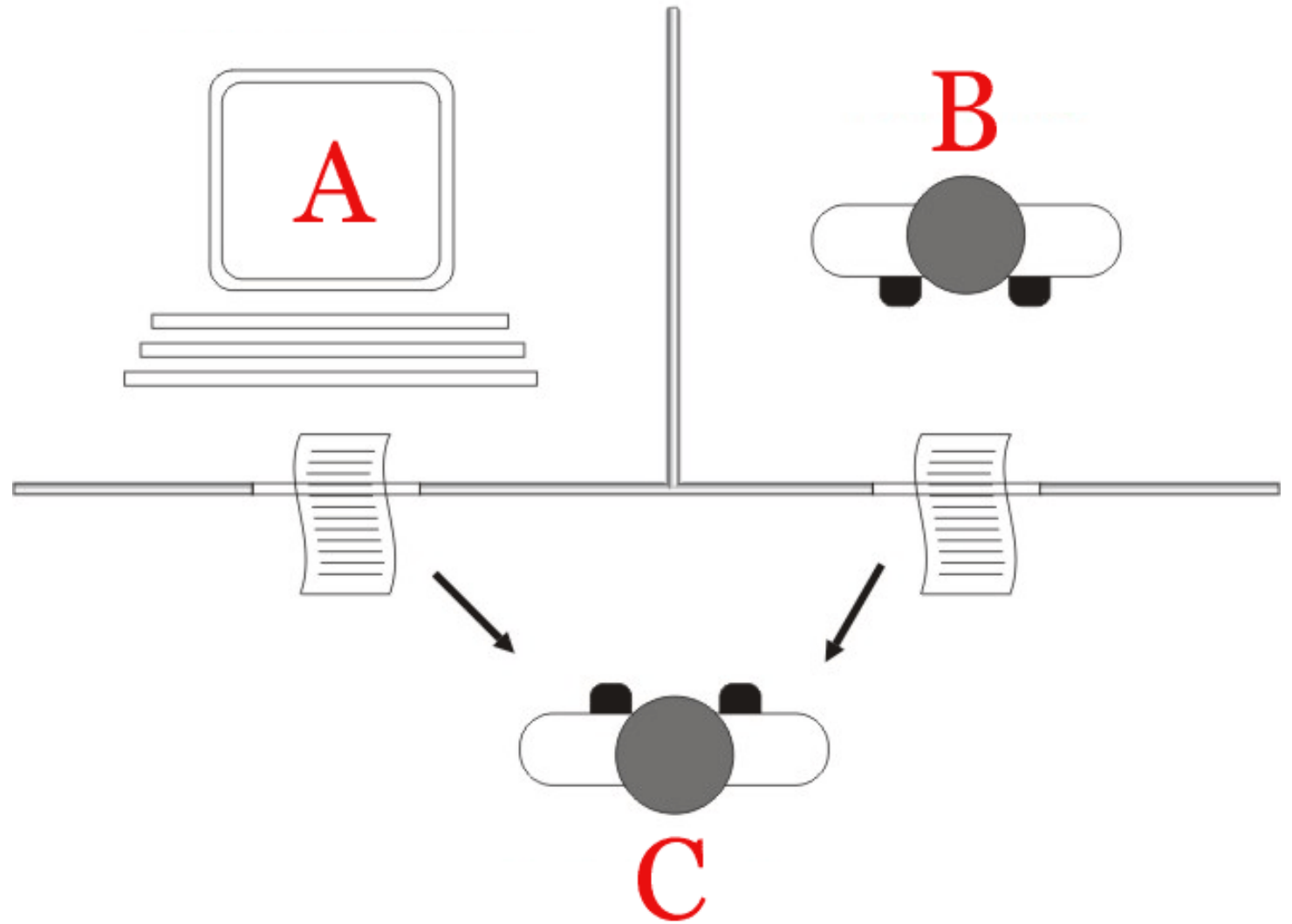


## About Charles

- 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





# The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Huggingface.co

The screenshot displays the Hugging Face website interface. At the top, there are navigation tabs for 'Tasks', 'Libraries', 'Datasets', 'Languages', 'Licenses', and 'Other'. Below these is a search bar labeled 'Filter Tasks by name'. The main content area is organized into several categories:

- Multimodal:** Text-to-Image, Image-to-Text, Text-to-Video, Visual Question Answering, Document Question Answering, Graph Machine Learning.
- Computer Vision:** Depth Estimation, Image Classification, Object Detection, Image Segmentation, Image-to-Image, Unconditional Image Generation, Video Classification, Zero-Shot Image Classification.
- Natural Language Processing:** Text Classification, Token Classification, Table Question Answering, Question Answering, Zero-Shot Classification, Translation, Summarization, Conversational, Text Generation, Text2Text Generation, Sentence Similarity.
- Audio:** Text-to-Speech, Automatic Speech Recognition.
- Reinforcement Learning:** Reinforcement Learning, Robotics.

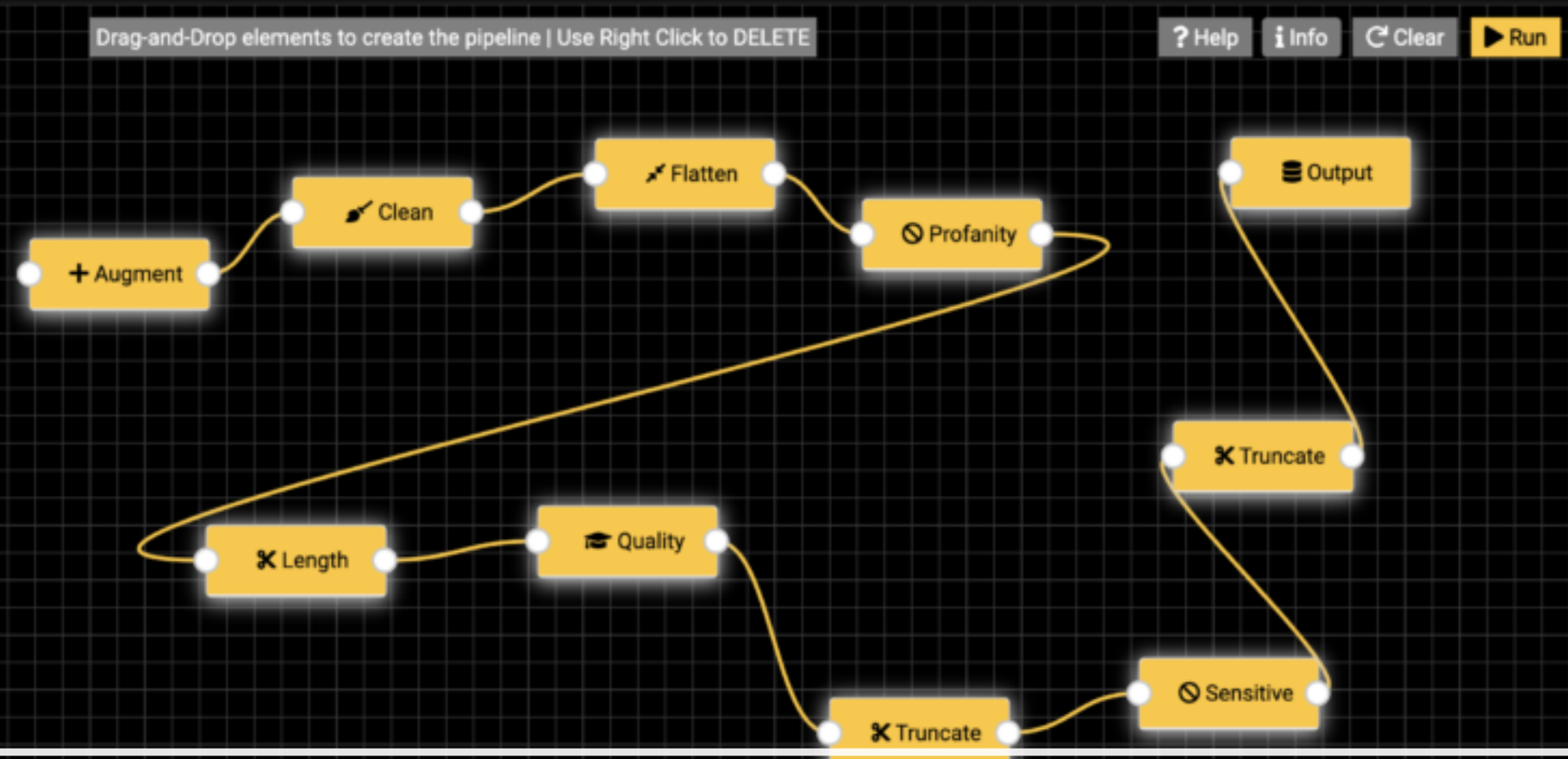
On the right side, there is a 'Models' section with 469,541 models. A search bar 'Filter by name' is present. Several model cards are visible, including:

- meta-llama/Llama-2-70b (Text Generation, Updated 4 days ago, 25.2k downloads)
- stabilityai/stable-diffusion-xl-bas (Updated 6 days ago, 2.01k downloads, 393 likes)
- openchat/openchat (Text Generation, Updated 2 days ago, 1.3k downloads, 1 like)
- llyyasviel/ControlNet-v1-1 (Updated Apr 26, 1.87k likes)
- cerspense/zeroscope\_v2\_XL (Updated 3 days ago, 2.66k downloads, 334 likes)
- meta-llama/Llama-2-13b (Text Generation, Updated 4 days ago, 328 downloads, 64 likes)
- tiiuae/falcon-40b-instruct (Text Generation, Updated 27 days ago, 288k downloads, 899 likes)
- WizardLM/WizardCoder-15B-V1.0 (Text Generation, Updated 3 days ago, 12.5k downloads, 332 likes)
- CompVis/stable-diffusion-v1-4 (Text-to-Image, Updated about 17 hours ago, 1.1k downloads, 3.7k likes)
- stabilityai/stable-diffusion-2-1 (Text-to-Image, Updated about 17 hours ago, 782k downloads, 2.81k likes)
- Salesforce/xgen-7b-8k-inst (Text Generation, Updated 4 days ago, 6.18k downloads, 57 likes)

# OpenAssistant Conversations / Workflow

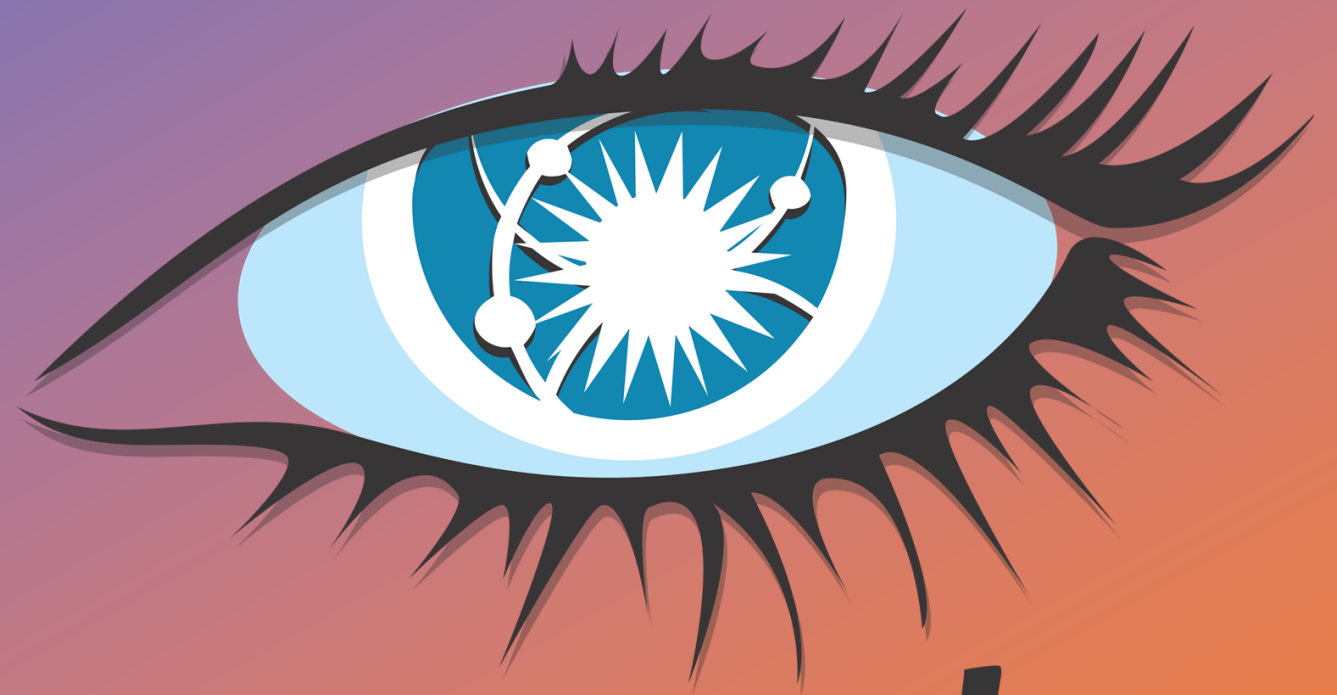
- 1. Ingestion
- 2. Workflow
- 3. Configuration
- 4. Review
- 5. Output
- 6. Insights

- Source
- + Augment
- Clean
- Flatten
- Profanity
- Detoxify
- Quality
- Length
- Valid Ques
- Compress
- EndToken
- Sensitive
- Language
- Dedupe
- Pad
- Truncate
- Output



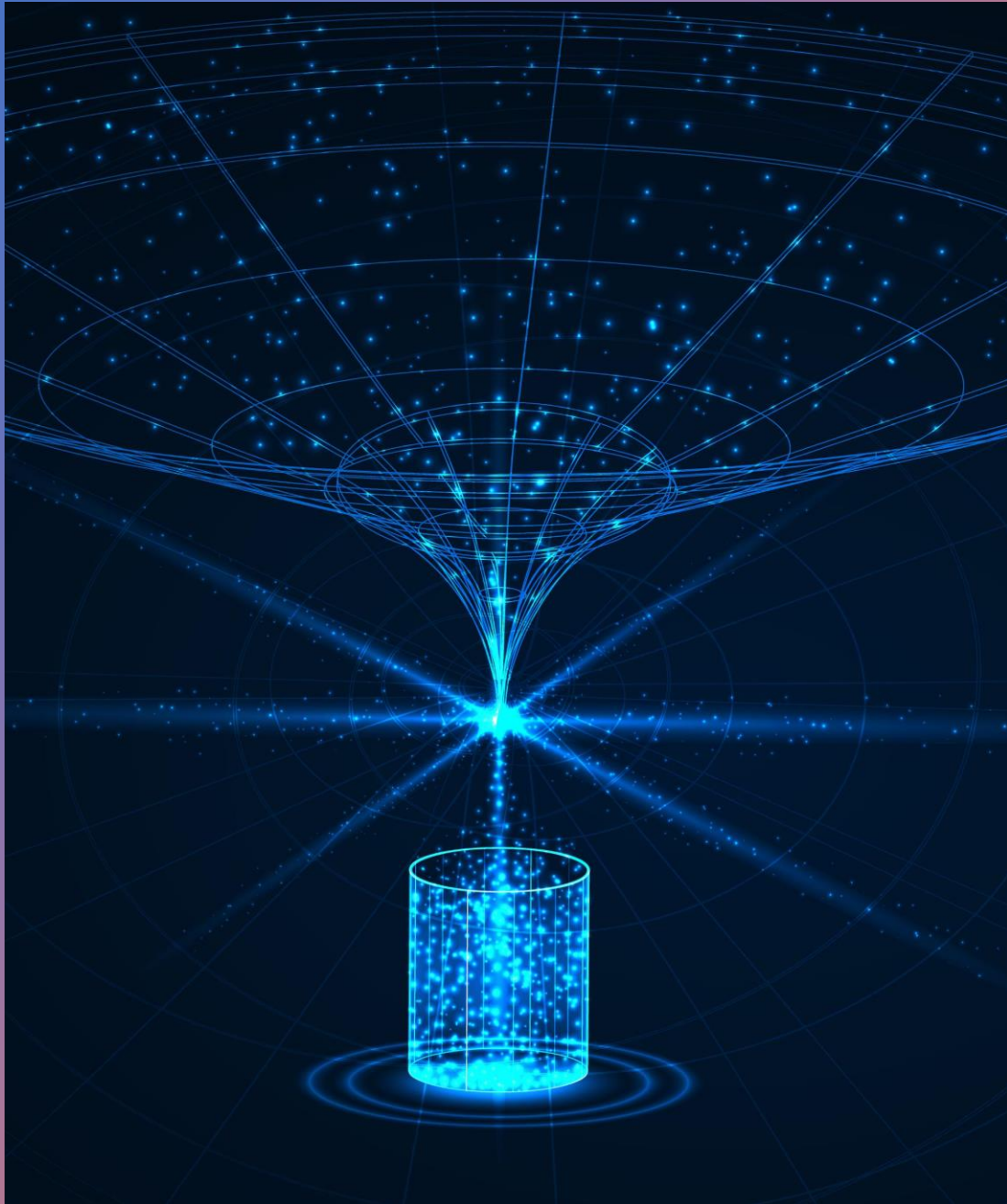
h2o.ai

Big Data  
Vector  
NoSQL  
Database



***cassandra***

---

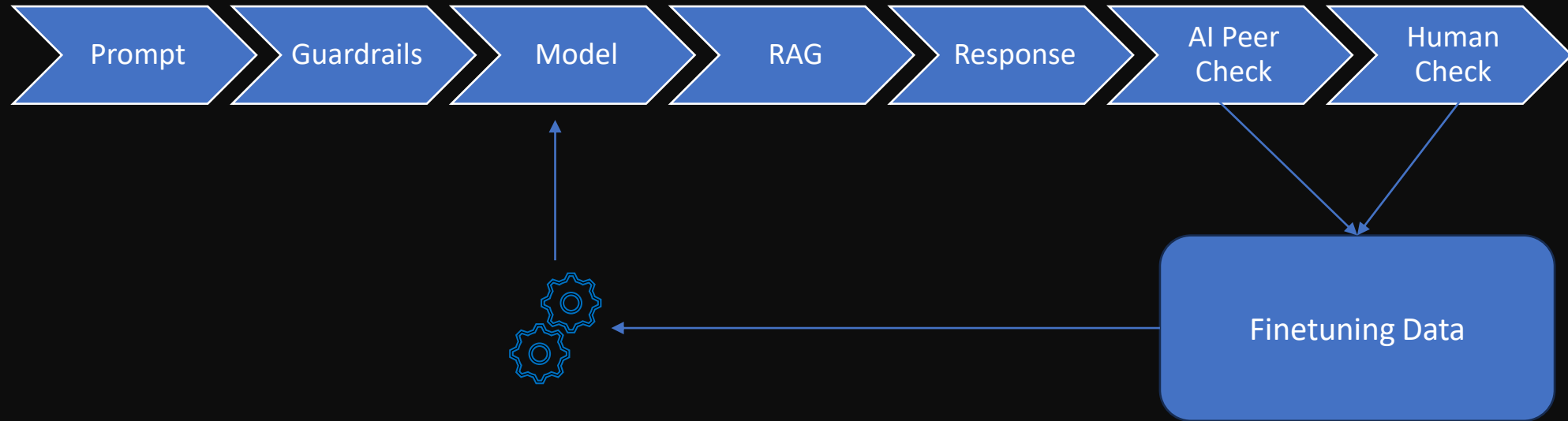


# Vector Data

- 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](http://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](mailto:Charles@WitFoo.com)

CharlesHerring.com

@charlesherring