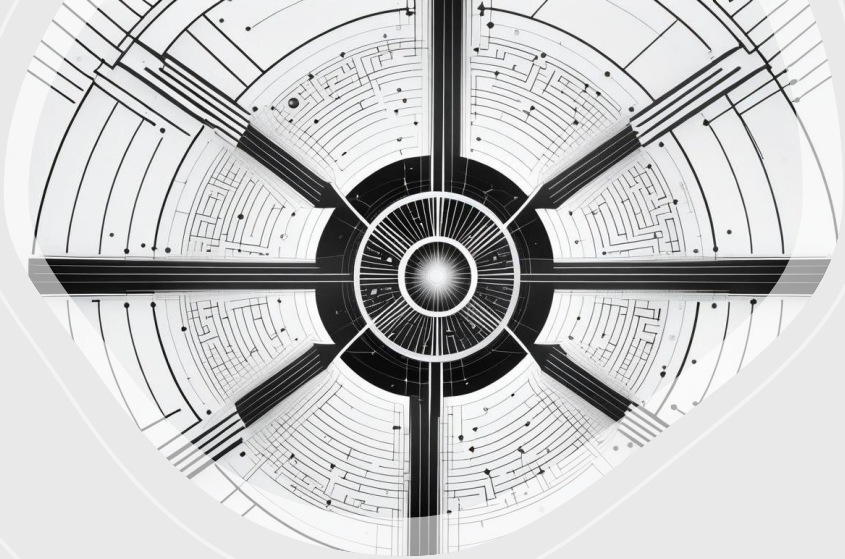
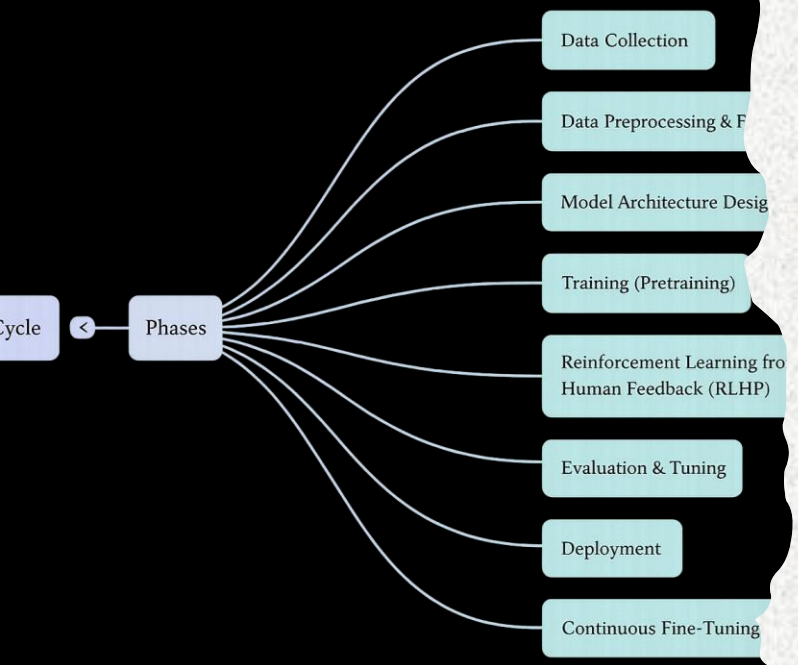


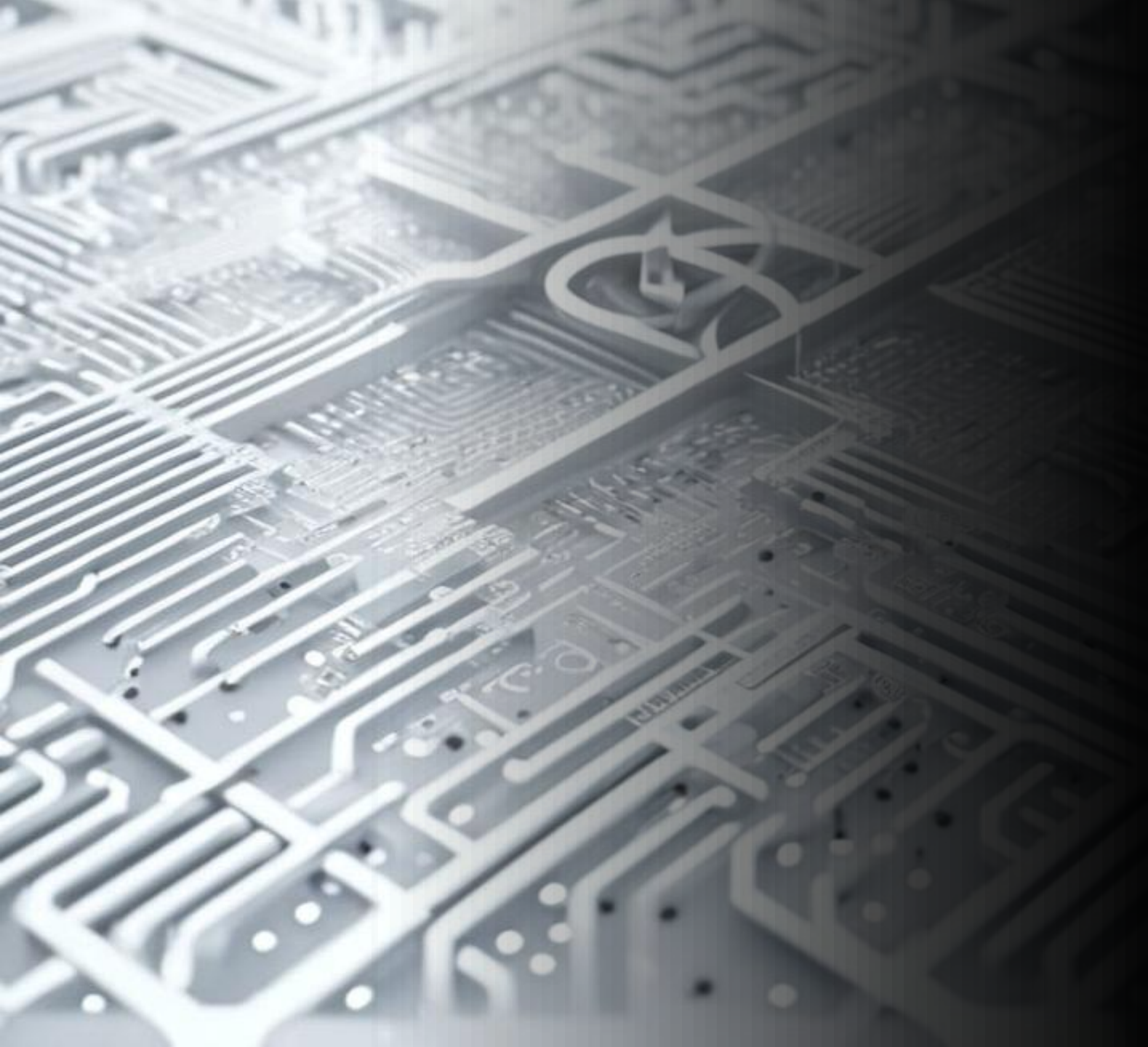
Inside the Mind of a Machine: How Generative AI Models Are Trained

- Robert McCoy
- June 24,2025

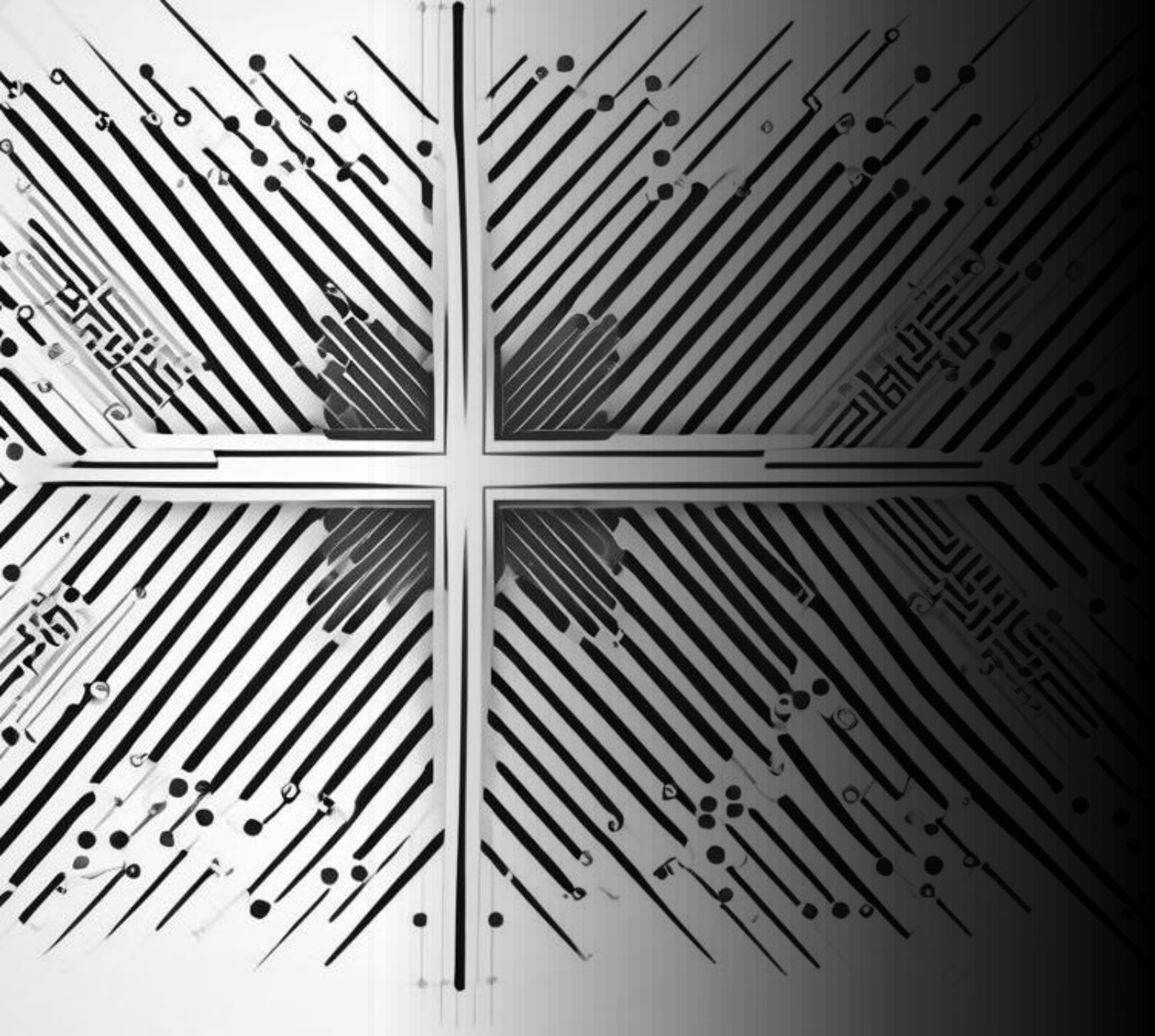
Overview of the Training Lifecycle



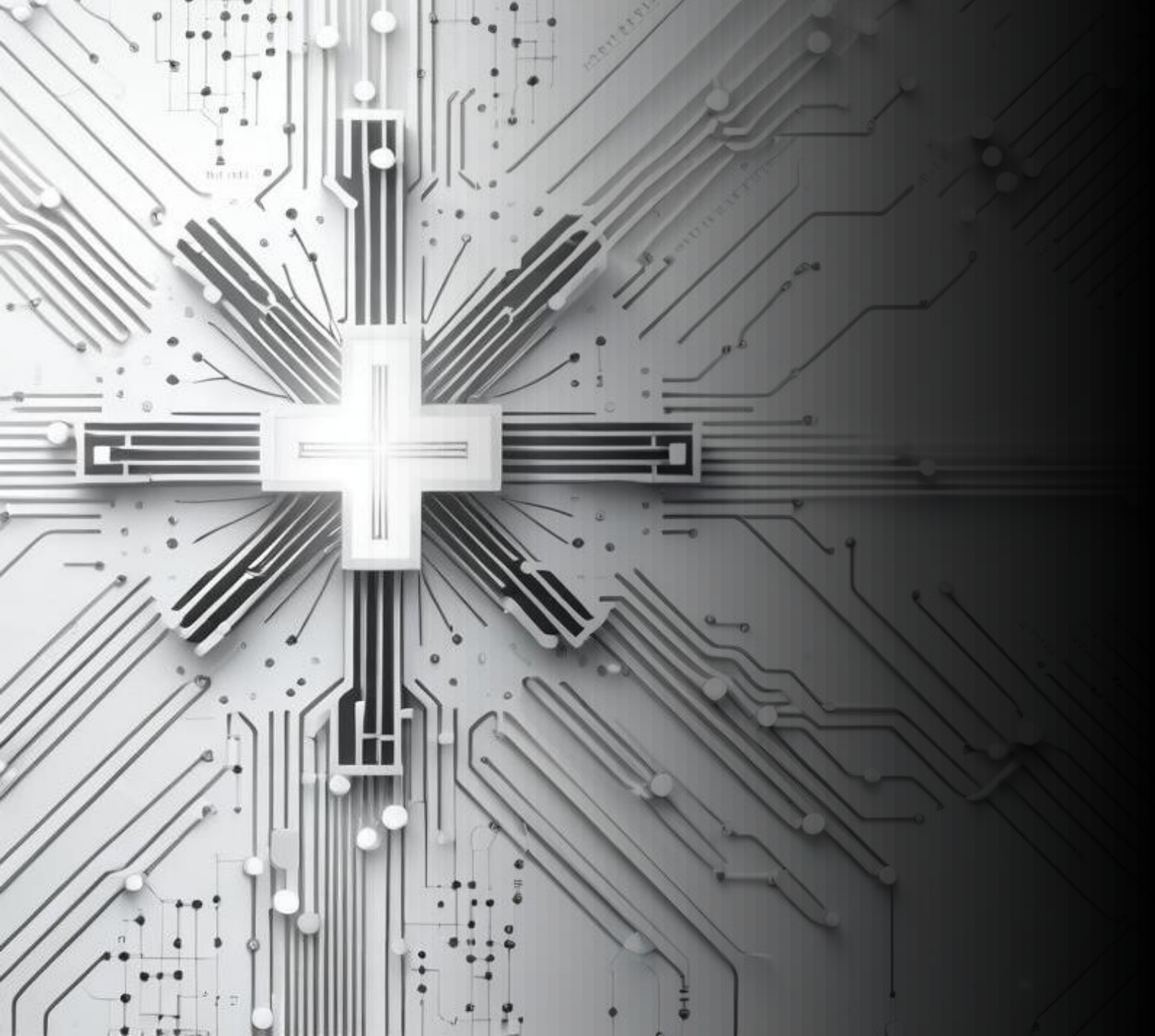
- 1 Training-Data Sources
- OpenAI lists three input channels (OpenAI, 2025a).
  - Public internet text – Common Crawl snapshots, English Wikipedia, WebText2 (Reddit-linked pages), and other open sites.
  - Licensed / partner corpora – large book collections and proprietary datasets.
  - User-generated content – conversations, code samples, and demonstrations contributed by crowd workers and ordinary ChatGPT users (unless they disable *Improve the model*) (OpenAI, 2025b).



- **Data Processing and Filtering**
  - Cleans raw text data (removes duplicates, spam, noise)
  - Normalizes formatting (punctuation, casing, tokenization)
  - Filters out toxic, biased, or low-quality content
  - Converts input into tokens (numerical form)
  - Ensures dataset balance across topics, languages, and styles

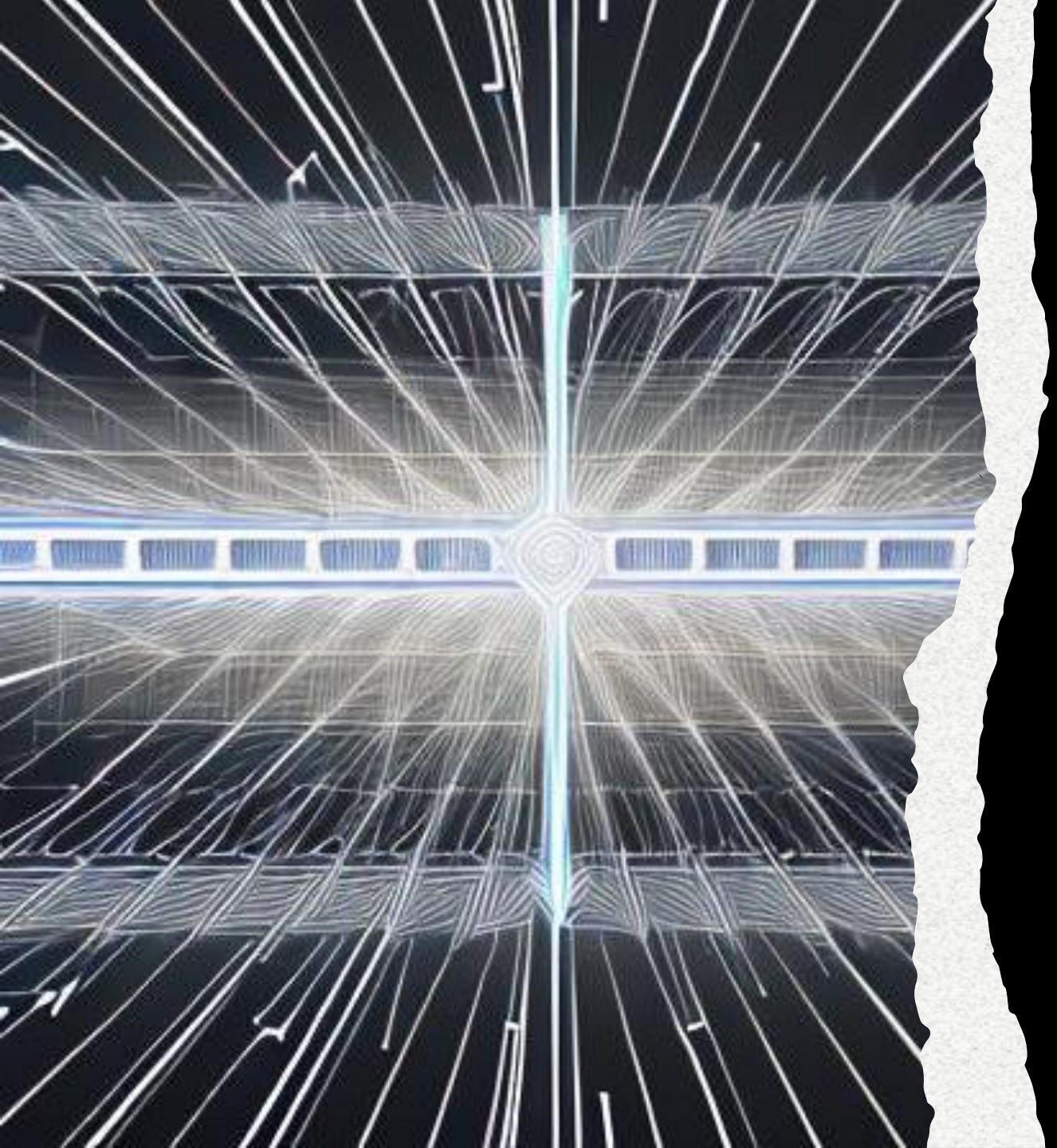


- **Model Architecture Design**
- Choose neural network type (e.g., Transformer)
- Set hyperparameters (layers, attention heads, embeddings)
- Define input/output format (token-level, sentence-level)
- Optimize for performance, scalability, and memory use
- Plan for parallelization across compute clusters



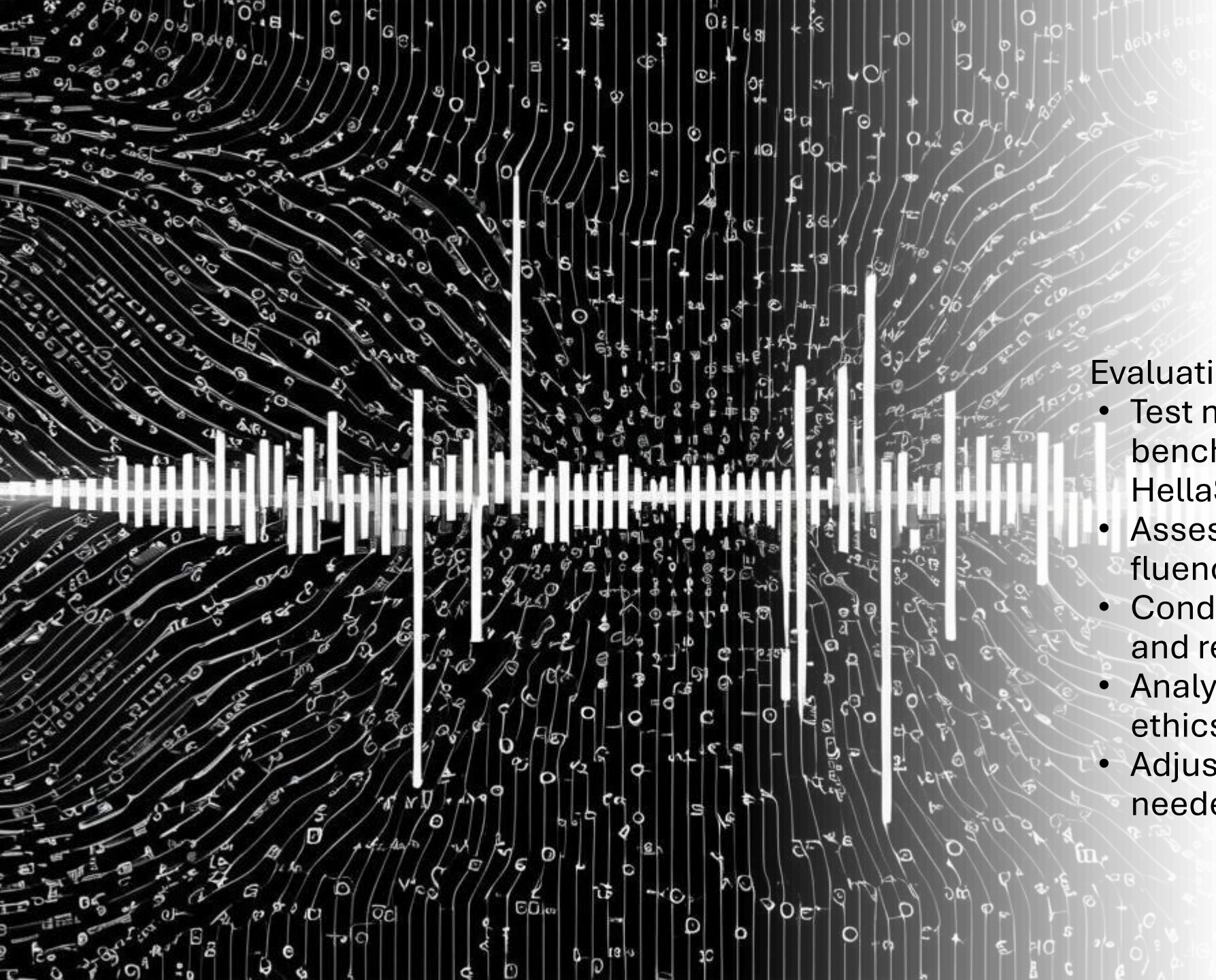
Training (Pretraining)

- Feed massive token datasets into the model
- Adjust weights using gradient descent and backpropagation
- Use masked or next-token prediction objectives
- Train over multiple epochs on large GPU/TPU clusters
- Monitor loss metrics and convergence behavior



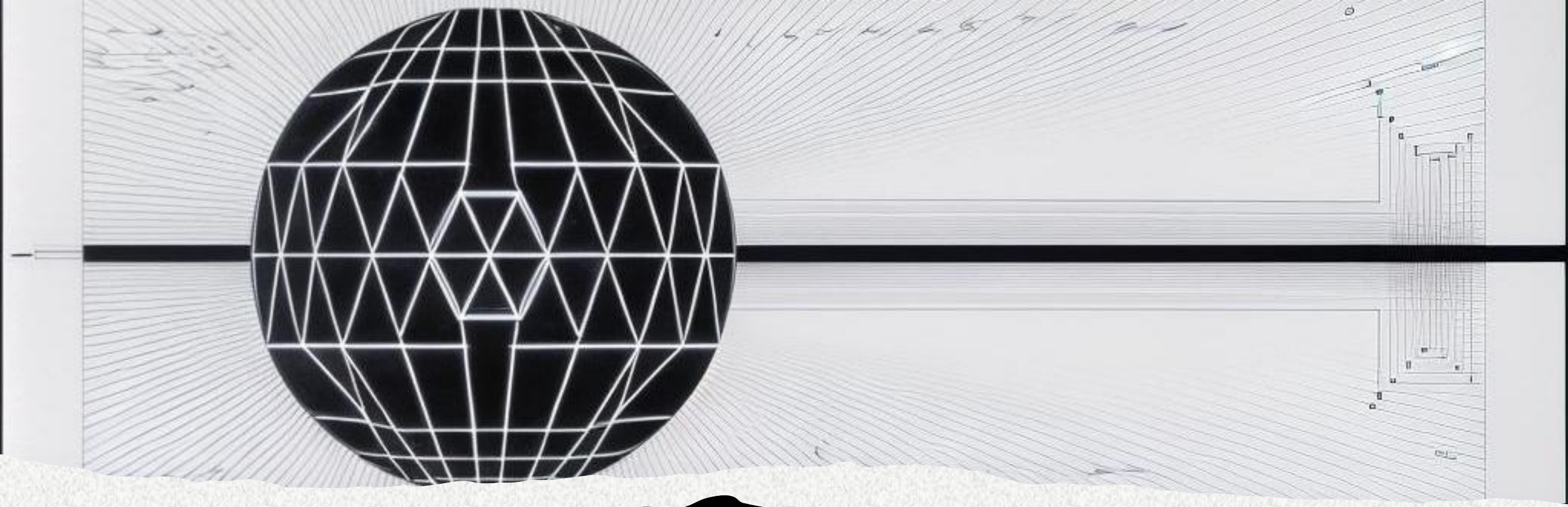
Reinforcement Learning from Human Feedback (RLHF)

- Collect human preference data through ranking tasks
- Train a reward model on human-labeled responses
- Fine-tune the model using Proximal Policy Optimization (PPO)
- Reduce harmful or misleading outputs
- Align model responses with human values and expectations



Evaluation & Tuning

- Test model on standard benchmarks (e.g., MMLU, HellaSwag)
- Assess accuracy, bias, fluency, and coherence
- Conduct adversarial testing and red-teaming
- Analyze outputs for safety, ethics, and performance gaps
- Adjust architecture or data as needed



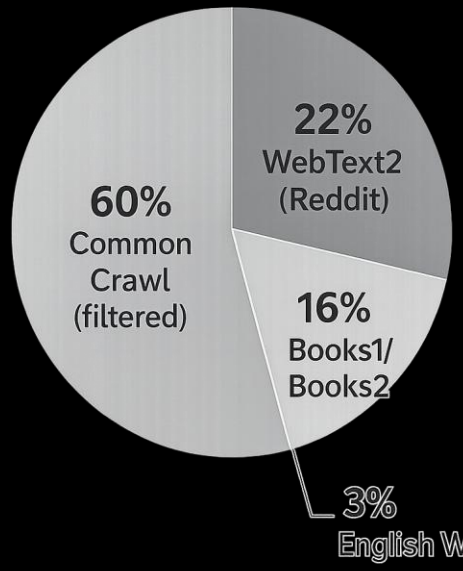
Deployment

- Integrate model into production environments (API, app, chatbot)
- Apply latency, scaling, and throughput optimizations
- Monitor real-world usage for bugs or failure modes
- Establish access policies and safety filters
- Ensure uptime, versioning, and resource efficiency

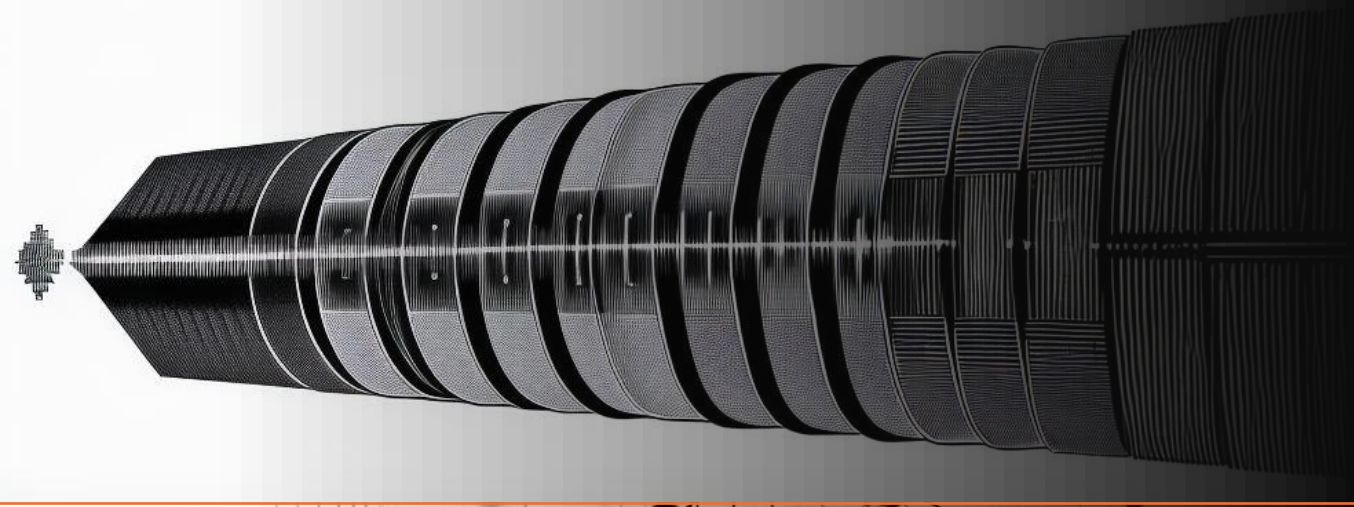
- Continuous Fine-Tuning
  - Incorporate new feedback and evolving data trends
  - Fix hallucinations or bad behavior via targeted updates
  - Adapt to new topics, languages, or regulatory standards
  - Perform lightweight updates without full retraining
  - Support personalization and domain specialization

Dataset Sources

(Brown, 2020, Layton, 2023)

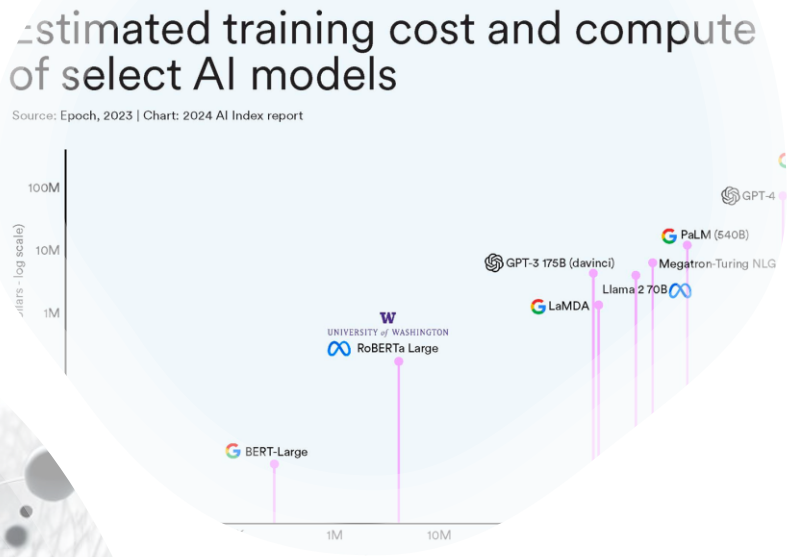


Dataset size
≈ 570GB
(high quality)
from 45TB raw text



Training Infrastructure

- Hardware: TPUs/GPUs (e.g., NVIDIA A100s)
- Training Time: Weeks to months
- Energy Consumption: Tens of megawatt-hours per model
- Cloud Providers: Azure (GPT), AWS (Anthropic), Meta’s in-house clusters

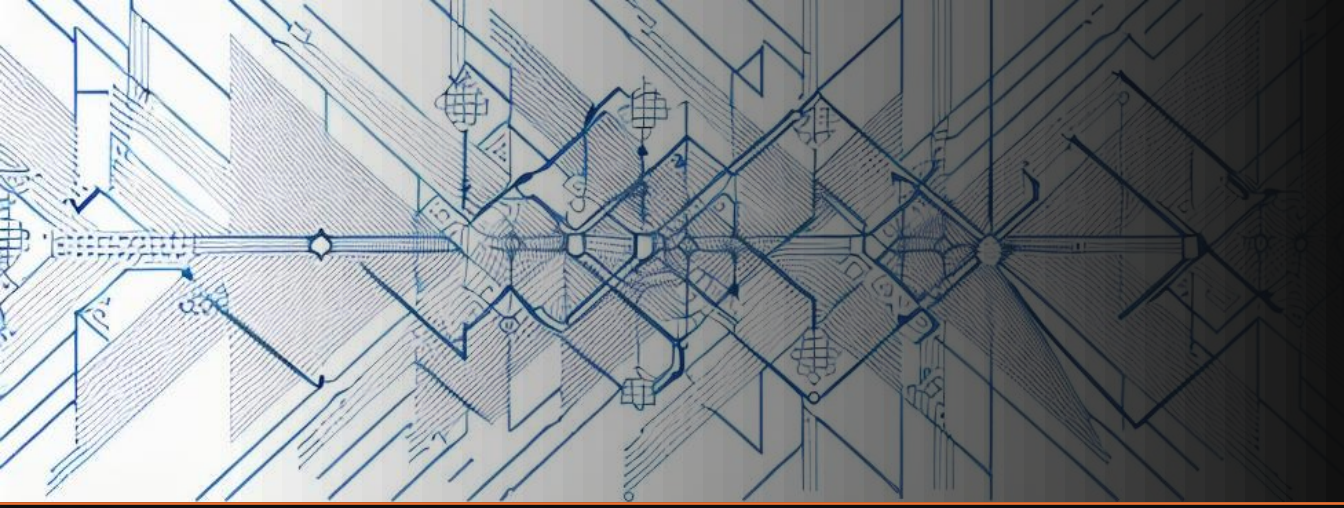


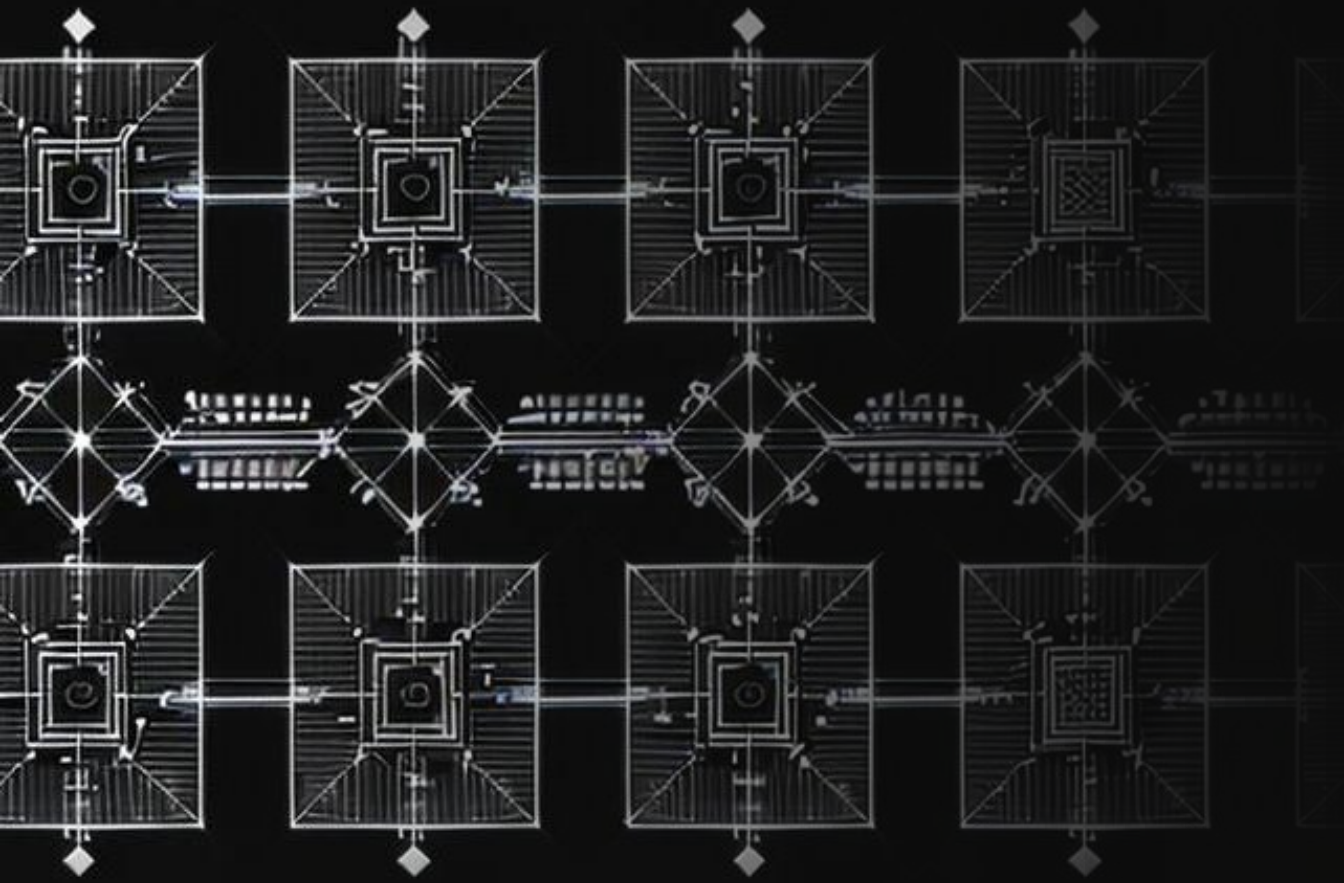
https://www.cudocompute.com/blog/what-is-the-cost-of-training-large-language-models

Cost ~ $100M

Resource Impact on Model Performance

- Data Diversity → Reduces Bias
- Compute Power → Enables Depth & Fluency
- Energy Use → Environmental Cost
- Time → *Governs* Model Staleness
- Bias (Language and Culture)
- Privacy





Final Takeaways

- LLM training is data-hungry, power-intensive, and ethically complex.
- Every design choice impacts model capability, bias, and cost.
- Deployment demands ongoing audits and public transparency.

References

Brown et al., 2020 – *Language Models Are Few-Shot Learners* arXiv:2005.14165

Layton, D. (2023) – Medium post: *ChatGPT—* Medium article

OpenAI. (2025a). *How ChatGPT and our foundation models are developed*. OpenAI Help Center. https://help.openai.com/articles/7842364

OpenAI. (2025b). *How your data is used to improve model performance*. OpenAI Help Center. https://help.openai.com/articles/5722486