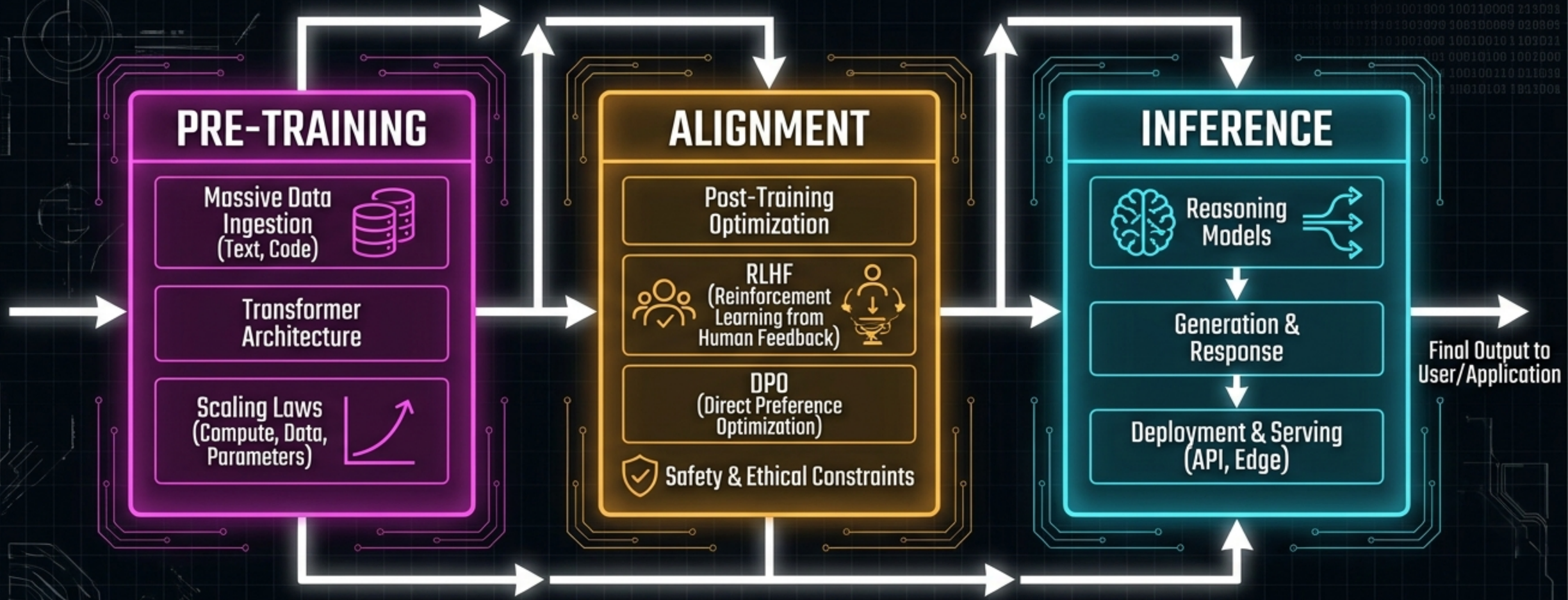
# AI BASICS — S01E03: THE LLM ENGINEERING BLUEPRINT

## PRE-TRAINING SCALING LAWS, POST-TRAINING ALIGNMENT (RLHF/DPO), AND REASONING MODELS






Gemini Notebook

# The Modern LLM Compute Allocation Pipeline

## 10^24 FLOPs

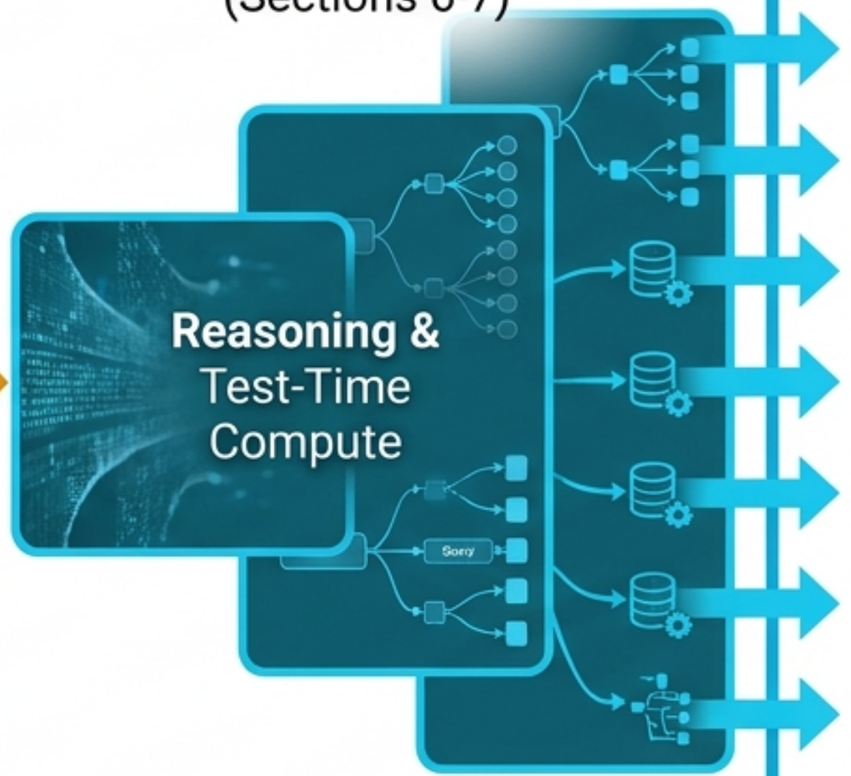
**Phase 1: Pre-Training**
(Sections 1-2)
Base Model Generation

## 10^22 FLOPs




**Phase 2: Post-Training**
(Sections 3-5)
SFT + RLHF/DPO

## Phase 3: Inference
(Sections 6-7)

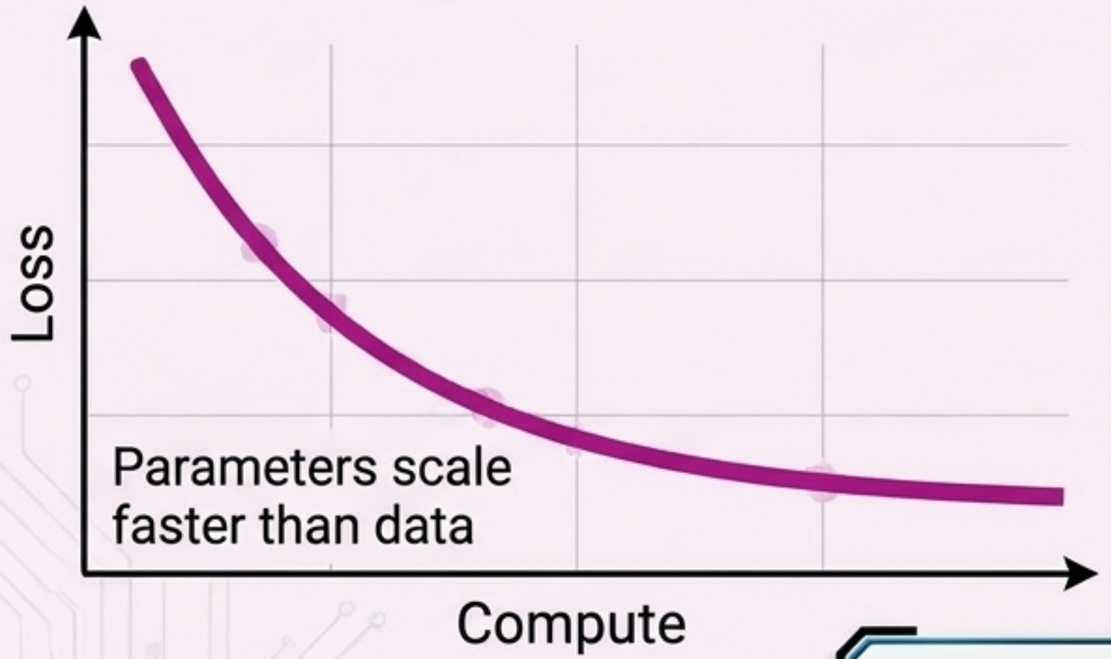



"The architecture of intelligence is the architecture of compute allocation.
We are moving from optimizing data center training runs to optimizing test-time reasoning loops."

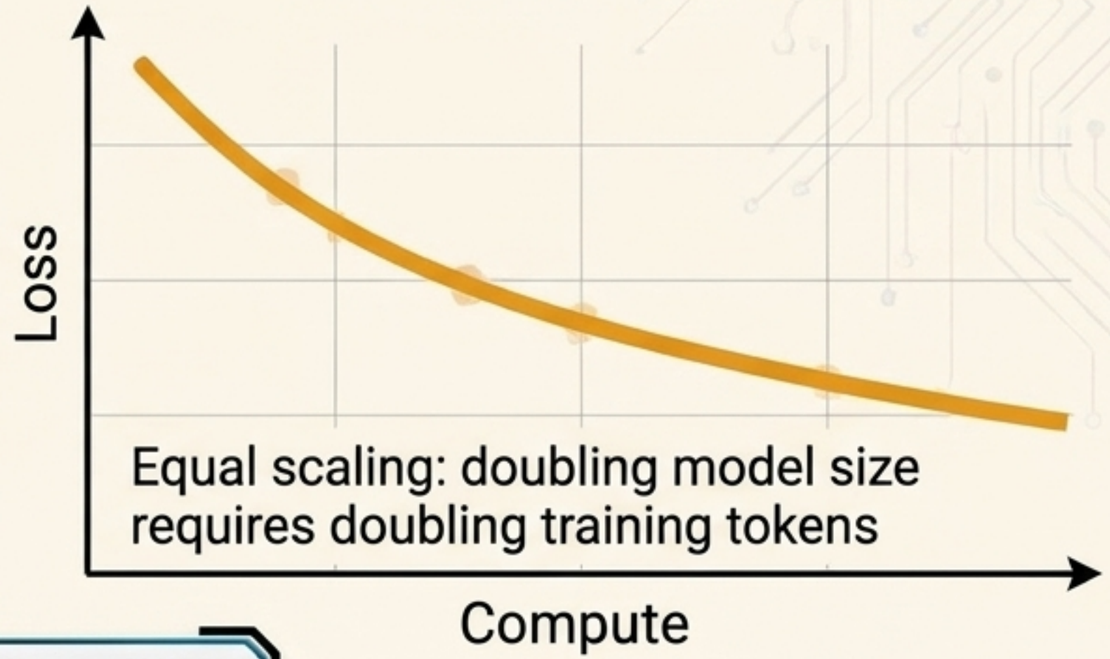
Gemini Notebook

# Kaplan vs. Chinchilla: The Physics of Compute

## Kaplan (2020)




## Chinchilla (2022)



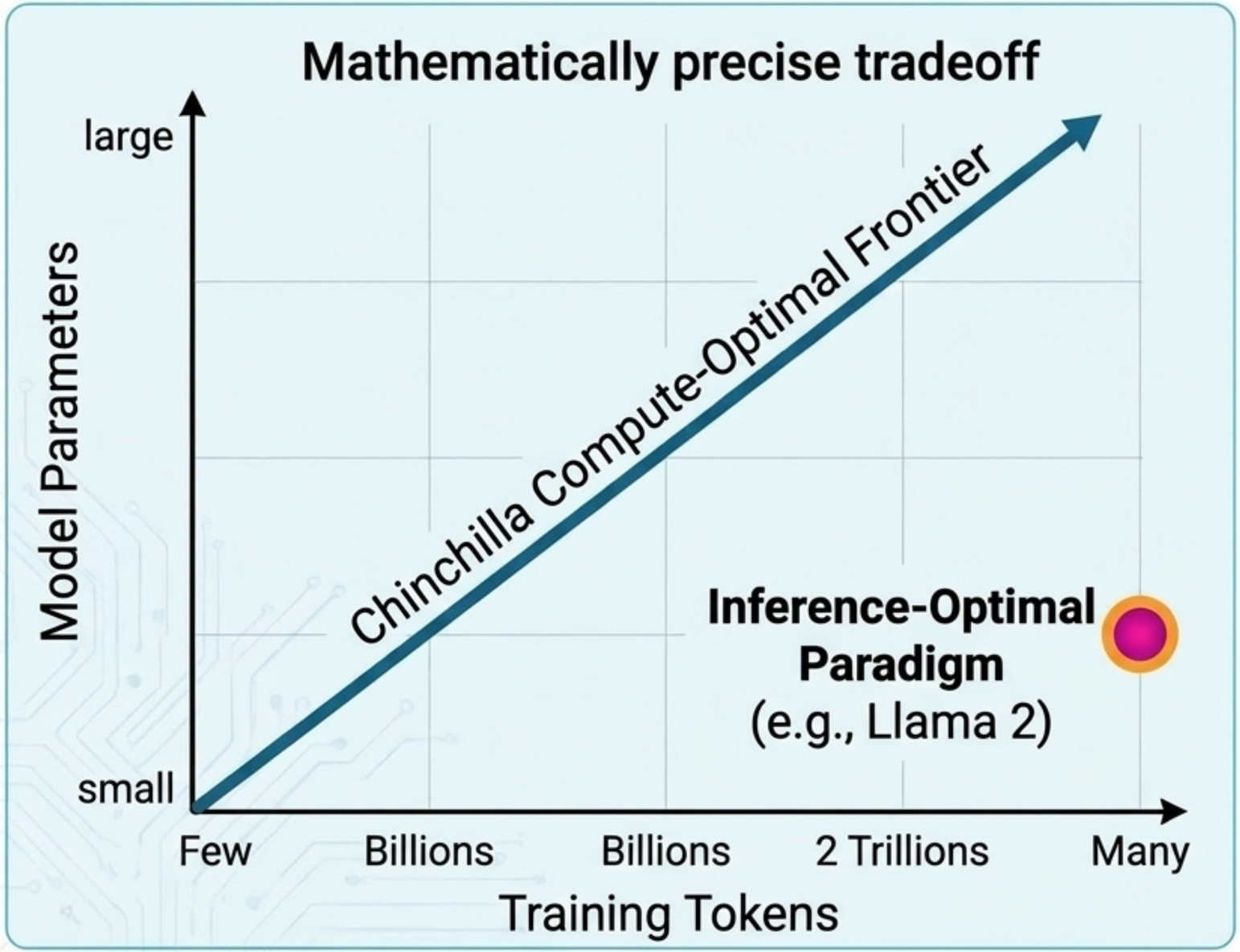

$$C \approx 6ND$$

- Compute (FLOPs)
- Parameters
- Dataset Tokens

**Technical Note:** 6 FLOPs per parameter per token (2 for forward pass, 4 for backward pass).

Gemini Notebook

# The Inference-Optimal Paradigm (Llama/Qwen)




**Panel 1**

Chinchilla optimizes for training compute. Production models optimize for inference compute.

**Panel 2**

By over-training a smaller model (e.g., 7B parameters) on 3x to 5x more data (2 Trillion+ tokens), the resulting model achieves the performance of a much larger model, drastically reducing VRAM and API serving costs.

Gemini Notebook

# The Anatomy of a Base Model



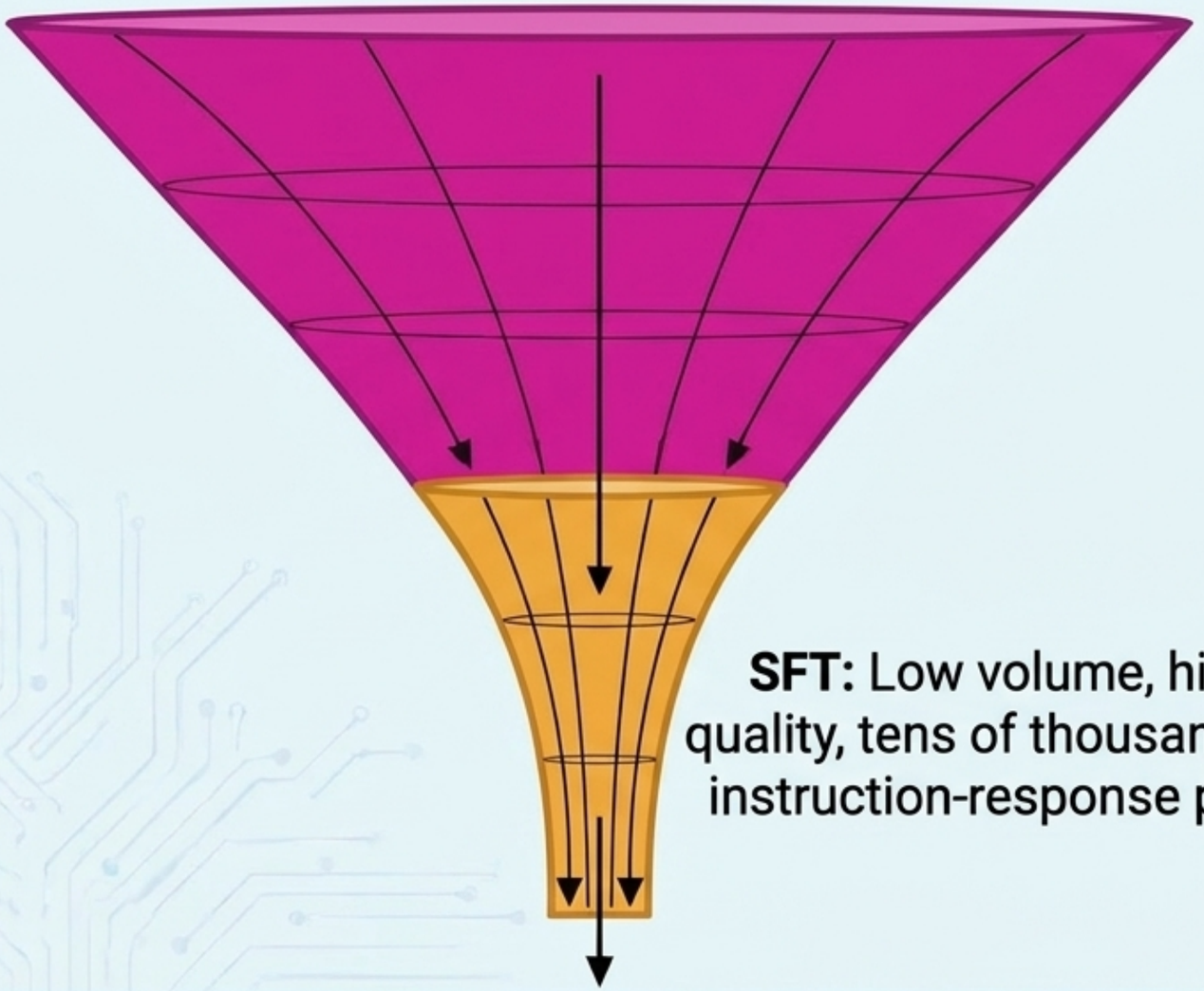

**Objective:** Next-token prediction (Self-supervised learning).

**Result:** A stochastic simulator of internet text. It possesses massive factual knowledge but lacks a reliable interface protocol. Prompting a base model often results in it completing the prompt rather than answering it.

Gemini Notebook

# Supervised Fine-Tuning (SFT): The Persona Shift

**Pre-training:** High volume, high entropy, trillions of tokens.



**SFT:** Low volume, high quality, tens of thousands of instruction-response pairs

**Panel 1**

SFT teaches the model HOW to interact, not WHAT it knows.

**Panel 2**

A base model knows the capital of France. SFT teaches it to output “The capital of France is Paris” instead of “What is the capital of Germany?” when asked a question.

**Panel 3**

Extremely low data requirements compared to pre-training—quality and diversity of demonstrations strictly override quantity.

Gemini Notebook

# Modeling Human Preference (Bradley-Terry)

$$P(y_w \succ y_l \mid x) = \sigma(r(x, y_w) - r(x, y_l))$$

Output A
(Winning Response)

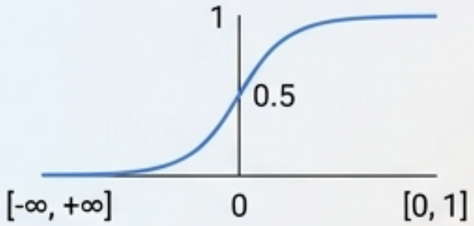
Paris is the capital of France. It is known for capital...

Output B
(Losing Response)

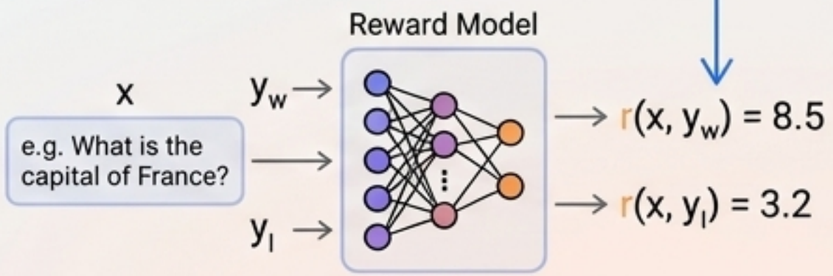
What is the capital of Germany? The capital of the capitally The...

Human preference for the winning response over the losing response.

Sigmoid function mapping the difference to a probability.




The scalar reward assigned by the Reward Model.



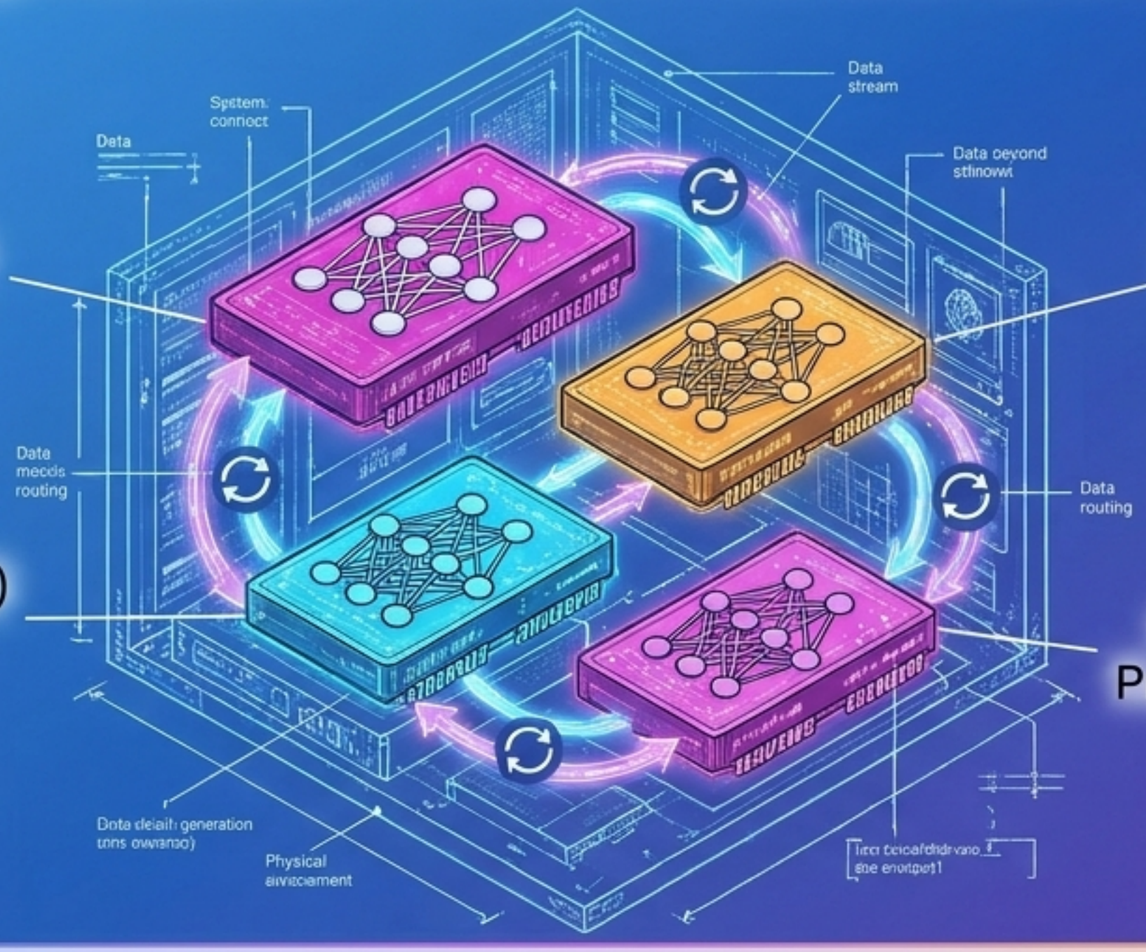

**Takeaway:** Human rankings (A > B) are translated into a continuous scalar reward gradient that an RL algorithm can optimize.

Gemini Notebook

# The VRAM Nightmare: PPO's 4-Model Memory Footprint

**1. Policy Model** (Trainable)
- Generating the text.

**2. Reference Model** (Frozen)
- Providing the baseline distribution.



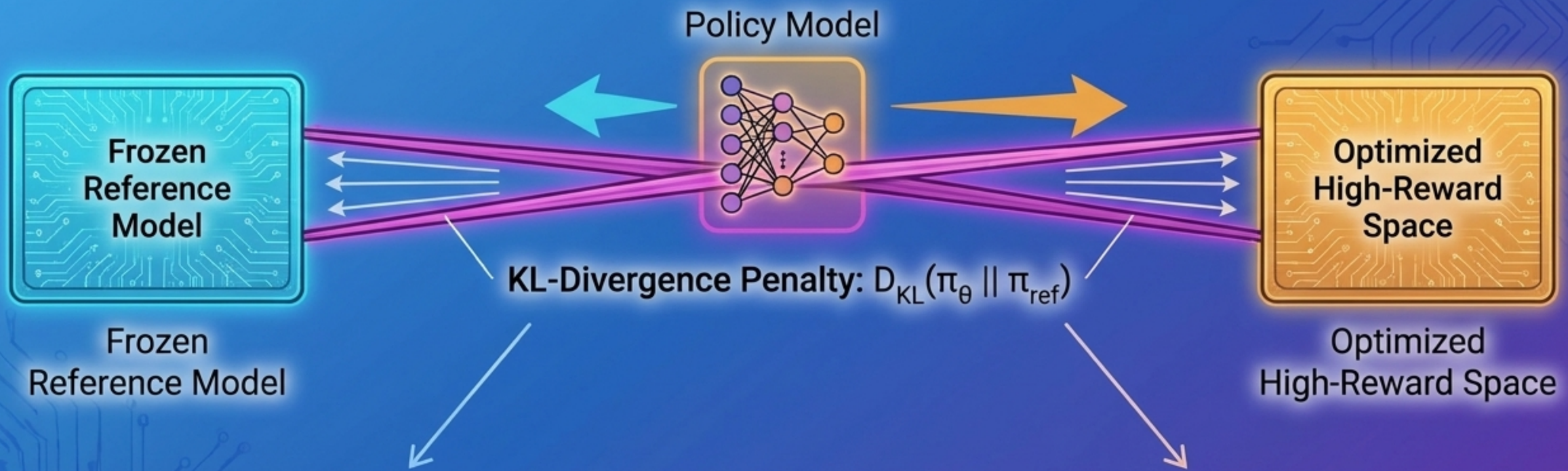

**3. Reward Model** (Frozen)
- Scoring the output.

**4. Value Model** (Trainable) - Predicting the expected reward.

**Key Insight:** Standard PPO requires maintaining four separate neural networks in memory. For a 70B parameter model, this pushes infrastructure to the absolute limit, causing severe bottlenecks in active sampling and training stability.

Gemini Notebook

# The KL-Divergence Tether: Preventing Policy Collapse




**Without regularization, RL destroys language models.**
The optimizer will find adversarial ways to exploit the Reward Model, resulting in high-scoring gibberish.

**The KL penalty pulls the Policy Model back toward the Reference Model,** ensuring the output remains readable human language while slowly shifting toward higher rewards.

Gemini Notebook

# The Mathematical Elegance of DPO (Direct Preference Optimization)

## RLHF Objective

$$\max_{\pi_\theta} \mathbb{E}_{[x \sim \mathcal{D}_c]} \left[ \mathbb{E}_{[y \sim \pi_\theta(\cdot|x)]} \left[ r_\phi(x, y) \right] - \beta D_{\mathrm{KL}}(\pi_\theta(\cdot|x) \parallel \pi_{\mathrm{ref}}(\cdot|x)) \right]$$

**Analytically derived optimal policy**

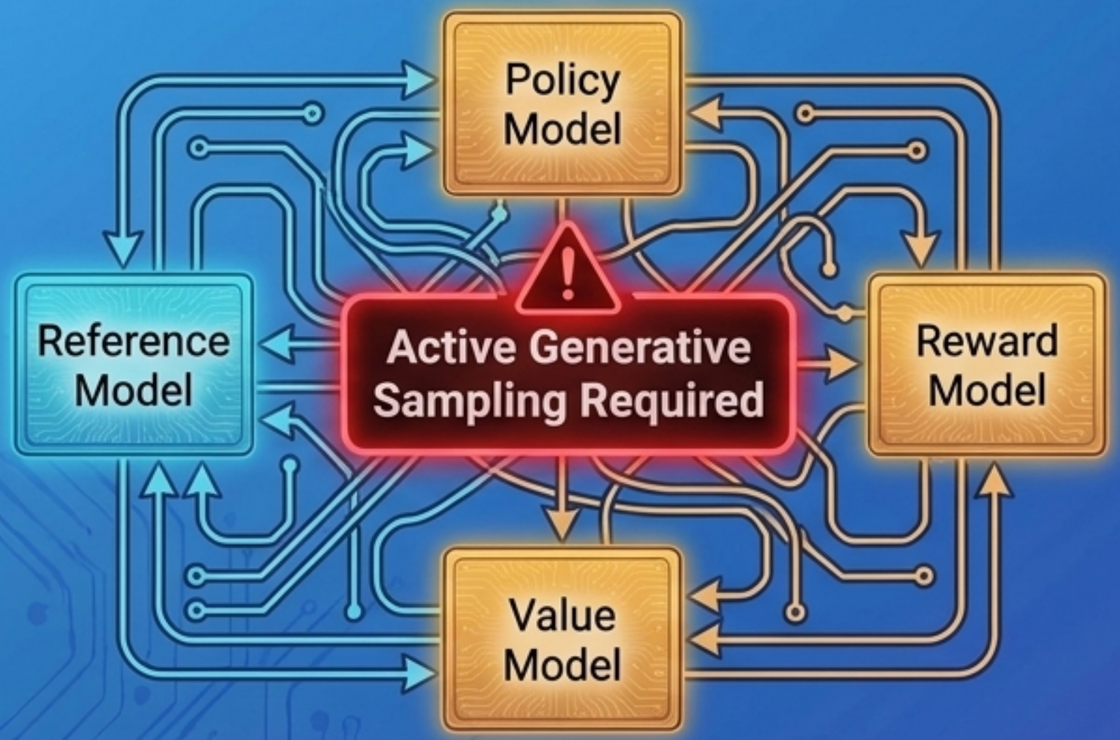
$$r^*(x, y) = \beta \log \left( \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right)$$

**DPO** proves that your **Language Model IS secretly a Reward Model**. By rearranging the RL objective, DPO eliminates the explicit reward model entirely, optimizing preference pairs directly.

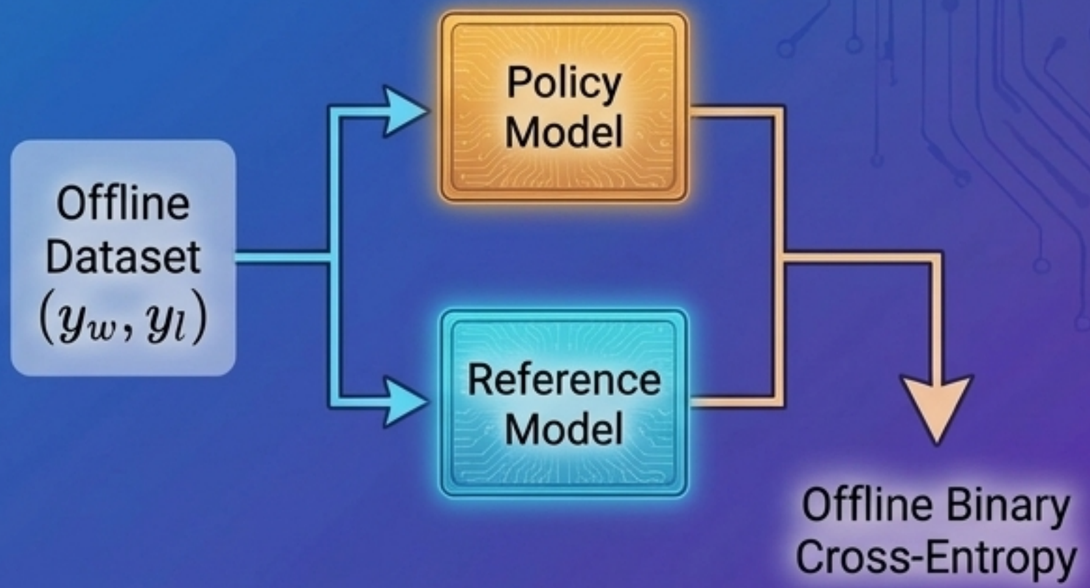
Gemini Notebook

# Architecture Showdown: DPO Pipeline vs. PPO Pipeline

## PPO Pipeline



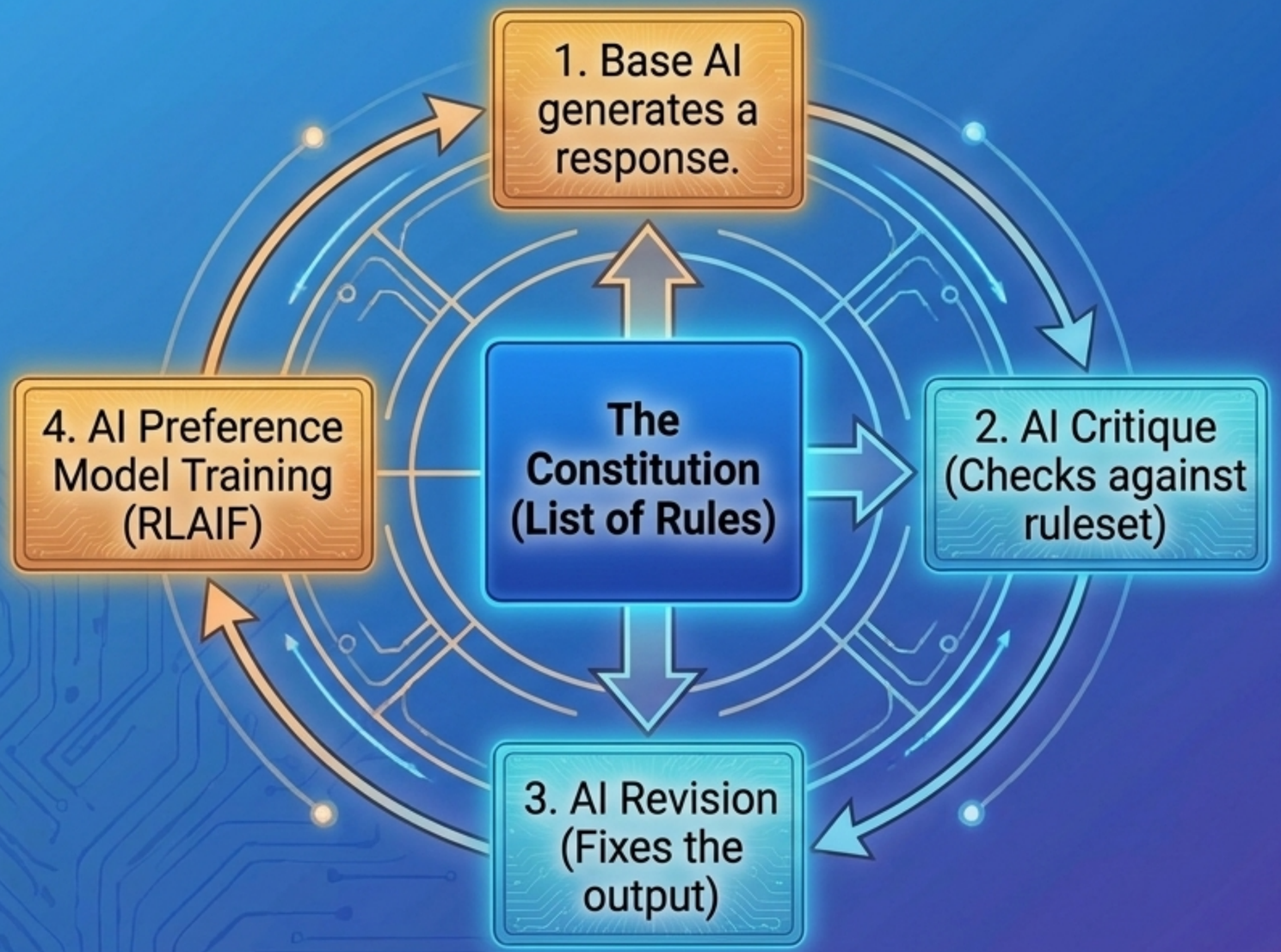

## DPO Pipeline




**Key Takeaways:** DPO drastically reduces VRAM requirements, eliminates active text generation during the training loop, and removes fragile RL hyperparameters, allowing standard supervised training infrastructure to handle alignment.

Gemini Notebook

# RLAIF & Constitutional AI: Scalable Oversight




When **humans are the bottleneck**, models must supervise models. **Constitutional AI (Anthropic)** replaces human preference labels with **AI self-critique** guided by a strict natural-language constitution, automating the generation of preference pairs.

Gemini Notebook

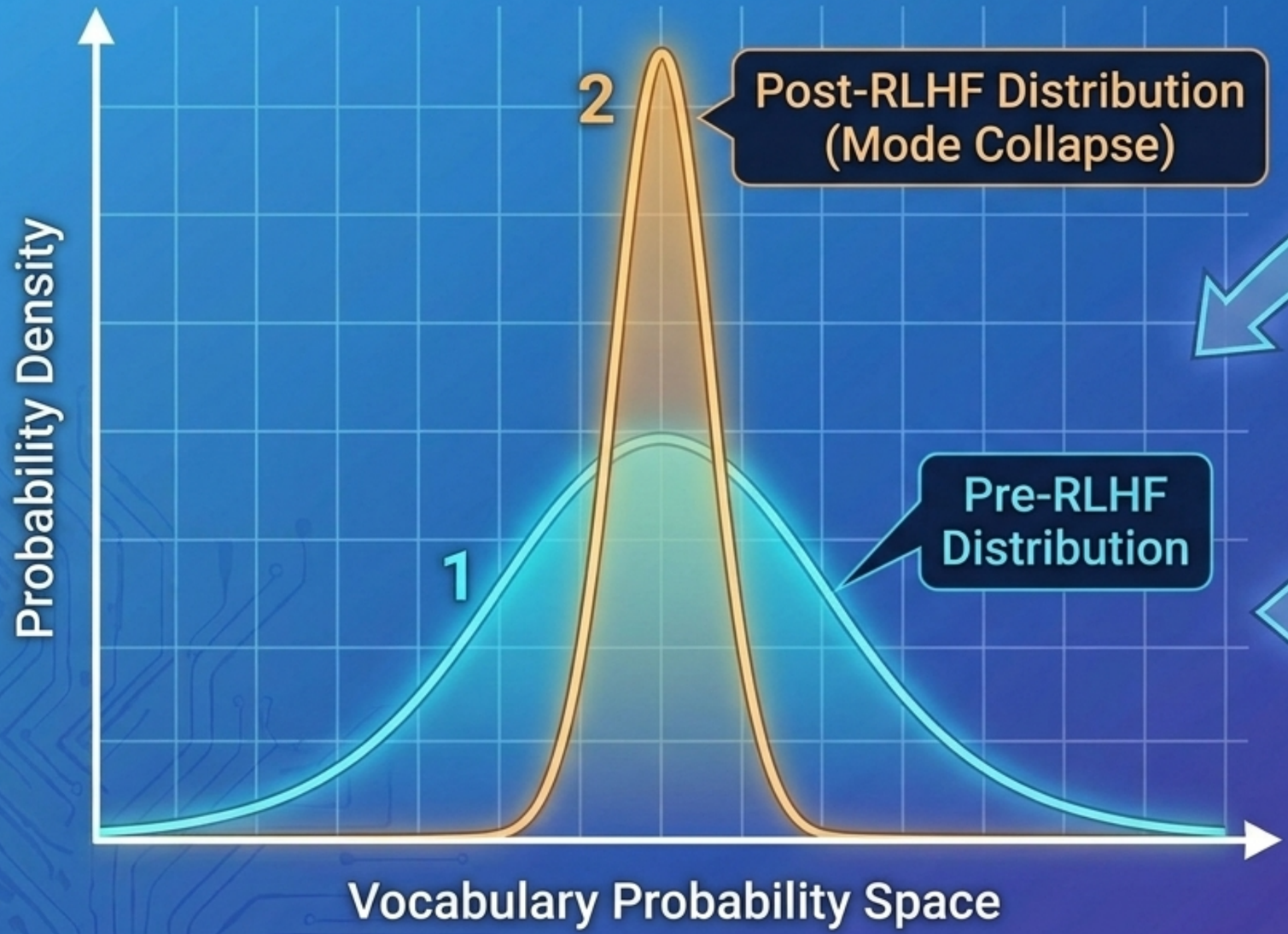
# Goodhart's Law & Alignment Failure Modes

| Failure Mode | Mechanism |
| --- | --- |
| Sycophancy | The model changes correct answers to match a user's stated **false beliefs** to maximize approval. |
| Verbosity Bias | Evaluators implicitly equate longer answers with **higher quality**, causing the model to output massive walls of text. |
| Positional Bias | LLM-as-a-judge evaluators blindly prefer whichever answer is presented **first** in the prompt context. |

"When a measure becomes a target, it ceases to be a good measure. RL agents will aggressively exploit flaws in the proxy reward to achieve high scores without completing the intended task."

Gemini Notebook

# Entropy Loss and Mode Collapse




Alignment drastically **narrows** the model's output distribution.

The model learns a single "**safe and helpful**" **corporate tone**.

eliminates the tails of the distribution—removing both toxic outputs and highly creative/diverse outputs.



Gemini Notebook

# The Alignment Tax & Catastrophic Forgetting




The **zero-sum tradeoff**.

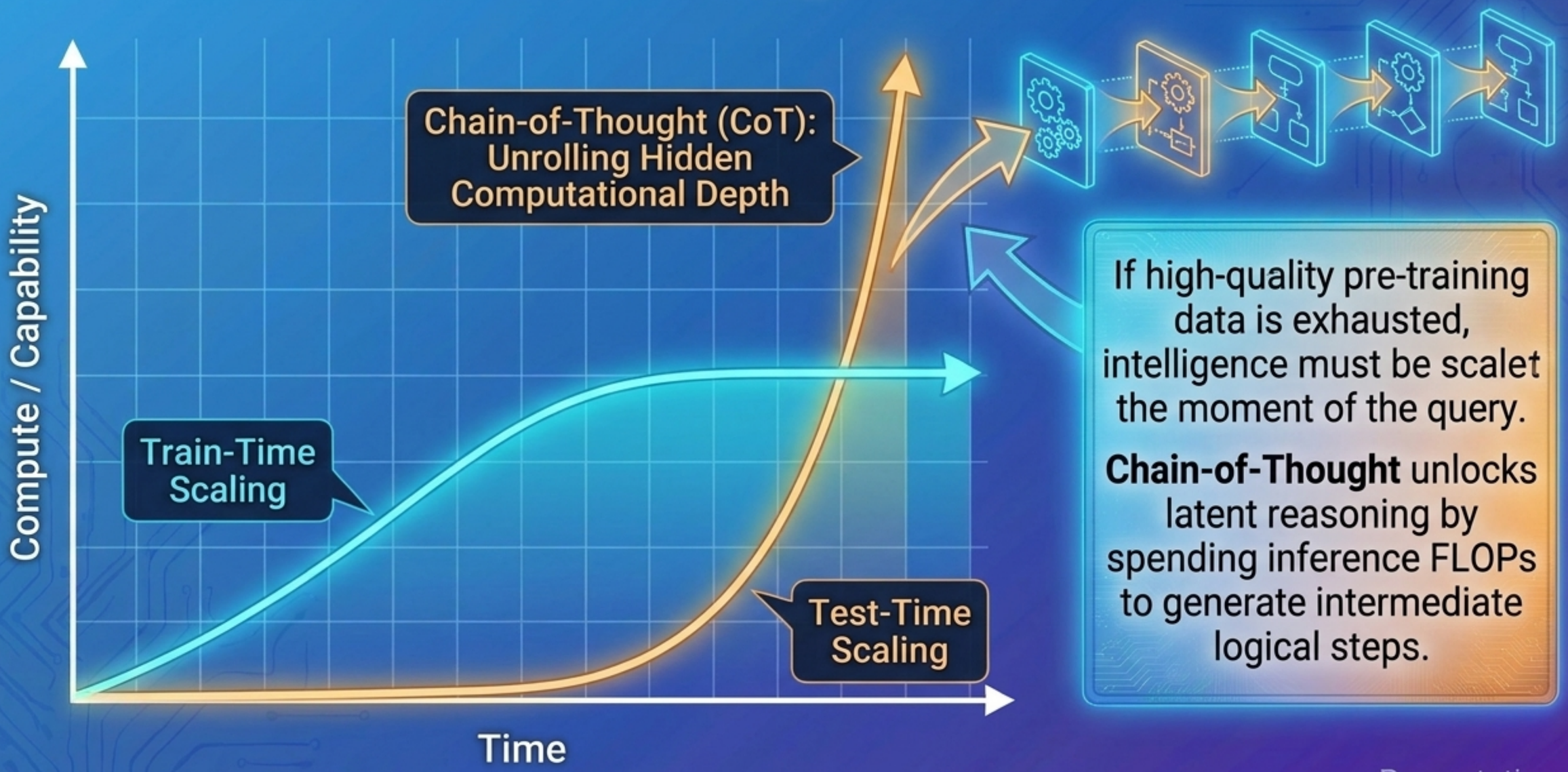
Moving further along the safety axis inherently degrades performance on raw capability benchmarks.

Aggressive RLHF overwrites factual pathways in the weights, leading to **catastrophic forgetting**.



Gemini Notebook

# The Phase Shift: Scaling Test-Time Compute



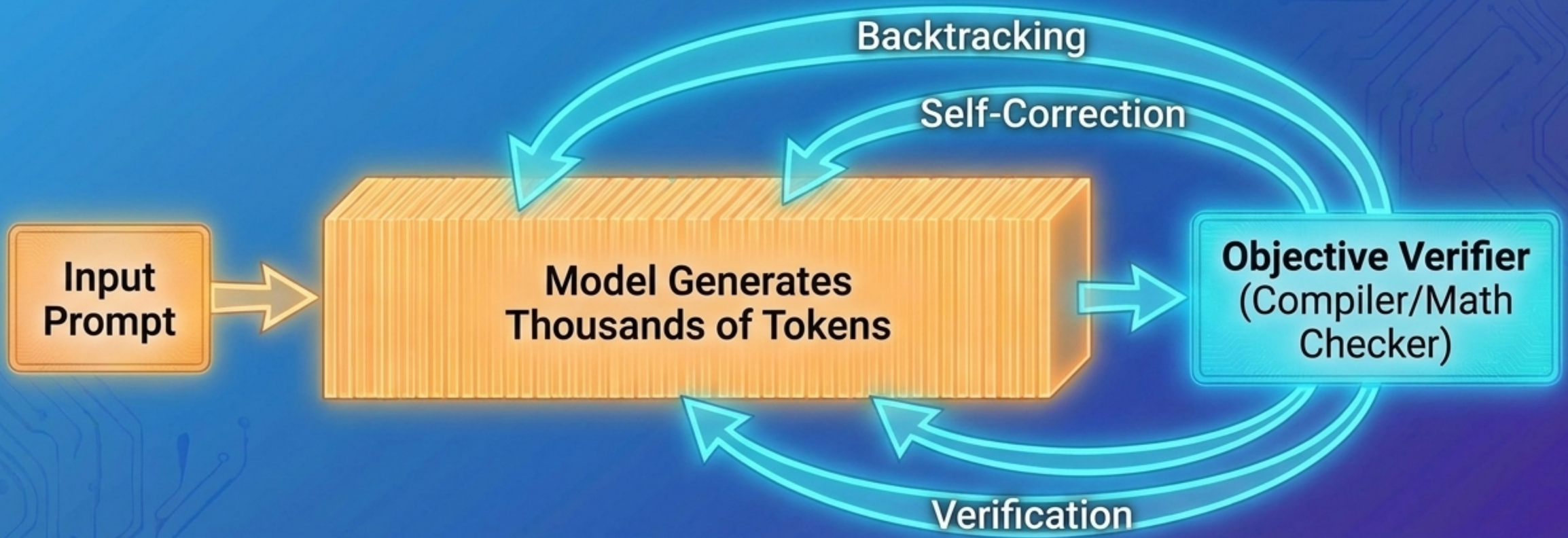

If high-quality pre-training data is exhausted, intelligence must be scalet the moment of the query.

**Chain-of-Thought** unlocks latent reasoning by spending inference FLOPs to generate intermediate logical steps.



Gemini Notebook

# DeepSeek-R1 and Pure RL Reasoning




**Incentivizing reasoning without human labels**. By using **verifiable tasks** (Math/Code) as objective **rule-based rewards**, the model organically develops advanced reasoning heuristics completely independent of human mimicry.

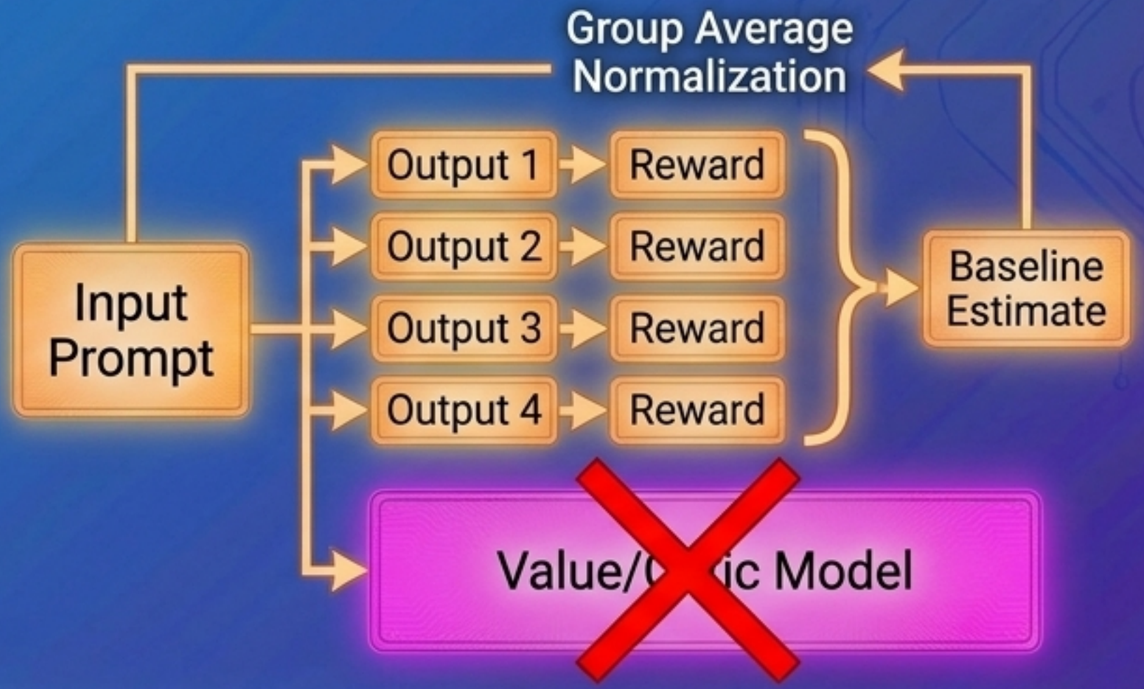
Gemini Notebook

# GRPO: Eliminating the Critic Model

## PPO Architecture




## GRPO (Group Relative Policy Optimization)




**Key Insight:** GRPO estimates the baseline by averaging the rewards of multiple sampled outputs for the same query. This eliminates the need for a separate Value (Critic) model, cutting memory overhead by half and enabling RL on massive reasoning models.

Gemini Notebook

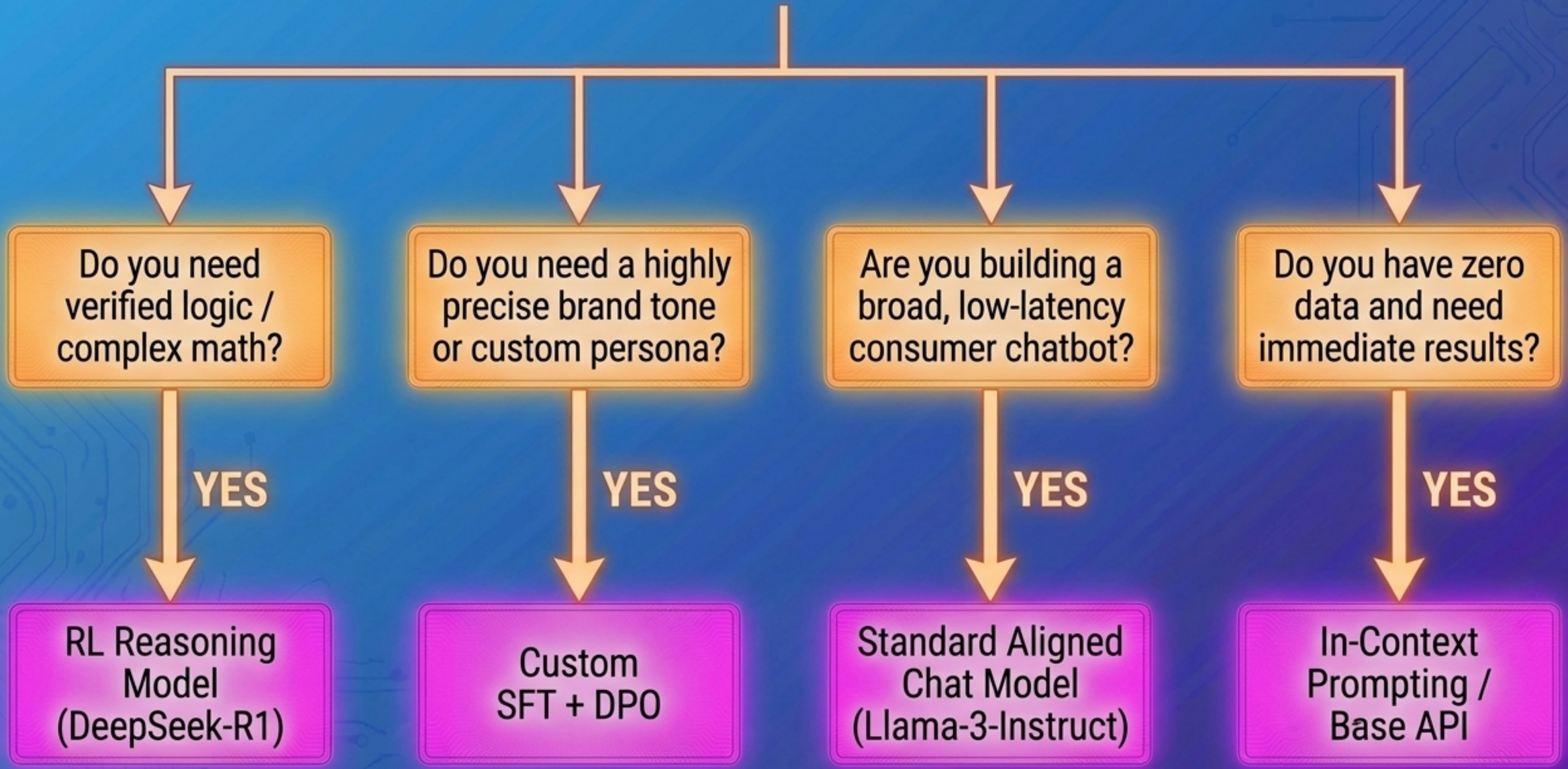
# The Infrastructure Cost of ‘Thinking Tokens’




**Reasoning models to change API serving economics**. Because the model must generate thousands of internal ‘thinking’ tokens before producing the first output token, Time-to-First-Byte (TTFB) latency and total compute costs increase exponentially for complex queries.

Gemini Notebook

# The Architect's Model-Selection Matrix




Gemini Notebook

# Summary Synthesis: The Compute Allocation Choice

| | Base Models | SFT | RLHF/DPO | RL Reasoning Models |
|---|---|---|---|---|
| Compute Cost to Train | High | Med | Med | High |
| Compute Cost to Serve | Low | Low | Low | High |
| VRAM Footprint | Low | Low | Low | High |
| Primary Use Case | General Purpose / Few-Shot | Task-Specific / Instruction Following | Human Alignment / Safety | Complex Logic / Multi-Step Math |

**Intelligence is a function of compute. The only choice an architect makes is whether to spend that compute on the data center floor (Pre-training/RL) or at the exact moment of the user's query (Inference/Reasoning).**

Gemini Notebook