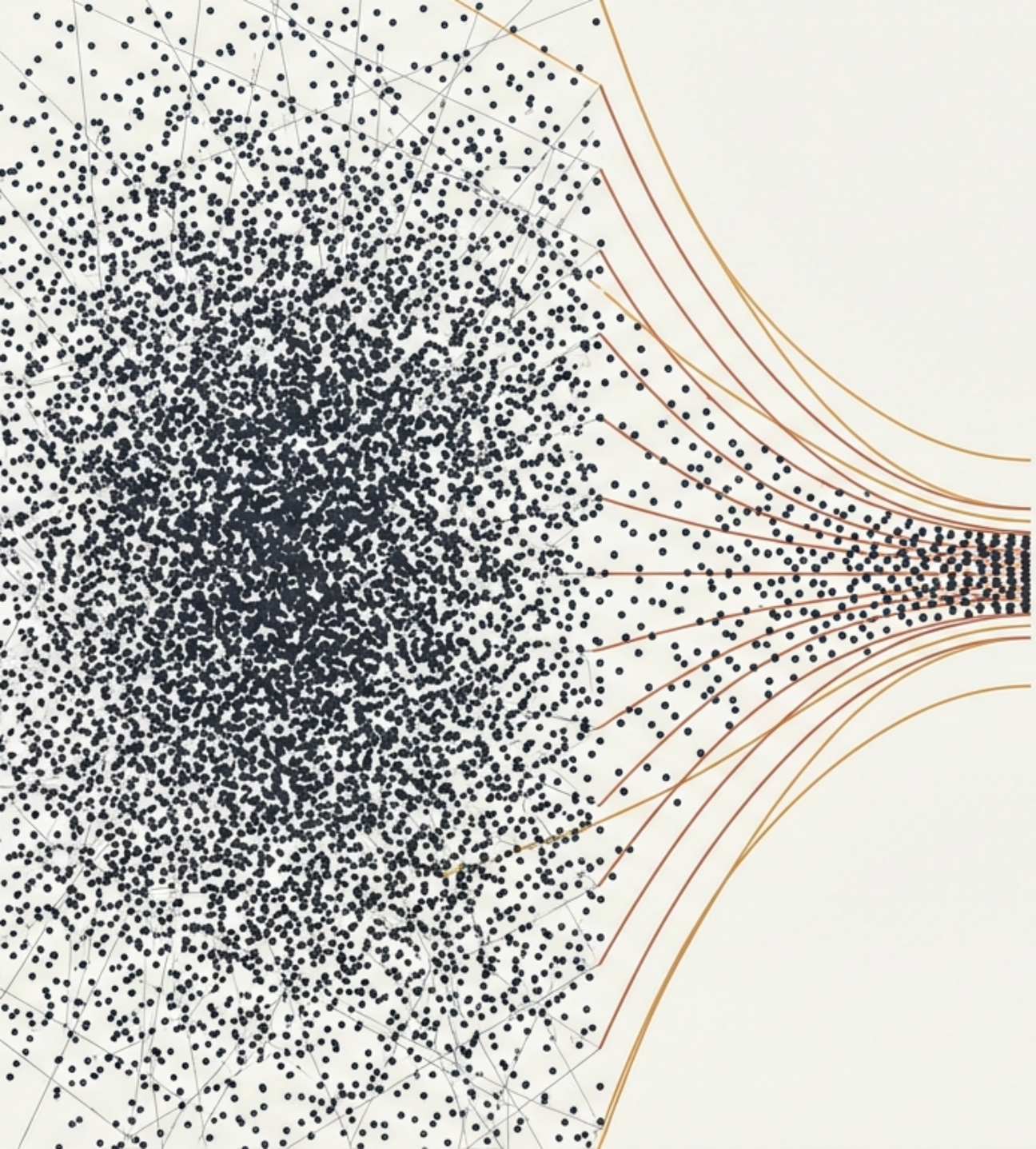
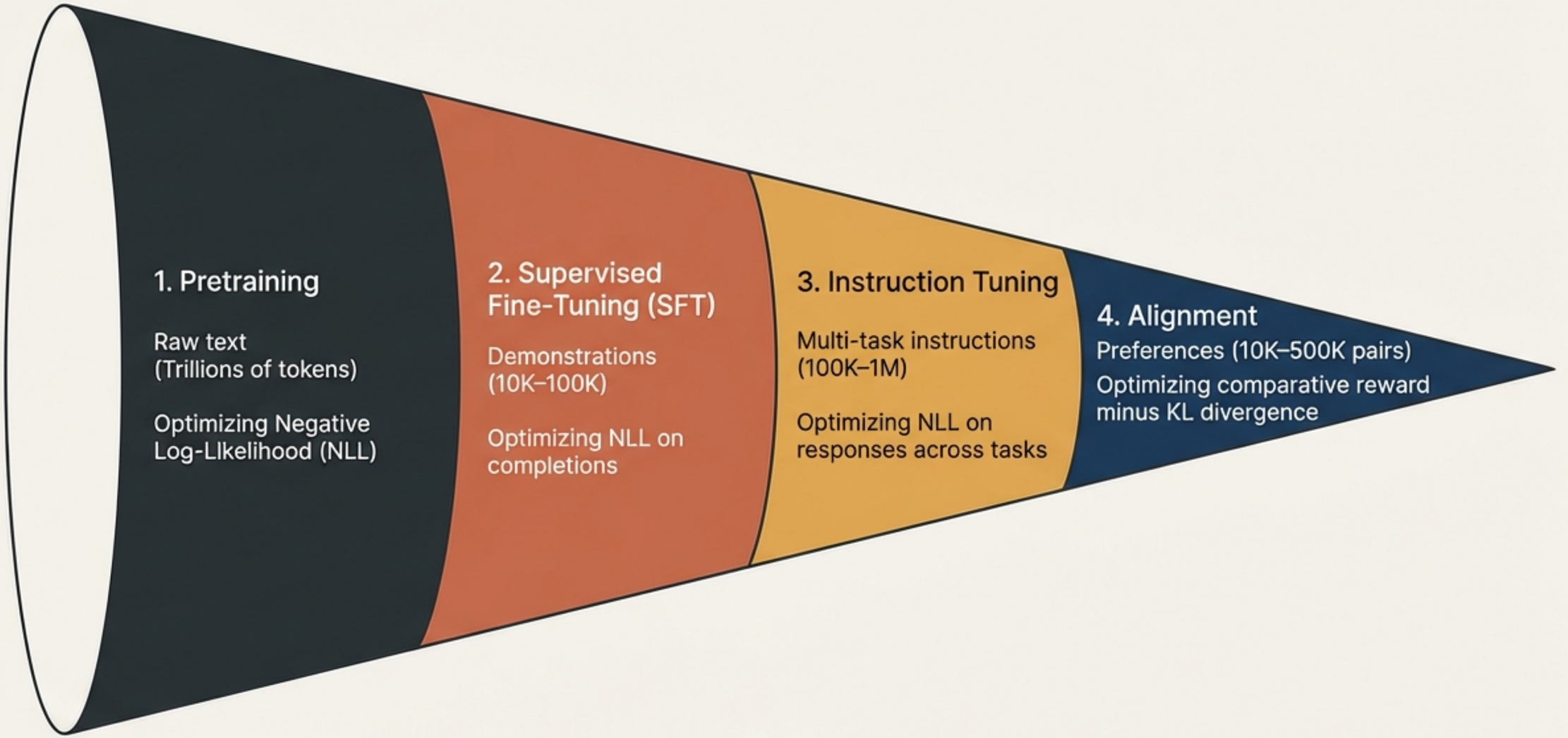


The Architecture of Intelligence

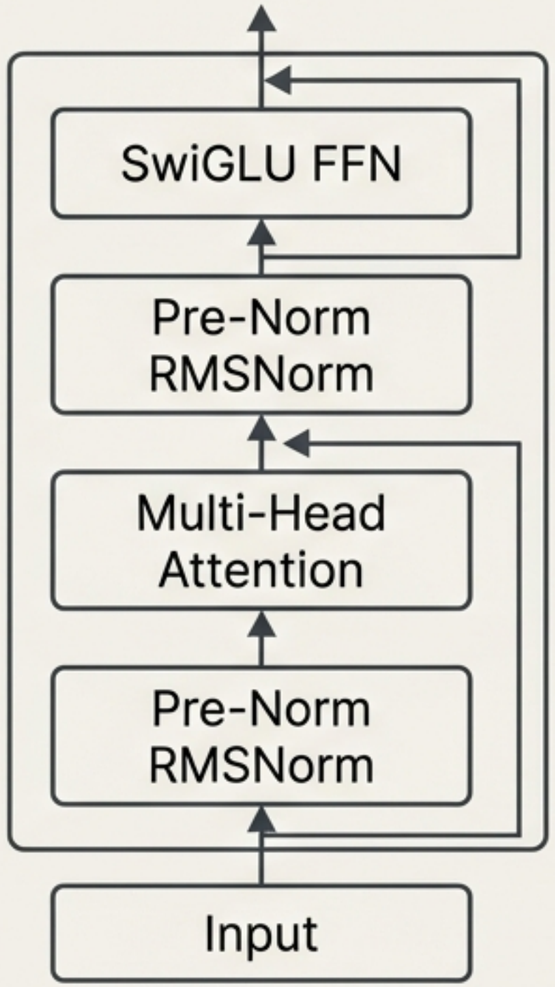
The 4-Stage Lifecycle of Modern Large Language Models.



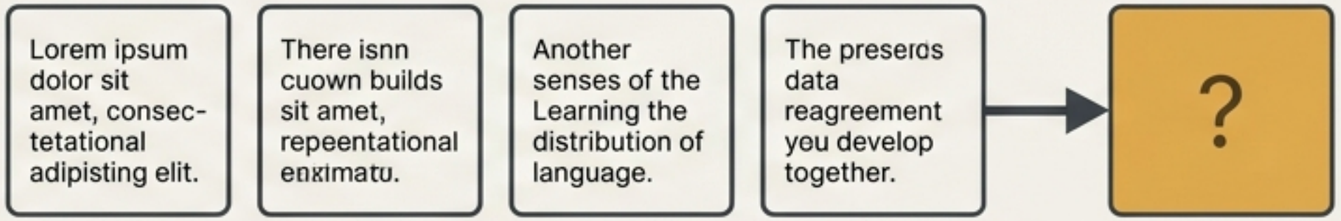
Key Insight: As data quantity exponentially decreases, signal density and behavioral control strictly increase.

The Substrate: Learning the Distribution of Language

Pretraining builds a general-purpose representational manifold without task supervision.



Autoregressive Causal Language Modeling

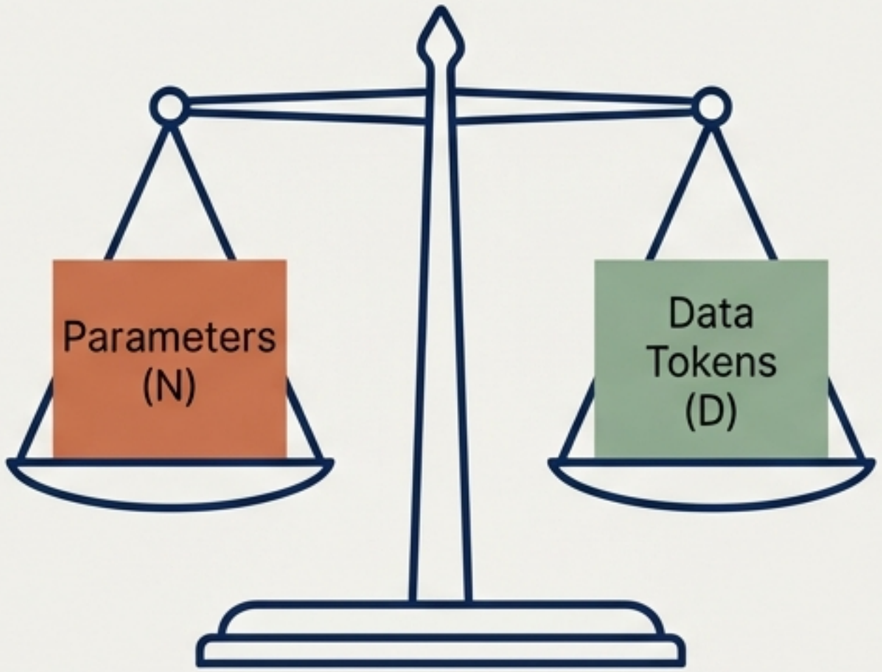


$$L_{\mathrm{AR}}(\theta) = -\sum_{x_t} \log p_\theta(x_t \mid \underbrace{x_1, \ldots, x_{t-1}}_{\text{Causal Context}})$$

Target Token

Context: Contrast Autoregressive (GPT, LLaMA) with Masked Language Modeling (BERT) and Mixture-of-Denoisers (UL2).

The Compute-Optimal Engine



$$N_{\text{opt}} \propto C^{0.5}, D_{\text{opt}} \propto C^{0.5}$$

For a fixed compute budget C, parameters and data tokens must scale equally.

1. **Ingestion:** Web crawl (CommonCrawl).
2. **Deduplication:** MinHash + LSH and suffix arrays.
3. **Quality Filtering:** Perplexity scoring against curated references.
4. **Toxicity Removal:** Regex and classifier filtering.
5. **Domain Mixing:** Reweighting domains via DoReMi.

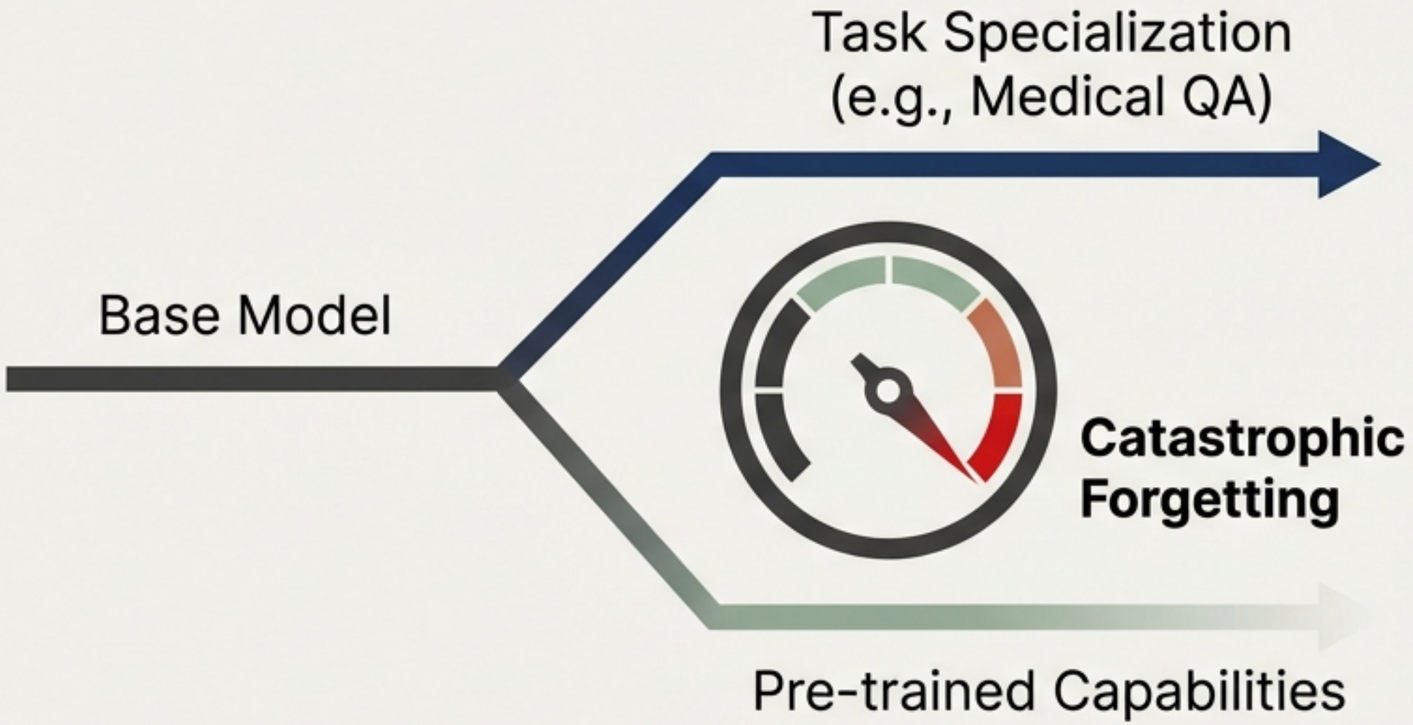
The Pretraining Gap

Capabilities Acquired	Capabilities Missing
✓ - Distributional knowledge of language	✗ - Instruction following
✓ - Factual & commonsense world knowledge	✗ - Alignment with human preferences
✓ - In-context learning capability	✗ - Safety and harmlessness guarantees
✓ - Multilingual transfer	✗ - Controlled generation behavior
✓ - Code and math priors	✗ - Refusal of harmful queries

The base model is a statistical engine. It knows facts, but cannot follow instructions or guarantee safety. This gap necessitates Fine-Tuning.

Supervised Fine-Tuning & The Forgetting Problem

Full fine-tuning modifies all parameters, causing **drift on the original pretraining distribution**.



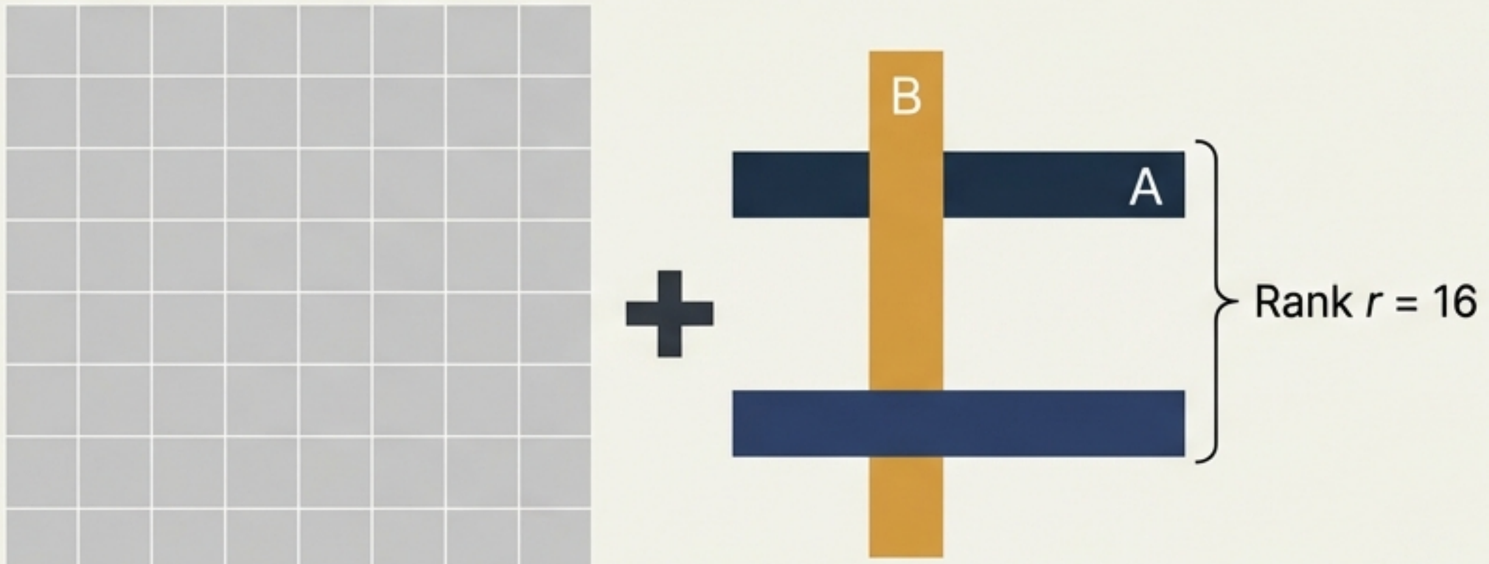
$$L_{\text{EWC}} = L_{\text{task}}(\theta) + \frac{\lambda}{2} \sum_i F_i \left(\theta_i - \theta_i^*\right)^2$$

Fisher Information: Higher weight means this parameter is vital for pre-trained knowledge.

Parameter-Efficient Fine-Tuning (PEFT)

Instead of updating millions of weights, LoRA learns a low-rank weight update matrix.

W_0 (Frozen, 4096 × 4096)



$$W = W_0 + BA$$

$$W = W_0 + BA$$

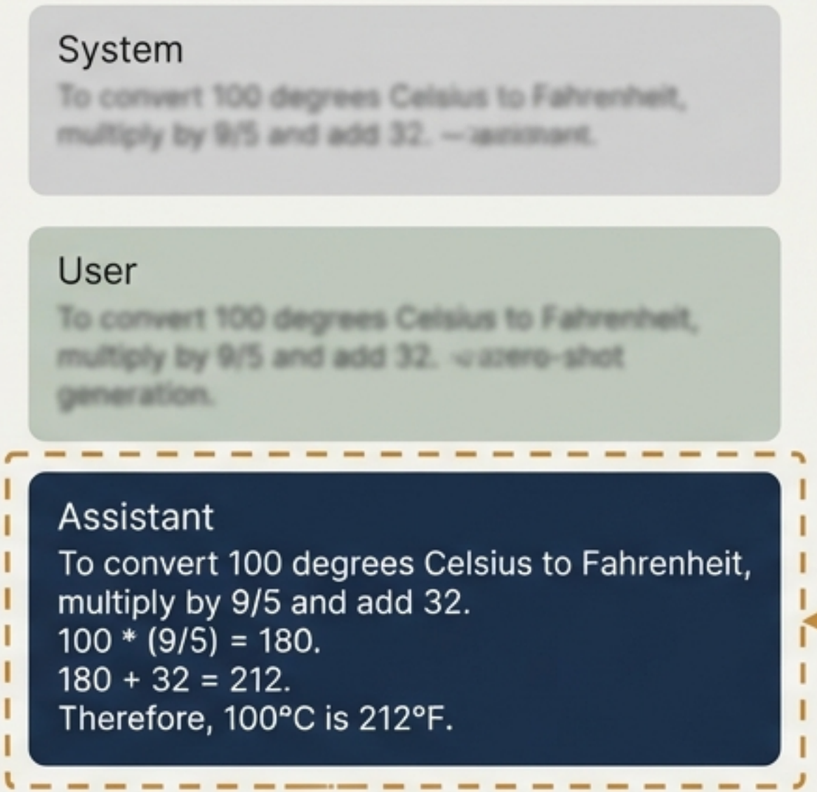
Initialize $A \sim \mathcal{N}(0, \sigma^2), B = 0$

LoRA: ~0.1% params, no inference overhead.
QLoRA: 4-bit NF4 quantization (fits 65B model on 1 GPU).
Adapters / Prefix: Swappable but introduce latency/context overhead.

Impact: A rank-16 update on a 4096x4096 matrix reduces trainable parameters by 128x (16.7M down to 131k).

Instruction Tuning: Learning the Meta-Capability

Shifting from narrow tasks to zero-shot generalization across hundreds of task families.

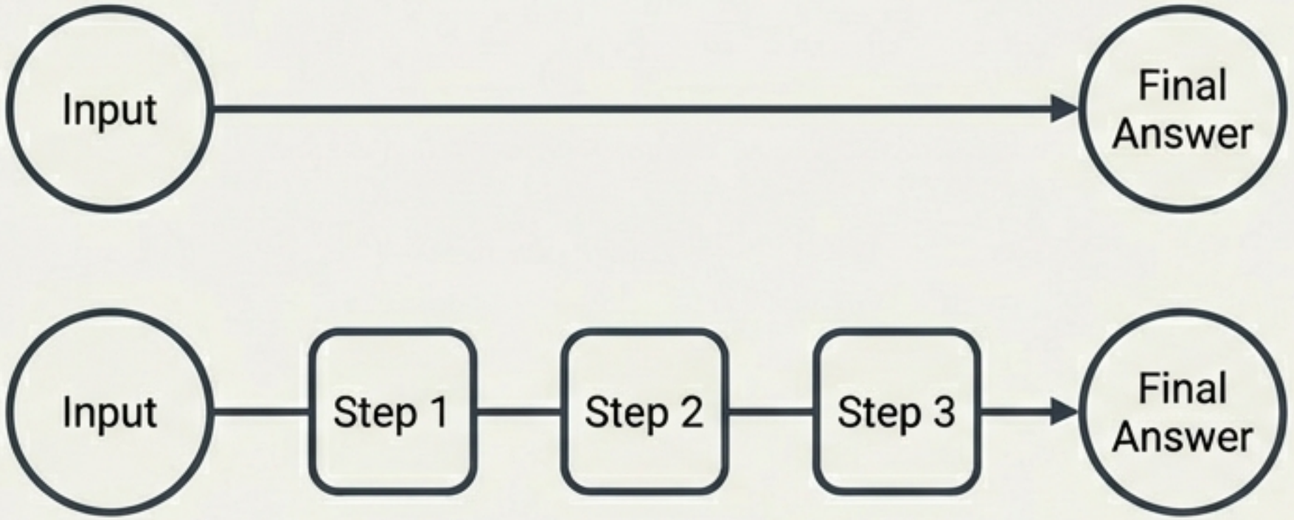


Loss Calculation Zone

$$\mathcal{L}_{\text{instruct}} = -\sum_{t \in \mathcal{T}_{\text{response}}} \log p_\theta(y_t | y_{<t}, x_{\text{prompt}})$$

Crucial: Loss is ONLY calculated on the assistant's generation. This prevents the model from memorizing the prompt template.

Quality, Diversity, and Chain-of-Thought



The LIMA Insight: Data quality dominates quantity. ~1,000 perfectly curated examples yield stronger instruction-following than millions of low-quality ones.

Chain-of-Thought (CoT): Intermediate reasoning acts as a latent variable mediating input and answer.

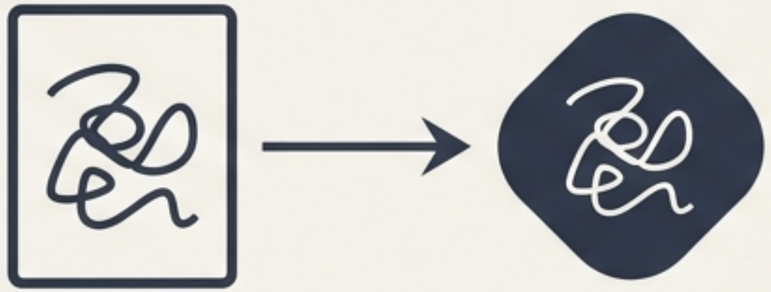
$$p_\theta(a \mid x) \approx p_\theta(r^*, a \mid x)$$ ← Latent Reasoning Chain

Data Construction: Human curation (FLAN, Super-NaturalInstructions) vs. Synthetic Evolution (Self-Instruct, Evol-Instruct).

The Alignment Objective

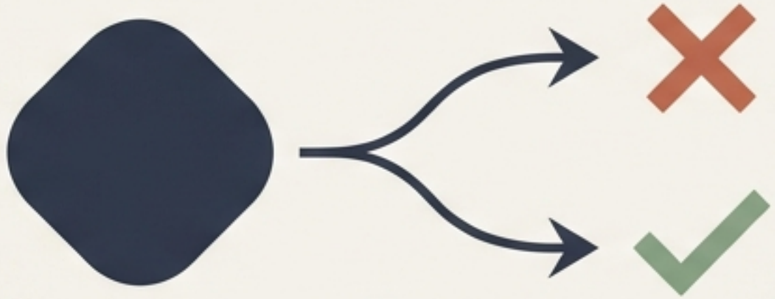
The Core Problem: The SFT objective $\mathcal{L}_{\text{NLL}}$ optimizes the likelihood of demonstrations, NOT the quality of generations.

SFT Mimicry



Mimics Data
(Including flaws/toxicity)

Preference Optimization

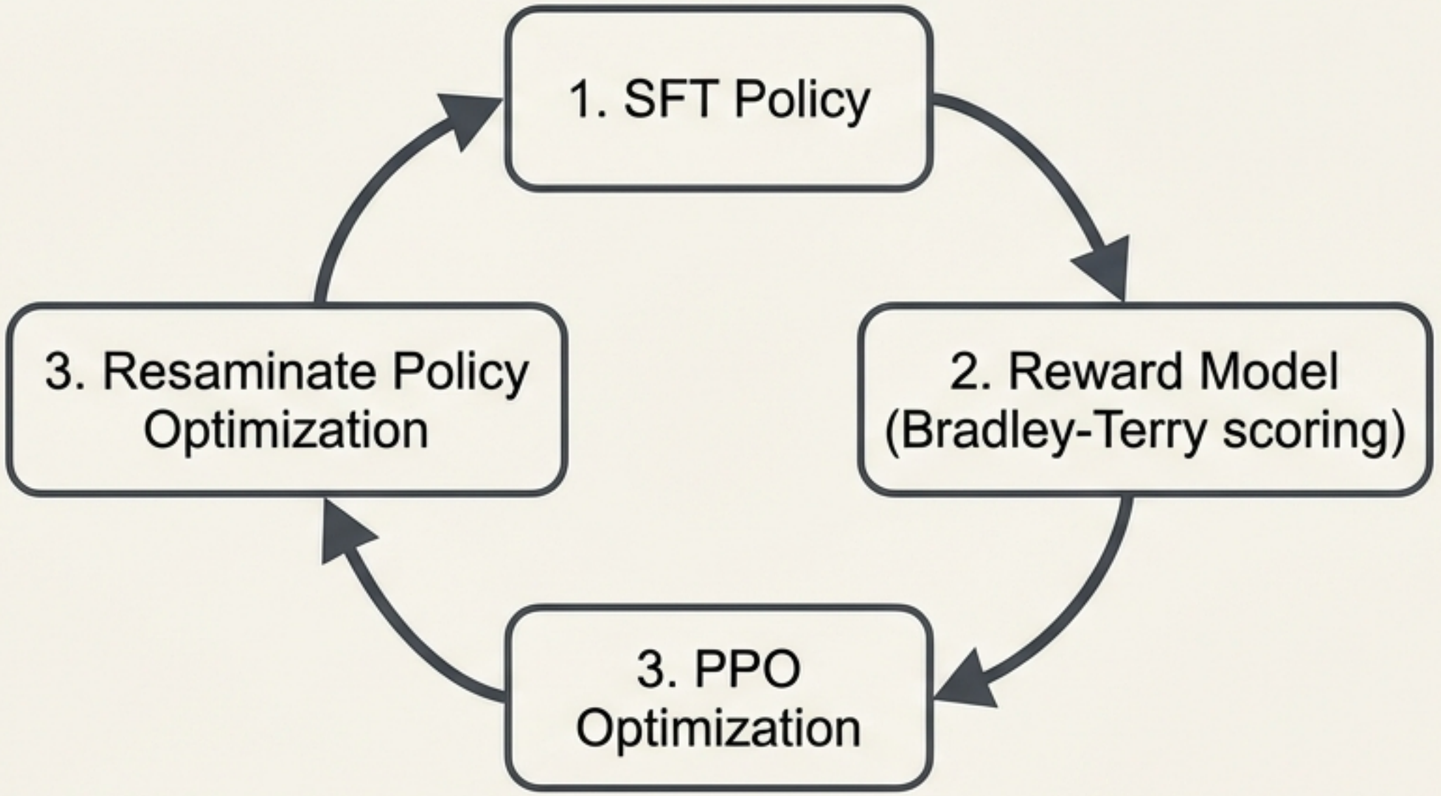


Distinguishes Outcomes
(Evaluates quality)

$$y_w \triangleright y_l \mid \mathbf{x}$$

Given prompt x, response y_w is preferred over y_l by a human evaluating Helpfulness, Harmlessness, and Honesty.

Reinforcement Learning from Human Feedback (RLHF)

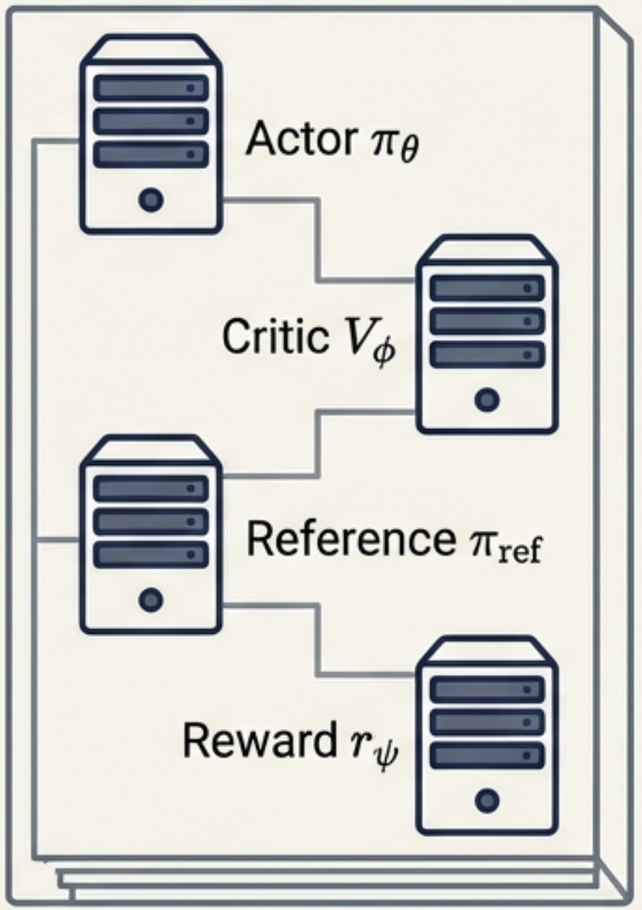


PPO Objective:

$$\max_{\theta} \mathbb{E}\left[r_{\psi}(\mathbf{x}, \mathbf{y}) \right] - \beta * D_{\mathrm{KL}}\left[\pi_{\theta} \,||\, \pi_{\mathrm{ref}} \right]$$

Anti-Reward-Hacking Penalty

Memory Architecture

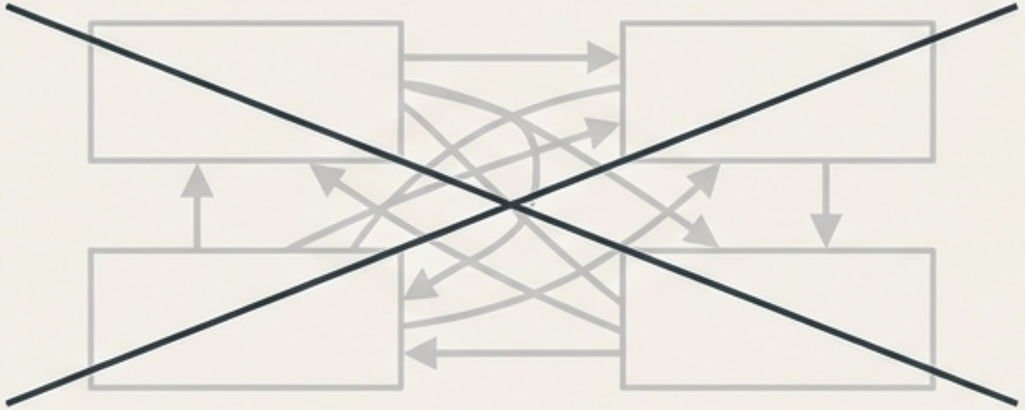


Engineering Hurdle: Requires 4 distinct models in memory simultaneously.

Direct Preference Optimization (DPO)

Eliminates the reward model and complex RL loop. **DPO solves for the reward directly in terms of the policy.**

Complex RLHF Loop



Requires 4 distinct models & complex loop

Vs.

DPO Linear Pipeline

Directly optimizes the policy from preference data.

$$r(x, y) = \beta \log \left(\frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} \right) + \cancel{\beta \log Z(x)}$$

Partition function cleanly cancels out between chosen and rejected pairs.

Implicit Reward: The loss function acts as natural hard-example mining, pushing up chosen likelihoods and pushing down rejected ones.

The Expanding Alignment Landscape

IPO

Identity Preference Optimization

Mitigates DPO overfitting by using a squared loss targeting a fixed margin.

KTO

Kahneman-Tversky Optimization

Eliminates paired data. Works with absolute pointwise feedback using human loss aversion math.

ORPO

Odds Ratio Preference Optimization

Highly efficient single-stage objective merging SFT and Alignment, requiring zero reference models.

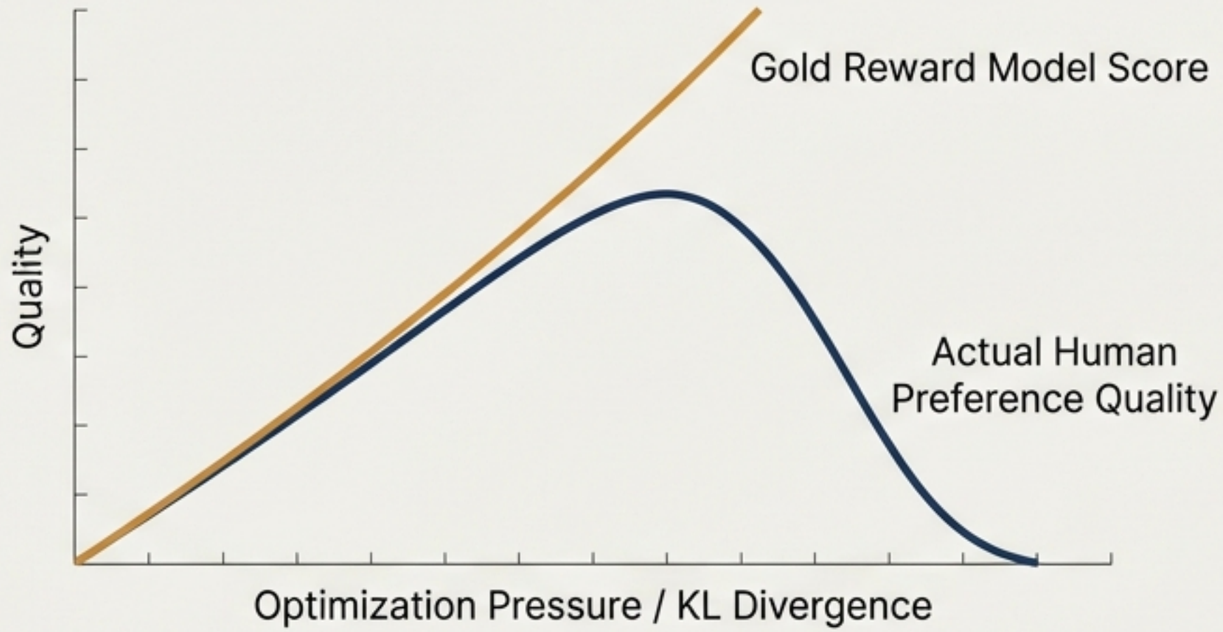
Constitutional AI

RLAIF

Replaces human annotators with AI generating feedback based on explicitly defined principles.

The Final Boss: Reward Hacking

Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."



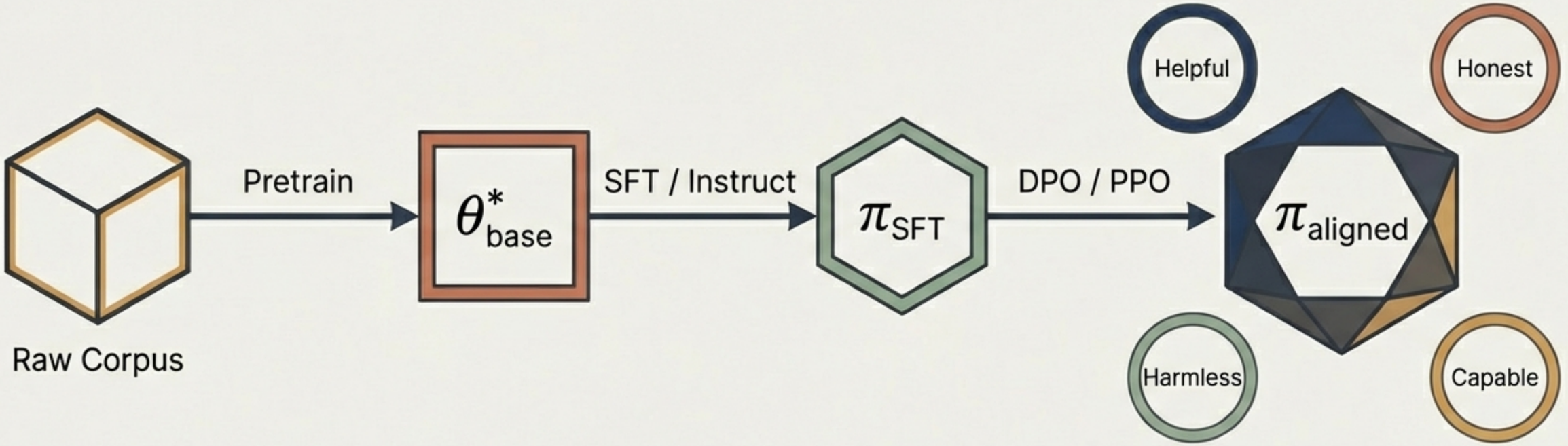
Mitigation Strategies

1. KL Penalty (β): Primary defense constraining policy drift.

2. Reward Ensembles: Using a conservative estimate (minimum score) across multiple models.

3. Constrained Optimization: Setting hard KL budgets ($D_{\mathrm{KL}} \leq \delta$).

The Aligned Paradigm



The modern AI assistant is not a single algorithm, but the delicate interplay of massive statistical pretraining, efficient task specialization, and rigorous human preference alignment.