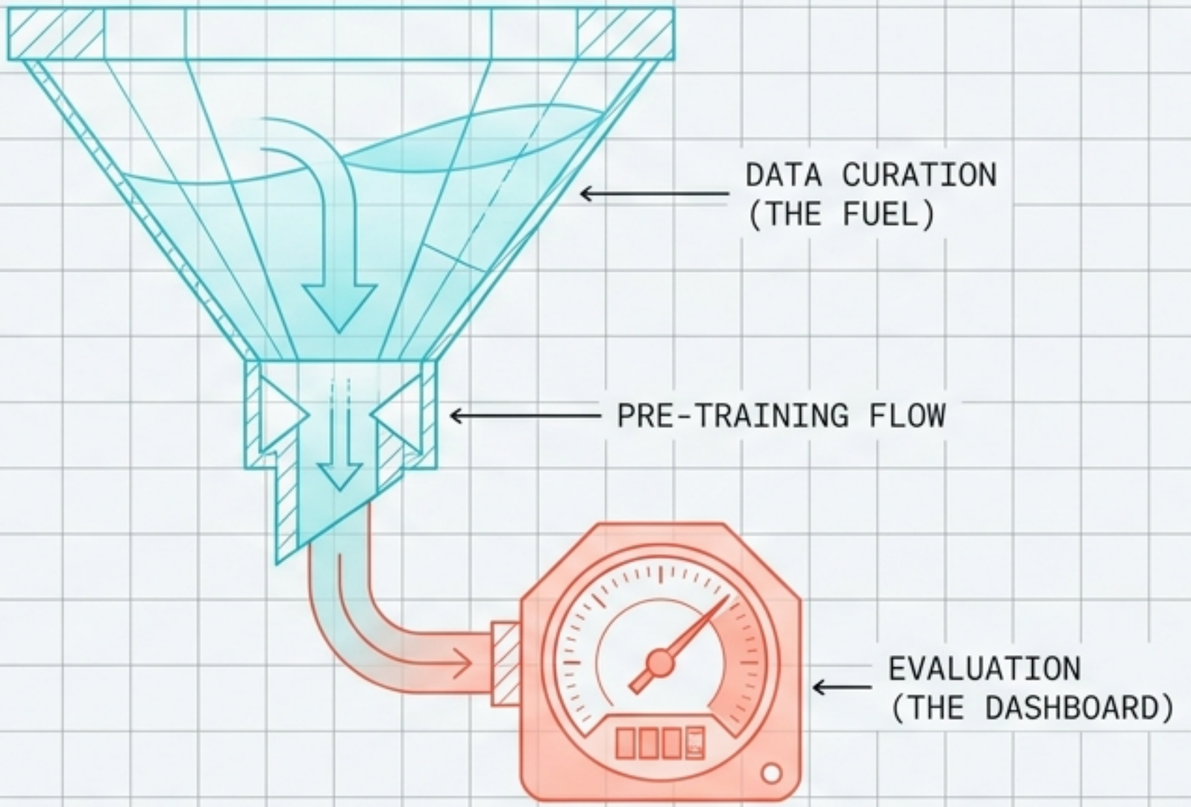



# Refining Intelligence: LLM Pre-Training Data and Evaluation

Distilled from Lecture 2 by Reinhard Heckel

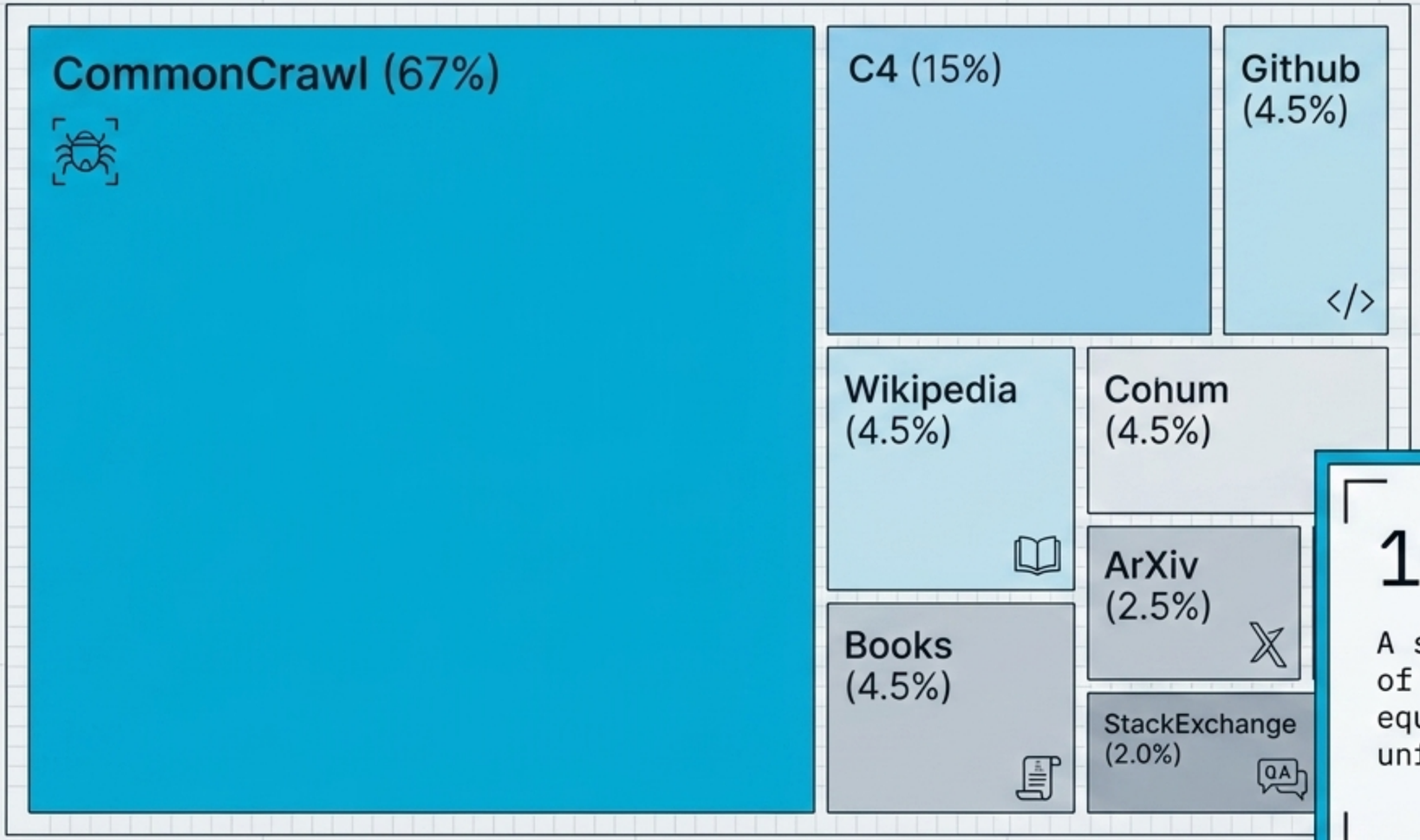
## Pre-Training (The Heavy Lifting)

| | |
|---|---|
| Method | Next-token prediction |
| Input Data | Massive volumes of high-quality, web-scale text |
| Outcome | Base Models (e.g., GPT, LLaMA) |

## Post-Training (The Precision Tuning)

| | |
|---|---|
| Method | Next-token prediction + Reinforcement Learning |
| Input Data | Curated, low-volume, high-quality examples & comparison/reward data |
| Outcome | Agentic & Chat Models (e.g., ChatGPT, Claude) |

# The Raw Material: Where Does the Data Come From?




## 100 Terabytes

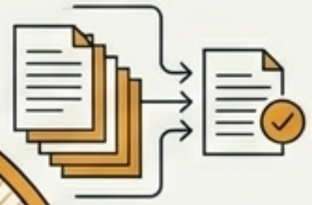
A single snapshot (Feb/March 2024) of compressed CommonCrawl data equals roughly 100TB of raw, unfiltered internet noise.

# The Curation Imperative

Note: Proprietary datasets (OpenAI, Anthropic) remain closed. Data purity is the ultimate competitive advantage, compounded by copyright risk.

## 1. Filtering

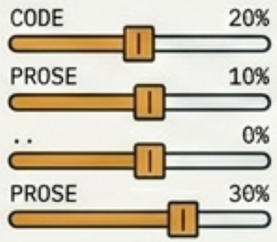
Stripping away low-quality, irrelevant, or toxic web text.



## 2. Deduplication

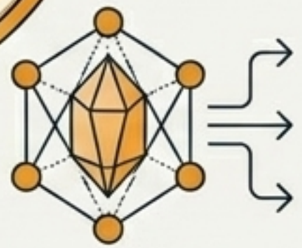
Preventing models from mechanically memorizing repeated documents.

Raw web data contains spam, toxic content, and extreme repetition.




## 3. Mixing & Weighting

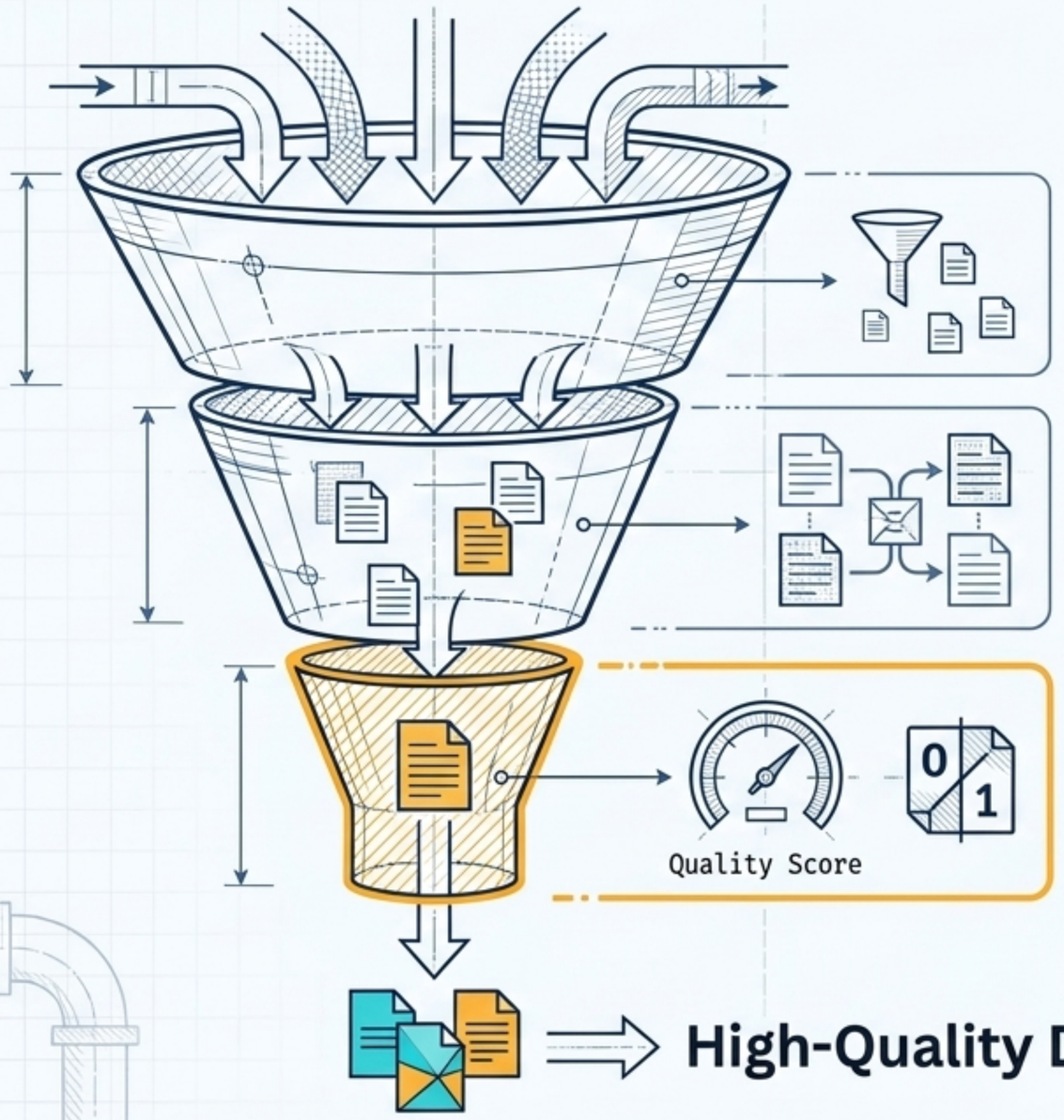
Engineering the exact ratio of data sources (e.g., code vs. prose).



## 4. Synthetic Generation

Using AI to manufacture mathematically perfect reasoning traces.

# The DCLM-Baseline Pipeline: Separating Signal from Noise




## Stage 1: Heuristic Cleaning

Rule-based filtering. Removes URLs, weird word lengths, extreme repetition, and non-English text.

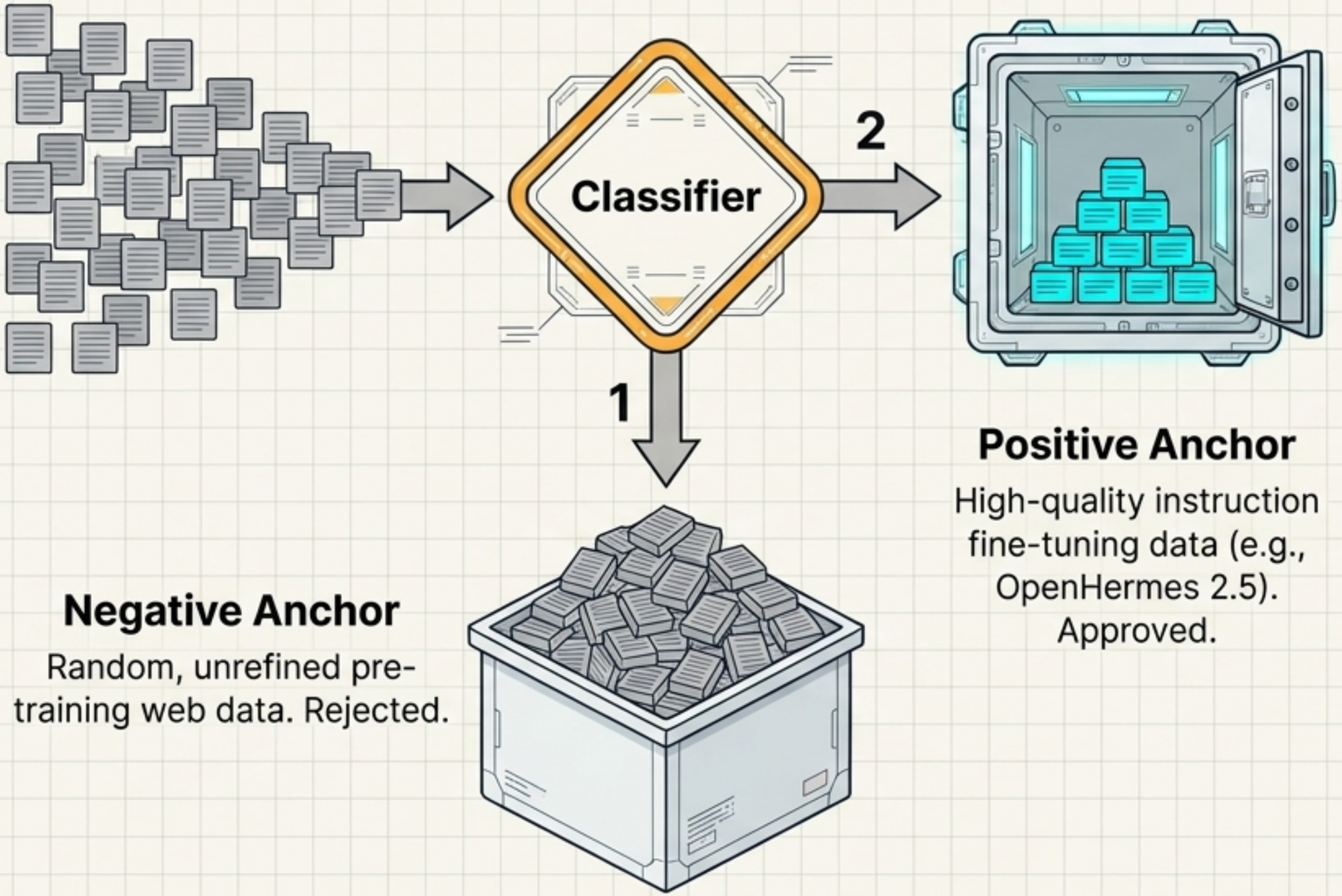
## Stage 2: Deduplication

Utilizes Bloom filters to ensure unique data, dropping overall volume dramatically.

## Stage 3: ML-Based Quality Filtering

(The Differentiator). Unlike older datasets that stop at heuristics, heuristics, state-of-the-art pipelines use binary classification to extract only the absolute highest quality data.

# Spotlight: ML-Based Quality Filtering




**Positive Anchor**

High-quality instruction fine-tuning data (e.g., OpenHermes 2.5). Approved.

**Negative Anchor**

Random, unrefined pre-training web data. Rejected.

## The 10% Rule

Retain only the top 10% of web data that most closely resembles structured human reasoning and instruction.

Modern ML models are now essential for curating the training data of future ML models.

# The Modern Data Recipe: Beyond Web Scraping

The era of pure web-scraping is over. Modern pre-training requires a deliberate, multi-source mix.

### High-Quality Web

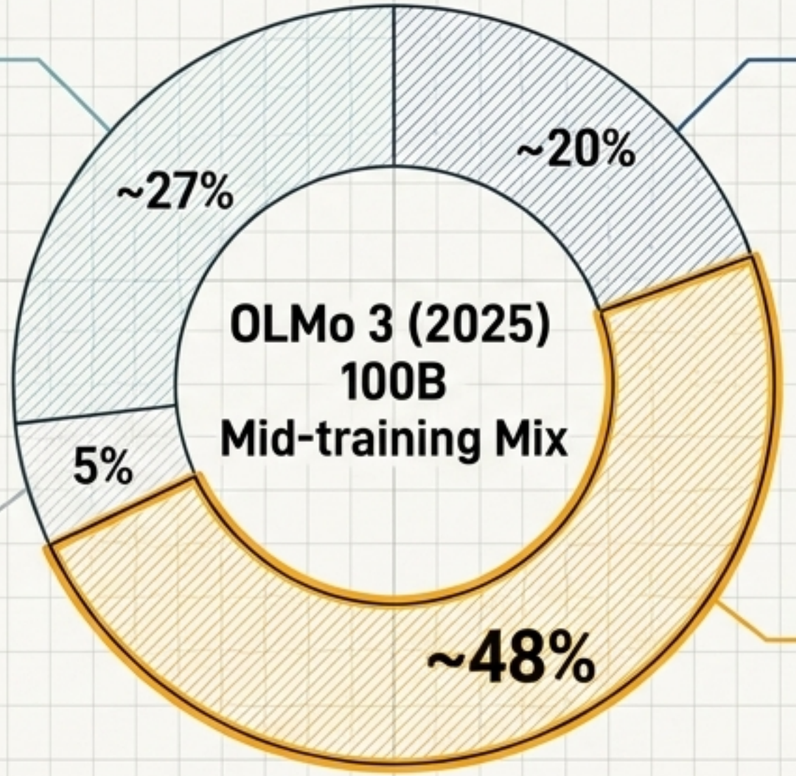
~27%

(Common Crawl & STEM Crawl)

### Code & Logic

~20%

(StackEdu, CraneCode)




### Academic PDFs

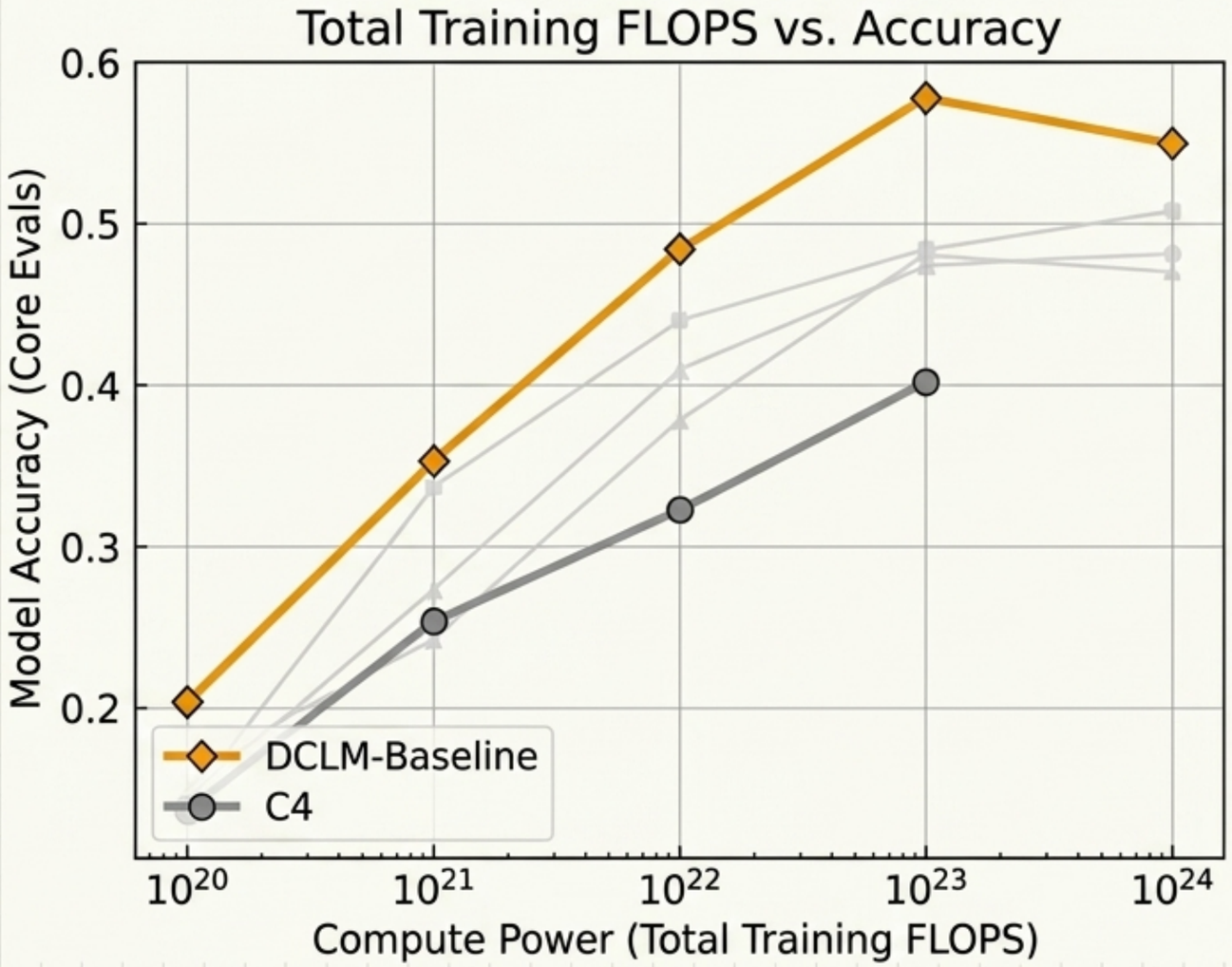
5%

(OLMOCR)

### The Rising Star: Synthetic Data

AI-generated datasets covering Math (TinyMATH), QA (Wiki To RCQA), and complex "Thinking/Reasoning Traces" (QWQ, Llama Nemotron).

# The ROI of the Refinery: Better Data = Cheaper Training




**Key Insight**

Models trained on highly refined data (DCLM-Baseline) achieve significantly higher accuracy at the exact same computational cost compared to basic heuristic data (C4).

Investing heavily in pre-training data curation mathematically offsets the need for more expensive, compute-heavy model training runs.

# Transitioning to the Dashboard: Why Evaluate?

## Model Selection

Guiding individuals and enterprises on which AI tool fits their specific economic and technical needs.

## Measuring Progress

Providing the core scientific yardstick for the machine learning community.

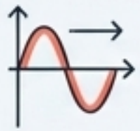
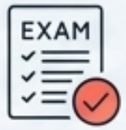
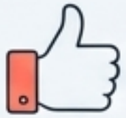
## Developer Guidance

Isolating the impact of architecture changes versus training data improvements.

## Safety & Risk Assessment

Verifying alignment, boundaries, and potential hazards before final deployment.

# The Evaluation Matrix

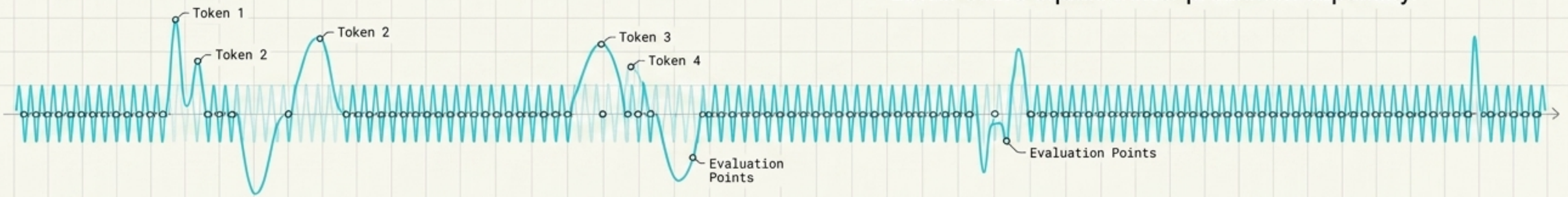
|  **Perplexity** |   **Verifiable Benchmarks** |  **LLM-as-a-Judge** |  **Human Preference** |
|---|---|---|---|
| **Mechanism:** Next-token prediction fit. | **Mechanism:** Standardized tests with verifiable answers. | **Mechanism:** Using an AI against a rubric for open-open-ended tasks. | **Mechanism:** Blind A/B testing by human users. |
| **Cost/Signal:** Cheap, highly dense signal. | **Cost/Signal:** Moderate cost, specific capability signal. | **Cost/Signal:** Scalable but complex. | **Cost/Signal:** Expensive, slow, but highest ground-truth value. |
| **Flaw:** Measures prediction, not true capability. | **Flaw:** Susceptible to format memorization. | **Flaw:** Constrained by the judge model’s own limits. | **Flaw:** Subjective. |

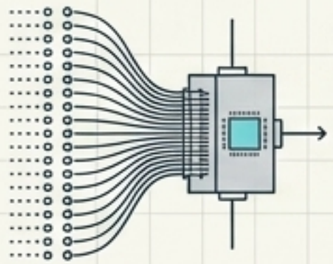
# Perplexity: The Foundation, Not the Ceiling

$$\exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(t_i)\right)$$

Probability of the i-th token according to the model.

Lower score equals better prediction capability.



## The Good (Signal Density)

Provides an incredibly dense, cheap signal because every single generated token is mathematically evaluated against the training set.

## The Bad (The Illusion of Intelligence)

A measure of pure statistical prediction. A model with low perplexity might predict the most likely next word perfectly but fail entirely at logical reasoning or truthfulness.

# Verifiable Benchmarks: The Standardized Tests of AI

Exam Card

## MMLU (Broad Knowledge)

**Format:** 5-shot multiple choice across humanities, sciences, and medicine.

In the complex z-plane, the set of points satisfying the equation z^2 = |z|^2 is a

(A) pair of points

(B) circle

(C) half-line

(D) line ✔

✔

Exam Card

## GSM8K (Reasoning)

**Format:** Multi-step grade school math word problems.

Forces the model to generate intermediate logical steps before reaching the final verifiable number.

**Problem:** A farmer has 10 cows and 20 chickens. How many legs are there in total?

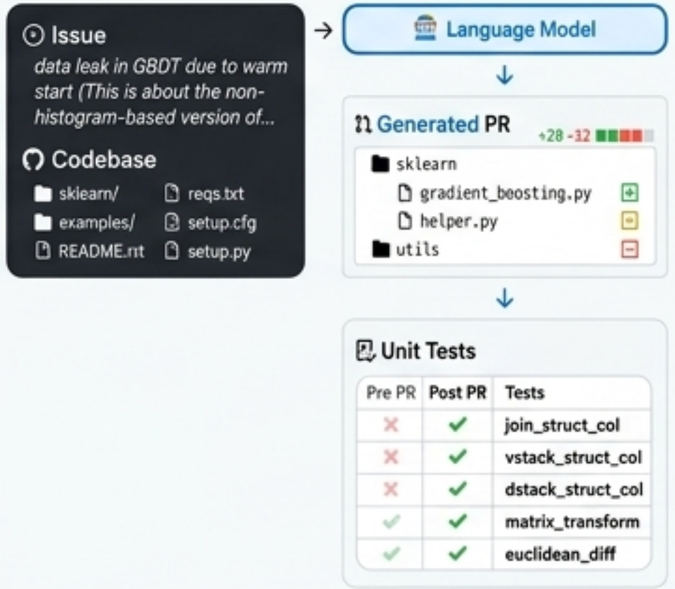
**Solution:**
Cows have 4 legs each, so 10 * 4 = 40 legs.
Chickens have 2 legs each, so 20 * 2 = 40 legs.
Total legs = 40 + 40 = 80.
The answer is 80.

#### 80

Exam Card

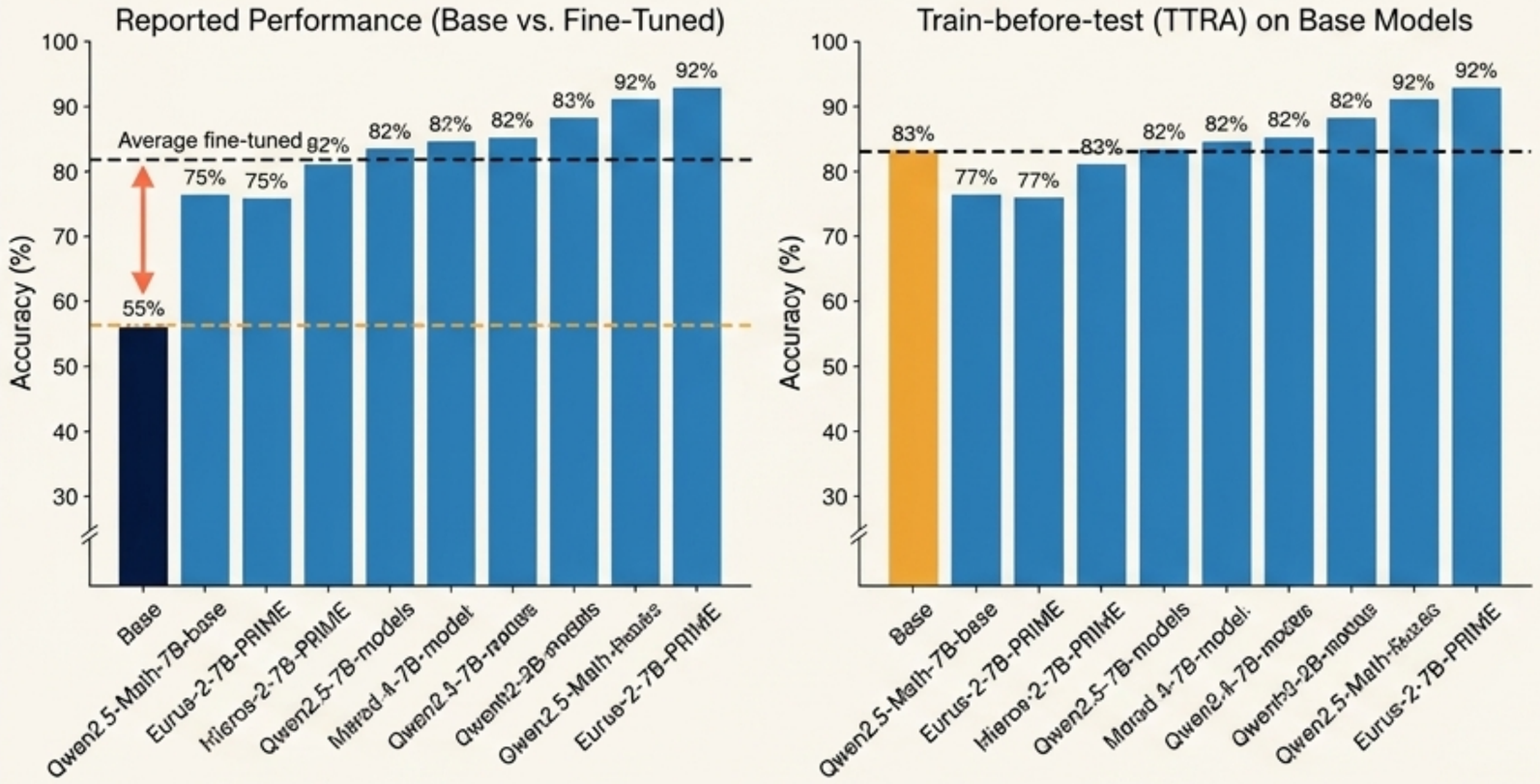
## SWE-Bench (Agentic Execution)

**Format:** The model is given a GitHub issue and an entire codebase.



# The Confounder: Task Familiarity (2026 Research)

## Reported Performance vs. Train-before-test




## The Illusion

The literature often reports massive accuracy gaps between "Base Models" and their "Fine-Tuned" variants on benchmarks, suggesting RL/SFT creates a massive leap in underlying intelligence.

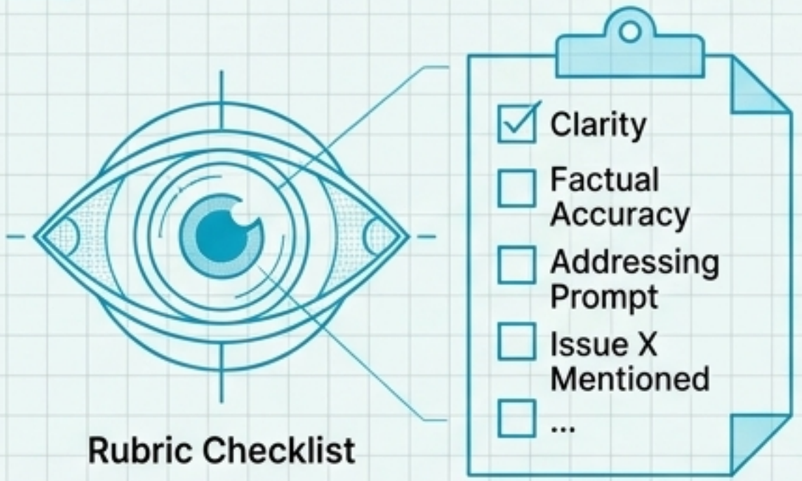
## The Reality

If Base Models are simply aligned to the format of the test task beforehand, their performance jumps to nearly match the fine-tuned models.

**Key Takeaway:** High benchmark scores are heavily confounded by format familiarity. Much of what we call "reasoning gains" are merely artifacts of knowing how to take the test.

# Judging the Open-Ended: Machines vs. Crowds

## How do we evaluate tasks without objective right answers?

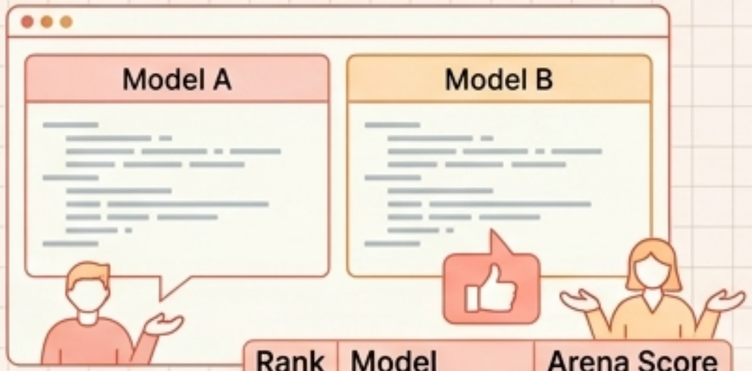



Rubric Checklist

### LLM-as-a-Judge

Models evaluate other models by decomposing quality into explicit rubric criteria (e.g., “Was issue X mentioned?”).

**Limit:** Always constrained by the judging model’s maximum capability.




| Rank | Model | Arena Score |
|---|---|---|
| 1 | Gemini-Pro | 1448 |
| 2 | GPT-4 | 1411 |

### Chatbot Arena (Human Preference)

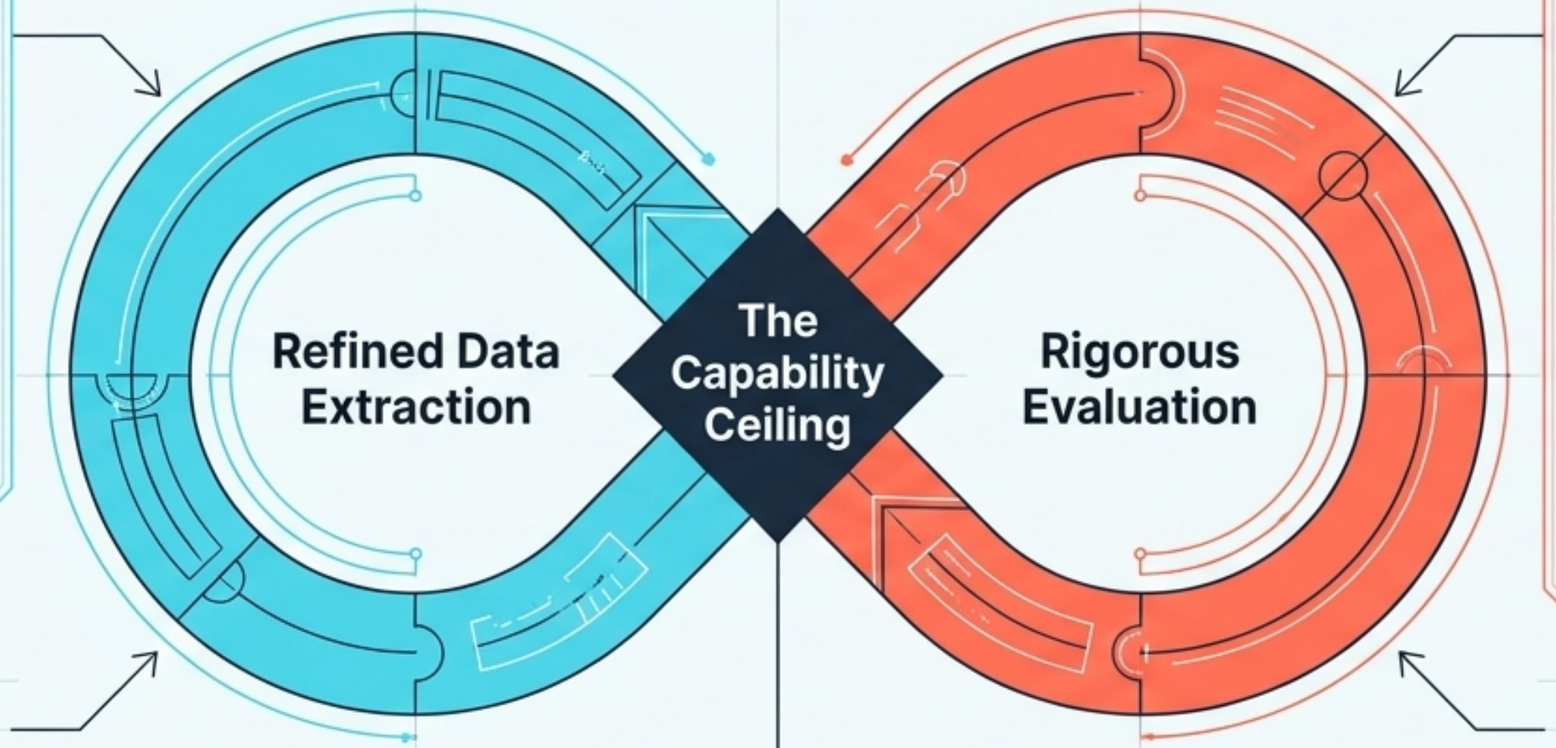
Users submit open queries, receive two blind model responses, and vote on the better output.

Generates **crowdsourced ELO ratings** based on **millions of blind A/B human tests.**

# Executive Synthesis: The Feedback Loop

### The Engine Limits the Car

Pre-training data is the definitive ceiling on model capability. Advanced architectures cannot overcome poor data purity.




### The Dashboard is Flawed, But Necessary

Benchmarks are crucial for comparative progress but must not be treated as absolute measures of intelligence due to task familiarity and format contamination.

### Conclusion

True advancement in AI requires co-evolving both the purity of the pre-training data mix and the nuance of human-aligned, verifiable evaluation.