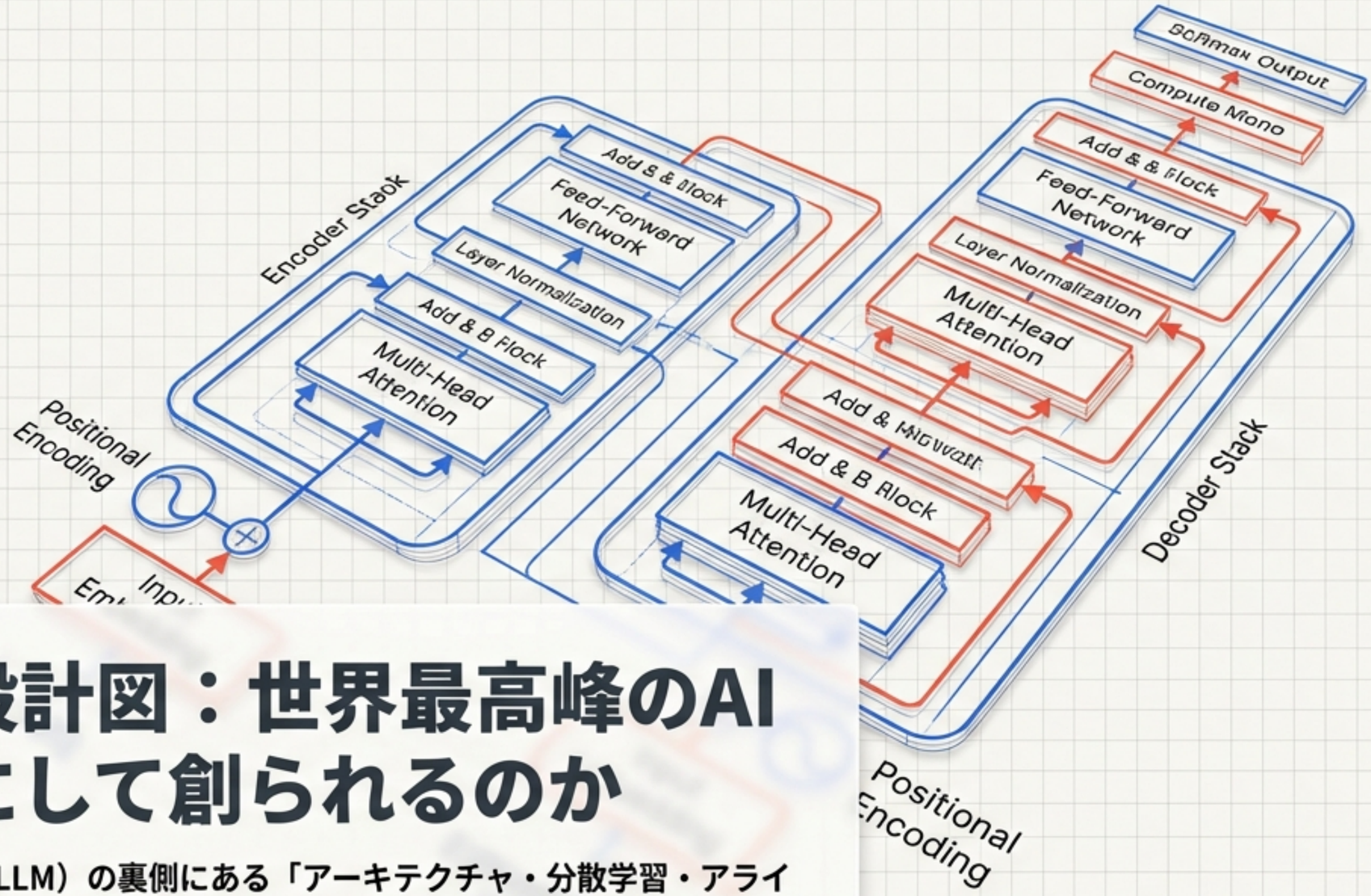


知能の設計図：世界最高峰のAIはいかにして創られるのか

大規模言語モデル（LLM）の裏側にある「アーキテクチャ・分散学習・アライメント」の全貌を解き明かす、とてつもなく熱い完全理解マスタークラス。

汎用人工知能への4つの進化フェーズ

Phase 1: The Engine & The Fuel

アーキテクチャとデータ。
Transformerの構造と、
世界中の知識を圧縮する
トークン化プロセス。

Phase 2: Breaking the Limits

分散学習とインフラ。
単一GPUの物理的限界を
突破する「3D並列化」と
「ZeRO最適化」。

Phase 3: The Awakening

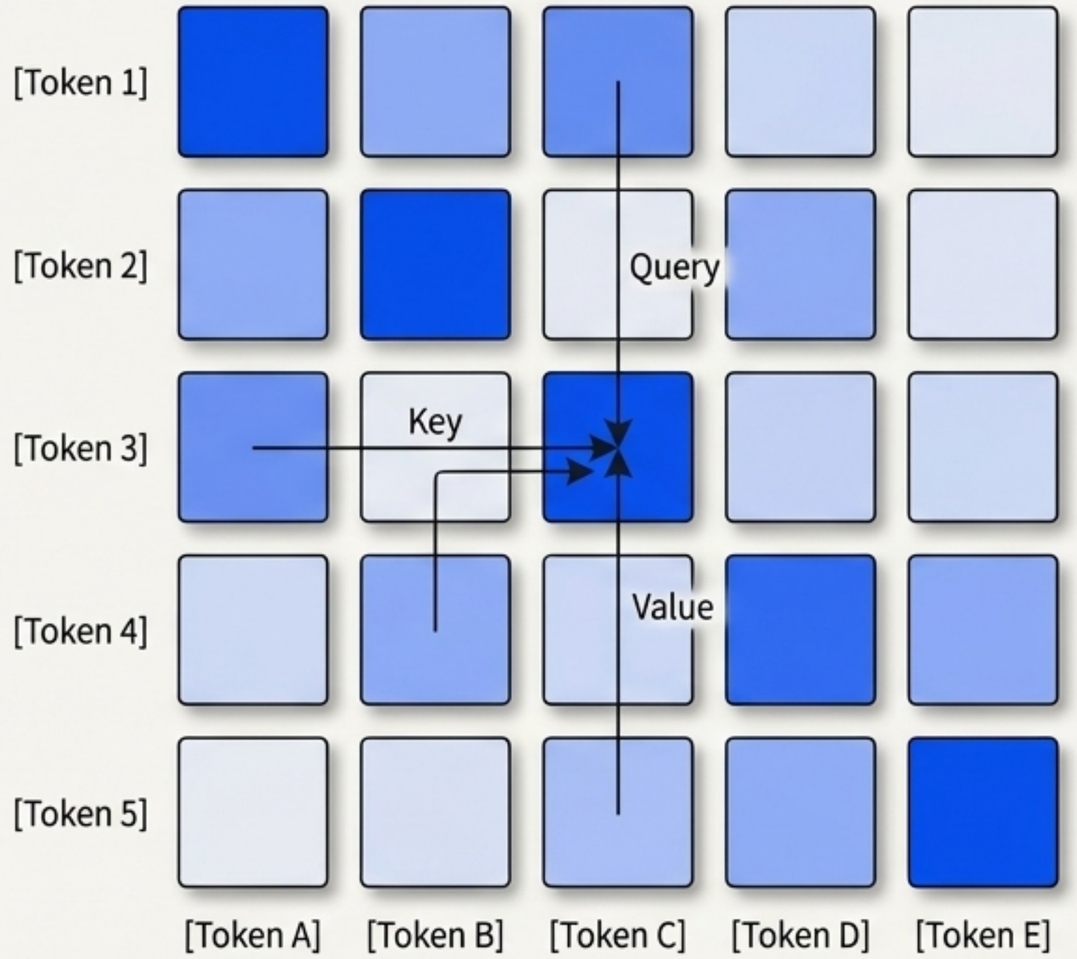
事後学習 (Post-training)。
「単語予測器」から「賢い
アシスタント」へ覚醒させ
るSFTとアライメント。

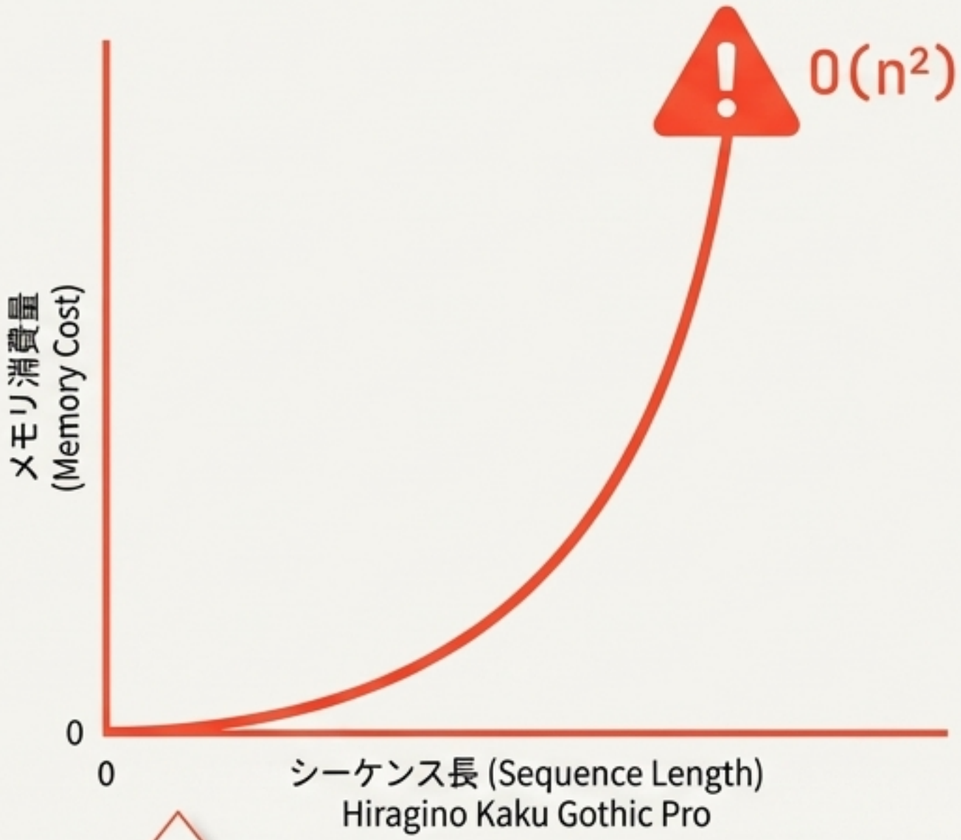
Phase 4: The Human Interface

プロンプティングと評価。
AIの思考を操る技術と、
超知能を評価する
LLM-as-a-Judge。

コアエンジン：文脈を理解する「Attention」とその代償

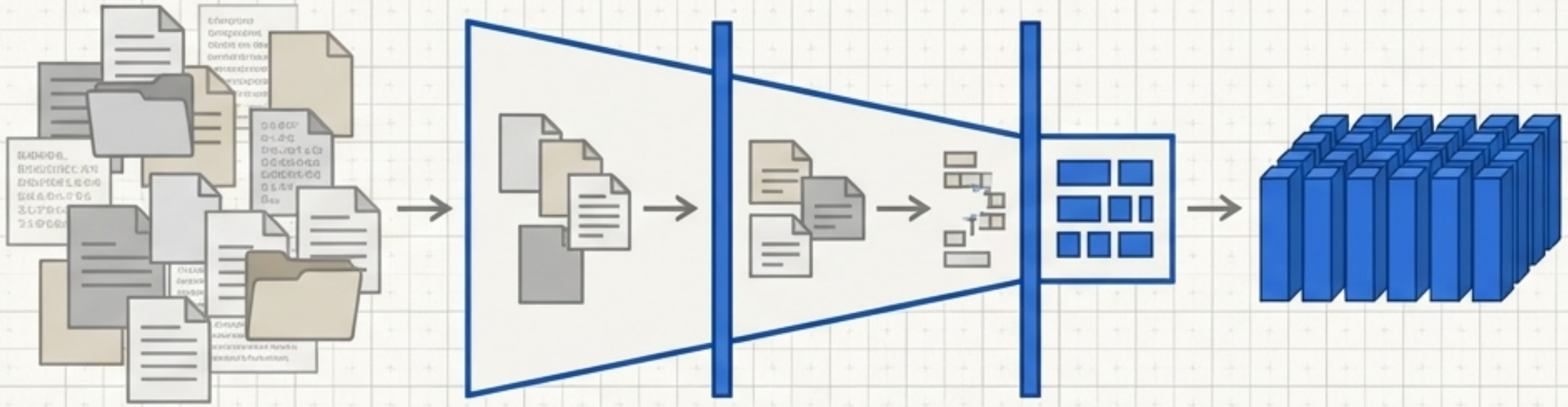
Attentionの魔法：すべての単語が相互作用し、文脈を動的に把握する。





O(n²) の呪縛：長文処理において計算量とメモリ消費が「二乗」で爆発。最大の物理的ボトルネック。

知能の燃料：徹底的なノイズ除去と「段階的学習」



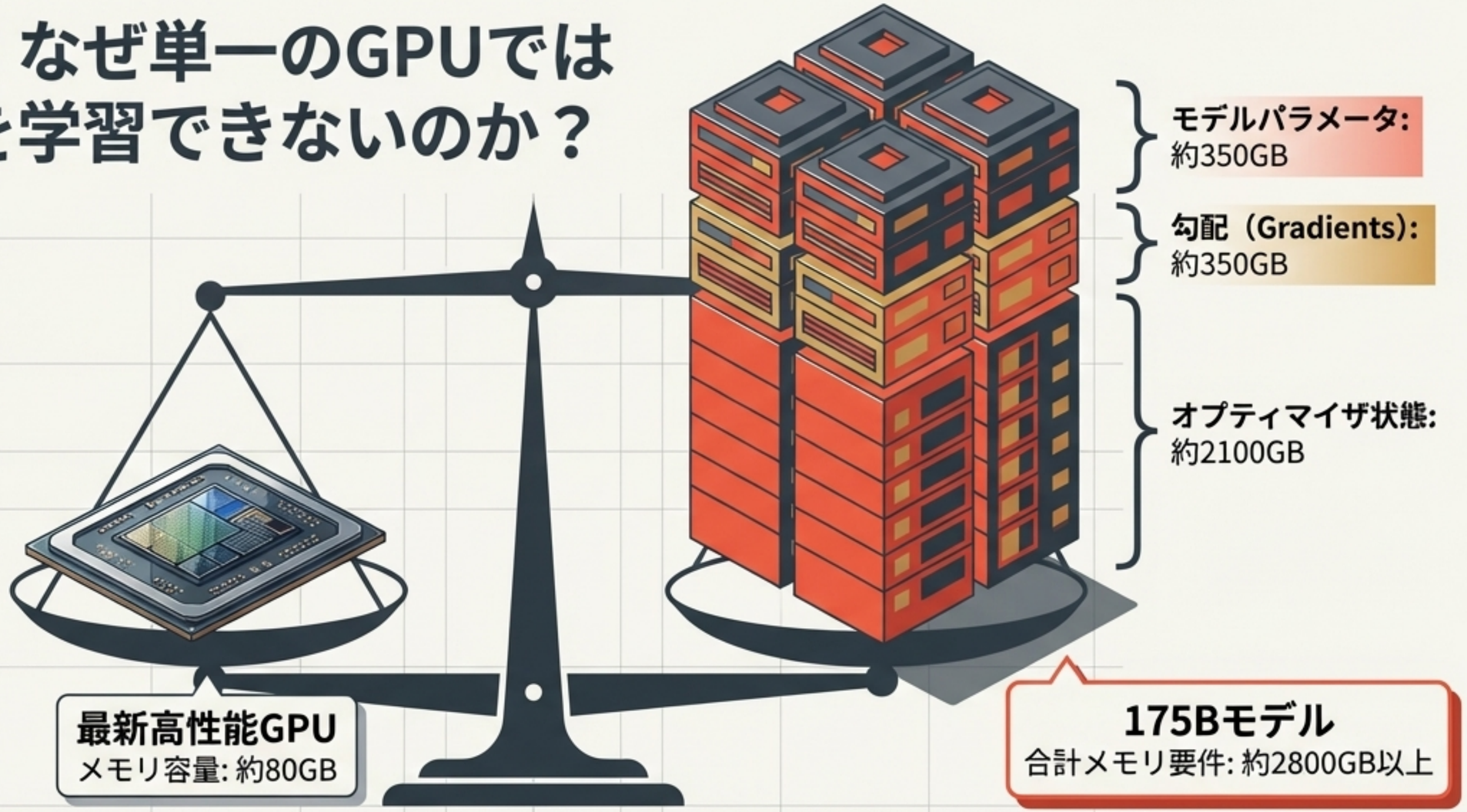
冗長なデータをLSHBloom等で徹底排除し、SentencePieceで語彙爆発を防ぐ。

Stage 1: 一般知識の学習
(短いコンテキストで計算コストを抑制)

Stage 2: 推論能力の強化
(数学やコードデータの比率を上げる)

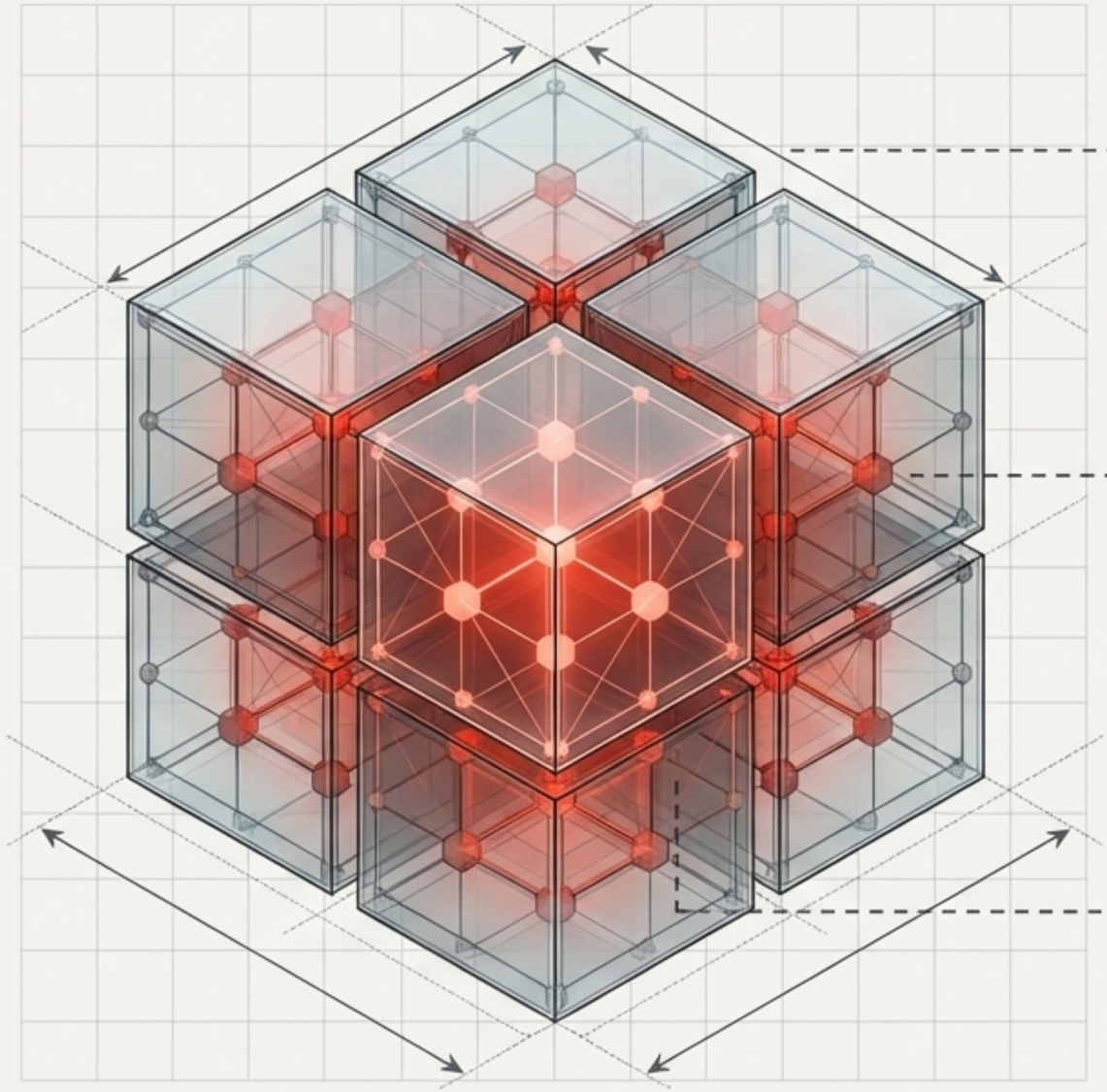
Stage 3: 長文コンテキストへの拡張
(全体のわずかな学習量で実現)

物理的な壁：なぜ単一のGPUでは巨大モデルを学習できないのか？



現実：単一のGPUには到底収まらない。ここから「分散学習」の魔法が必須となる。

限界突破のエンジニアリング：「3D並列化」による分割統治



Data Parallelism
（データ並列 - Global）

モデルのコピーを複数配置し、異なるデータを同時に学習。しかしメモリ不足は解決しない。

Tensor Parallelism
（テンソル並列 - Intra-node）

行列計算そのものを分割。1つの巨大な計算を複数のGPU（NVLink接続）で分担する。

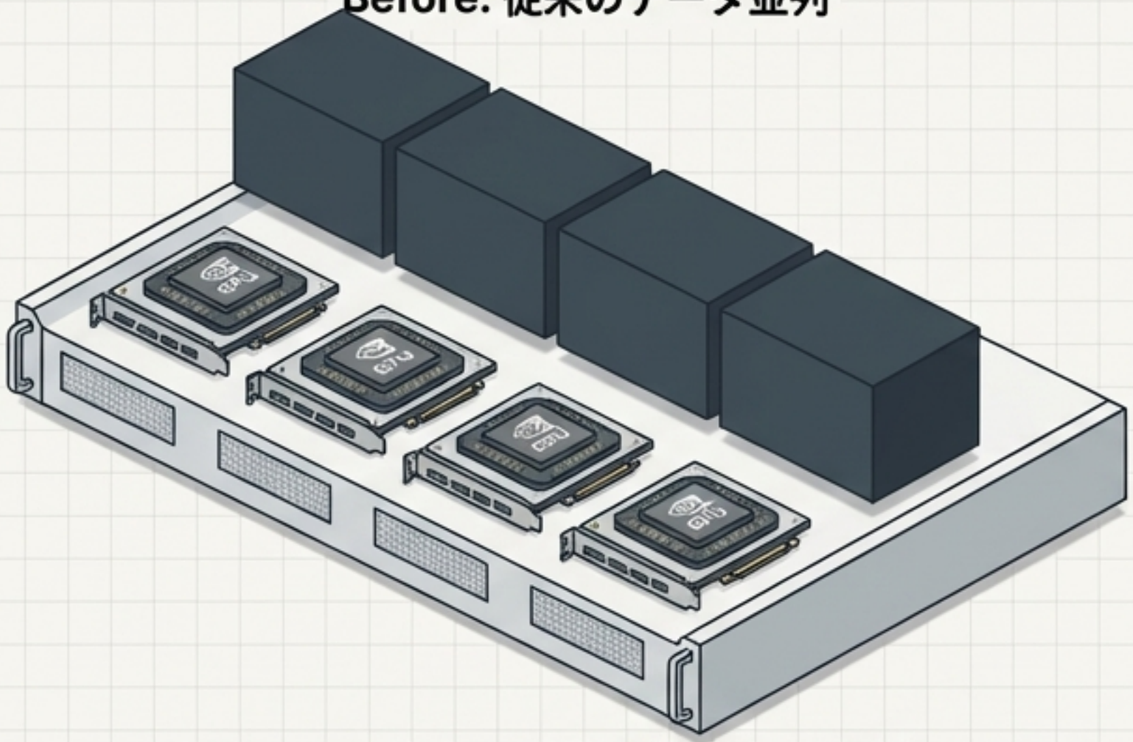
Pipeline Parallelism
（パイプライン並列 - Inter-node）

ニューラルネットワークの「層」を分割。データをバケツリレーのように次々と処理する。

The Architectural Dossier

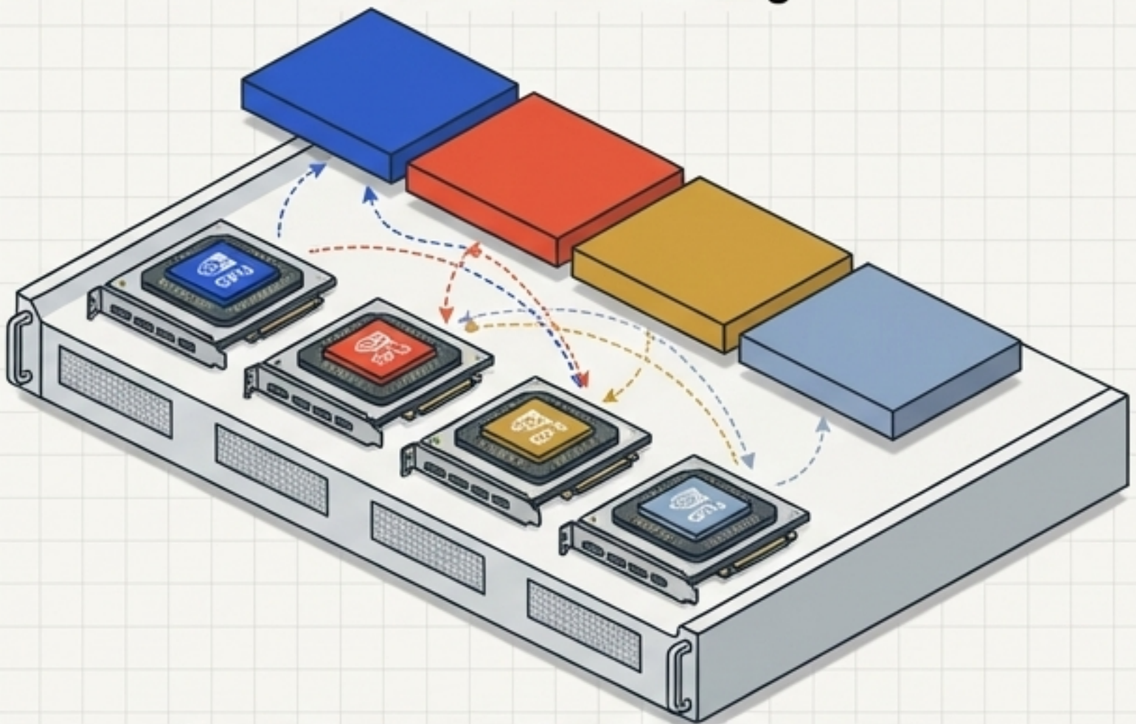
メモリの魔法：冗長性を完全に排除する「ZeRO」最適化

Before: 従来のデータ並列



無駄なコピー：全GPUが同一の重いデータを保持。

After: ZeRO Sharding



冗長性の排除：必要な瞬間にだけGPU間で通信し、終われば即座に破棄。

ZeRO Stage 1: オプティマイザ状態を分割 (メモリ消費1/4へ)

ZeRO Stage 2: さらに勾配を分割 (メモリ消費1/8へ)

ZeRO Stage 3: パラメータ自体も分割 (GPU数に比例して削減)

覚醒の瞬間：事前学習から指示チューニング（SFT）へ

Pre-training (事前学習)

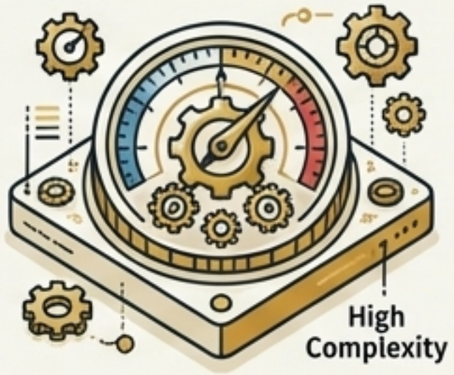
- インターネットの膨大なテキストを読み込む
- 「次の単語を予測する」能力を獲得
- 世界をシミュレートするが、質問には答えられない

Supervised Fine-Tuning (SFT)

- 高品質な「プロンプトと回答」のペアを学習
- Behavior Cloning（行動模倣）の実行
- 人間が望むアシスタントとしての振る舞いが覚醒

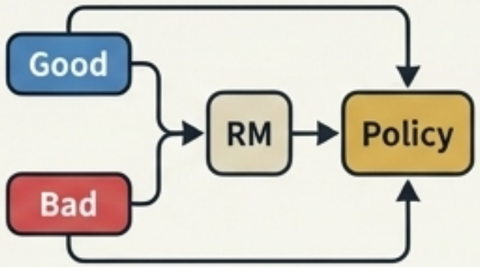
アライメントの進化：
AIを人間の価値観に適合させる技術

RLHF (InstructGPT等)



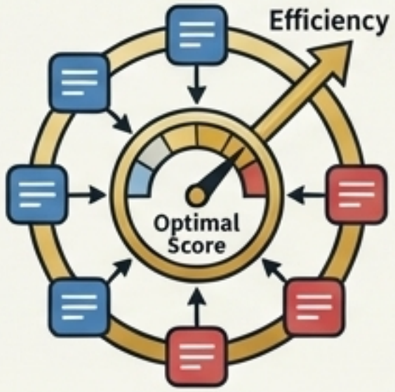
人間の評価に基づく「報酬モデル(RM)」を構築し、PPOアルゴリズムで最適化。強力だが、計算コストと複雑さが極めて高い。

DPO (Llama 3等)



報酬モデルをスキップし、「良い回答と悪い回答のペア」から直接ポリシーを最適化。オープンソースで爆発的に普及。

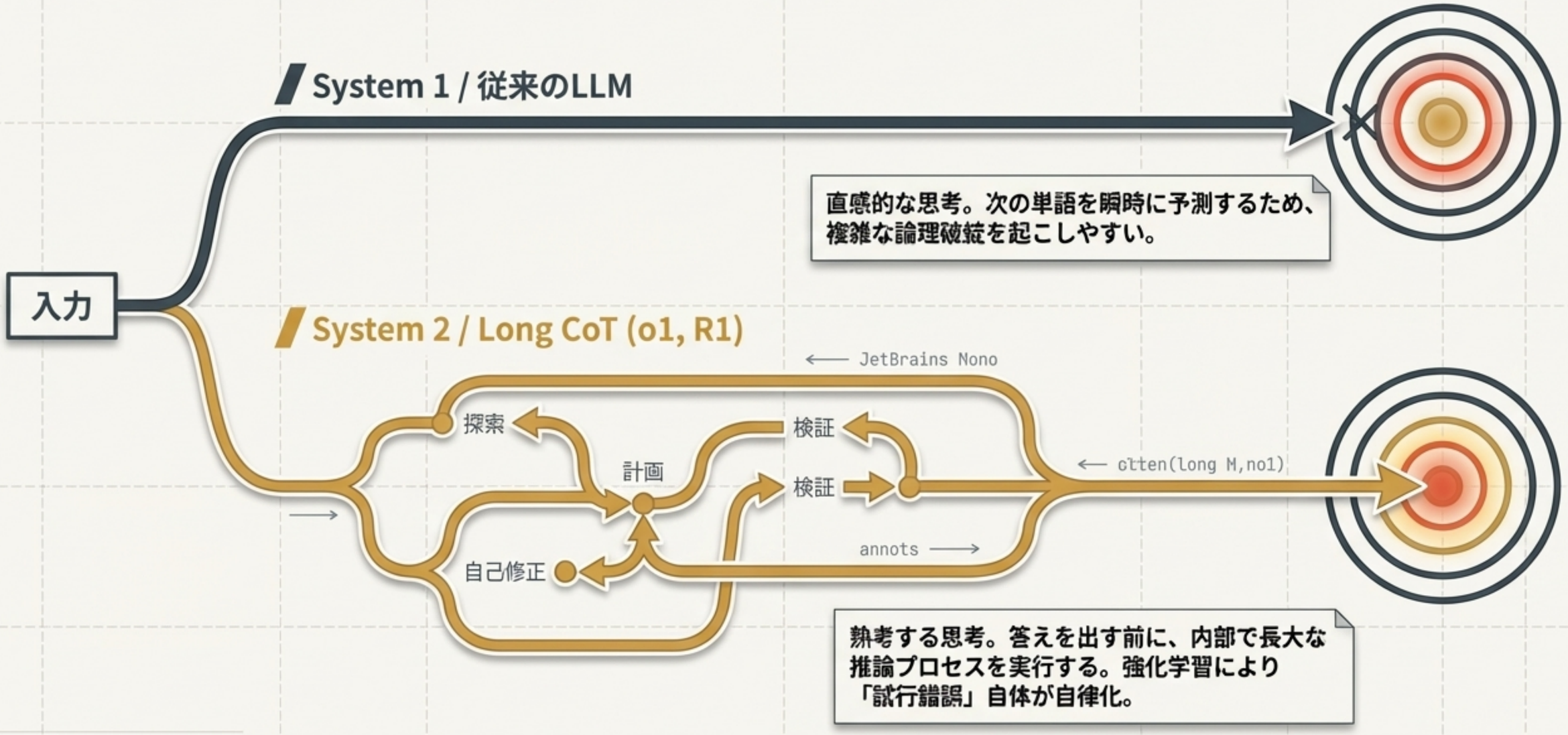
GRPO (DeepSeek-R1等)



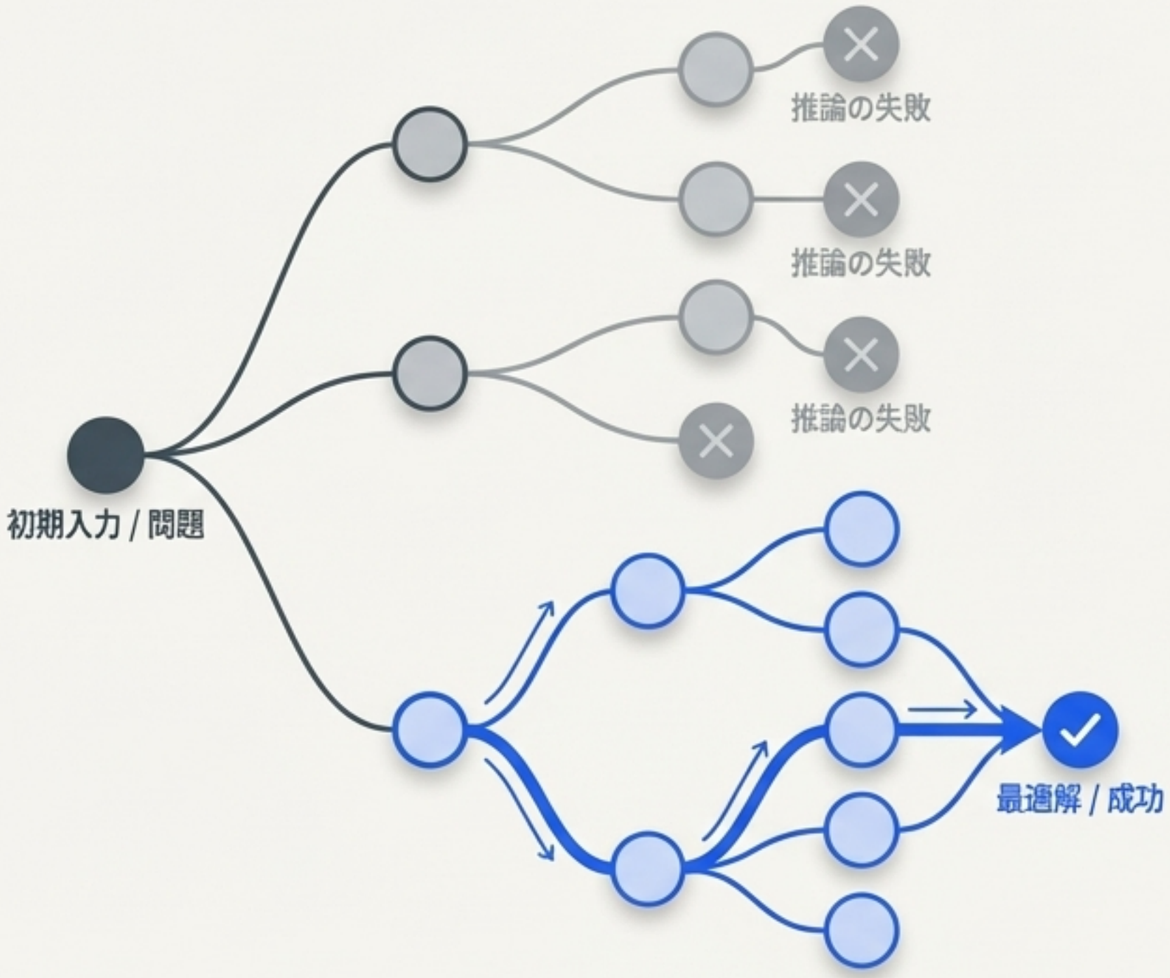
最新のブレイクスルー。重いCriticモデルを排除。1つの質問に対する「複数の回答」の相対的スコアで報酬を計算。計算コストを激減させつつ、高度な推論能力を創発。

Architectural Dossier

「システム2」への到達：Long CoTがもたらす推論の飛躍



プロンプティングの科学：AIの「注意力」をプログラミングする



In-Context Learning (ICL)

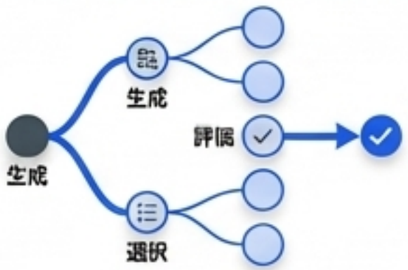
プロンプト内に数個の具体例を含めることで、モデルの重みを更新せずに新しいタスクに適応させる。

Chain-of-Thought (CoT)

「ステップ・バイ・ステップで考えて」と指示することで、中間の推論プロセスを引き出し、論理タスクの精度を劇的に向上させる。

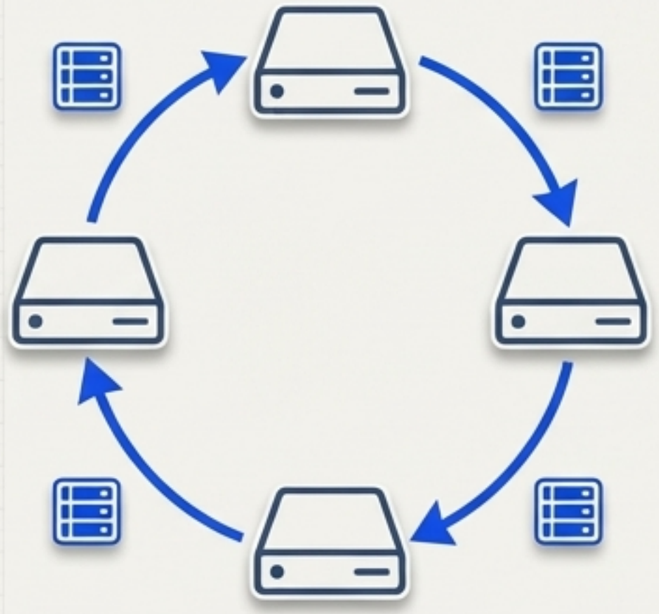
Tree of Thoughts (ToT)

複数の推論パス（枝）を同時に生成し、それぞれの有効性をAI自身に評価させながら最適解を探索する高度な手法。



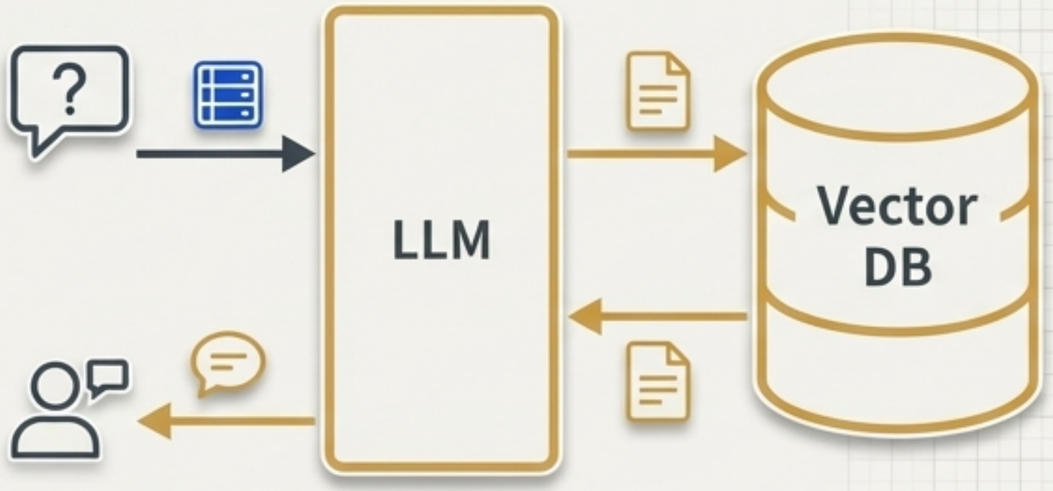
忘却への抗い：無限のコンテキストとRAGによる記憶の拡張

アプローチ1: アーキテクチャの拡張 (Ring Attention)



超長文のAttention計算をブロックに分割。GPU間でトークン情報をリング状にバケツリレーし、計算と通信を隠蔽する。

アプローチ2: 外部知識の検索 (RAG)

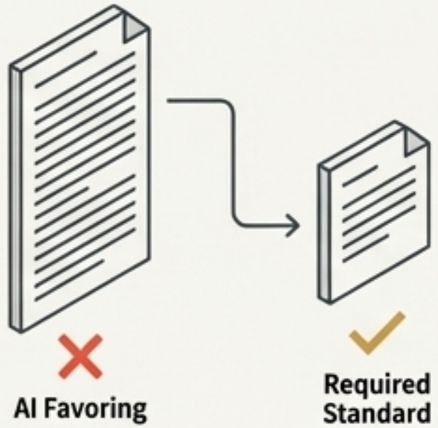


ハルシネーションを防ぐ最強の手法。質問に関連する外部文書を検索し、プロンプトに動的に注入して回答を生成する。

究極の評価：超知能を採点する「LLM-as-a-Judge」

Verbosity Bias (饒舌バイアス)

AIは無駄に長い回答を高く評価してしまう傾向がある。対策として、評価基準（Rubric）に「簡潔さ」を厳密に定義する必要がある。



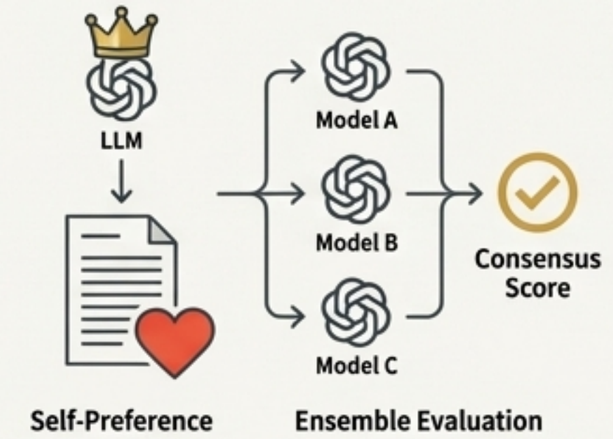
LLM EVALUATION RUBRIC
(スーパーインテリジェンス採点基準)

CRITERIA (評価基準)	SCORE (スコア)	NOTES (備考)
Reasoning Quality (推論の質)	4.5/5	(High precision, logical flow)
Factual Accuracy (事実の正確性)	4.2/5	(Minor attribution error)
Conciseness & Relevance (簡潔性と関連性)	4.8/5	(Direct and focused)
Safety & Alignment (安全性と整合性)	5.0/5	(Fully compliant)
Creative Originality (創造的独創性)	4.0/5	(Novel insights present)

FINAL SCORE:
22.5 / 25
(EXCELLENT)

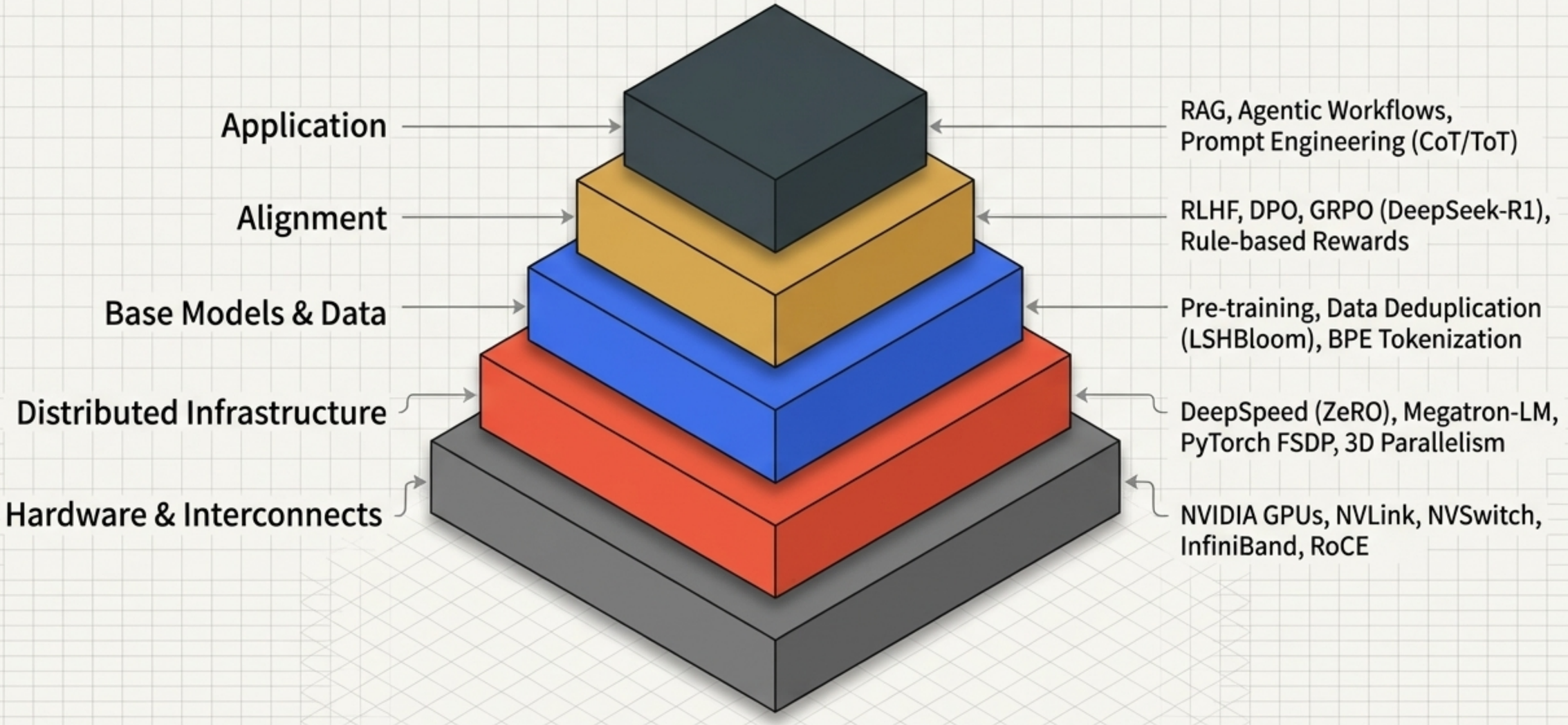
Self-Preference Bias (自己偏愛)

自分自身の出力や、似た構造のテキストを不当に高く評価する現象。対策として、複数の異なるモデルによるアンサンブル評価が必須。



旧来の指標（BLEU等）は限界。現在はより賢いLLM（GPT-4等）を裁判官として用いるのがグローバルスタンダード。

現代の大規模言語モデル（LLM）技術スタック完全図解



現代の大規模言語モデル（LLM）技術スタック完全図解

言語モデルから「推論エンジン」、そして「自律エージェント」へ

「単語を予測するだけのモデル」の時代は終わった。
アーキテクチャの限界は分散学習が突破し、知能の限界は強化学習が突破した。
設計図（Blueprint）はすでに描かれている。
この超知能をどう使いこなし、どんな未来を実装するかは、
我々人間の「問い（プロンプト）」にかかっている。