The Engine of Modern AI: Evolution of Language Model Pretraining

> The foundational statistical premise rests on the **distributional hypothesis**: linguistic units occurring in similar contexts share semantic properties. Pretraining operationalises this by maximising the likelihood of observed token co-occurrence patterns.

Mathematical Foundations of Pretraining Objectives

Autoregressive (Causal)

$$\mathcal{L}_{\mathrm{AR}}(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P_\theta \left(x_t^{(i)} \mid x_{<t}^{(i)} \right)$$

$h_t^{(L)} \in \mathbb{R}^d$: Top-layer hidden state at position t.

Masked Language Modelling (Bidirectional)

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = \sum_{i=1}^{N} \sum_{t \in \mathcal{M}^{(i)}} \log P_\theta \left(x_t^{(i)} \mid \tilde{x}^{(i)} \right)$$

$\mathcal{M}$: Subset of masked positions.

$\tilde{x}^{(i)}$: The corrupted sequence resulting from corruption function $\mathcal{C}$.

Sequence-to-Sequence (Span)

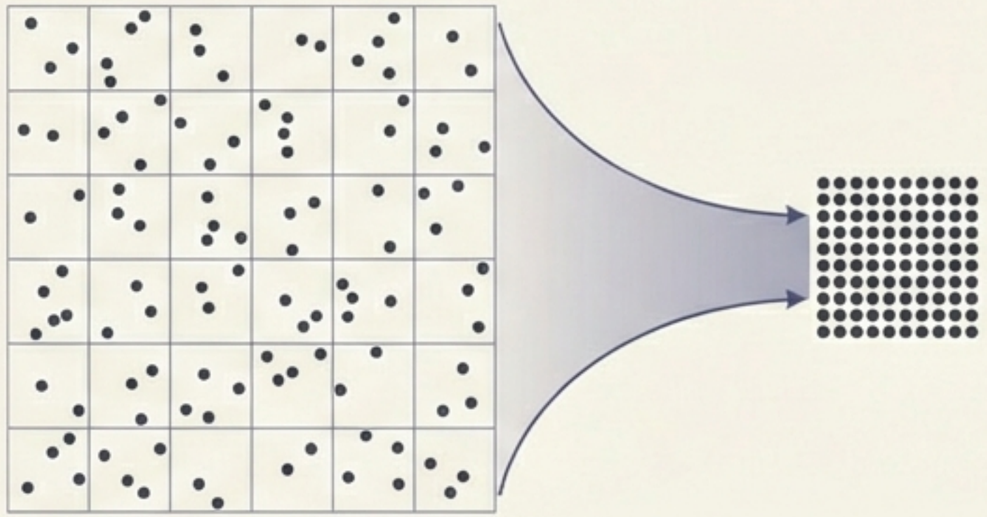
$$\mathcal{L}_{\mathrm{S2S}}(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{S}^{(i)}|} \sum_{k=1}^{|s_j^{(i)}|} \log P_\theta \left(s_{j,k}^{(i)} \mid \tilde{x}^{(i)}, s_{j,<k}^{(i)} \right)$$

The Information-Theoretic Perspective

$$H\left(\hat{P}_{data}, P_\theta\right) = H\left(\hat{P}_{data}\right) + D_{KL}\left(\hat{P}_{data} \,\|\, P_\theta\right)$$

Minimising Cross-Entropy

Minimising KL Divergence

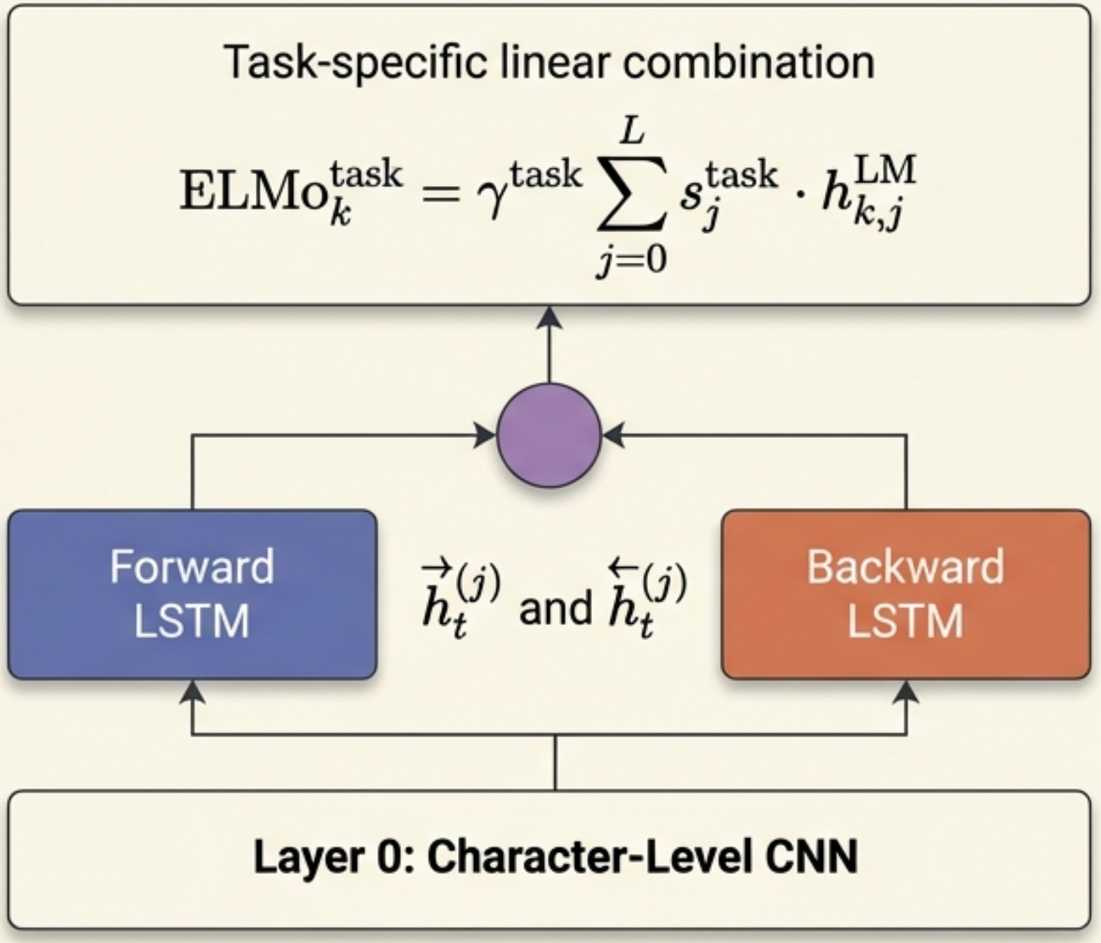


Perplexity (PPL)

$$\mathrm{PPL}(\theta) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P_\theta(x_t \mid x_{<t})\right)$$

Lower perplexity implies better compression. Under the minimum description length principle, this corresponds to capturing more of the underlying linguistic structure. Minimising KL divergence drives the model toward the true data-generating process.

Era 1: ELMo and Deep Contextualisation



Produces context-independent token representation $x_t^{(0)}$

Layer-Wise Linguistic Analysis

Layer 0 (Char CNN)	Morphology	Character n-grams and word shape
Layer 1 (biLSTM-1)	Syntax	Dependency relations and parsing
Layer 2 (biLSTM-2)	Semantics	Word sense disambiguation and coreference

Era 2: The BERT Revolution

Input Representation

Token Embedding

+

Segment Embedding

+

Position Embedding

↓

Multi-Head Self-Attention (MHSA)

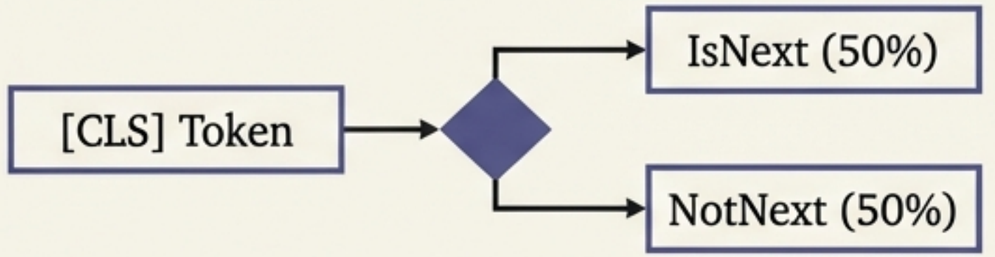
$$\mathbf{x}_t = \mathbf{E}_{\text{token}}[w_t] + \mathbf{E}_{\text{segment}}[s_t] + \mathbf{E}_{\text{position}}[t]$$

Split-Pane Objectives

Masked Language Modelling (MLM)

Mitigates the pretrain-finetune discrepancy.

Next Sentence Prediction (NSP)



Architectural Paradigm Shift: Feature Extraction to Fine-Tuning

	ELMo	BERT
Backbone	biLSTM (recurrent)	Transformer encoder (attention)
Bidirectionality	Shallow (independent L→R and R→L)	Deep (joint bidirectional at every layer)
Transfer Mechanism	Feature extraction (frozen embeddings)	Fine-tuning (all parameters updated end-to-end)
Tokenisation	Word-level + Char CNN	Subword (WordPiece)

Subsequent Innovations: ELECTRA

Replaced Token Detection (RTD) via generator-discriminator.

$$\mathcal{L}_{\text{RTD}} = -\sum_{t=1}^{T} \Big[\mathbf{1}[x_t = x_t^{\text{orig}}] \log D_\theta(x_t, t) + \mathbf{1}[x_t \neq x_t^{\text{orig}}] \log(1 - D_\theta(x_t, t))\Big]$$

Impact: Computes loss over all T tokens, making it 15× more sample-efficient than MLM.

Measuring Intelligence Through Standardised Benchmarks

Rigorous evaluation demands construct validity, discriminative power, resistance to artifacts, and reproducibility.

Sentence NLU

GLUE / SuperGLUE

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Reading Comprehension

SQuAD 1.1 / 2.0

- **Exact Match (EM)**: Strictly requires the predicted contiguous span to perfectly match the ground truth.
- **Token-level F1**: Measures partial overlap.

Code Generation

HumanEval

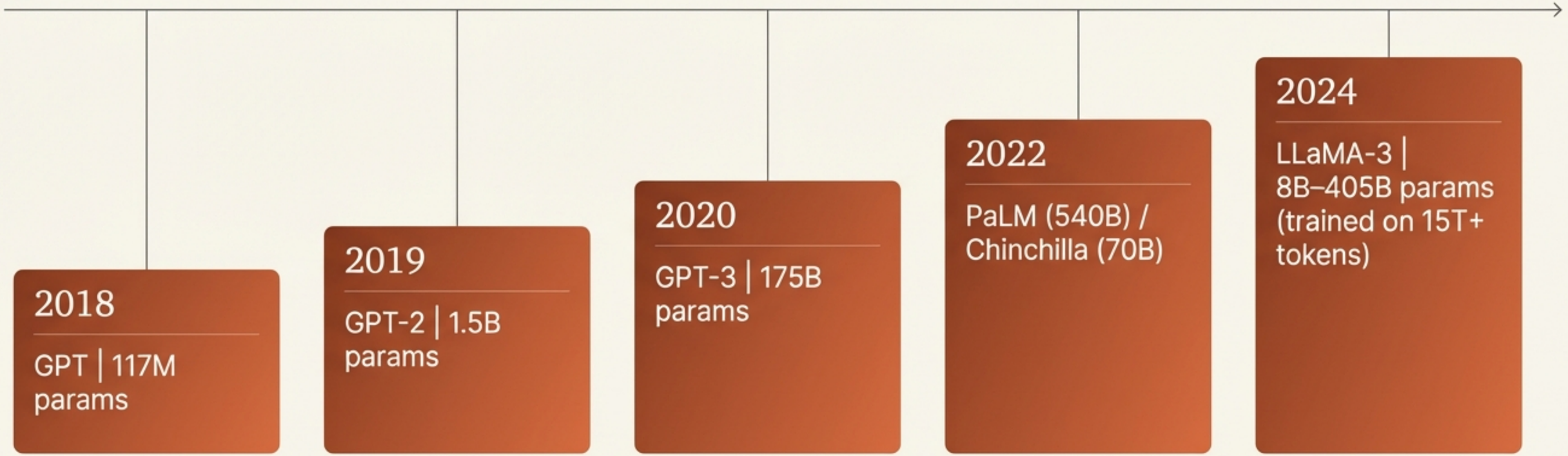
$$\mathrm{pass@}k = \mathop{\mathbb{E}}_{\text{problems}} \left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \right]$$

Massive Multitask

MMLU

57 diverse subjects evaluating zero-shot and few-shot knowledge without task-specific fine-tuning.

Era 3: Scale and the Emergence of Foundation Models



Emergent Capabilities

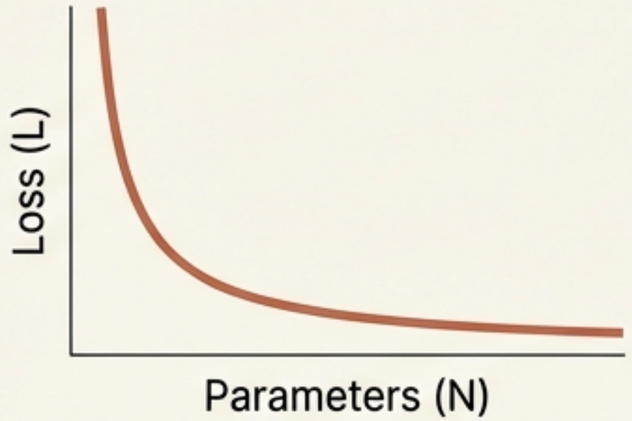
Definition: Qualitatively new behaviours that are absent in smaller models of the same family, appearing discontinuously as a function of model scale, compute, or data volume.

Impact: Shifts AI from pretrain-finetune pipelines to zero-shot task execution via natural language instructions alone.

The Science of Neural Scaling Laws

Kaplan Scaling Laws (2020)

$$L(N) = \left(\frac{N_c}{N}\right)^{\alpha_N}$$



Emphasises power-law decay where scaling parameters (N) was prioritised over data (D).

Chinchilla Scaling Laws (2022)

$$C \approx 6ND$$

Compute-optimal training dictates that parameters and data must scale equally.

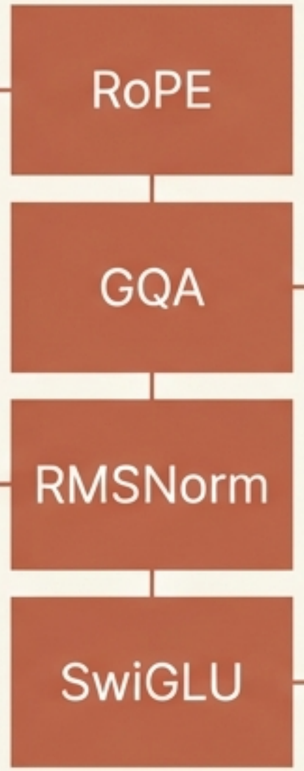
Practical Implication: A 70B parameter model should be trained on ~1.4T tokens. Earlier models like GPT-3 (175B parameters, 300B tokens) were severely under-trained by Chinchilla standards.

Anatomy of a Modern LLM Decoder

Rotary Position Embedding

Encodes position via rotation in 2D subspaces. Dot product depends only on relative position:

$$f_q(x_m, m)^\top f_k(x_n, n) = x_m^\top W_q^\top R_{\Theta, n-m} W_k x_n$$



Grouped Query Attention

Partitions A query heads into G groups. Reduces KV cache memory by factor A/G with minimal quality degradation.

Replaces LayerNorm

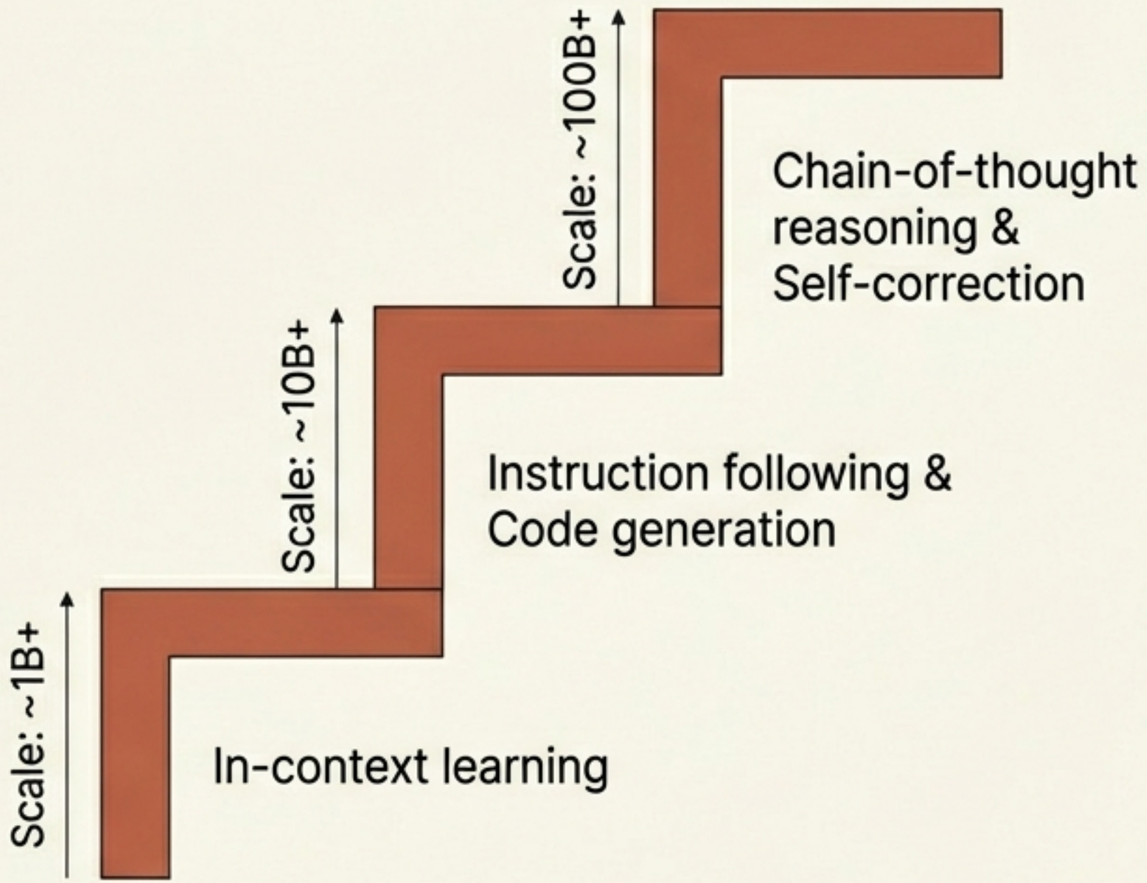
Eliminates mean-centring to improve training stability and efficiency.

Activation

Replaces standard ReLU/GELU.

$$\text{Swish}_\beta(x) = x \cdot \sigma(\beta x)$$

Mapping the Thresholds of Emergence



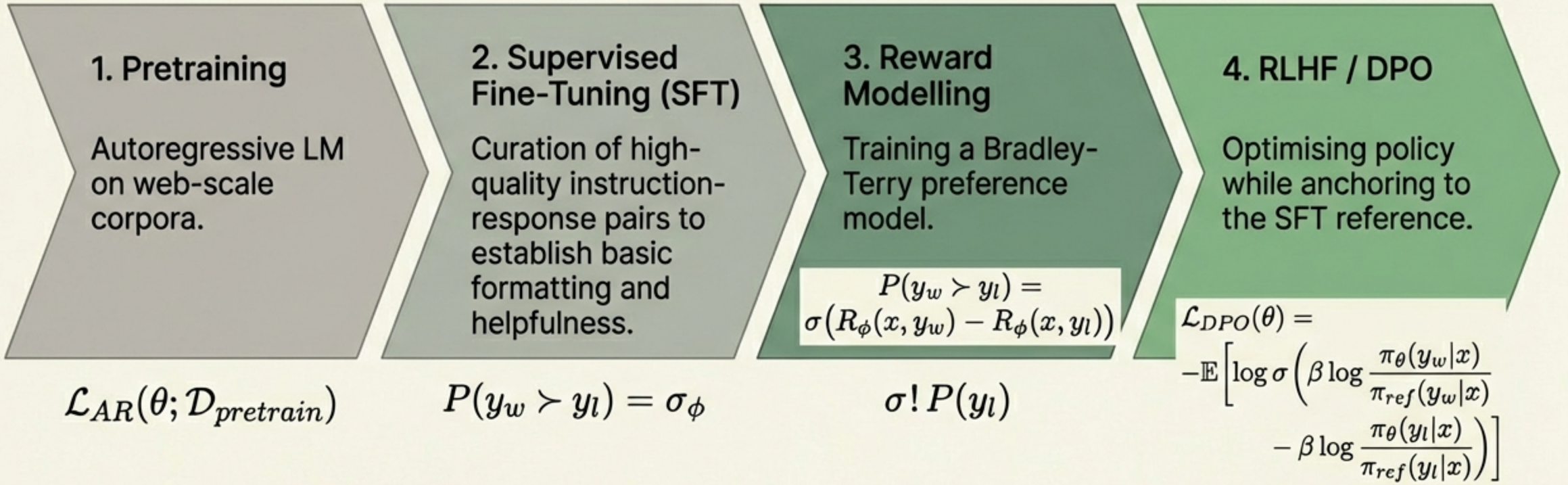
Theoretical Lens: In-Context Learning

In-Context Learning functions as Implicit Bayesian Inference.

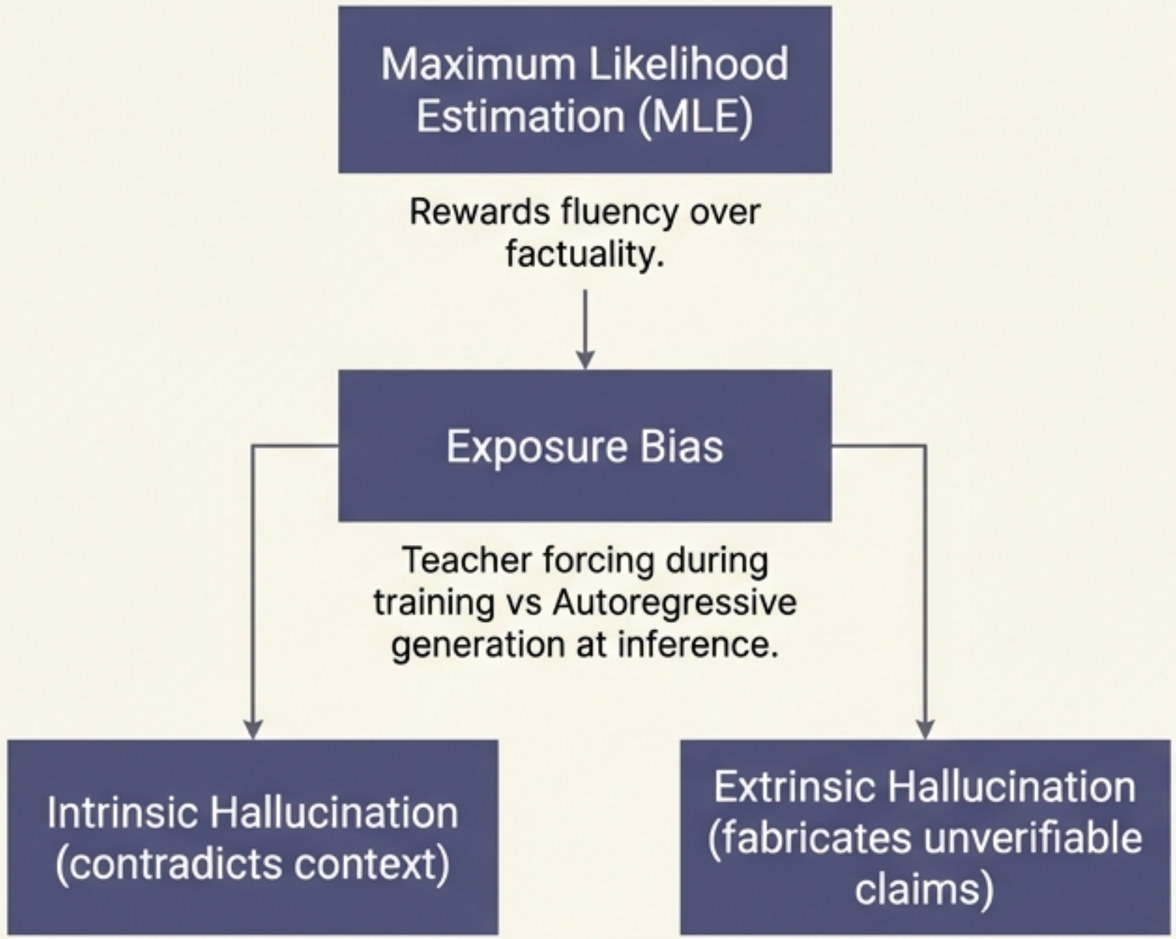
$$\hat{y}_{k+1} = \arg\max_{y} P_\theta(y \mid \mathcal{P})$$

Given a prompt with *k* demonstrations, the model performs implicit gradient descent within the forward pass, executing tasks without any parameter updates.

The End-to-End Alignment Pipeline



The Frontier: Objective Mismatch and Hallucination



Mathematical Disconnect

The fundamental flaw is that pretraining is a proxy objective.

$$\arg\min_\theta \mathcal{L}_{\text{pretrain}}(\theta) \neq$$

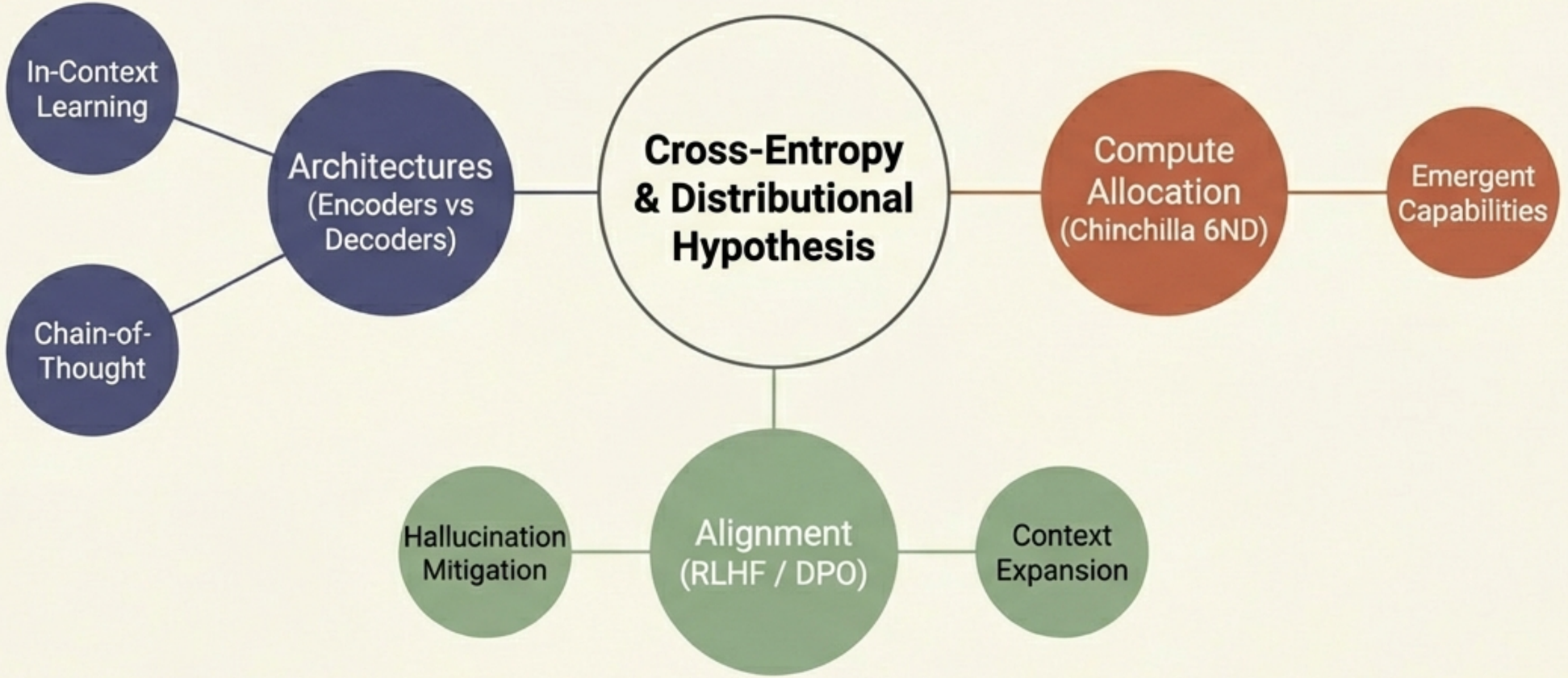
$$\arg\min_\theta \mathcal{L}_{\text{desired}}(\theta)$$

Takeaway: Pretraining produces an exceptional compressor of text, but **not necessarily** a reliable factual oracle or aligned agent.

Systemic Flaws and Algorithmic Mitigations

Problem	Root Cause	Equation/Metric	Mitigation
Catastrophic Forgetting	Fine-tuning overwrites weights	Fisher Information Matrix	Elastic Weight Consolidation ($\mathcal{L}_{\text{EWC}}$) or LoRA
Static Knowledge	Temporal cutoff in training data	Accuracy drop on post-cutoff facts	Retrieval-Augmented Generation (RAG) / Tool use
Poor Calibration	Softmax overconfidence	$\text{ECE} = \sum \frac{\lvert S_b \rvert}{N} \lvert \text{acc} - \text{conf} \rvert$	Temperature scaling / Conformal prediction
Context Limitation	Attention complexity is $\text{O}(T^2 d)$	Memory bounding constraints	Efficient attention, RoPE extrapolation (YaRN), Ring Attention

The Pretraining Dependency Architecture



Pretraining remains the undeniable engine of artificial intelligence. While architectural epochs transition from feature extraction to massive generative scale, the fundamental paradigm—compressing the world's knowledge through self-supervised distribution matching—continues to define the frontier of AI capabilities.