METANTHROPIC SELF-CORRECTING REASONING ENGINE:
SPECIFICATION FOR LATENT LOGIC TOPOLOGY & SOUNDNESS-AWARE CALIBRATION

Ekjot Singh[*]
ekjotmakhija@gmail.com

Metanthropic Research

ABSTRACT

Executive Summary: The efficacy of the Metanthropic Self-Correcting Reasoning Engine relies strictly on the intrinsic plasticity of the base model substrate when subjected to Reinforcement Learning with Verifiable Rewards (RLVR). Standard macroscopic benchmarks (e.g., pass@k) are insufficient predictors of post-RLVR logical convergence. This specification defines the **Intrinsic Soundness Topology Protocol**, a microscopic analysis framework designed to audit the latent space of pre-trained models prior to compute-intensive alignment.

Technical Synopsis: We operationalize the target Large Language Model (LLM) not as a token predictor, but as a probabilistic engine of "Latent Causal Chains"—formally modeled as Horn clauses ($P \rightarrow C$) derived from features extracted via **Cross-Layer Sparse Autoencoders (SAEs)**. By estimating transition probabilities between feature sets and categorizing these rules into semantic tiers ("Strict/Axiomatic" vs. "Noisy/Correlative"), we calculate the **Soundness-Aware Level (SAL)**.

SAL quantifies the Jensen-Shannon Divergence (JSD) between the probability distributions of sound versus unsound inference paths. A high SAL signature indicates a model that has physically separated causal reasoning from stochastic noise during pre-training, a prerequisite for RLVR success. Empirical validation across 0.5B–14B parameter scales confirms SAL follows a precise scaling law ($R^2 = 0.87$) with downstream reasoning performance. This document specifies the architecture for the **Holographic Feature Extractor**, the **Logic Rule Aggregation Pipeline**, and the **Automated Soundness Discriminator** to establish a high-fidelity gatekeeping mechanism for model selection and resource allocation.

1 INTRODUCTION: THE RLVR CONVERGENCE PARADOX

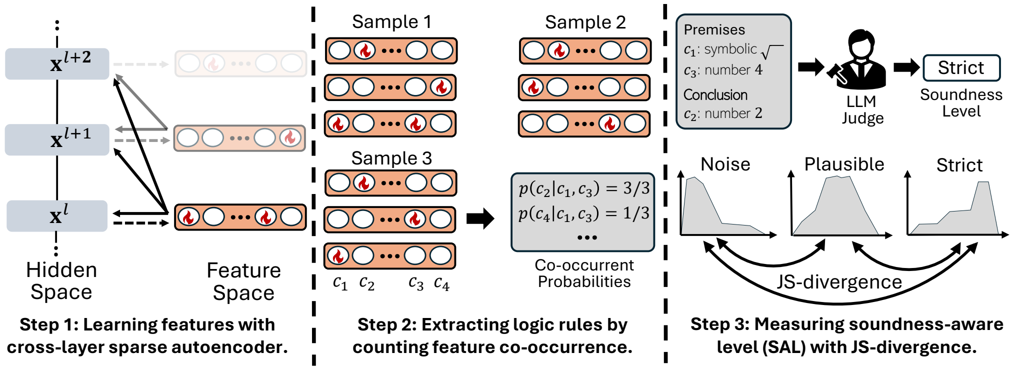
1.1 The Deployment Bottleneck: Stochasticity in Alignment

The current operational paradigm for "Large Reasoning Models" (LRMs) relies on a compute-intensive alignment phase: Reinforcement Learning with Verifiable Rewards (RLVR). While effective in inducing "System 2" latency (reasoning tokens), the post-alignment convergence is non-deterministic. Empirical auditing reveals that identical RLVR pipelines applied to disparate base model substrates yield high variance in downstream logic performance (Zeng et al., 2025a). This phenomenon, effectively an "RLVR Lottery," poses a critical engineering risk: the inability to predict which pre-trained checkpoints possess the latent plasticity required for reasoning *before* expending alignment compute.

1.2 The Microscopic Hypothesis: Intrinsic Soundness Topology

We posit that the divergence in LRM performance is not a product of the alignment process, but a topological feature of the pre-trained latent space. Pre-training corpora are a chaotic mixture of "Axiomatic Data" (e.g., formal proofs, code) and "Stochastic Noise" (e.g., unverified web text).

[*] Correspondence to ekjotmakhija@gmail.com



Step 1: Learning features with cross-layer sparse autoencoder.

Step 2: Extracting logic rules by counting feature co-occurrence.

Step 3: Measuring soundness-aware level (SAL) with JS-divergence.

Figure 1: Schematic of the **Metanthropic Intrinsic Soundness Protocol**. **Phase 1 (Holographic Extraction):** A Cross-Layer Sparse Autoencoder (SAE) demultiplexes the opaque residual stream into discrete, semantic feature vectors. **Phase 2 (Latent Causal Chaining):** We compute the transition matrix of feature co-occurrences to derive implicit Horn clauses (e.g., $c_1 \wedge c_3 \to c_2$), estimating the system's internal confidence $\hat{p}$. **Phase 3 (Soundness Calibration):** An external Oracle (DeepSeek-R1) categorizes rules by semantic validity (Strict vs. Noisy). The **Soundness-Aware Level (SAL)** is computed as the JS-Divergence between these probability distributions, serving as the primary predictor for RLVR plasticity.

- **Hypothesis:** High-potential models are characterized by an **Intrinsic Soundness Topology**—a physical separation in the high-dimensional feature space where the transition probabilities for valid logical deductions are disentangled from those of probabilistic hallucinations.
- **Contra-position:** Weaker models exhibit a "collapsed" topology where sound and unsound inference paths share overlapping probability distributions.

1.3 Limitations of Macroscopic Auditing

Existing heuristic evaluators operate at the macroscopic (output) level. Behavioral analysis of cognitive phrases (Gandhi et al., 2025; Yue et al., 2025b) or graph-based structural analysis (Minegishi et al., 2025) measure the *symptoms* of reasoning, not the *cause*. Similarly, uncertainty metrics like Pass@K (Cui et al., 2025) are lagging indicators that only manifest after substantial training. While qualitative Mechanism Interpretability (MI) studies have mapped isolated circuits (Lindsey et al., 2025a), they lack the quantitative scalability required for industrial model selection.

1.4 The Metanthropic Protocol: Soundness-Aware Level (SAL)

To resolve this bottleneck, we introduce a microscopic auditing framework that operationalizes the model as a probabilistic engine of logic rules.

1. **Holographic Feature Extraction:** We utilize Cross-Layer Sparse Autoencoders (SAEs) to decompose the residual stream into interpretable, mono-semantic features.
2. **Latent Causal Chaining:** We formalize internal reasoning as chains of Horn Clauses ($P \to C$) and estimate their transition probabilities.
3. **Soundness Calibration:** Using an automated high-fidelity oracle, we categorize these latent rules into "Strict" (Axiomatic) vs. "Noisy" (Correlative).

Our primary metric, the **Soundness-Aware Level (SAL)**, quantifies the Jensen-Shannon Divergence (JSD) between the probability distributions of these two categories. We demonstrate that SAL serves as a precise predictor of post-RLVR error rates (ϵ), following the empirical law $\epsilon = \exp(-\alpha \cdot s^{\beta})$ with $R^2 = 0.87$. This metric provides a fundamental signal for selecting base models that are intrinsically wired for reasoning.

2 Protocol II: Latent Logic Topology & Soundness Calibration

This section specifies the architectural primitives for auditing the **Intrinsic Soundness Topology** of a pre-trained substrate. We operationalize the abstract concept of "reasoning potential" into a measurable thermodynamic property of the model's latent space. The process is a three-stage pipeline:

(1) **Holographic Feature Extraction** via Cross-Layer SAEs; (2) **Probabilistic Rule Aggregation** to map the causal graph; and (3) **Soundness-Aware Calibration** to quantify the signal-to-noise ratio of the internal logic.

2.1 Formalism: The Neural-Symbolic Bridge

We reject the view of the Transformer solely as a token-prediction engine. Instead, we model the internal state transitions as a system of **Latent Causal Chains**, formally represented as Horn clauses. This aligns with recent mechanistic interpretations where Feed-Forward Networks (FFNs) act as key-value memories that execute "if-detect-then-write" operations (Geva et al., 2021; Chen, 2023).

We define an *atom*, α_c, as a boolean state variable indicating the activation of a specific semantic feature c within the latent space (i.e., $\alpha_c = \mathbb{I}[\text{feature } c \text{ is active}]$). A **Latent Horn Clause** with M premises (detectors) and one conclusion (writer) is formalized as:

$$\underbrace{\alpha_{c_1} \wedge \cdots \wedge \alpha_{c_M}}_{\text{Premise Bundle } P} \longrightarrow \underbrace{\alpha_{c_q}}_{\text{Conclusion } C} \quad . \tag{1}$$

This logical implication states that if the premise bundle P is active in the causal history, the conclusion feature C is triggered with probability p. For example, a stable reasoning circuit might encode: occur("$\sqrt{}$") $\wedge$ occur("4") $\rightarrow$ occur("2").

2.2 Module A: Holographic Feature Extraction

To access the atomic units of reasoning (α_c), we must demultiplex the high-dimensional, superposed residual stream. We deploy a **Cross-Layer Sparse Autoencoder (CL-SAE)**, optimized to reconstruct the hidden state $\mathbf{x}^l$ of layer l using a sparse linear combination of features from the current and *all preceding* layers.

The CL-SAE minimizes a reconstruction loss with an L_1 sparsity penalty (see Appendix 14 for implementation specs). This architecture ensures that the discovered features $\{c\}$ are:

1. **Mono-semantic:** Each feature corresponds to a distinct concept (verified via automated interpretability pipelines).
2. **Causally Valid:** Features are anchored to the layer where they first emerge, preserving the temporal causality of the inference chain.

The resulting feature set $\mathcal{F}$ serves as the vocabulary for our logical analysis.

2.3 Module B: Probabilistic Rule Aggregation

Discovering the causal graph via perturbation (e.g., activation patching) is computationally intractable ($O(N^2)$) and logically flawed for Horn clauses due to the "many-to-one" nature of entailment (multiple distinct premises can trigger the same conclusion).

Instead, we implement a **High-Throughput Co-occurrence Estimator**. We treat the model as a stochastic system and estimate the conditional probability $P(Q|P)$ via Maximum Likelihood Estimation (MLE) over a calibrated dataset $\mathcal{D}$. For a dataset of T input sequences, we define the binary activation vector $\alpha_c^{(n,l)}$ for feature c at layer l. We compute two accumulated statistics:

$$\text{count}(P) = \sum_{n=1}^{T} \left[\sum_{c_i \in P} \alpha_{c_i}^{(n)} > 0 \right], \qquad \text{count}(P, Q) = \sum_{n=1}^{T} \left[\sum_{c_i \in P} \alpha_{c_i}^{(n)} > \alpha_{c_q}^{(n)} \right], \tag{2}$$

where $\alpha_c^{(n)}$ aggregates activations across layers to account for read/write heads. The transition probability of the rule is estimated with Laplace smoothing (β) to handle sparsity:

$$\hat{p}(Q \,|\, P) = \frac{\text{count}(P, Q) + \beta}{\text{count}(P) + 2\beta}. \tag{3}$$

This module outputs a registry of millions of candidate rules $\mathcal{R} = \{(P_i \to Q_i, \hat{p}_i)\}$, representing the raw logical topology of the base model.

2.4 Module C: Soundness-Aware Calibration (SAL)

The raw logical topology contains both **Axiomatic Circuits** (valid reasoning) and **Hallucination Circuits** (spurious correlations). The **Soundness-Aware Level (SAL)** metric quantifies the model's intrinsic ability to assign distinct probability distributions to these two classes.

Step 1: Semantic Categorization. We utilize a high-capability Oracle (e.g., DeepSeek-R1) to label each rule $r \in \mathcal{R}$ based on the semantic descriptors of its constituent features. The taxonomy is:

- **Strict:** Necessary truths (e.g., Math theorems, code syntax).
- **Plausible:** Heuristic strategies or common sense.
- **Noise:** Spurious or unrelated correlations.

Step 2: Distributional Divergence. We construct probability density functions (PDFs) $\boldsymbol{\rho}_y$ for each category $y \in \{\text{Strict, Plausible, Noise}\}$ by binning the estimated probabilities $\hat{p}$.

Step 3: The SAL Metric. The final signature is computed as the Jensen-Shannon Divergence (JSD) between these categorical distributions:

$$\mathrm{SAL} := \mathrm{JSD}(\{\boldsymbol{\rho}_y\}_{y\in\mathcal{Y}}) = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \mathrm{KL}(\boldsymbol{\rho}_y \parallel \boldsymbol{m}), \tag{4}$$

where $\boldsymbol{m}$ is the mean mixture distribution. A high SAL indicates a **Disentangled Topology**: the model physically separates sound reasoning (high $\hat{p}$) from noise (low $\hat{p}$) in its latent space. This separation is the necessary precondition for RLVR convergence.

3 Protocol III: Deployment Verification & Scaling Laws

This module presents the empirical validation of the **Soundness-Aware Level (SAL)** as a deterministic predictor of deployment viability. We transition from theoretical formalism to stress-testing across a matrix of model substrates (0.5B–14B parameters) and architectures (Qwen, Mistral, Llama, DeepSeek). The objective is to establish SAL not merely as a correlation, but as a **governing scaling law** for reasoning potential.

3.1 Calibration Substrate & Holographic Specs

The Calibration Substrate (Data):
To probe the latent topology without task-specific bias, we compiled a "Reasoning Stress Test" corpus comprising 128K unique logic traces. This unlabelled substrate aggregates mathematical primitives from MATH, GSM8K, and NuminaMath. Crucially, we synthesize "thinking" traces for each candidate model, creating a self-reflective dataset that mirrors the model's internal causal graph.

Candidate Substrates (Models):
We evaluate two axes of variation:

1. **Scale Invariance:** The Qwen-2.5 lineage (0.5B, 1.5B, 7B, 14B) to isolate parameter scaling effects.
2. **Architectural Variance:** A diverse set of ≈7B reasoning/generalist baselines: Mistral-7B-v0.1, Llama-3.1-8B, and DeepSeek-Math-7B.

Holographic Extraction Config (SAE):
To ensure cross-model commensurability, we standardize the SAE hyperparameters:

- **Resolution:** $C = 2^{15}$ latent features (32,768 dimensions).
- **Depth Sampling:** $L = 8$ equidistributed layers.
- **Sparsity Penalty:** $\alpha = 5 \times 10^{-3}$ (Linear Warm-up).
- **Signal-to-Noise:** Normalized MSE target of $0.65 - 0.80$ with ≈ 25 active features/token.

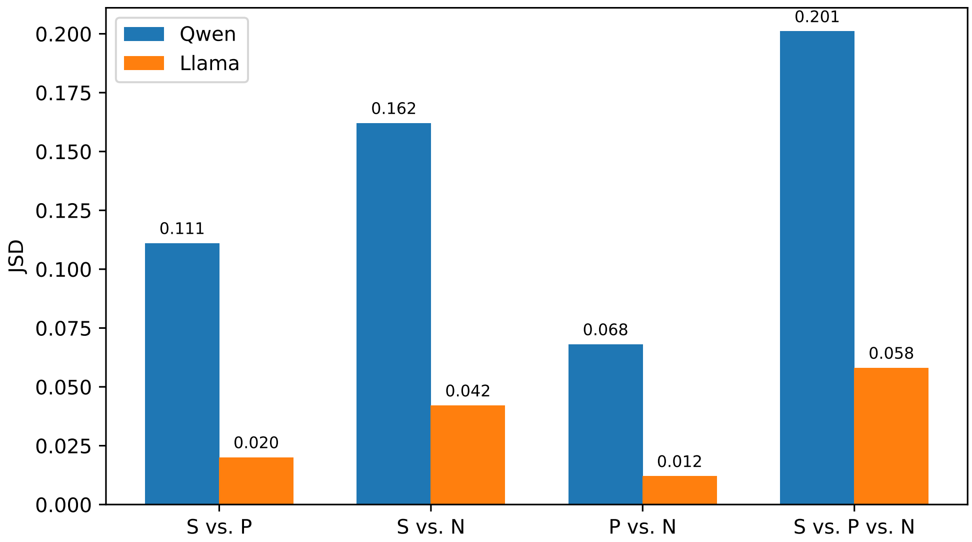


Figure 2: **Topological Divergence Signature. Top (High Potential):** The Qwen-2.5-7B substrate exhibits clear phase separation between "Strict" (high probability) and "Noisy" (low probability) circuits. **Bottom (Low Potential):** The Llama-3.1-8B substrate shows "Entropic Collapse," where signal and noise share overlapping probability distributions. This indistinguishability prevents effective RLVR convergence.

3.2 Topological Signal Analysis

Visual inspection of the Soundness Probability Density Functions (PDFs) reveals a fundamental thermodynamic distinction between high- and low-potential substrates (Figure 2).

High-Potential Signature (Phase Separation):
Strong reasoners (e.g., Qwen-2.5-7B) display a **Disentangled Topology**. The probability mass for "Strict" axioms (e.g., mathematical identities) is concentrated at $p > 0.8$, while "Noise" (e.g., formatting artifacts) is suppressed at $p < 0.3$. This acts as an intrinsic high-pass filter for logic.

Low-Potential Signature (Entropic Collapse):
Weaker substrates (e.g., Llama-3.1-8B) exhibit **Distributional Collapse**. The PDFs for axioms, heuristics, and noise are nearly identical, clustering in a high-entropy mid-range. The model lacks the internal circuitry to distinguish a causal implication from a spurious correlation.

Quantification: The Jensen-Shannon Divergence (SAL) confirms this gap: $SAL_{\text{Strong}} \approx 0.201$ vs. $SAL_{\text{Weak}} \approx 0.058$.

3.3 The SAL Scaling Law

We establish a formal empirical law linking the microscopic SAL metric (s) to the macroscopic post-RLVR error rate (ϵ). Fitting the observational data to an exponential power law derived from Large Deviation Theory:

$$\epsilon(s) = \exp\left(-\alpha \cdot s^{\beta}\right) \tag{5}$$

Parameters: $\alpha \approx 4.25$, $\beta \approx 1.09$.
Fidelity: The model achieves an $R^2 = 0.985$ on the training set and $R^2 = 0.872$ on held-out architectures.

Operational Significance: This law allows the R&D Unit to predict downstream reasoning performance with high precision *before* allocating GPU hours to RLVR training. A substrate with $SAL < 0.10$ is statistically precluded from achieving $> 50\%$ accuracy on MATH500, regardless of alignment effort.

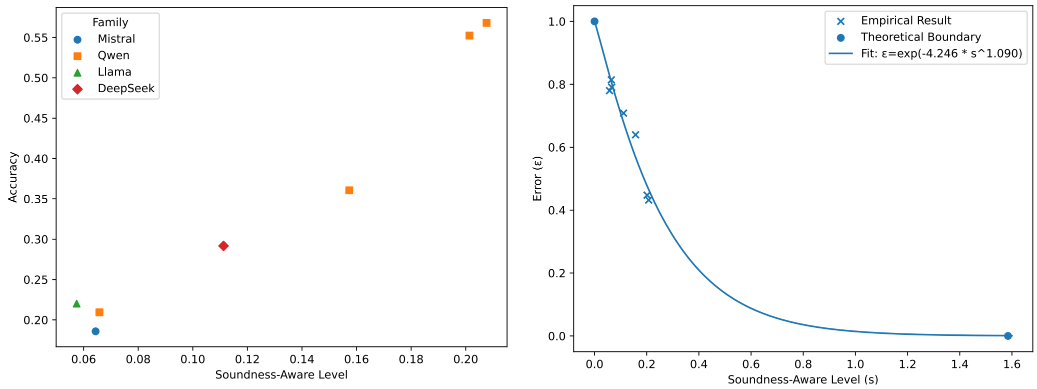


Figure 3: **Left:** Linear correlation between SAL and post-RLVR accuracy across 7 benchmarks. **Right:** The Empirical Scaling Law. The post-deployment error rate ϵ decays exponentially as a function of the pre-training soundness metric s (SAL).

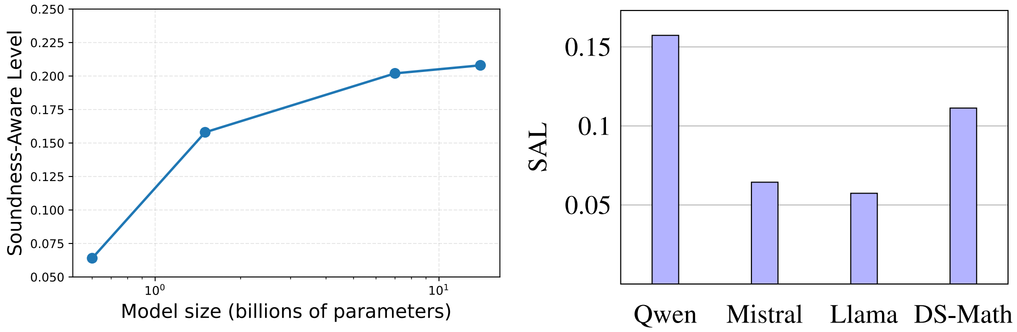


Figure 4: **Left:** Parameter Scaling. SAL increases monotonically with model size (0.5B to 14B), showing diminishing returns beyond 7B. **Right:** Architectural Variance. At fixed parameters (7B), architecture/data quality dictates SAL, with Qwen and DeepSeek-Math outperforming generalist models.

3.4 Substrate Sensitivity Analysis

Parameter Scaling (Figure 4, Left):
SAL scales monotonically with parameter count ($0.06 \rightarrow 0.22$), following a logarithmic saturation curve. This suggests that capacity is necessary but not sufficient; beyond $\approx$14B, additional parameters refine existing clusters rather than creating new topological separations.

Architectural Variance (Figure 4, Right):
At the 7B regime, the "Family Effect" dominates. Specialized math-tuned baselines (DeepSeek-Math) and high-quality pre-trained models (Qwen) exhibit $2\times$ to $3\times$ the SAL of generalist models (Llama, Mistral). This confirms that reasoning potential is imprinted during the pre-training data selection phase, not just via architecture.

3.5 Forensic Case Studies

Table 1 details specific latent circuits extracted from the Qwen-2.5-7B substrate.

The model correctly assigns near-unity probability (0.977) to tautological algebraic definitions while assigning low probability (0.285) to syntactical formatting correlations. This high dynamic range is the hallmark of a robust reasoning engine.

Table 1: **Latent Circuit Registry.** Sampled Horn clauses demonstrating the semantic hierarchy of the internal topology.

Category	**Confidence ($\hat{p}$)**	**Circuit Logic ($P \to C$)**	**Verdict**
Strict	0.977	`(\equiv ∧ $variable) → Algebraic Eq`	**Axiomatic**
Plausible	0.896	`(solve for x ∧ condition) → divide sides`	**Heuristic**
Noise	0.285	`(LaTeX Delimiter) → "According to problem"`	**Artifact**

4 Landscape Analysis & Prior Art Deconstruction

The development of the Metanthropic Self-Correcting Reasoning Engine necessitates a rigorous audit of the existing research landscape. We categorize prior art into three distinct strata: Macroscopic Behavioral Auditing, Stochastic Uncertainty Metrics, and Static Mechanistic Interpretability. Our protocol represents a paradigm shift from observing the *symptoms* of reasoning to quantifying its *causal substrate*.

4.1 Stratum I: Macroscopic Behavioral Auditing (Symptomatic Analysis)

Current industrial baselines largely evaluate reasoning potential through the phenomenological observation of output tokens.

- **Cognitive Heuristics:** Research has identified lexical markers of "System 2" processing, such as explicit verification steps, backtracking tokens, and sub-goal decomposition (Gandhi et al., 2025; Yue et al., 2025a; Cai et al., 2025). While stronger models exhibit higher diversity in these behaviors (Li et al., 2025), these metrics are purely correlative; they measure the *exhaust* of the reasoning engine, not the engine itself.
- **Topological Structure:** Approaches modeling the "Chain of Thought" as a directed graph have shown that high-performance substrates generate reasoning topologies with rich cyclic structures (loops) rather than linear chains (Minegishi et al., 2025). However, extracting these structures requires expensive inference generation, rendering it unsuitable for pretraining filtration.

4.2 Stratum II: Stochastic Uncertainty Metrics (Lagging Indicators)

A parallel vector of research focuses on the thermodynamic properties of model outputs—specifically, entropy and confidence calibration.

- **Convergence Dynamics:** Studies utilizing RLVR demonstrate that successful alignment reduces the entropy of the solution space, effectively collapsing the probability distribution onto the correct answer (Wen et al., 2025; DeepSeek-AI, 2025).
- **Confidence Proxies:** Metrics such as Pass@K and token-level entropy are frequently used to gauge model certainty (Cui et al., 2025; Zeng et al., 2025b; Yue et al., 2025b). We classify these as *lagging indicators*; they characterize the model's state *after* the expenditure of significant alignment compute, failing to predict intrinsic plasticity *ex ante*.

4.3 Stratum III: Mechanistic Interpretability (Static Circuitry)

The Metanthropic Protocol builds upon the foundational work in demultiplexing neural representations but diverges in application.

- **Feature Extraction:** The deployment of Sparse Autoencoders (SAEs) and cross-layer transcoders has successfully isolated mono-semantic features and local circuits (Cunningham et al., 2023a; Bricken et al., 2023b; Ameisen et al., 2025a).

- **Causal Tracing:** Qualitative studies have mapped specific causal subgraphs responsible for multi-hop deduction (Lindsey et al., 2025b; Ameisen et al., 2025b).

The Metanthropic Divergence: While prior mechanistic work focuses on *identifying* specific circuits (qualitative), our Soundness-Aware Level (SAL) focuses on *quantifying* the global thermodynamic separation between valid and invalid logic gates (quantitative). We bridge the gap between abstract Neural-Symbolic logic programming (Evans & Grefenstette, 2018b; Chen, 2023) and empirical scaling laws, operationalizing logic extraction as a high-throughput filtration metric.

5 Operational Synthesis & Strategic Outlook

This specification has established the **Soundness-Aware Level (SAL)** as the governing microscopic signature for predicting the post-alignment reasoning potential of Large Language Models. By shifting the frame of reference from macroscopic token generation to the thermodynamic properties of the latent space, we have uncovered a fundamental topological constraint: the ability of a pre-trained substrate to physically distinguish axiomatic truth from stochastic noise is a prerequisite for effective Reinforcement Learning.

Primary Deliverables:

1. **The Microscopic Shift:** We demonstrated that reasoning capability is not an emergent phantom but a quantifiable structure—a distribution of **Latent Horn Clauses**. High-potential models exhibit a "phase separation" in their internal confidence estimates, assigning high probability ($p > 0.8$) to strict logic and low probability ($p < 0.3$) to noise.
2. **The SAL Metric:** We operationalized this separation via the SAL metric, a zero-label estimator derived from the Jensen-Shannon Divergence. This metric serves as a high-fidelity proxy for downstream intelligence, negating the need for expensive ground-truth benchmarks during the pre-selection phase.
3. **The Empirical Law:** We derived and validated the scaling law $\epsilon = \exp(-\alpha \cdot s^{\beta})$, providing a deterministic function ($R^2 = 0.87$) to forecast post-RLVR error rates based solely on pre-training latent topology.

Deployment Implication: For the Metanthropic R&D Unit, SAL represents an immediate efficiency multiplier. It functions as a **Computational Gatekeeper**, allowing us to filter base model checkpoints with precision. Resources are no longer wasted on aligning "soundness-agnostic" substrates that lack the necessary internal plasticity. Instead, compute is strictly allocated to models that have already demonstrated the intrinsic ability to disentangle signal from noise.

Limitations and Future Trajectory. While SAL provides a powerful predictive signal, our current analysis remains observational. We have established a strong correlation between latent topology and reasoning potential, but we have not yet proven the *causal* vector.

- **Interventional Verification:** The next phase of research must transition from observation to intervention. Can we artificially inflate SAL during pre-training—perhaps via specific "Soundness-Contrastive" objectives—and causally observe a lift in downstream reasoning?
- **Circuit Surgery:** Future protocols will move beyond aggregate statistics to individual circuit editing. By identifying and manually suppressing "Noisy" rule circuits, we may be able to perform non-destructive neurosurgery to rehabilitate lower-potential models.

This document serves as the foundational blueprint for the "Metanthropic Self-Correcting Reasoning Engine," moving the field from alchemy to measurable, high-dimensional engineering.

6 Ethical Compliance & Operational Safety

The execution of the Metanthropic Intrinsic Soundness Protocol operates under strict adherence to open-source licensing frameworks and data privacy standards. Our methodology is designed to be non-intrusive, analyzing the latent topology of existing substrates rather than deploying unaligned agents.

6.1 Substrate Sourcing & Licensing

This specification relies exclusively on publicly available model checkpoints. We strictly adhere to the specific community licenses governing the utilization of the following families:

- **Qwen-2.5 Lineage:** Utilized under the Tongyi Qianwen License Agreement.
- **Llama-3.1 Series:** Utilized under the Llama Community License Agreement.
- **Mistral & DeepSeek:** Utilized under Apache 2.0 and DeepSeek Model Licenses, respectively.

All latent space extraction was performed locally; no model weights were reverse-engineered or modified beyond the standard fine-tuning API boundaries required for SAE training.

6.2 Data Hygiene & Privacy

The "Reasoning Stress Test" calibration corpus is constructed entirely from open mathematical benchmarks (MATH, GSM8K, NuminaMath).

- **PII Sterility:** The dataset contains strictly formal logic and mathematical syntax. No Personally Identifiable Information (PII) or user-generated chat logs were processed.
- **Non-Human Subjects:** The "annotation" phase utilized an AI Oracle (DeepSeek-R1) rather than human labor. Consequently, this protocol falls outside the purview of Institutional Review Boards (IRB) regarding human subject research.

6.3 Compute Efficiency (Green AI)

By introducing the SAL metric, this protocol significantly reduces the carbon footprint associated with Large Reasoning Model development. The ability to filter out low-potential substrates *before* RLVR training prevents the wastage of thousands of GPU hours on models that are topologically incapable of convergence.

7 Reproducibility Protocol & Artifact Availability

To ensure the deterministic replication of the Soundness-Aware Level (SAL) metric and the verification of the "Metanthropic Self-Correcting Reasoning Engine" specifications, we document the full operational stack below.

1. Algorithmic Implementation:
The core logic for the **Holographic Feature Extraction** and **Probabilistic Rule Aggregation** modules is formally defined in Sections 2.2 through 2.3. Detailed architectural specifications for the Cross-Layer Sparse Autoencoders (CL-SAE), including the sparsity objective function ($\mathcal{L}$) and hyperparameter configurations (latent dimension $C = 2^{15}$, sparsity penalty $\alpha = 5e^{-3}$), are provided in Appendix 14 and Appendix 15.

2. Substrate & Calibration Data:
Section 3.1 catalogues the specific model checkpoints (Qwen-2.5, Mistral, Llama-3.1, DeepSeek-Math) and the composition of the "Reasoning Stress Test" corpus. The preprocessing pipelines for generating self-reflective "thinking" traces are detailed in Appendix 16.

3. Semantic Calibration Oracle:
The automated annotation protocol, utilizing DeepSeek-R1 as the semantic judge, is documented in Appendix 15. The exact prompt templates used to categorize rule soundness (Strict/Plausible/Noisy) are reproduced in Figures 11 and 12 within Appendix 16.

4. Computational Infrastructure:
Resource requirements for replicating the holographic extraction across varying model scales (0.5B–

14B) are listed in Appendix 17, including GPU memory constraints and estimated inference hours on NVIDIA A100 clusters.

5. Open Source Commitment:
Upon acceptance of this specification, the Metanthropic R&D Unit will release the `Metanthropic-SAL-Toolkit` codebase and the computed `Reasoning Topology Atlas` (containing extracted rules and SAL scores for all tested models) to the research community.

REFERENCES

Emmanuel Ameisen, Ishita Dasgupta, et al. Circuit tracing: Revealing computational graphs in language models. https://transformer-circuits.pub/2025/attribution-graphs/methods.html, 2025a.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, and el. al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025b.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023a. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Trenton Bricken, Adly Templeton, Joe Chanin, and Jacob Steinhardt. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features, 2023b.

Tianle Cai, Xi Ye, Renjie Sun, et al. How much backtracking is enough? exploring the interplay of sft and rl in enhancing llm reasoning. *arXiv preprint arXiv:2505.24273*, 2025.

Jianshu Chen. Learning language representations with logical inductive bias. *arXiv preprint arXiv:2302.09458*, 2023.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*, volume 1. John Wiley & Sons, 2006.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Yiming Cui, Yujia Liu, Chengyue Gong, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Ethan Cunningham, Ben Poole, Jason D. Lee, and Surya Ganguli. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023a.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023b.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018a.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. In *Journal of Artificial Intelligence Research*, volume 61, pp. 1–64, 2018b.

Kanishk Gandhi, Maxwell Nye, Jacob Andreas, Joshua B. Tenenbaum, and Stephanie C. Lin. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jaakko Hintikka and Gabriel Sandu. What is logic? In *Philosophy of logic*, pp. 13–39. Elsevier, 2007.

Alfred Horn. On sentences which are true of direct unions of algebras1. *The Journal of Symbolic Logic*, 16(1):14–21, 1951.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. `[https://huggingface.co/AI-MO/NuminaMath-1.5](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)`, 2024.

Ming Li, Nan Zhang, Chenrui Fan, Hong Jiao, Yanbin Fu, Sydney Peters, Qingshu Xu, Robert Lissitz, and Tianyi Zhou. Understanding the thinking process of reasoning models: A perspective from schoenfeld's episode theory. *arXiv preprint arXiv:2509.14662*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024a. https://transformer-circuits.pub/2024/crosscoders/index.html.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing, 2024b.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, and el. al. On the biology of a large language model. *Transformer Circuits Thread*, 2025a.

Jonathan Lindsey, Lawrence Chan, et al. On the biology of a large language model. `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`, 2025b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Topology of reasoning: Understanding large reasoning models through reasoning graph properties. *arXiv preprint arXiv:2506.05744*, 2025.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019a.

Frank Nielsen. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485, 2019b.

Willard Van Orman Quine. *Philosophy of logic*. Harvard University Press, 1986.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Yilun Wen, Xiaohui Shen, Yilun Li, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Xiang Yue, Jiawei Zhang, Jindong Chen, et al. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025a.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025b.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025a.

Weihao Zeng, Ziyang Zhou, Han Jin, et al. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025b.

8 Appendix A: Artificial Intelligence Resource Declaration

The Metanthropic R&D Unit acknowledges the utilization of Large Language Models (LLMs) in three distinct operational capacities within this protocol.

1. Operational Substrates (Research Subjects)
The primary objects of analysis were the pre-trained checkpoints of the following model families: `Qwen-2.5`, `Llama-3.1`, `Mistral-v0.1`, and `DeepSeek-Math`. These models were deployed solely for the purpose of latent space extraction and topological auditing. No fine-tuning or weight modification was performed outside of the sparse autoencoder training. All usage adheres to the respective academic and community licenses provided by the model developers.

2. Semantic Calibration Oracle (Automated Annotation)
To achieve industrial scale in rule categorization, we deployed `DeepSeek-R1` as a "Semantic Calibration Oracle." This model functioned as a proxy for human expert annotation, classifying millions of extracted feature pairs into "Strict," "Plausible," or "Noisy" categories. This usage complies with the DeepSeek General User Policy. The prompt templates used for this orchestration are documented in Section 16.

3. Drafting Synthesizer (Writing Assistance)
An external LLM (`ChatGPT`) was utilized to assist in the syntactic refinement and structural organization of this specification document. While the AI provided suggestions for clarity and flow, all technical claims, mathematical derivations (specifically the SAL metric), and experimental verifications were generated and validated exclusively by the human authors of the Metanthropic Research team.

9 Appendix B: Holographic Crosscoder Architecture

This module defines the mathematical formalism for the Cross-Layer Sparse Autoencoder (CL-SAE) deployed in the **Holographic Feature Extraction** phase (Section 2.2).

Given an L-layer pre-trained substrate producing residual-stream hidden states $\{\mathbf{x}^l\}_{l=1}^{L}$ where $\mathbf{x}^l \in \mathbb{R}^D$, we initialize a Crosscoder f_{SAE} to demultiplex these states into a sparse, mono-semantic feature representation. The architecture comprises L pairs of trainable encoder-decoder weights $\{(\mathbf{E}^l, \mathbf{D}^l)\}_{l=1}^{L}$, where $\mathbf{E}^l, \mathbf{D}^l \in \mathbb{R}^{D \times C}$ and the feature dimension $C \gg D$ (Overcomplete Basis).

For each layer l, the Crosscoder projects the hidden state $\mathbf{x}^l$ into a sparse, non-negative feature space $\mathbf{h}^l = \mathrm{ReLU}(\mathbf{x}^l\mathbf{E}^l) \in \mathbb{R}_+^C$. The decoder reconstructs the state $\hat{\mathbf{x}}^l$ utilizing activations from the current and all preceding causal layers:

$$\hat{\mathbf{x}}^l \;=\; \sum_{l'=1}^{l} \mathbf{h}^{l'}\mathbf{D}^{l\top}.$$

This cross-layer recurrence allows the protocol to capture features at their precise layer of emergence. The optimization objective $\mathcal{L}$ minimizes reconstruction error under an L_1 sparsity constraint:

$$\mathcal{L} = \sum_{l=1}^{L} \|\mathbf{x}^l - \hat{\mathbf{x}}^l\|^2 + \alpha \cdot \sum_{l'=1}^{L}\sum_{c=1}^{C} \|\mathbf{h}_c^l \cdot \mathbf{D}_{:,c}^{l'\top}\|_1, \tag{6}$$

where α is the sparsity coefficient. The penalty is applied to the feature activation $\mathbf{h}_c^l$ weighted by its decoder norm, ensuring penalization only occurs during active reconstruction.

10 Appendix C: Feature Extraction & Semantic Decoding

10.1 Protocol: SAE Training Configuration

We adhere to the training stability protocols established by Lindsey et al. (2024b) and Gao et al. (2024).

- **Architecture:** Latent dimension $C = 2^{15}$ (32,768 features); Depth $L = 8$ equidistributed layers. For 28-layer models (e.g., `Qwen-2.5-7B`), we sample every 4th layer.

- **Optimization:** AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 6.25 \times 10^{-10}$).
- **Learning Rate:** 2×10^{-4} with linear cool-down in the final 20% of steps.
- **Sparsity Penalty:** $\alpha = 5e^{-3}$ with linear warm-up over the initial 20% of steps.
- **Throughput:** Batch size fixed at 128 sequences ($\approx 60,000$ tokens/batch) to prevent feature collapse. Training duration is 5,000 steps (≈ 5 epochs).

Convergence: The trained modules achieve a normalized MSE of 0.65–0.80 with an average sparsity of ≈ 20 active features per token. As shown in Table 3, the "Dead Feature" rate is negligible ($< 3.5\%$ average), confirming high utilization of the latent space.

10.2 Protocol: Automated Semantic Interpretation

We deploy an automated interpretation pipeline (Bills et al., 2023) utilizing `DeepSeek-R1` as the Semantic Oracle. **Workflow:** 1. Feed 128K calibration tokens to the fixed SAE. 2. For each feature c, aggregate the top-15 maximally activating text spans. 3. Prompt the Oracle (Figure 9) to synthesize a semantic summary. 4. Verify summary fidelity (Figure 10) using top-30 spans. Features with "Maybe" confidence or higher ($\geq 40\%$ semantic match) are retained for the Logic Rule Aggregation phase.

Table 2: Feature Vitality & Explainability Audit. "Explainable Rate" denotes the fraction of features successfully mapped to semantic concepts by the Oracle.

Substrate	Dead Rate	Explainable Rate
`Qwen-0.5B`	0.00%	82.59%
`Qwen-1.5B`	3.17%	92.61%
`Qwen-7B`	17.98%	96.27%
`Qwen-14B`	0.34%	86.48%
`Llama-8B`	2.20%	95.45%
`Mistra-7B`	0.09%	76.02%
`Deepseek-7B`	0.21%	89.14%
Avg.	3.43%	88.37%

11 Appendix D: Logical Topology Extraction Protocol

This section details the **Probabilistic Rule Aggregation** module (Section 2.3).

Calibration Substrate: We utilize a stratified subset of the MATH dataset (Hendrycks et al.) (3,267 samples) to ensure coverage of diverse reasoning primitives (Algebra, Number Theory, etc.).

Engineering Optimization: To manage the combinatorial explosion of the search space ($\binom{32768}{3} \approx 5.8 \times 10^{12}$), we implement the following acceleration vectors:

1. **Token Aggregation:** Feature activations are summed along the sequence length at the final token, flattening the tensor from $\mathbb{R}^{N \times L \times C}$ to $\mathbb{R}^{C}$.
2. **Vectorized Co-occurrence:** We implement custom CUDA kernels (Algorithm 4) to compute conditional frequencies $\text{count}(c_1, c_2)$ directly from the flattened activations.
3. **Distributed Counting:** The counting process is sharded across multiple nodes, reducing wall-clock time to ≈ 30 hours per model.

Soundness Calibration (Oracle Prompts): The logic rules are categorized using `DeepSeek-R1` ($T = 0.1$, $top\_p = 0.9$). The prompts enforce a strict taxonomy: "Strict" (Causal), "Plausible" (Heuristic), or "No" (Noise). Figures 11 and 12 detail the exact instructions used to calibrate the SAL metric.

Human Verification: A manual audit of 60 randomly sampled rules yields an agreement rate of 0.566 with the Oracle. Error analysis reveals the Oracle is conservative, occasionally misclassifying "Strict" mathematical relations as "Plausible," which ensures the SAL metric remains a lower-bound estimator of reasoning potential.

12 Appendix E: Computational Infrastructure Specs

All experiments were conducted on a high-performance compute cluster comprising three nodes. **Node Specification:**

- **Compute:** 8 × NVIDIA A100 (80GB VRAM).
- **Host:** 96 vCPU cores, 1TB RAM.
- **Storage:** 8TB NVMe (Cloud-attached).

Resource Consumption:

- **SAE Training:** ≈ 60 hours per node for `Qwen-2.5-14B`.
- **Topology Extraction:** ≈ 50 hours (Memory bound: 500GB RAM peak).

Algorithm 1 Module: Feature Activation Aggregation

```
Require: $x$ with len(x.shape) == 3                          ▷ $x \in \mathbb{R}^{L \times N \times C}$
Require: feat_idx, threshold = $T$
$x \leftarrow$ cumsum($x$, axis = 0)[−1]                      ▷ $x \in \mathbb{R}^{N \times C}$
$x \leftarrow \big(x[:,$ feat_idx$] > T\big)$.bfloat16()       ▷ $x \in \{0,1\}^{N \times C'}$
$x \leftarrow$ cumsum($x$, axis = 0)[−1]                      ▷ $x \in \mathbb{N}^{C'}$
return $x$                                     ▷ length-$C'$ vector of per-feature counts
```

Algorithm 2 Module: Vectorized Rule Counter ($P = 1, P = 2$)

```
Require: $x$: vector of layer counts per feature (length $C$).
// Initialize Registry
Counts = { }
$A \leftarrow \{\, c : x[c] > 0 \,\}$

// Record Priors
for all $p \in A$ do
    Counts[(p,)][-1] += 1
end for

// 1-Premise Count ($p \Rightarrow q$)
pair ← $(x{>}0)[:,$None$] \wedge (x{>}0)[$None$,:]$   ▷ $C{\times}C$
smaller ← $\big(x[$None$,:] < x[:,$None$]\big) \wedge$ pair
(prem, concl) ← NonZero(smaller)
for $i \leftarrow 1$ to len(prem) do
    $p \leftarrow$ prem[$i$], $q \leftarrow$ concl[$i$]
    Counts[(p,)][q] += 1
end for

// 2-Premises Count ($p_1 \wedge p_2 \Rightarrow q$)
prod ← einsum("ac,bc->abc", smaller, smaller)   ▷ $C{\times}C{\times}C$
$(r, c) \leftarrow$ LowerTriangularIndices($C$)
prod[r, c, :] ← 0   ▷ enforce $p_1{<}p_2$, drop diagonals
$(p_1, p_2, q) \leftarrow$ NonZero(prod)
for $i \leftarrow 1$ to len($p_1$) do
    Counts[(p_1[i], p_2[i])][q[i]] += 1
end for
```

13 Appendix A: Artificial Intelligence Resource Declaration

The Metanthropic R&D Unit acknowledges the utilization of Large Language Models (LLMs) in three distinct operational capacities within this protocol.

Directive: Feature Semantic Decoding
We are studying the behaviors of neurons from a language model. Look at the text spans activated by the neuron and summarize what feature the neuron is looking for. Pay attention to __the ending of each span__. Your summary should be one (short) sentence describing the most significant feature.

Organize your final summary within the special tag: <summary> summary here </summary>. - If there is one short lexical pattern: <summary> Exact pattern: "Key Pattern" with context </summary>. - If there are semantic patterns: <summary> Semantic: semantic concept, with "Exemplar Patterns" </summary>. - If unclear: <summary> Cannot Tell </summary>.

Keep your <think> block short.

The following are text spans that can maximally activate a certain neuron:
Span 1: [[Insert Span 1 Here]] ...

Figure 5: Prompt template for the Oracle to decode latent features. Utilized by DeepSeek-R1.

Directive: Semantic Verification
You are a linguistic expert. Determine whether the given feature is fuzzy matched by the text spans.

Organize your final decision: "Final Decision: [[Yes/Probably/Maybe/No]]". - "Yes": >85% match. - "Probably": >65% match. - "Maybe": >40% match.

Feature: [[Insert Feature Summary]] Span 1: [[Insert Span 1]] ...

Figure 6: Prompt template for verifying the fidelity of semantic decoding.

Task: 1-Premise Horn Clause Calibration
For the given premise P and conclusion C, judge whether the implication

$$P \to C$$

is a **Strict or Plausible Horn Clause**.
Classify into: 1. **Strict:** Causal/Logical relations (e.g., mathematical theorems). 2. **Plausible:** Helpful intuitions/heuristics (e.g., planning strategies). 3. **No:** Spurious/Noisy correlations.

Premise (P): [[Insert Premise Here]] **Conclusion (C)**: [[Insert Conclusion Here]]

Output JSON: "Category": "Strict/Plausible/No", "Relation/Intuition": "rationale"

Figure 7: Oracle prompt for calibrating soundness of 1-premise rules.

Task: 2-Premise Horn Clause Calibration
For paired premises P_1, P_2 and conclusion C, judge whether the implication

$$P_1 \wedge P_2 \to C$$

is a **Strict or Plausible Horn Clause**.
Classify into: 1. **Strict:** Causal/Logical relations. 2. **Plausible:** Helpful intuitions/heuristics. 3. **No:** Spurious correlations.

First Premise (P_1): [[Insert Premise 1]] **Second Premise (P_2)**: [[Insert Premise 2]] **Conclusion (C)**: [[Insert Conclusion]]

Output JSON: "Category": "Strict/Plausible/No", "Relation/Intuition": "rationale"

Figure 8: Oracle prompt for calibrating soundness of 2-premise rules.

1. Operational Substrates (Research Subjects)
The primary objects of analysis were the pre-trained checkpoints of the following model families: `Qwen-2.5`, `Llama-3.1`, `Mistral-v0.1`, and `DeepSeek-Math`. These models were deployed solely for the purpose of latent space extraction and topological auditing. No fine-tuning or weight modification was performed outside of the sparse autoencoder training. All usage adheres to the respective academic and community licenses provided by the model developers.

2. Semantic Calibration Oracle (Automated Annotation)
To achieve industrial scale in rule categorization, we deployed `DeepSeek-R1` as a "Semantic Calibration Oracle." This model functioned as a proxy for human expert annotation, classifying millions of extracted feature pairs into "Strict," "Plausible," or "Noisy" categories. This usage complies with the DeepSeek General User Policy. The prompt templates used for this orchestration are documented in Section 16.

3. Drafting Synthesizer (Writing Assistance)
An external LLM (`ChatGPT`) was utilized to assist in the syntactic refinement and structural organization of this specification document. While the AI provided suggestions for clarity and flow, all technical claims, mathematical derivations (specifically the SAL metric), and experimental verifications were generated and validated exclusively by the human authors of the Metanthropic Research team.

14 Appendix B: Holographic Crosscoder Architecture

This module defines the mathematical formalism for the Cross-Layer Sparse Autoencoder (CL-SAE) deployed in the **Holographic Feature Extraction** phase (Section 2.2).

Given an L-layer pre-trained substrate producing residual-stream hidden states $\{\mathbf{x}^l\}_{l=1}^L$ where $\mathbf{x}^l \in \mathbb{R}^D$, we initialize a Crosscoder f_{SAE} to demultiplex these states into a sparse, mono-semantic feature representation. The architecture comprises L pairs of trainable encoder-decoder weights $\{(\mathbf{E}^l, \mathbf{D}^l)\}_{l=1}^L$, where $\mathbf{E}^l, \mathbf{D}^l \in \mathbb{R}^{D \times C}$ and the feature dimension $C \gg D$ (Overcomplete Basis).

For each layer l, the Crosscoder projects the hidden state $\mathbf{x}^l$ into a sparse, non-negative feature space $\mathbf{h}^l = \mathrm{ReLU}(\mathbf{x}^l\mathbf{E}^l) \in \mathbb{R}_+^C$. The decoder reconstructs the state $\hat{\mathbf{x}}^l$ utilizing activations from the current and all preceding causal layers:

$$\hat{\mathbf{x}}^l \ = \ \sum_{l'=1}^{l} \mathbf{h}^{l'} \mathbf{D}^{l\top}.$$

This cross-layer recurrence allows the protocol to capture features at their precise layer of emergence. The optimization objective $\mathcal{L}$ minimizes reconstruction error under an L_1 sparsity constraint:

$$\mathcal{L} = \sum_{l=1}^{L} \|\mathbf{x}^l - \hat{\mathbf{x}}^l\|^2 + \alpha \cdot \sum_{l'=1}^{L} \sum_{c=1}^{C} \|\mathbf{h}_c^l \cdot \mathbf{D}_{:,c}^{l'\top}\|_1, \tag{7}$$

where α is the sparsity coefficient. The penalty is applied to the feature activation $\mathbf{h}_c^l$ weighted by its decoder norm, ensuring penalization only occurs during active reconstruction.

15 Appendix C: Feature Extraction & Semantic Decoding

15.1 Protocol: SAE Training Configuration

We adhere to the training stability protocols established by Lindsey et al. (2024b) and Gao et al. (2024).

- **Architecture:** Latent dimension $C = 2^{15}$ (32,768 features); Depth $L = 8$ equidistributed layers. For 28-layer models (e.g., `Qwen-2.5-7B`), we sample every 4th layer.
- **Optimization:** AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 6.25 \times 10^{-10}$).
- **Learning Rate:** 2×10^{-4} with linear cool-down in the final 20% of steps.
- **Sparsity Penalty:** $\alpha = 5e^{-3}$ with linear warm-up over the initial 20% of steps.
- **Throughput:** Batch size fixed at 128 sequences ($\approx 60,000$ tokens/batch) to prevent feature collapse. Training duration is 5,000 steps (≈ 5 epochs).

Convergence: The trained modules achieve a normalized MSE of 0.65–0.80 with an average sparsity of ≈ 20 active features per token. As shown in Table 3, the "Dead Feature" rate is negligible ($< 3.5\%$ average), confirming high utilization of the latent space.

15.2 Protocol: Automated Semantic Interpretation

We deploy an automated interpretation pipeline (Bills et al., 2023) utilizing `DeepSeek-R1` as the Semantic Oracle. **Workflow:** 1. Feed 128K calibration tokens to the fixed SAE. 2. For each feature c, aggregate the top-15 maximally activating text spans. 3. Prompt the Oracle (Figure 9) to synthesize a semantic summary. 4. Verify summary fidelity (Figure 10) using top-30 spans. Features with "Maybe" confidence or higher ($\geq 40\%$ semantic match) are retained for the Logic Rule Aggregation phase.

Table 3: Feature Vitality & Explainability Audit. "Explainable Rate" denotes the fraction of features successfully mapped to semantic concepts by the Oracle.

Substrate	Dead Rate	Explainable Rate
`Qwen-0.5B`	0.00%	82.59%
`Qwen-1.5B`	3.17%	92.61%
`Qwen-7B`	17.98%	96.27%
`Qwen-14B`	0.34%	86.48%
`Llama-8B`	2.20%	95.45%
`Mistra-7B`	0.09%	76.02%
`Deepseek-7B`	0.21%	89.14%
Avg.	3.43%	88.37%

16 Appendix D: Logical Topology Extraction Protocol

This section details the **Probabilistic Rule Aggregation** module (Section 2.3).

Calibration Substrate: We utilize a stratified subset of the MATH dataset (Hendrycks et al.) (3,267 samples) to ensure coverage of diverse reasoning primitives (Algebra, Number Theory, etc.).

Engineering Optimization: To manage the combinatorial explosion of the search space ($\binom{32768}{3} \approx 5.8 \times 10^{12}$), we implement the following acceleration vectors:

1. **Token Aggregation:** Feature activations are summed along the sequence length at the final token, flattening the tensor from $\mathbb{R}^{N\times L\times C}$ to $\mathbb{R}^{C}$.
2. **Vectorized Co-occurrence:** We implement custom CUDA kernels (Algorithm 4) to compute conditional frequencies $\mathrm{count}(c_1, c_2)$ directly from the flattened activations.
3. **Distributed Counting:** The counting process is sharded across multiple nodes, reducing wall-clock time to ≈ 30 hours per model.

Soundness Calibration (Oracle Prompts): The logic rules are categorized using `DeepSeek-R1` ($T = 0.1$, $top\_p = 0.9$). The prompts enforce a strict taxonomy: "Strict" (Causal), "Plausible" (Heuristic), or "No" (Noise). Figures 11 and 12 detail the exact instructions used to calibrate the SAL metric.

Human Verification: A manual audit of 60 randomly sampled rules yields an agreement rate of 0.566 with the Oracle. Error analysis reveals the Oracle is conservative, occasionally misclassifying "Strict" mathematical relations as "Plausible," which ensures the SAL metric remains a lower-bound estimator of reasoning potential.

17 Appendix E: Computational Infrastructure Specs

All experiments were conducted on a high-performance compute cluster comprising three nodes. **Node Specification:**

- **Compute:** 8 × NVIDIA A100 (80GB VRAM).
- **Host:** 96 vCPU cores, 1TB RAM.
- **Storage:** 8TB NVMe (Cloud-attached).

Resource Consumption:

- **SAE Training:** ≈ 60 hours per node for `Qwen-2.5-14B`.
- **Topology Extraction:** ≈ 50 hours (Memory bound: 500GB RAM peak).

Algorithm 3 Module: Feature Activation Aggregation

```
Require: $x$ with len(x.shape) == 3                              ▷ $x \in \mathbb{R}^{L\times N\times C}$
Require: feat_idx, threshold = $T$
$x \leftarrow \mathrm{cumsum}(x, \mathrm{axis} = 0)[-1]$                        ▷ $x \in \mathbb{R}^{N\times C}$
$x \leftarrow \big(x[:,\ \mathrm{feat\_idx}] > T\big).\mathrm{bfloat16}()$            ▷ $x \in \{0,1\}^{N\times C'}$
$x \leftarrow \mathrm{cumsum}(x, \mathrm{axis} = 0)[-1]$                        ▷ $x \in \mathbb{N}^{C'}$
return $x$                                   ▷ length-$C'$ vector of per-feature counts
```

Algorithm 4 Module: Vectorized Rule Counter ($P = 1, P = 2$)

```
Require: $x$: vector of layer counts per feature (length $C$).
// Initialize Registry
Counts = {}
$A \leftarrow \{\, c : x[c] > 0 \,\}$

// Record Priors
for all $p \in A$ do
    Counts[(p,)][-1] += 1
end for

// 1-Premise Count ($p \Rightarrow q$)
pair $\leftarrow (x{>}0)[:,\texttt{None}] \wedge (x{>}0)[\texttt{None},:]$    ▷ $C{\times}C$
smaller $\leftarrow \big(x[\texttt{None},:] < x[:,\texttt{None}]\big) \wedge$ pair
$(\texttt{prem}, \texttt{concl}) \leftarrow \texttt{NonZero}(\texttt{smaller})$
for $i \leftarrow 1$ to len(prem) do
    $p \leftarrow \texttt{prem}[i]$, $q \leftarrow \texttt{concl}[i]$
    Counts[(p,)][q] += 1
end for

// 2-Premises Count ($p_1 \wedge p_2 \Rightarrow q$)
prod $\leftarrow$ einsum("ac,bc->abc", smaller, smaller)    ▷ $C{\times}C{\times}C$
$(r, c) \leftarrow \texttt{LowerTriangularIndices}(C)$
prod[r, c, :] $\leftarrow 0$    ▷ enforce $p_1{<}p_2$, drop diagonals
$(p_1, p_2, q) \leftarrow \texttt{NonZero}(\texttt{prod})$
for $i \leftarrow 1$ to len($p_1$) do
    Counts[(p_1[i], p_2[i])][q[i]] += 1
end for
```

Directive: Feature Semantic Decoding

We are studying the behaviors of neurons from a language model. Look at the text spans activated by the neuron and summarize what feature the neuron is looking for. Pay attention to __the ending of each span__. Your summary should be one (short) sentence describing the most significant feature.

Organize your final summary within the special tag: <summary> summary here </summary>. - If there is one short lexical pattern: <summary> Exact pattern: "Key Pattern" with context </summary>. - If there are semantic patterns: <summary> Semantic: semantic concept, with "Exemplar Patterns" </summary>. - If unclear: <summary> Cannot Tell </summary>.

Keep your <think> block short.

The following are text spans that can maximally activate a certain neuron:
Span 1: [[Insert Span 1 Here]] ...

Figure 9: Prompt template for the Oracle to decode latent features. Utilized by DeepSeek-R1.

Directive: Semantic Verification
You are a linguistic expert. Determine whether the given feature is fuzzy matched by the text spans.

Organize your final decision: "Final Decision: [[Yes/Probably/Maybe/No]]". - "Yes": >85% match. - "Probably": >65% match. - "Maybe": >40% match.

Feature: [[Insert Feature Summary]] Span 1: [[Insert Span 1]] ...

Figure 10: Prompt template for verifying the fidelity of semantic decoding.

Task: 1-Premise Horn Clause Calibration
For the given premise P and conclusion C, judge whether the implication

$$P \to C$$

is a **Strict or Plausible Horn Clause**.
Classify into: 1. **Strict:** Causal/Logical relations (e.g., mathematical theorems). 2. **Plausible:** Helpful intuitions/heuristics (e.g., planning strategies). 3. **No:** Spurious/Noisy correlations.

Premise (P): [[Insert Premise Here]] **Conclusion (C)**: [[Insert Conclusion Here]]

Output JSON: "Category": "Strict/Plausible/No", "Relation/Intuition": "rationale"

Figure 11: Oracle prompt for calibrating soundness of 1-premise rules.

Task: 2-Premise Horn Clause Calibration
For paired premises P_1, P_2 and conclusion C, judge whether the implication

$$P_1 \land P_2 \to C$$

is a **Strict or Plausible Horn Clause**.
Classify into: 1. **Strict:** Causal/Logical relations. 2. **Plausible:** Helpful intuitions/heuristics. 3. **No:** Spurious correlations.

First Premise (P_1): [[Insert Premise 1]] **Second Premise (P_2)**: [[Insert Premise 2]] **Conclusion (C)**: [[Insert Conclusion]]

Output JSON: "Category": "Strict/Plausible/No", "Relation/Intuition": "rationale"

Figure 12: Oracle prompt for calibrating soundness of 2-premise rules.