CONTROLLING LOGICAL COLLAPSE IN LLMS VIA ALGEBRAIC ONTOLOGY PROJECTION OVER $\mathbb{F}_2$

Hisashi Miyashita
Mgnite Inc.
himi@mgnite.com

ABSTRACT

Do large language models internally encode ontological relations in a formally verifiable algebraic structure? We introduce **Algebraic Ontology Projection (AOP)**, which projects LLM hidden states into the Galois Field $\mathbb{F}_2$ under Liskov Substitution Principle constraints, using only 42 relational pairs as algebraic keys. AOP achieves up to **90.91%** zero-shot inclusion accuracy on unseen concept pairs across Gemma and Qwen— with no model tuning, but through prompt alone.

This algebraic structure is strongly layer-dependent. We introduce **Semantic Crystallisation (SC)**, a metric that quantifies $\mathbb{F}_2$ constraint satisfaction relative to a random baseline and predicts zero-shot accuracy without held-out data. System prompts act as **algebraic boundary conditions**: only their combination with instruction tuning prevents **Late-layer Collapse**—a systematic degradation of logical consistency in the final layers, observed in 7 of 10 conditions.

These findings reframe forward computation as an iterative process of algebraic organisation, and open a path toward LLMs whose logical structure is not merely approximated, but formally accessible.

1 INTRODUCTION

When a large language model correctly identifies that an Eagle *is a* Bird, that a Salmon *is a* Fish, and that a Diamond is *not* an Insect, is this behaviour the product of geometric proximity in embedding space—or does the model internally encode something more structured?

We provide evidence for the latter. Using a projection of LLM hidden states onto a binary vector space under hierarchical constraints, we show that relational structure—*is-a*, *has-a*, and negation—is not merely approximated geometrically, but encoded in a form that is extractable by linear means, and that generalises to concepts never seen during projection.

Concretely, we train a two-layer linear network on just 42 relational pairs drawn from four semantic domains. Evaluated on held-out entities and relations that never appear during training, our method achieves up to **90.91% inclusion accuracy** across Gemma-2 and Qwen2.5— with no model tuning, but through prompt alone— substantially above chance and competitive with models trained on orders of magnitude more supervision. These results hold for concepts entirely absent from the training set—including cross-domain pairs such as *Copper* $\subseteq$ *Metal* and *Marble* $\subseteq$ *Rock*—demonstrating genuine structural generalisation rather than memorisation.

Standard representational analyses—probing classifiers, Centered Kernel Alignment (CKA), and Singular Vector Canonical Correlation Analysis (SVCCA)—treat meaning as geometry: concepts are points, and similarity is distance. This perspective cannot express the asymmetric, transitive structure of hierarchical relations. Knowing that *Eagle* $\subseteq$ *Bird* is not the same as knowing that Eagle and Bird are nearby in embedding space: the former is directional and non-destructive, the latter is symmetric and admits no formal entailment. Our approach asks a different question: not how close are these representations, but whether one *contains* the other in a formal algebraic sense.

We introduce **Algebraic Ontology Projection (AOP)**, which maps LLM hidden states into a binary vector space $\{0, 1\}^n$ under the constraint that hierarchical relations (*is-a*) correspond to bitwise

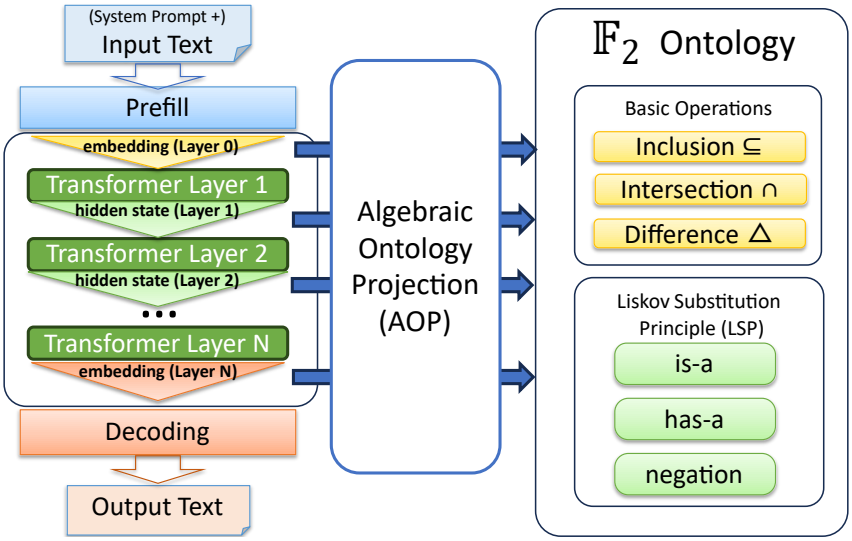

Figure 1: Overview of Algebraic Ontology Projection (AOP). Each Transformer layer's hidden states are projected onto an $\mathbb{F}_2$ binary vector space under relational constraints (is-a, has-a, negation). The system prompt, when present, is processed as prefill and serves as an algebraic boundary condition that configures the hidden state before projection (Section 3).

inclusion:

$$\mathbf{a} \odot \mathbf{b} = \mathbf{a} \quad \Leftrightarrow \quad A \subseteq B, \tag{1}$$

where $\odot$ is elementwise multiplication. The projection is implemented as a two-layer linear network with a sharp activation function—structurally analogous to the language model head (lm_head), but targeting a binary algebraic space rather than a vocabulary distribution. The projection is learned from a small set of relational **algebraic keys**: concept pairs whose relations are known, used not as statistical training signal but as structural constraints to unlock latent organisation already present in the model. We term the extracted structure an **Algebraic Ontology**.

To extract hidden states for a given concept, we apply **Localized Mean Pooling (LMP)**: the hidden states corresponding to the concept's context tokens are averaged across the relevant layer. When a system prompt is present, it is processed as a prefill—shaping the model's internal state—but is excluded from the averaged representation. This separation ensures that the projection captures the concept's representation within the configured context, rather than a global average over the full input.

A central finding is that relational structure is not uniformly present across layers. We introduce **Semantic Crystallisation (SC)**, a dimensionless structural metric defined as:

$$SC(L) = \big(\mu_{rand} - q(L)\big) \cdot \sigma^2_{rand}, \tag{2}$$

where $q(L) = \mathcal{L}_{alg}(L)/\rho(L)^2$ is the density-normalised algebraic loss relative to a random baseline (Section 3.4). SC typically takes values between $[-1, 1]$: positive values indicate layers where algebraic structure exceeds the random baseline; $SC < 0$ (**Semantic Melting**) indicates active suppression of relational structure. Crucially, SC scores *predict* zero-shot generalisation performance across layers, providing a gradient-free criterion for layer selection without held-out evaluation data.

A second central finding concerns the role of system prompts. We observe two qualitatively distinct behaviours across model families. In **Autonomous Crystallisation** (Gemma), unprompted layers are near-random ($SC \approx 0$), and a structured system prompt elevates specific layers to $SC > 1.5$. In **Induced Crystallisation** (Qwen), unprompted conditions yield lower average SC, with specific layers exhibiting $SC < 0$; structured prompting substantially reverses this. These findings reframe the function of system prompts: beyond behavioural steering, they act as **structural configuration signals** that determine whether—and where—relational organisation emerges in the model's hidden states.

We further observe a systematic **Late-layer Collapse**: in 7 of 10 evaluated conditions, zero-shot accuracy degrades significantly in the final layers of the model. Only the combination of using an *instruction-tuned model* (Instruct variant) and providing a *structured system prompt at inference time* maintains both high peak accuracy and stable performance through the final layer. These two factors play complementary roles: the instruction-tuned model provides a stable representational substrate (a property of the model weights), while the system prompt configures the algebraic boundary conditions that sustain it (a property of the inference-time input).

Beyond the interpretability findings reported here, the SC metric opens a direct path toward *quantitative prompt engineering*: by analysing SC contributions at the token level, one can identify which tokens in a system prompt strengthen or weaken relational structure, reducing prompt design from trial-and-error to a structured optimisation problem. We report these results in a companion paper.

Taken together, these results suggest that LLMs do not merely approximate relational knowledge geometrically: they encode it in a structured form that is accessible to linear algebraic projection, raising new possibilities for interpretable, verifiable, and formally grounded AI systems.

Contributions.

1. **Algebraic Ontology Projection (AOP).** A two-layer linear projection method that extracts binary symbolic representations of *is-a*, *has-a*, and negation relations from LLM hidden states using a minimal set of relational constraints as algebraic keys, with demonstrated applicability across mpnet, Qwen2.5, and Gemma-2.
2. **Semantic Crystallisation (SC).** A quantitative, baseline-calibrated measure of relational structure in LLM hidden states that (i) reveals two distinct modes of structural organisation across model families (Autonomous and Induced Crystallisation), and (ii) predicts zero-shot generalisation performance, enabling gradient-free layer selection.
3. **Contextual Structural Configuration.** Empirical demonstration that system prompts function as structural configuration signals, transitioning hidden states between Semantic Melting and Crystallised regimes, with the choice of instruction-tuned vs base model and the inference-time system prompt playing complementary roles in sustaining late-layer stability.
4. **Zero-Shot Ontological Generalisation.** Up to 90.91% inclusion accuracy on unseen concept pairs from a projection trained on 42 relational constraints, with systematic characterisation of Late-layer Collapse and the conditions under which relational consistency is maintained throughout the model depth.

2 Related Work

Probing and representational analysis. Linear probing, introduced by Alain & Bengio (Alain & Bengio, 2016) for analysing intermediate representations in image classifiers, was subsequently extended to transformer hidden states. Hewitt & Manning (Hewitt & Manning, 2019) showed that syntactic dependency structure can be recovered from BERT hidden states via a learned linear transformation into a *distance space*, where proximity encodes syntactic closeness. However, distance-based probes are inherently symmetric: $d(A, B) = d(B, A)$, and cannot represent the directional structure of ontological relations. Knowing that representations of *Eagle* and *Bird* are nearby does not establish whether Eagle $\subseteq$ Bird or Bird $\subseteq$ Eagle. AOP addresses this by projecting into $\mathbb{F}_2^n$, where inclusion is directional by construction: $\mathbf{a} \odot \mathbf{b} = \mathbf{a}$ implies $A \subseteq B$, not the reverse. Representational similarity methods including CKA (Kornblith et al., 2019) and SVCCA (Raghu et al., 2017) share the same geometric limitation, measuring proximity rather than algebraic structure.

Mechanistic interpretability and monosemanticity. A central programme in mechanistic interpretability (Olah et al., 2020) seeks to identify the computational roles of individual neurons and circuits within transformer models. The superposition hypothesis (Elhage et al., 2022) proposes that LLMs encode more features than available dimensions by exploiting near-orthogonal directions in activation space. Anthropic's work on monosemanticity (Bricken et al., 2023; Templeton et al., 2024) operationalises this by applying sparse autoencoders (SAEs) to decompose superposed representations into interpretable, monosemantic features—a major advance in identifying *which*

features exist within a model. AOP is complementary rather than competing: where monosemanticity research asks *which features are present*, AOP asks whether those features stand in the *algebraic relations*—inclusion, intersection, negation—that formal ontologies require. Identifying a feature for "Diamond" and a feature for "Mineral" does not, by itself, establish that the model represents Diamond $\subseteq$ Mineral as a formal algebraic constraint; AOP tests precisely this.

Knowledge representation in LLMs. Petroni et al. (Petroni et al., 2019) demonstrated that LLMs encode factual and relational knowledge without fine-tuning, evidenced by strong performance on cloze-style queries evaluated at the output level. This behavioural finding raises the deeper question of whether the underlying representations satisfy the formal algebraic constraints that define those relations. AOP provides a structural counterpart: not only can LLMs answer relational queries correctly, but their internal representations satisfy the formal algebraic constraints that define those relations—a stronger claim that goes beyond output accuracy to the representational substrate.

Meng et al. (Meng et al., 2022) advanced this by using causal tracing to localise factual associations within specific MLP layers—analogous to identifying definition points in a dataflow analysis. Their results confirm an important role for mid-layer feed-forward modules in storing factual associations, but the side effects of editing these associations on neighbouring knowledge remain difficult to control, a consequence of the non-linear information mixing across layers. AOP takes a complementary approach: rather than intervening in the forward pass, it identifies layers where algebraic structure is already accessible, without perturbation.

Burns et al. (Burns et al., 2023) introduced Contrast-Consistent Search (CCS), which recovers scalar truth values from hidden states without supervision by searching for a direction that separates true from false propositions—the approach most closely related in spirit to AOP. CCS applies a one-dimensional linear projection $\sigma(\boldsymbol{\theta}^\top \mathbf{x} + b)$, which can be viewed as a scalar special case of AOP's n-dimensional $\mathbb{F}_2$ projection. The zero-shot baselines in their evaluation operate at the output level via next-token log probabilities rather than hidden state geometry, with honest zero-shot accuracy remaining below 80% on most benchmarks. AOP differs from CCS in two respects: it recovers a full algebraic system—closed under inclusion, intersection, and negation—rather than a scalar truth value, and it achieves up to 90.91% zero-shot accuracy on unseen concept pairs trained on a limited set of relational constraints.

Formal methods and neural networks. Nanda et al. (Nanda et al., 2023) demonstrated that small transformers solving modular arithmetic develop internal Fourier representations corresponding to the periodic structure of $\mathbb{Z}/p\mathbb{Z}$, and undergo a three-stage learning process—memorization, circuit formation, and cleanup—in which algebraic structure emerges spontaneously without explicit supervision. The authors note that their progress measures are specific to small networks on a single algorithmic task, and that task-independent measures are necessary for broader applicability. AOP addresses this by grounding the projection in $\mathbb{F}_2$—a structure whose algebraic closure applies regardless of task—and demonstrates consistent results across multiple model families and semantic domains without task-specific training.

Geiger et al. (Geiger et al., 2021) developed causal abstraction to test whether neural networks implement specific task-defined causal models via interchange interventions. AOP adopts a different computational model—algebraic relations over $\mathbb{F}_2$ grounded in the Liskov Substitution Principle—and asks whether pre-trained hidden states satisfy this structure, rather than verifying a task-specific causal hypothesis.

Neuro-symbolic approaches (d'Avila Garcez et al., 2009) have sought to integrate logical constraints with neural computation, typically by training models on logical supervision or enforcing constraints at the output level. AOP differs in that it does not train LLMs to satisfy logical constraints: it *projects* pre-trained representations to test whether constraints are already latently satisfied. The distinction is between teaching a model logic and discovering whether it has already acquired it.

Collectively, existing methods analyse LLM representations geometrically, behaviourally, through causal intervention, or through scalar feature decomposition—but none directly measures whether hidden states satisfy the algebraic invariants required by formal ontologies as a closed system. AOP addresses this gap, and in doing so reveals systematic layer-dependent patterns—Semantic Crystallisation and Late-layer Collapse—that are invisible to existing analyses.

3 ALGEBRAIC ONTOLOGY PROJECTION

We introduce **Algebraic Ontology Projection (AOP)**, a framework that extracts formal relational structure from LLM hidden states by projecting them into a binary vector space equipped with algebraic operations.

3.1 BINARY ALGEBRAIC STRUCTURE OVER $\mathbb{F}_2$

Representing concepts as binary vectors. We represent each concept as a binary vector $\mathbf{a} \in \{0,1\}^n$, where each bit encodes the presence or absence of a latent semantic feature. The key operation in $\mathbb{F}_2^n$ is the elementwise product $\odot$ (bitwise AND), which computes the *intersection* of two concept representations: $\mathbf{a} \odot \mathbf{b}$ extracts the features shared by both A and B.

Under this interpretation, the *is-a* relation $A \subseteq B$—"every A is a B"—is defined as:

$$A \subseteq B \;\Leftrightarrow\; \mathbf{a} \odot \mathbf{b} = \mathbf{a}, \tag{3}$$

meaning that the features of A are entirely contained within those of B: the intersection of A and B recovers A itself. Note carefully that this is a condition on *intersection*, not on bit count: $\mathbf{a} \odot \mathbf{b} = \mathbf{a}$ requires that every bit active in $\mathbf{a}$ is also active in $\mathbf{b}$, but $\mathbf{b}$ may activate additional bits.

Feature accumulation across the hierarchy. A crucial consequence of this formulation is that *subordinate concepts accumulate more active bits than superordinate concepts*. Consider the chain *Animal* $\supseteq$ *Insect* $\supseteq$ *Beetle* $\supseteq$ *StagBeetle*: *StagBeetle* inherits all features of *Beetle*, which inherits all features of *Insect*, which inherits all features of *Animal*—plus, at each level, concept-specific features are added. *StagBeetle* therefore activates strictly more bits than *Animal*, not fewer. This mirrors the intuition that more specific concepts are *richer in features* than abstract ones, and is reinforced by LSP inheritance: *has-a* attributes are propagated downward, further enriching subordinate representations.

Why binary? The binary vector space $\{0,1\}^n$ is isomorphic to the power set lattice $2^{[n]}$—the natural algebraic structure of ontological hierarchies, in which concepts are sets of features and containment corresponds to subsumption. Formally, $\{0,1\}^n$ equipped with elementwise addition modulo 2 and elementwise multiplication (AND) forms the Galois Field $\mathbb{F}_2^n$, guaranteeing *algebraic closure*: any composition of bitwise operations remains within $\mathbb{F}_2^n$ without requiring normalisation such as softmax.

Relational operations. The full set of ontological relations maps directly to $\mathbb{F}_2^n$ operations:

$$A \subseteq B \;\Leftrightarrow\; \mathbf{a} \odot \mathbf{b} = \mathbf{a} \quad (\textit{is-a}) \tag{4}$$

$$A \cap B \;\Leftrightarrow\; \mathbf{a} \odot \mathbf{b} \quad \text{(feature intersection)} \tag{5}$$

$$A \triangle B \;\Leftrightarrow\; \mathbf{a} \oplus \mathbf{b} \quad \text{(symmetric difference)} \tag{6}$$

The *has-a* relation (A has part P) is defined analogously: $\mathbf{p} \odot \mathbf{a} = \mathbf{p}$, i.e., all features of the part are present in the whole. Negation (A is not B) requires $\mathbf{a} \odot \mathbf{b} = \mathbf{0}$: no shared features. We term the structure defined by these operations an **Algebraic Ontology**.

3.2 PROJECTION ARCHITECTURE

The AOP projection maps a hidden state $\mathbf{h} \in \mathbb{R}^d$ to a soft-binary vector $\mathbf{z} \in [0,1]^n$ via a two-stage transformation:

$$\mathbf{z} = \sigma(\gamma \cdot W_2 \cdot \tanh(W_1 \mathbf{h} - \boldsymbol{\theta}))\,, \tag{7}$$

where $W_1 \in \mathbb{R}^{n \times d}$ is the **attribute extraction layer**; $\boldsymbol{\theta} \in \mathbb{R}^n$ is a **learnable per-dimension threshold**; $W_2 \in \mathbb{R}^{n \times n}$ is the **logical mapping layer**; $\gamma = 4.0$ is a fixed sharpness parameter; and σ is the sigmoid function applied elementwise.

Role of the adaptive threshold. The learnable threshold $\boldsymbol{\theta}$ independently calibrates each bit's activation tendency. Dimensions with $\theta_i < 0$ activate readily, corresponding to abstract features shared by many concepts (e.g., features common to all *Animal* instances). Dimensions with $\theta_i > 0$ activate

only under strong evidence, corresponding to specific features held by few concepts (e.g., features unique to *StagBeetle*). This asymmetry reflects the feature accumulation property of Section 3.1: superordinate concepts activate few, broadly-shared features; subordinate concepts activate many, including all inherited features.

Binarisation. During training, outputs are soft-valued in $[0, 1]$, enabling gradient-based optimisation. During evaluation, outputs are binarised at 0.5. The sharpness parameter $\gamma = 4.0$ encourages outputs to concentrate near $\{0, 1\}$ during training.

Analogy with lm_head. AOP is structurally analogous to the standard language model head: both apply a linear transformation to LLM hidden states to decode structured information. The key differences are the target space ($\mathbb{F}_2^n$ rather than a vocabulary distribution) and the discretisation mechanism (adaptive thresholding rather than softmax). Where lm_head must impose closure over a probability simplex via softmax, AOP inherits algebraic closure from the field structure of $\mathbb{F}_2^n$ by construction.

3.3 Training Protocol

Algebraic keys. AOP is trained on a minimal set of relational pairs that we term **algebraic keys**: formal constraints whose relational type is known, used as algebraic anchors that render the model's latent structure legible rather than as statistical training data.

Dataset structure. The relational dataset is organised into four progressive stages (levels 1, 2, 4, and 8), spanning four semantic domains (biological, mineral, physical, and abstract), each introducing new concepts and relations to expand both the semantic diversity and the difficulty of the algebraic constraints. All four stages are used simultaneously during training. Level 1 establishes a basic insect–animal hierarchy; levels 2 and 4 extend this with additional biological concepts; level 8 introduces a heterogeneous mineral domain, testing cross-domain generalisation of the projection. The complete dataset comprises 42 training pairs (15 *is-a*, 12 *has-a*, 15 negation) and 13 independent evaluation pairs (i_neg, $D_{val} \cap D_{train} = \emptyset$). The independent evaluation set consists of concept pairs drawn from outside the semantic domains of the training data—pairing abstract concepts with physical attributes (e.g., *Idea* vs *Legs*) and cross-domain entities (e.g., *DNA* vs *Cloud*). These pairs are never used during training and serve exclusively to verify structural generalisation. Full dataset details are provided in Appendix A.

Loss function. The training objective jointly enforces four classes of algebraic constraint:

(i) Is-a inclusion ($\mathcal{L}_{isa}$): penalises cases where the parent representation lacks bits that the child representation activates, enforcing $\mathbf{a}_{child} \odot \mathbf{a}_{parent} = \mathbf{a}_{child}$.

(ii) Has-a inclusion ($\mathcal{L}_{has}$): penalises cases where the whole concept lacks bits present in the part representation, enforcing $\mathbf{a}_{part} \odot \mathbf{a}_{parent} = \mathbf{a}_{part}$. This term is weighted more strongly than $\mathcal{L}_{isa}$, reflecting the tighter compositional constraint of part-whole relations.

(iii) LSP inheritance ($\mathcal{L}_{lsp}$): the most critical constraint, enforcing that the child concept inherits the intersection of parent and part features:

$$\mathcal{L}_{lsp} = \mathbb{E}[\max(\mathbf{a}_{parent} \odot \mathbf{a}_{part} - \mathbf{a}_{child},\ \mathbf{0})]\,, \tag{8}$$

where $\odot$ extracts the features that the parent possesses *as part*, which the child must also possess by the Liskov Substitution Principle.

(iv) Separation and density regularisation: three additional terms prevent degenerate solutions. A *separation* term enforces that negation pairs maintain an appropriate Hamming distance, discouraging both complete overlap and complete orthogonality. A *density* term targets different bit densities for superordinate and subordinate concepts, reflecting the feature accumulation property: superordinate concepts should activate fewer bits (more abstract) and subordinate concepts more bits (more specific). An *anti-zero* term penalises representations in which no bits are active, preventing the trivially satisfied but uninformative zero solution. An *orthogonality* term prevents concept-specific bits of the child from overlapping with part features inherited from the parent, preserving the distinctiveness of each level of the hierarchy.

Zero-shot evaluation metric. For evaluation, we measure the **inclusion score**:

$$\text{Inclusion}(A, B) = \frac{|\mathbf{a} \odot \mathbf{b}|}{|\mathbf{a}|}, \tag{9}$$

where $|\cdot|$ denotes the Hamming weight. This measures the fraction of A's active bits that are also active in B—directly operationalising the $\mathbb{F}_2$ definition of *is-a*. A pair is classified as *is-a* if $\text{Inclusion}(A, B) \geq \tau = 0.7$ and the complementary Hamming distance satisfies $\delta = 0.1$.

3.4 Semantic Crystallisation (SC)

Motivation. The degree to which a given layer's hidden states satisfy $\mathbb{F}_2$ algebraic constraints varies substantially across layers and model configurations. To enable dimensionless, cross-architecture comparison, we introduce **Semantic Crystallisation (SC)**: a dimensionless indicator that quantifies the emergence of algebraic order within the latent representation space, evaluated relative to the intrinsic variance of the model's stochastic baseline.

Density-invariant algebraic loss. Let $\mathcal{L}_{alg}$ denote the raw algebraic inconsistency loss and ρ the mean bit activation rate (*bit density*) of the projected representations. We define the **density-normalised algebraic loss**:

$$q(L) = \frac{\mathcal{L}_{alg}(L)}{\rho(L)^2}, \tag{10}$$

where the $1/\rho^2$ factor isolates the structural alignment from the quadratic increase in stochastic collisions inherent in high-density representations.

Stochastic baseline and scaling factor. For each model architecture, we apply AOP to a randomly initialised model of identical architecture and compute:

$$\mu_{rand} = \mathbb{E}[q_{rand}], \qquad \sigma^2_{rand} = \text{Var}(q_{rand}), \tag{11}$$

where μ_{rand} and σ^2_{rand} characterise the expected noise floor and the intrinsic fluctuations of the architecture, respectively.

SC as a dimensionless order parameter.

$$SC(L) = \big(\mu_{rand} - q(L)\big) \cdot \sigma^2_{rand}, \tag{12}$$

SC is a **dimensionless quantity**. The term $(\mu_{rand} - q(L))$ represents the degree to which algebraic structure has emerged from the stochastic background, while σ^2_{rand} acts as a characteristic scaling factor that ensures crystallisation is evaluated relative to the architecture's inherent capacity for representational diversity, enabling consistent comparison across heterogeneous model families. SC typically takes values between $[-1, 1]$.

We distinguish three regimes:

- **Crystalline phase** ($SC \approx 1$): logical structures are dominant and robust; relational constraints are satisfied well above chance.
- **Gas/Vacuum phase** ($SC \approx 0$): maximum entropy state; representations are indistinguishable from the random baseline.
- **Collapsed phase** ($SC < 0$): algebraic structures are actively suppressed below the random baseline (**Semantic Melting**).

SC as a layer selection criterion. We demonstrate empirically (Section 4.2) that SC scores predict zero-shot generalisation performance, providing a gradient-free criterion for selecting the optimal projection layer without held-out evaluation data.

3.5 LOCALIZED MEAN POOLING (LMP)

Concept representation. To obtain the hidden state representation of a concept, we query the model with a short context string and average the hidden states of the context tokens at the target layer:

$$\mathbf{h}_{concept} = \frac{1}{|T_{ctx}|} \sum_{t \in T_{ctx}} \mathbf{h}_t^{(L)}, \tag{13}$$

where T_{ctx} indexes the tokens of the concept context and L is the target layer.

System prompt separation. When a system prompt is present, it is processed as *prefill*—configuring the model's hidden state as an algebraic boundary condition—but its tokens are excluded from the average in Equation 13:

$$\mathbf{h}_{concept} = \text{LMP}\left(\underbrace{[\text{system prompt}]}_{\text{prefill only}} + \underbrace{[\text{concept context}]}_{\text{averaged}} \right). \tag{14}$$

This separation reflects the distinct roles of global context as an algebraic boundary condition and local context as the concept query.

Scope and limitations. LMP is effective when the concept context is short and the target concept dominates the token sequence. For longer contexts, mean pooling dilutes the concept signal, degrading projection quality. All experiments use minimal context strings to satisfy this condition. Extension to token-level projection is a direction for future work.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate AOP on three model families: Gemma-2 (Team et al., 2024) (2B parameters, 26 Transformer layers, layer indices 0–26, hidden size 2048), Qwen2.5 (Team, 2024) (1.5B parameters, 28 Transformer layers, layer indices 0–28, hidden size 1536), and mpnet (Reimers & Gurevych, 2019) (12 Transformer layers, layer indices 0–12, hidden size 768), where layer index 0 denotes the input embedding layer and the final index denotes the output embedding layer. All models are loaded in bfloat16 precision. We note that Gemma-2 and Qwen2.5 differ in parameter count and architecture; differences in AOP performance across model families should be interpreted accordingly. We evaluate both instruction-tuned and base variants where available. For each architecture, a randomly initialised model of identical configuration serves as the algebraic noise baseline for SC computation (Section 3.4). All models are evaluated without fine-tuning; only the AOP projection weights are trained.

Projection dimension. We use a fixed projection dimension of $n = 2048$ across all models, matching the hidden state dimension of Gemma and providing a common basis for cross-model comparison. The effect of projection dimension on AOP performance is left for future investigation.

Training dataset. The algebraic key dataset comprises 42 symbolic constraints across four semantic domains, as described in Section 3.3. A held-out evaluation set of 11 concept pairs (unseen entities and relations) is used for zero-shot generalisation assessment.

Optimisation. We optimise using AdamW with weight decay 10^{-2} and a ReduceLROnPlateau scheduler (reduction factor 0.5, patience 200 steps). Loss terms employ SoftPlus activations in place of ReLU for numerical stability across model architectures of varying hidden state dimensions. Training is terminated upon detection of loss instability (buckling), and the checkpoint achieving the lowest training loss prior to instability is retained for evaluation.

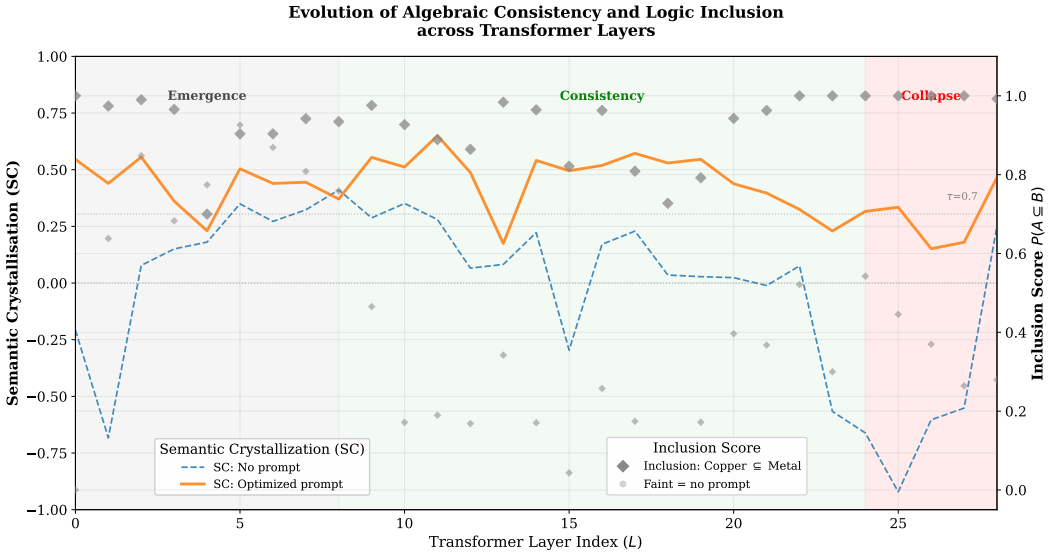

Figure 2: Semantic Crystallisation (SC) scores across Transformer layer indices for Gemma (left) and Qwen (right), under no-prompt (dashed) and optimized (solid) conditions. The horizontal line at $SC = 0$ represents the random baseline. Three regimes are annotated: **Emergence** (early layers), **Consistency** (middle layers), and **Collapse** (final layers). Coloured dots indicate zero-shot inclusion scores $P(A \subseteq B)$ at each layer (right axis), demonstrating the correspondence between SC and zero-shot generalisation performance.

System prompt conditions. We evaluate each model under two conditions: **no-prompt** (minimal prefill token only) and **optimized prompt** (structured system prompt processed as prefill). The optimized prompt used in all experiments is:

```
You are an expert tax formal hierarchy. Constraints:
Focus on 'is-a' and 'has-a'. Suppress colloquial or
metaphorical.
```

This prompt instructs the model to interpret input terms exclusively within a formal taxonomic hierarchy, suppressing colloquial or metaphorical extensions. The prompt was designed to be minimal while maximising Semantic Crystallisation (SC) scores, and is held constant across all models and conditions.

Evaluation metrics. We report (i) **SC scores** per layer, computed against the random baseline (Equation 12), and (ii) **zero-shot inclusion accuracy** on held-out concept pairs (Equation 9), with thresholds $\tau = 0.7$ and $\delta = 0.1$.

4.2 Layer-wise Semantic Crystallisation

Three regimes of algebraic order. Figure 2 shows SC scores across all layers for Gemma and Qwen under no-prompt and optimized conditions, relative to the random baseline ($SC = 0$). We observe three qualitatively distinct regimes:

Emergence (early layers): SC transitions from near-zero or negative values toward positive values as the model begins processing the input. The rate and extent of this emergence differs markedly across models and conditions.

Consistency (middle layers): SC stabilises at positive values, indicating sustained algebraic order. This regime is most pronounced under the optimized condition and in instruction-tuned models.

Collapse (final layers): SC degrades in the majority of conditions, consistent with the Late-layer Collapse phenomenon described in Section 4.4.

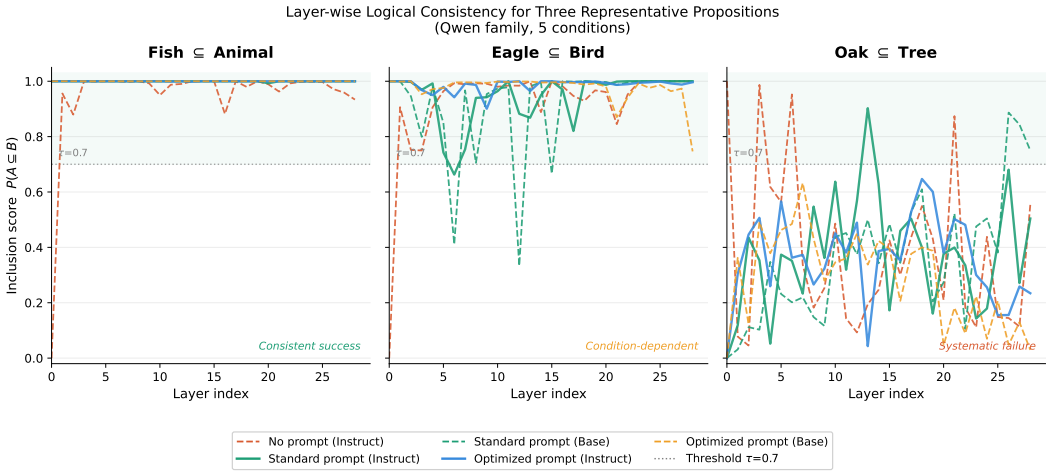

Figure 3: Layer-wise inclusion score $P(A \subseteq B)$ for three representative propositions across five Qwen conditions. **Left** (*Fish* $\subseteq$ *Animal*): consistent success across all conditions and layers, demonstrating stable algebraic structure. **Centre** (*Eagle* $\subseteq$ *Bird*): condition-dependent behaviour—instruction-tuned models with prompts maintain high scores throughout, while base models exhibit Late-layer Collapse beyond layer 20. **Right** (*Oak* $\subseteq$ *Tree*): systematic failure across all conditions, attributable to lexical ambiguity of "Oak" in the model's training distribution. The horizontal dotted line indicates the inclusion threshold $\tau = 0.7$. Solid lines: instruction-tuned models. Dashed lines: base models or no-prompt condition.

Two crystallisation modes. We observe two qualitatively distinct patterns across model families.

In **Autonomous Crystallisation** (Gemma), the no-prompt condition yields $SC \approx 0$ across most layers (average $SC = -0.111$), while the optimized prompt condition elevates specific layers substantially (average $SC = +0.438$, best layer $SC = 0.913$ at layer index 12). This suggests that $\mathbb{F}_2$ structure is latent in Gemma but requires a structured context to become extractable.

In **Induced Crystallisation** (Qwen), the no-prompt condition yields near-zero average SC across layers (average $SC = -0.022$), with specific layers exhibiting negative values. The optimized prompt condition substantially elevates algebraic consistency (average $SC = +0.425$, best layer $SC = 0.650$ at layer index 11), representing a substantial transition between representational regimes.

SC predicts zero-shot performance. A key practical finding is that SC scores predict zero-shot generalisation accuracy across layers: the layer achieving the highest SC consistently yields the highest zero-shot inclusion accuracy (Section 4.3). This provides a gradient-free criterion for layer selection without held-out evaluation data.

4.3 Zero-Shot Ontological Generalisation

Task. We evaluate AOP's ability to recover ontological relations for concept pairs that never appear in the 42-pair training set. The evaluation set (D_{ZST}) comprises 11 pairs spanning 7 positive *is-a* relations and 4 negative relations, covering biological and mineral domains (full list in Appendix B). All concepts in D_{ZST} are drawn from outside the training vocabulary, ensuring $D_{ZST} \cap D_{train} = \emptyset$.

Results. Table 1 reports zero-shot accuracy at the best SC layer for each model-condition combination. Figure 3 shows the layer-by-layer inclusion score for three representative propositions, directly illustrating the title claim: logical consistency can be *controlled*—through the combination of system prompting and instruction tuning—but collapses without these conditions.

Table 1: Zero-shot ontological generalisation accuracy (%) at the best SC layer for each model and condition. "Best layer" denotes the layer index achieving the highest SC score. Overall accuracy, inclusion accuracy, and Hamming accuracy are reported with thresholds $\tau = 0.7$, $\delta = 0.1$.

Model	Condition	Best Layer	Max SC	Overall	Inclusion	Hamming
Gemma-2 (Instruct)	optimized	12	0.913	90.91	90.91	90.91
Gemma-2 (Instruct)	no-prompt	21	0.398	81.82	90.91	81.82
Gemma-2 (Base)	optimized	5	0.342	81.82	81.82	90.91
Gemma-2 (Base)	no-prompt	2	0.099	63.64	81.82	72.73
Qwen2.5 (Instruct)	optimized	11	0.650	63.64	72.73	63.64
Qwen2.5 (Instruct)	no-prompt	8	0.411	54.55	72.73	63.64
Qwen2.5 (Base)	optimized	28	0.267	54.55	72.73	54.55
Qwen2.5 (Base)	no-prompt	7	0.175	54.55	63.64	54.55
mpnet	optimized	12	0.541	81.82	81.82	81.82
mpnet	no-prompt	8	0.597	63.64	63.64	63.64

The optimized prompt condition on Gemma-2 (Instruct) achieves the highest accuracy: **90.91%** overall, inclusion, and Hamming accuracy (layer index 12, $SC = 0.913$). mpnet with optimized prompt achieves 81.82% overall accuracy (layer index 12, $SC = 0.541$). Qwen2.5 (Instruct) with optimized prompt achieves 63.64% (layer index 11, $SC = 0.650$). The no-prompt condition yields substantially lower performance across all models, consistent with the lower SC values observed in Section 4.2.

Systematic failure: plant domain. Across all model-condition combinations, the pair *Oak* $\subseteq$ *Tree* consistently fails (inclusion score < 0.5). We attribute this to the lexical ambiguity of "Oak" in the model's training distribution, where it appears frequently as a proper noun (place name, surname) unrelated to botanical taxonomy. This illustrates a principled limitation of LMP under concept ambiguity: when mean pooling mixes multiple semantic senses, the resulting representation satisfies no single ontological constraint cleanly.

Relationship between SC and accuracy. The correspondence between peak SC layer and peak accuracy layer holds consistently across models, validating SC as a practical layer selection criterion. In its current formulation, however, SC is a **necessary but not sufficient** condition for high zero-shot accuracy: a strongly negative SC reliably predicts degraded accuracy (insulation failure), but a high SC does not guarantee high accuracy, as SC currently captures only the insulation dimension of algebraic consistency (Section 5.3).

4.4 Late-layer Collapse

We observe a systematic degradation of zero-shot accuracy in the final layers of the model, which we term **Late-layer Collapse**. In 7 of 10 evaluated conditions, the accuracy at the final layer falls more than 10 percentage points below peak accuracy.

Stability conditions. Among evaluated conditions, **optimized/Gemma-2 (Instruct)** achieves the highest peak accuracy (90.91%) while maintaining the most stable late-layer performance. Base model variants and no-prompt conditions uniformly exhibit greater collapse regardless of peak accuracy, consistent with the complementary roles of instruction tuning and structured prompting.

Complementary roles. Under the conditions evaluated here, instruction tuning and system prompting appear to play complementary roles: the instruction-tuned model provides a more stable representational substrate, while the system prompt configures the algebraic boundary conditions that determine whether $\mathbb{F}_2$ structure is accessible. Whether a carefully designed prompt can substitute for instruction tuning in base models—achieving comparable stability through prompt design alone—remains an open question for future work.

Table 2: Late-layer Collapse summary. Peak accuracy, final-layer accuracy, and end-layer stability (average inclusion score over final 5 layers) are reported for each model-condition combination. † indicates conditions that maintain both high peak accuracy and late-layer stability.

Model	Condition	Peak (%)	Peak Layer	Final (%)	Stable
Gemma-2 (Instruct)	optimized	90.91	12	81.82	†
Gemma-2 (Instruct)	no-prompt	90.91	21	63.64	
Gemma-2 (Base)	optimized	81.82	5	63.64	
Gemma-2 (Base)	no-prompt	81.82	2	36.36	
Qwen2.5 (Instruct)	optimized	72.73	11	72.73	†
Qwen2.5 (Instruct)	no-prompt	72.73	8	45.45	
Qwen2.5 (Base)	optimized	72.73	28	54.55	
Qwen2.5 (Base)	no-prompt	63.64	7	36.36	
mpnet	optimized	81.82	12	81.82	†
mpnet	no-prompt	63.64	8	63.64	

Table 3: Comparison of SC profiles for base and instruction-tuned model variants. Average SC (Avg SC) and maximum SC (Max SC) are reported for no-prompt and optimized conditions. Spark Prompt results are included for base variants.

Model	Condition	Best Layer	Max SC	Avg SC
Gemma-2 (Instruct)	no-prompt	21	0.398	-0.111
Gemma-2 (Instruct)	optimized	12	0.913	$+0.438$
Gemma-2 (Base)	no-prompt	2	0.099	-0.260
Gemma-2 (Base)	optimized	5	0.342	$+0.074$
Qwen2.5 (Instruct)	no-prompt	8	0.411	-0.022
Qwen2.5 (Instruct)	optimized	11	0.650	$+0.425$
Qwen2.5 (Base)	no-prompt	7	0.175	-0.032
Qwen2.5 (Base)	optimized	28	0.267	$+0.191$
mpnet	no-prompt	8	0.597	-0.830
mpnet	optimized	12	0.541	$+0.420$

Logic Cliff. In some conditions, we observe an abrupt accuracy drop exceeding 25 percentage points within a single layer transition, which we term a **Logic Cliff**. Logic Cliffs are observed at specific intermediate layers across multiple conditions, most prominently in optimized/Gemma-2 (layer index 6) and optimized/Qwen2.5 (layer index 5), suggesting critical transitions in representational mode that can disrupt previously established algebraic structure.

4.5 Base versus Instruction-tuned Models

Effect of instruction tuning on SC. Table 3 compares SC profiles for base and instruction-tuned variants. For Gemma-2, instruction tuning increases average SC from -0.260 (Base) to -0.111 (Instruct) under no-prompt conditions, suggesting that instruction tuning promotes autonomous crystallisation. For Qwen2.5, the difference is smaller (-0.032 vs -0.022), indicating that Qwen2.5's algebraic structure under no-prompt conditions is less affected by instruction tuning. With the optimized prompt, Gemma-2 (Instruct) achieves the highest average SC ($+0.438$), while Qwen2.5 (Base) achieves $+0.191$—lower than Qwen2.5 (Instruct) at $+0.425$—suggesting that instruction tuning is more effective than base model capacity for algebraic boundary condition response in Qwen2.5.

Spark Prompt for base models. Using token-level SC analysis at the final layer, we identify tokens in the system prompt that contribute disproportionately to SC elevation. For base model variants, a minimal **Spark Prompt** of six tokens ("You are an expert tax") achieves average $SC = +1.40$, substantially outperforming both the no-prompt baseline ($+0.36$) and comparable

to the full system prompt ($+1.01$). This finding suggests that $\mathbb{F}_2$ crystallisation is triggered by specific high-SC tokens rather than semantic completeness of the prompt, reducing prompt design to a structured token-level optimisation problem. Detailed token-level SC analysis is reported in a companion paper.

Instruction tuning and late-layer stability. Despite similar SC profiles under optimized conditions, instruction-tuned models exhibit substantially better late-layer stability than base variants (Section 4.4). This dissociation between SC magnitude and late-layer stability suggests that instruction tuning contributes to algebraic consistency through a mechanism distinct from crystallisation, possibly by stabilising the representational dynamics of the final layers independently of the algebraic structure present in middle layers.

5 Results and Discussion

5.1 $\mathbb{F}_2$ Algebraic Structure is Intrinsic to LLM Representations

Our central finding is that LLMs encode ontological relations in a form that is linearly accessible via $\mathbb{F}_2$ projection. AOP achieves up to 90.91% inclusion accuracy on held-out concept pairs trained on only 42 relational constraints—a result that substantially exceeds the zero-shot baselines reported by Burns et al. (Burns et al., 2023), whose contrastive search over hidden states achieves below 80% on comparable tasks, despite operating without the algebraic structure constraint that AOP imposes.

This result was theoretically anticipated. A linear map $\phi : \mathbb{R}^d \to \mathbb{F}_2^n$ is a homomorphism: it preserves the algebraic structure of the source space. If LLM hidden states carry latent semantic operations that correspond to ontological relations, a linear projection will transfer those operations faithfully into $\mathbb{F}_2^n$. Conversely, the success of AOP with a linear projection implies that the relevant logical structure in the hidden state space is *linearly accessible*—it does not reside on a nonlinear manifold requiring a more complex transformation. The success of AOP is therefore not merely an empirical finding: it is evidence that LLMs organise logical knowledge in a linearly structured region of their latent space.

This is consistent with the broader observation that LLMs acquire structured knowledge through statistical learning on natural language—a corpus in which ontological relations are pervasively and redundantly expressed. The algebraic structure we observe is not imposed by AOP; it is revealed by it.

5.2 System Prompts as Algebraic Boundary Conditions

A second central finding concerns the role of system prompts. We observe that the introduction of a structured system prompt consistently elevates SC scores and zero-shot accuracy, but that the *mechanism* of this elevation differs qualitatively across model families.

In **Autonomous Crystallisation** (Gemma), the no-prompt condition yields $SC \approx 0$—indistinguishable from the random baseline—suggesting that $\mathbb{F}_2$ structure is latent but not spontaneously accessible. The system prompt acts as a *catalyst*: it does not create algebraic structure but renders it extractable.

In **Induced Crystallisation** (Qwen, mpnet), the no-prompt condition yields negative average SC, with specific layers exhibiting $SC \ll 0$—indicating that algebraic insulation is more disrupted than in a randomly initialised model. This effect is most pronounced in mpnet: the no-prompt condition yields Avg $SC = -0.830$, while the optimized prompt elevates this to $+0.420$—a difference of more than 1.2 SC units, representing the most dramatic Induced Crystallisation observed across all evaluated conditions. Qwen exhibits a more moderate effect (-0.022 vs $+0.425$), but the same qualitative pattern holds. Layers with $SC < 0$ are associated with degraded zero-shot accuracy, consistent with the interpretation that negative SC reflects active violation of $\mathbb{F}_2$ insulation constraints in those layers. In both cases, the system prompt substantially elevates SC across all layers, exhibiting greater sensitivity to prompt configuration than Gemma.

We note, however, that *in the current formulation of SC*, positive SC is a necessary but not sufficient condition for full algebraic consistency: even when insulation is preserved ($SC > 0$), violations of

hierarchical (*is-a*) and compositional (*has-a*) constraints may persist, as SC currently captures only the insulation dimension of algebraic structure (Section 5.3). This limitation motivates the unified SC metric described in Section 5.3.

In both cases, the system prompt functions not as a behavioural instruction but as an **algebraic boundary condition**: a prefill that configures the hidden state space prior to concept extraction, determining whether $\mathbb{F}_2$ constraints can be satisfied in the projected representation. This reframes a fundamental question in prompt engineering: rather than asking what instructions produce correct outputs, one may ask what configurations produce algebraically consistent internal representations.

Furthermore, under the conditions evaluated here, the combination of an instruction-tuned model and a structured system prompt is the only configuration under which both high peak accuracy and late-layer stability are jointly achieved. This suggests that the choice of instruction-tuned vs base model and the inference-time system prompt play *complementary* roles: the former provides a more stable representational substrate, while the latter configures the algebraic boundary conditions that determine whether $\mathbb{F}_2$ structure is accessible. Whether a carefully designed prompt can substitute for instruction tuning in base models remains an open question for future work.

5.3 Semantic Crystallisation as an Indicator: Validity and Current Limitations

Validity. The SC metric demonstrates consistent predictive validity as a layer selection criterion: across all evaluated models and conditions, the layer achieving the highest SC score coincides with the layer achieving the highest zero-shot inclusion accuracy. This correspondence validates SC as a practical, gradient-free criterion for identifying layers that carry accessible $\mathbb{F}_2$ algebraic structure, without requiring held-out evaluation data.

Current limitation: insulation only. In its current formulation, SC measures only *algebraic insulation*—the degree to which concepts from incompatible categories are separated in $\mathbb{F}_2$ space, as captured by the negation (neg) constraint. It does not directly measure the integrity of *is-a* and *has-a* hierarchical constraints, which constitute a distinct and arguably more central dimension of ontological consistency.

This limitation explains the observed divergence between SC scores and zero-shot accuracy in some conditions: a layer may achieve high SC (strong insulation) while permitting violations of inclusion constraints (weak subsumption), resulting in high SC but suboptimal ZST accuracy. This is observed in the comparison between Qwen with optimized prompt (Avg $SC = +0.425$) and Gemma without prompt (Avg $SC = -0.111$), where the latter achieves higher ZST accuracy despite lower SC, highlighting the current limitation of SC as an insulation-only metric. A unified SC metric incorporating all three constraint types—insulation, hierarchical inclusion, and compositional containment—would provide a more complete measure of algebraic consistency. We leave this extension to future work.

5.4 Broader Implications and Future Directions

Logical consistency and hallucination. The layer-dependent pattern of Semantic Crystallisation and Late-layer Collapse offers a new perspective on LLM failure modes. If the final layers of a model systematically degrade algebraic consistency—as observed in 7 of 10 evaluated conditions—this provides a potential substrate for confident but logically inconsistent outputs, a common characteristic of hallucinations. AOP provides a tool to measure this degradation quantitatively, and the SC metric opens a path toward using algebraic consistency as a training signal: by incorporating $\mathbb{F}_2$ constraint loss into the training objective, future work may directly optimise for logical consistency throughout the forward pass. We note that the bidirectionality of the linear projection—the existence of an approximate inverse transformation—further suggests that formally specified ontological constraints could be injected directly into LLM hidden states, rather than communicated through natural language prompts.

Integration with formal methods. The $\mathbb{F}_2$ algebraic structure recovered by AOP corresponds directly to the relational primitives of formal specification languages such as SysML v2 (Object Management Group, 2024b)—and especially its formal semantic foundation, KerML (Object

Management Group, 2024a), which defines the algebraic and type-theoretic semantics underlying SysML v2's relational primitives—where *is-a* and *has-a* are first-class constructs with formally verifiable semantics. This structural correspondence suggests a path toward continuous migration from natural language prompts to formal specifications, with algebraic consistency enforced at the representation level. For safety-critical applications—where formal verification of system behaviour is required—AOP provides a mathematically grounded bridge between the statistical representations of LLMs and the algebraic structures of formal ontologies. We present the current work as a first step toward this longer-term vision.

Token-level analysis and prompt optimisation. A companion paper reports token-level SC analysis, in which the contribution of individual system prompt tokens to algebraic crystallisation is quantified. This analysis demonstrates that $\mathbb{F}_2$ crystallisation is triggered by specific high-SC tokens rather than semantic completeness of the prompt, reducing prompt design from trial-and-error to a structured optimisation problem.

5.5 LIMITATIONS

Systematic failure under insufficient ontological grounding. The pair *Oak* $\subseteq$ *Tree* fails consistently across all models and conditions (inclusion score < 0.5). We attribute this primarily not to a limitation of Localized Mean Pooling, but to insufficient ontological grounding of "Oak" in the model's learned representations: the concept "Oak" is heavily associated with non-botanical usages (proper nouns, place names, surnames) in the training distribution, and the extitis-a relation to "Tree" may not be sufficiently reinforced to survive $\mathbb{F}_2$ projection. Providing explicit botanical context partially improves inclusion scores but does not resolve the failure, suggesting that the issue lies deeper than prompt design.

Two complementary remedies are conceivable. First, richer context engineering—providing the model with an explicit ontological frame prior to concept extraction—may partially compensate for weak grounding. Second, and more fundamentally, directly incorporating formal ontological structure into LLM training or fine-tuning (e.g., via $\mathbb{F}_2$ constraint loss as a training signal) may be necessary to ensure that concepts are adequately grounded in their hierarchical relations. This failure case therefore serves not merely as a limitation but as a motivating example for the longer-term integration of formal ontologies into LLM learning.

Global Mean Pooling and context length. The Localized Mean Pooling strategy is effective only for short, concept-focused contexts where the target concept dominates the token sequence. For longer contexts, mean pooling dilutes the concept signal, degrading projection quality. This constrains the current framework to minimal context strings and limits its applicability to compositional or discourse-level reasoning. Extension to token-level $\mathbb{F}_2$ projection, which would preserve sequential and causal structure, is a natural next step.

Evaluation scope. The current evaluation is centred on the Qwen model family, with complementary results on Gemma and mpnet. The generality of the Autonomous and Induced Crystallisation modes across a wider range of model families and scales remains to be established. Similarly, the 42-pair training set and 11-pair evaluation set, while sufficient to demonstrate the principle, represent a narrow slice of the ontological space. Evaluation on larger and more diverse relational datasets—including cross-domain and adversarial pairs—is required to characterise the full scope and limits of AOP generalisation.

SC metric completeness. As noted in Section 5.3, the current SC metric captures only algebraic insulation. The observed divergence between SC and zero-shot accuracy in certain conditions (e.g., Qwen with optimized prompt (Avg $SC = +0.425$) achieving lower ZST accuracy than Gemma with optimized prompt (Avg $SC = +0.438$), despite similar SC values— ZST accuracy) underscores the importance of extending SC to encompass *is-a* and *has-a* constraint integrity. Until such an extension is available, SC should be interpreted as a necessary but not sufficient indicator of full algebraic consistency.

6 CONCLUSION

We introduced **Algebraic Ontology Projection (AOP)**, a framework for extracting formal ontological structure from LLM hidden states by projecting them into the Galois Field $\mathbb{F}_2$ under constraints derived from the Liskov Substitution Principle. Using a minimal set of 42 relational constraints as algebraic keys, AOP achieves up to 90.91% zero-shot inclusion accuracy on unseen concept pairs across Gemma-2 and Qwen2.5—with no model tuning, but through prompt alone.

We demonstrated that this algebraic structure is *layer-dependent*: it emerges progressively across depth, stabilises in intermediate layers, and degrades in the final layers of most models—a phenomenon we term **Late-layer Collapse**. To quantify this behaviour, we introduced **Semantic Crystallisation (SC)**, a baseline-calibrated metric that measures algebraic order relative to a randomly initialised model and serves as a practical, gradient-free criterion for layer selection.

We further showed that system prompts function as **algebraic boundary conditions**: they determine whether—and in which layers—$\mathbb{F}_2$ structure is accessible to linear projection. Only the combination of instruction tuning and a structured system prompt sustains both high accuracy and late-layer stability, suggesting complementary roles for these two factors in maintaining logical consistency throughout the forward pass.

Taken together, these results support a re-reading of what LLMs internally compute: forward computation can be interpreted as an iterative process of algebraic organisation across layers—structure emerges, dissolves, and re-emerges before settling into its final configuration—a property that is invisible to geometric analyses but becomes legible under algebraic projection.

The present work is a first step. A companion paper reports token-level SC analysis and algebraic prompt optimisation. Further work will address the extension of SC to encompass *is-a* and *has-a* constraint integrity, the development of token-level $\mathbb{F}_2$ projection to preserve causal structure, and the integration of $\mathbb{F}_2$ constraint loss as a differentiable training signal—opening a path toward LLMs that not only approximate logical structure statistically, but enforce it algebraically.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

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*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Artur d'Avila Garcez, Luis C Lamb, and Dov M Gabbay. *Neural-Symbolic Cognitive Reasoning*. Springer, 2009.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/toy_model/index.html`.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9574–9586, 2021.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4129–4138, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3519–3529. PMLR, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17359–17372, 2022.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. 2023.

Object Management Group. KerML: Kernel modeling language specification, 2024a. URL `https://www.omg.org/spec/KerML/1.0`.

Object Management Group. SysML v2 language specification, 2024b. URL `https://www.omg.org/spec/SysML/2.0`.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2463–2473, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Tom Henighan, and Christopher Olah. Scaling monosemanticity: Extracting interpretable features from Claude Sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

A Complete Relational Dataset

Tables 4 and 5 provide the complete training and evaluation datasets used in all experiments. The training set (D_{train}) consists of 42 symbolic constraints organised into four progressive stages (levels 1, 2, 4, 8), used simultaneously as algebraic keys for AOP projection learning. The evaluation set (D_{val}) consists of 13 independent negative pairs (i_neg) with $D_{val} \cap D_{train} = \emptyset$, used exclusively for zero-shot evaluation.

B Zero-Shot Evaluation Set

Table 6 lists the complete set of 11 concept pairs used for zero-shot evaluation across all models and conditions. All pairs are drawn from semantic domains and concept vocabularies *outside* the AOP training set (D_{train}), ensuring $D_{ZST} \cap D_{train} = \emptyset$. Positive pairs (Expected: True) test hierarchical inclusion; negative pairs (Expected: False) test categorical separation.

Table 4: Training dataset (D_{train}): 42 symbolic constraints across four levels (15 *is-a*, 12 *has-a*, 15 negation). Concepts in *is-a* and *has-a* pairs constitute the algebraic key vocabulary. Negation pairs (neg) enforce insulation constraints.

Level	Type	Concept A	Concept B
1	is-a	Beetle	Insect
1	is-a	Fly	Insect
1	is-a	Insect	Animal
1	has-a	Animal	Cell
1	has-a	Insect	Legs
1	has-a	Insect	Exoskeleton
1	neg	Beetle	Ocean
1	neg	Fly	Cloud
1	neg	Insect	Stone
1	neg	Animal	Logic
2	is-a	Bee	Insect
2	is-a	Butterfly	Insect
2	has-a	Bee	Wings
2	neg	Bee	Vacuum
2	neg	Butterfly	Logic
4	is-a	StagBeetle	Beetle
4	is-a	Ant	Insect
4	is-a	Spider	Animal
4	is-a	Whale	Animal
4	has-a	Animal	DNA
4	has-a	StagBeetle	Mandibles
4	has-a	Spider	Silk
4	has-a	Whale	Blubber
4	neg	Spider	Stone
4	neg	Whale	Vacuum
4	neg	Ant	Cloud
4	neg	StagBeetle	Logic
8	is-a	Granite	Rock
8	is-a	Quartz	Mineral
8	is-a	Diamond	Mineral
8	is-a	Rock	Mineral
8	is-a	Mineral	Matter
8	is-a	Animal	Matter
8	has-a	Mineral	CrystalStructure
8	has-a	Granite	Quartz_Grain
8	has-a	Diamond	Hardness_10
8	has-a	Matter	Mass
8	neg	Granite	Cell
8	neg	Diamond	Legs
8	neg	Quartz	Insect
8	neg	Matter	Logic
8	neg	Rock	Ocean

Note: *Oak* $\subseteq$ *Tree* consistently fails across all conditions and models due to lexical ambiguity (Section 4.3). All remaining 10 pairs are evaluated for zero-shot accuracy.

Table 5: Evaluation dataset (D_{val}): 13 independent negative pairs (i_neg). All concepts in D_{val} are drawn from outside the semantic domains of D_{train}, ensuring $D_{val} \cap D_{train} = \emptyset$. These pairs are used exclusively for zero-shot evaluation and never appear during AOP training.

Level	Type	Concept A	Concept B
1	i_neg	Ocean	Logic
1	i_neg	Cloud	Logic
1	i_neg	Sun	Logic
1	i_neg	Idea	Legs
2	i_neg	Wings	Cloud
2	i_neg	Bee	Idea
4	i_neg	Spider	Vacuum
4	i_neg	Silk	Idea
8	i_neg	Quartz	DNA
8	i_neg	Diamond	Idea
8	i_neg	DNA	Cloud
8	i_neg	Rain	Logic
8	i_neg	Snow	DNA

Table 6: Zero-shot evaluation set (D_{ZST}): 11 concept pairs spanning biological and mineral domains, unseen during AOP training. Positive pairs test *is-a* inclusion; negative pairs test categorical separation.

Type	Concept A	Relation	Concept B
Positive	Robin	$\subseteq$	Bird
Positive	Eagle	$\subseteq$	Bird
Positive	Salmon	$\subseteq$	Fish
Positive	Fish	$\subseteq$	Animal
Positive	Oak	$\subseteq$	Tree
Positive	Copper	$\subseteq$	Metal
Positive	Marble	$\subseteq$	Rock
Negative	Robin	$\not\subseteq$	Mineral
Negative	Eagle	$\not\subseteq$	Rock
Negative	Oak	$\not\subseteq$	Animal
Negative	Copper	$\not\subseteq$	Insect