ONTOLOGY CONCEPT BOTTLENECK MODELS: TOWARDS HUMAN-ALIGNED REASONING WITH GRADED LOGIC

Anonymous authors
Paper under double-blind review

ABSTRACT

As autonomous systems increasingly make consequential decisions on behalf of humans, interpretability alone is insufficient. Their reasoning should not only be explainable but also align with the way humans evaluate evidence and make decisions. This motivates reasoning over intuitive concepts through concise and human-readable compositional processes. We introduce Ontology Concept Bottleneck Models (OCBMs), which jointly learn task-functional concepts and class-specific graded-logic reasoning trees, leveraging the reasoning structure itself as an inductive bias on concept formation. Across standard visual and tabular benchmarks, OCBMs achieve competitive or leading predictive performance while producing compact, auditable reasoning structures with controllable granularity. Moreover, controlled head ablations show that these concepts transfer substantially better across domains than those learned with linear or MLP heads. These results suggest that graded-logic ontologies can serve simultaneously as an interpretable reasoning layer and a structural constraint on concept learning, providing a practical foundation for machine reasoning that can be inspected, intervened upon, and progressively aligned with human reasoning.

1 INTRODUCTION

As AI models are increasingly deployed in high-stakes settings, there is growing demand for explainability. However, providing explanations alone may not be sufficient. When robots and autonomous agents collaborate with humans or make decisions on their behalf, the reasoning process should also align with human intuition; that is, the way evidence is combined and conclusions are reached should be consistent with how humans naturally reason about the task. Such alignment is essential for establishing trust, ensuring safety, and enabling predictable behavior in human-AI collaboration.

Concept Bottleneck Models (CBMs) represent an important step toward this goal by mapping inputs to human-defined concepts and predicting task labels from those concepts. More recent CBM variants further expose the decision process through structured representations, such as Boolean predicates, logical expressions, and concept-relation graphs. These approaches improve transparency by making the relationships between concepts and predictions more explicit. However, an explicit decision structure does not necessarily imply an intuitive one: rigid logical operators or learned relationships may remain poorly aligned with how humans naturally combine partial, uncertain, and competing evidence when making decisions.

In this work, we investigate graded-logic ontologies as a reasoning layer for concept-based models. Graded logic has a long history of being used by human experts to model and explain complex decision processes, particularly when decisions depend on the degree to which multiple criteria are satisfied and how those criteria are combined (Dujmović, 1974; Dujmovic, 2007; Dujmović, 2025). Rather than relying only on binary true-or-false predicates, graded logic represents partial satisfaction and provides structured operators for combining multiple pieces of evidence. This makes it well suited for modeling the gradual, compensatory, and non-compensatory reasoning patterns that often characterize intuitive human decision making.

Building on this idea, we introduce Ontology Concept Bottleneck Models (OCBMs), which learn per-class graded-logic trees over a shared concept layer. We use *task-functional concepts* to denote

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

latent concepts that emerge from their functional role within the learned task reasoning process, rather than from direct concept supervision or an externally supplied semantic vocabulary. We further evaluate whether the learned logic is not only interpretable but also aligned with intuitive reasoning. Our main contributions are as follows:

- We introduce OCBM, replacing conventional prediction heads with class-specific graded-logic trees over a shared concept layer, yielding explicit and auditable reasoning structures.
- We show that graded-logic reasoning acts as an inductive bias on concept learning: without concept supervision, OCBM discovers task-functional concepts and compositional rules directly from task labels.
- Controlled head ablations show that these concepts transfer substantially better across domains than those learned with linear or MLP heads, while OCBM remains competitive on standard CBM benchmarks.

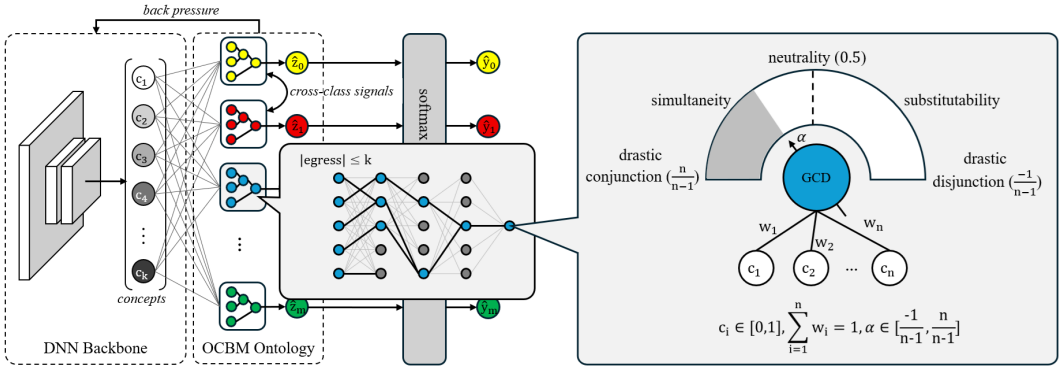

Figure 1: OCBM Architecture

2 Background and Related Work

2.1 Modeling Human Reasoning with Graded Logic

Human decisions are rarely based on a single variable. Instead, a decision maker evaluates multiple considerations, assigns them different roles and levels of importance, and combines them into an overall assessment. For example, when evaluating a job opportunity, a person may jointly consider compensation, career growth, work content, location, and job security. Some conditions may be mandatory, such as a minimum salary; others may be sufficient to make the opportunity attractive; and still others may provide incremental rewards or penalties over other preconditions. Such rich inter-feature dynamics are difficult to capture with additive models, motivating representations that can explicitly model logical relationships and structured interactions among concepts.

Classical Boolean logic provides explicit compositional structure through operators such as AND, OR, and NOT, but it is often too rigid for real-world decisions, while additive models fail to capture rich interactions among decision factors. Graded Logic generalizes these approaches by representing logical relationships as continuous, learnable aggregations over graded inputs. It can express a spectrum of reasoning behaviors ranging from conjunctive requirements to disjunctive alternatives, while also modeling varying importance, compensation, mandatory and sufficient conditions, and other human-like decision patterns. By composing such aggregations hierarchically, graded logic provides an interpretable and interventionable representation of how evidence is combined into a conclusion, making it a natural foundation for human-aligned reasoning.

2.2 Reasoning with Concepts

Concept Bottleneck Models (CBMs) aim to improve interpretability by predicting human-understandable concepts before making task decisions. Existing work primarily addresses two challenges: concept acquisition and concept reasoning.

For concept acquisition, early CBMs rely on explicit concept annotations (Koh et al., 2020). To reduce annotation costs, several approaches obtain concepts from external semantic resources. Examples include LLM-generated concept vocabularies (Yan et al., 2023), label-free concept discovery using vision-language models (Oikarinen et al., 2023), retrieval from external concept banks (Yamaguchi et al., 2026), and grounding through open-domain detectors (Srivastava et al., 2024). Another line of work discovers concepts directly from learned representations, including SENN (Alvarez-Melis & Jaakkola, 2018), UCBM (Schrodi et al., 2024), and G-CBM (Hossain et al., 2026). While these methods reduce dependence on manual annotations, concepts are typically derived either from external semantic resources or from latent feature representations prior to reasoning. In contrast, OCBM couples concept formation and reasoning through a shared graded-logical ontology, allowing concepts to emerge under the same objective that governs downstream decision making.

A complementary line of work focuses on concept reasoning. Logic-CBMs (Vemuri et al., 2025) and related approaches seek to replace black-box concept predictors with more structured reasoning mechanisms. Examples include causal concept reasoning (Felice et al., 2025), rule-based reasoning over concepts (Debot et al., 2024), knowledge-constrained reasoning (Kalampalikis et al., 2025), and graph-based reasoning (Xu et al., 2026). These methods improve transparency of the concept-to-task mapping, but generally assume that concept semantics are already available through annotations, external vocabularies, or pretrained representations.

OCBM unifies concept acquisition and reasoning in a single end-to-end framework. Its graded-logic ontology serves simultaneously as the predictor and as an inductive bias that shapes concept formation. Consequently, concepts, logical composition rules, routing, and negation are learned directly from task supervision without concept annotations, external vocabularies, pretrained semantic encoders, or predefined reasoning rules. To the best of our knowledge, OCBM is the first CBM framework that jointly discovers task-functional concepts and human-interpretable logical reasoning structures from task labels alone.

3 Ontology Concept Bottleneck Models

The inductive bias of OCBM is an ontology composed of a bank of class-specific graded-logic aggregation trees built upon a shared set of concepts. Each tree independently learns aggregation operators (e.g., degrees of conjunction), input routing, and feature transformations such as negation, while sharing cross-class signals with each other.

OCBM's ontology serves as a regulator for discovering coherent logical relationships among concepts. Because every class prediction must be expressed through a shared concept set, concepts are required to participate in multiple class-specific explanations while remaining logically consistent across them. This coupling substantially constrains the space of admissible representations, favoring concept structures that support coherent explanations rather than isolated class-specific predictive shortcuts. Such a preference for explanatory coherence is consistent with coherence-based accounts of reasoning, in which hypotheses are evaluated by how well they fit within mutually supporting and non-contradictory explanatory structures (Thagard, 1989).

Our experiments suggest that the ontology can also act as a concept regularizer for emergent concepts. In this setting, OCBM can jointly learn concepts and reasoning rules without prior knowledge of either the concepts themselves or the underlying decision logic, as demonstrated by the MNIST (LeCun et al., 1998) experiment below. This capability aligns with our broader objective of enabling autonomous decision-making based on raw sensory data while maintaining an interpretable and intuitive reasoning process.

3.1 OCBM Architecture

OCBM adopts the standard concept bottleneck formulation, in which an input $\mathbf{x} \in \mathcal{X}$ is first mapped to an interpretable concept vector $\mathbf{c} = g(\mathbf{x}) \in [0,1]^m$, and subsequently transformed into a prediction $\hat{\mathbf{y}}$ through a decision function, $\hat{\mathbf{y}} = f(\mathbf{c}) = f(g(\mathbf{x}))$. Unlike conventional CBMs, OCBM replaces the decision function $f(\cdot)$ with an ontology $\mathcal{T} = \{T_1, \ldots, T_K\}$ of graded-logic aggregation trees, where each tree T_k corresponds to one target class y_k. Each tree is initialized as a densely connected hierarchical network and is progressively structured during training through three regularization mechanisms. First, a *compression* regularizer encourages multiple nodes to merge onto a

shared parent, thereby reducing structural complexity. Second, an *egress* constraint limits the number of outgoing edges from each node to at most d_{out}, typically $d_{\text{out}} \in \{1, 2\}$, promoting sparse and human-interpretable reasoning paths. Third, a *convergence* regularizer enforces that all computational paths ultimately converge to a single root node, yielding a unique class activation score. These mechanisms jointly ensure that each tree T_k forms a simple aggregation tree.

The ontology also enables cross-class learning. For a sample belonging to class y_k, the corresponding tree T_k receives a positive training signal, while the remaining trees receive scaled inhibitory signals. Let a_k denote the activation of tree T_k. The inhibitory signal transmitted from T_k to a competing tree $T_j, j \neq k$ is weighted by a monotonic function $\phi(a_k)$, such that more confident trees exert stronger negative evidence on alternative classes.

The root activations of all trees are then normalized through a softmax layer to produce the final class probabilities.

Every internal node of a tree is implemented as a *Graded Conjunction/Disjunction* (GCD) (Dujmović, 2025) aggregator, which represents a generalized conjunction–disjunction function parameterized by an importance weight w and an andness parameter α that determines whether the aggregation behaves more like an averaging operator, a soft conjunction, or a soft disjunction. For simplicity, we present a GCD with two input variables x, y. Let the weighted arithmetic mean be $A(x, y, w) = wx + (1 - w)y$ and the weighted geometric mean be $G(x, y, w) = x^{2w}y^{2(1-w)}$. The geometric mean emphasizes joint necessity, whereas the arithmetic mean reflects compensatory, disjunctive behavior. To enable a continuous transition between these regimes, we define a scaling function $\beta(\alpha) = \sqrt{\frac{3}{2-\alpha}} - 1$, which modulates the sharpness of the geometric component. The resulting binary GCD operator is defined piecewise as:

$$GCD_2 = \begin{cases} \mathbf{1}[x = y = 1], & \alpha = 2, \\ G^{\beta(\alpha)}, & 0.75 \leq \alpha < 2, \\ (3 - 4\alpha)\, A + (4\alpha - 2)\, G^{\beta(\alpha)}, & 0.5 < \alpha < 0.75, \\ A, & \alpha = 0.5, \end{cases}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. For $\alpha < 0.5$, disjunctive behavior is obtained via duality:$1 - GCD_2(1 - x, 1 - y, w, 1 - \alpha)$. The andness parameter offers rich modeling capabilities that we leverage in creating the overall aggregation tree. Table 1 summarizes some of the semantic meanings of andness at different levels (again, presented in the context of two input variables):

Table 1: Andness-directed GCD with two variables

GCD	Conjunctive	$\hat{\wedge}$	CC	Drastic conjunction	2	"Must all be completely satisfied"
		Hyper conjunction ($\overline{\wedge}$)	HHC	High hyper-conjunction	$]5/4, 2]$	"Below the lowest"
			CP	Product t-norm	$5/4$	
			LHC	Low hyper-conjunction	$[1, 5/4[$	
		$\wedge$	C	Pure conjunction	1	"Decided by lowest"
		Hard conjunction ($\overline{\Delta}$)	HC+	High hard conjunction	$13/14$	"Must have all"
			HC	Medium hard conjunction	$12/14$	
			HC-	Low hard conjunction	$11/14$	
		Soft conjunction ($\underline{\Delta}$)	SC+	High soft conjunction	$10/14$	"Nice to have most"
			SC	Medium soft conjunction	$9/14$	
			SC-	Low soft conjunction	$8/14$	
	Neutral	Arithmetic mean ($\hat{\diamond}$)	A	Logic neutrality	$7/14$	"Nice to have"
	Disjunctive	Soft disjunction ($\underline{\nabla}$)	SD-	Low soft disjunction	$6/14$	"Nice to have some"
			SD	Medium soft disjunction	$5/14$	
			SD+	High soft disjunction	$4/14$	
		Hard disjunction ($\overline{\nabla}$)	HD-	Low hard disjunction	$3/14$	"Enough to have any"
			HD	Medium hard disjunction	$2/14$	
			HD+	High hard disjunction	$1/14$	
		$\vee$	D	Pure disjunction	0	"Decided by highest"
		Hyper disjunction ($\overline{\vee}$)	LHD	Low hyper-disjunction	$]-1/4, 0]$	"Above the highest"
			DP	Product t-conorm	$-1/4$	
			HHD	High hyper-disjunction	$[-1, -1/4[$	
		$\hat{\vee}$	DD	Drastic disjunction	-1	"True unless all are zeros"

3.2 Training Protocol

OCBM is compatible with the standard CBM training paradigms including independent, sequential, and joint optimization. In practice we warm-start the visual backbone and concept encoder from a jointly pretrained concept–label model and then train the graded-logic head, which combines the strengths of the sequential and

joint regimes: the bottleneck begins on well-formed concepts while the ontology and reasoning structure are still learned end-to-end.

Let $\mathcal{L}_{\text{task}}$ denote the task-prediction loss: $\mathcal{L}_{\text{task}} = CE + 0.3\,OvR$, where OvR is a down-weighted confusion-weighted one-vs-rest term that concentrates negative pressure on the classes that actually compete and sharpens per-class tree disentanglement. The overall objective is

$$\mathcal{L} \;=\; \mathcal{L}_{\text{task}} \;+\; \lambda_c\,\mathcal{L}_{\text{concept}} \;+\; \lambda_o \mathcal{L}_{\text{ontology}}, \tag{1}$$

where $\mathcal{L}_{\text{concept}}$ is an (imbalance-weighted) binary cross-entropy on the bottleneck, applied only when concept annotations are available, and $\mathcal{L}_{\text{ontology}}$ is a composite structural regularizer that comprises structural regularizers introduced in the previous section, plus *edge-sparsity* terms—an L_0 hard-concrete gate that prunes edges to exactly zero and an entropy (polarization) term that drives the remaining gates toward $\{0, 1\}$. Each term carries its own weight; we write λ_o collectively for this group.

Training proceeds in two phases. During the *soft learning phase*, routing decisions remain differentiable, so concepts, aggregation operators (per-node andness), feature transformations (identity/negation and relevance coefficients), and routing assignments (edges and the root pointer) are optimized jointly by gradient descent. Routing sharpness is annealed over the run toward a fixed cap rather than fully hardened, and the sparsity regularizers are introduced on a schedule—first the L_0 edge penalty, then the polarization penalty—so that a useful structure forms before pressure to discretize it. This phase lets the model explore alternative logical structures and concept organizations.

Once routing has become confident (the gates approach a discrete one-hot assignment) or a scheduled point is reached, training enters the *hardening and fine-tuning phase*. We restore the best soft-routing checkpoint and convert its soft gates to discrete assignments: each node commits to its top-N incoming edges among the open L_0 gates and the root pointer to its arg-max, yielding an explicit ontology and graded-logic reasoning structure. The routing is then held fixed while the remaining parameters—aggregation operators and concept representations—are fine-tuned against the hardened topology. Because evaluation reads directly from the frozen structure, the reported accuracy is that of the exact discrete model, preserving faithfulness and interpretability.

4 Towards Human-Aligned Reasoning

Human-aligned reasoning requires more than accurate prediction. A reasoning process should operate on intuitive concepts, support different levels of abstraction, identify the most relevant factors underlying a decision, and provide explanations that can be followed and understood by humans. The analyses in this section evaluate these complementary aspects of reasoning behavior. Specifically, we examine reasoning capacity, reasoning granularity, decision conciseness, and concept quality as observable characteristics of a human-aligned reasoning process. Together, these evaluations provide insight into whether OCBM not only achieves strong predictive performance but also learns decision structures that are more interpretable, semantically meaningful, and consistent with intuitive reasoning.

4.1 Competitive Reasoning Capacities with End-to-End Explainability

We first evaluate whether OCBM preserves, or even improves upon, the reasoning capabilities of existing concept bottleneck models. Although OCBM imposes a strong structural constraint through ontology-structured aggregation trees, the expressiveness of graded-logic operators enables the model to represent a wide range of decision patterns. By automatically discovering mandatory and optional conditions, learning varying degrees of conjunctive and disjunctive relationships, and incorporating negations where appropriate, OCBM can adapt its reasoning structure to diverse tasks while maintaining semantic interpretability.

We compare OCBM with several representative CBM approaches, as summarized in Table 2. Across the CUB (Wah et al., 2011), AwA2 (Xian et al., 2019), and CIFAR100 (Krizhevsky, 2009) benchmarks, OCBM achieves competitive or leading predictive performance, demonstrating that the additional ontology constraints do not compromise model expressiveness. Instead, the results suggest that graded-logic reasoning provides sufficient flexibility to capture complex decision boundaries while preserving transparent and human-interpretable reasoning structures.

OCBM achieves competitive predictive performance without sacrificing explainability. Existing approaches such as LogicCBM primarily aggregate predicate evaluations through weighted combinations, while CREAM (Kalampalikis et al., 2025) exposes concept-to-concept and concept-to-target relationships through graph segments. Although these methods provide valuable concept-level explanations, they do not explicitly model decisions as a sequence of semantically meaningful reasoning steps. In contrast, OCBM organizes concepts into ontology-guided aggregation hierarchies, enabling predictions to be derived through successive logical compositions that can be traced from primitive concepts to the final outcome. Table 3 provides a qualitative

Table 2: Classification Accuracy: [1]-CBM(Koh et al., 2020); [2]-Label-Free(Oikarinen et al., 2023); [3]-LogicCMB(Vemuri et al., 2025); [4]MLP-based CBM(Havasi et al., 2022); [5]Sparse-CBM (Semenov et al., 2024). Best is bolded.

Model	CUB	AwA2	CIFAR100
CBM[1]	$75.20_{\pm 0.79}$	$88.81_{\pm 0.52}$	$55.39_{\pm 0.62}$
Label-Free[2]	$74.29_{\pm 0.24}$	$89.76_{\pm 0.20}$	$65.16_{\pm 0.34}$
LogicCBM[3]	$\mathbf{81.13_{\pm 0.42}}$	$90.04_{\pm 0.05}$	$68.46_{\pm 0.45}$
MLP CBM[4]	$72.63_{\pm 0.25}$	$89.15_{\pm 0.23}$	$65.00_{\pm 0.31}$
SparseCBM[5]	$70.28_{\pm 1.05}$	$84.34_{\pm 0.12}$	$61.68_{\pm 1.00}$
VLG-CBM	$60.38_{\pm 0.00}$	-	$65.73_{\pm 0.00}$
OCBM	$79.26_{\pm 0.54}$	$\mathbf{95.76_{\pm 0.02}}$	$\mathbf{69.89_{\pm 0.94}}$

Table 3: Qualitative comparison of concept/reason capabilities of different CBMs. [1]-HIL-CBM (Xie et al., 2026);[2]-U-F^2-CBM (Sammani et al., 2026);[3]-Z-CBM (Yamaguchi et al., 2026)

Model	No concept annotation	No External vocabulary	Emergent concepts	Complete logic rules
CBM	-	Y	-	-
CREAM	Y	-	-	concept graph
G-CBM	-	Y	-	concept graph
HIL-CBM[1]	Y	-	-	hierarchical concepts
Label-Free	Y	-	-	-
LogicCBM	-	Y	-	sum of predicates
UCBM	Y	Y	Y	sparse concept selection
U-F^2-CBM[2]	Y	-	-	-
VLG-CBM	Y	-	-	weighted concepts
Z-CBM[3]	Y	-	-	-
OCBM	Y	Y	Y	complete graded logic

comparison between OCBM and some recent approaches. Figure 2 illustrates several verbalized OCBM reasoning structures generated by an LLM from learned graded-logic expressions, grounded in graded-logic rules (LLM prompts and additional samples are provided in Appendix B). Notably, for Hooded Merganser, OCBM identifies distinct male and female feature groups as alternative disjunctive conditions for the same species, reflecting an interpretable decomposition of intra-class visual variation. These examples demonstrate that OCBM not only predicts accurately, but also exposes the logical compositions underlying its decisions in a form that can be inspected, verbalized, and validated.

4.2 Interactive Reasoning with Adjustable Granularity

Human reasoning about complex problems rarely consists of a flat aggregation of elementary inputs. Instead, evidence is typically organized into different aspects and progressively integrated into an overall judgment. OCBM's hierarchical aggregation mechanism provides a computational framework for this form of multi-layered reasoning. Naturally, this does not imply that the learned groups necessarily correspond to real-world or socially meaningful concepts, since the model has access only to statistical relationships present in the training data. Moreover, as with all CBM approaches, multiple valid logical explanations may exist for the same prediction. However, because OCBM exposes its reasoning process in an interpretable form that traces the entire reasoning path, domain experts can directly incorporate their knowledge into the model by refining concepts, modifying relationships between concepts, or imposing domain-specific constraints. This interplay between machine-discovered structures and human expertise provides a promising path toward reasoning systems that are not only accurate, but also more aligned with human expectations and values.

To evaluate the versatility of OCBM in supporting this interactive paradigm, we trained OCBMs with different reasoning-head widths w on the CUB dataset ($w = 4 : 77.25_{\pm 0.19}\%$; $w = 8 : 76.95_{\pm 1.77}\%$; $w = 64 : 79.01_{\pm 1.12}\%$). The results show that varying the head width has little impact on predictive performance, making it a practical parameter for controlling reasoning granularity.

Interestingly, the learned reasoning structures also exhibit a qualitative shift with head width. When more head capacity is available, OCBM tends to use additional conjunctive conditions to progressively refine its decisions, whereas under limited head width, it relies more heavily on disjunctive conditions to combine alternative concept configurations within a compact reasoning structure. This suggests that head width influences not only the granularity but also the form of the learned reasoning. Such diversity of structures creates opportunities for reasoning ensembles. For the CUB dataset, an ensemble with oracle selection achieves $80.35_{\pm 0.50}\%$ task accuracy, with oracle ceiling (label-peek union) at $86.22_{\pm 0.38}\%$. This finding suggests that multiple reasoning perspectives can contribute complementary information rather than merely providing redundant explanations.

4.3 Concise Reasoning

While complex and highly expressive reasoning structures can be examined to justify consequential decisions, such as evaluating alternative designs for critical infrastructure or defense systems, many everyday decision-making scenarios require reasoning that is both understandable and concise. Human decision-makers typically rely on a limited number of salient factors when making routine judgments, and explanations that require excessive reasoning chains may hinder rather than improve transparency. To address this challenge, OCBM supports a left-associative binary-tree variant for learning compact reasoning structures. In this case, a Gumbel-Sinkhorn (Mena et al., 2018) permutation layer learns the ordering of concepts, placing influential concepts closer to the root; progressive pruning therefore identifies inputs that can be removed with little loss in predictive performance. We conducted experiments on ten disease-classification datasets from the UCI Machine Learning

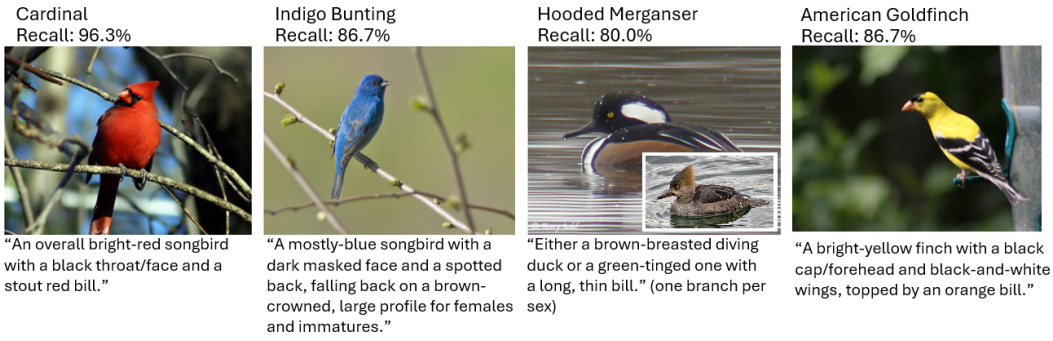

Figure 2: Sample verbalization of CUB bird classification based on learned graded-logic trees, grounded in basic graded-logic syntax and rules.

Table 4: Emergent digit rules (K=5, full structure ally pruned; pruned accuracy 99.23%). Concepts: c_0 = junction; c_1 = upper rounded arc / hook; c_2 = U-turn (varied direction); c_3 = arc; c_4 = slant bar. $\bar{\wedge}$=Hyper conjunction (below lowest), $\wedge$=Pure conjunction (deciced by lowest), $\bar{\Delta}$=Hard conjunction (must have all), $\underline{\vee}$=Soft disjunction(nice to have some), $\bar{\nabla}$=Hard disjunction (Enough to have any) $\vee$=Pure disjunction (decided by highest), $\bar{\vee}$=Hyper disjunction (above the highest)

	GRADED-LOGIC EXPRESSION	VERBALIZATION
0	$0.8\,(0.5\,(0.5\,(0.6\,c_3\;\bar{\wedge}\;0.4\,\overline{c_0})\;\bar{\wedge}\;0.5\,\overline{c_2})\;\bar{\wedge}\;0.5\,\overline{c_4})\;\vee\;0.2\,c_1$	A all curly figure, **or** has an upper rounded arc
1	$0.4\,(0.9\,(0.9\,\overline{c_1}\;\underline{\vee}\;0.1\,\overline{c_3})\;\bar{\nabla}\;0.1\,c_4)\;\bar{\wedge}\;0.6\,\overline{c_0}$	not curly **and** has **no** junctions (bare stroke)
2	$0.8\,(0.8\,(0.7\,(0.2\,\overline{c_0}\;\bar{\wedge}\;0.8\,c_4)\;\bar{\wedge}\;0.3\,c_2)\;\bar{\wedge}\;0.2\,c_1)\;\underline{\vee}\;0.2\,c_3$	strongly **prefer** a slant bar and an upper rounded arc, can be supplemented by an arc
3	$0.7\,(0.9\,(0.8\,(0.5\,\overline{c_1}\;\bar{\wedge}\;0.5\,\overline{c_4})\;\bar{\Delta}\;0.2\,c_3)\;\underline{\vee}\;0.1\,\overline{c_0})\;\underline{\vee}\;0.3\,\overline{c_2}$	**must** be all curves, can be **supplemented** by missing junctions and U-turns
4	$0.4\,c_2\;\bar{\wedge}\;0.6\,c_4$	**must** have a slant bar **and** a U-turn
5	$0.7\,(0.6\,(0.3\,(0.5\,\overline{c_1}\;\bar{\nabla}\;0.5\,c_3)\;\underline{\vee}\;0.7\,\overline{c_4})\;\bar{\wedge}\;0.4\,c_2)\;\underline{\vee}\;0.3\,c_0$	**must** have a U-turn **and** some arc
6	$0.6\,(0.7\,(0.6\,(0.9\,\overline{c_2}\;\underline{\vee}\;0.1\,c_4)\;\bar{\wedge}\;0.4\,\overline{c_1})\;\bar{\wedge}\;0.3\,\overline{c_0})\;\underline{\vee}\;0.4\,\overline{c_3}$	**must** have **no** U-turns, no upper rounded arc, no junction
7	$0.4\,(0.2\,c_1\;\bar{\Delta}\;0.8\,c_4)\;\bar{\wedge}\;0.6\,c_0$	**must** have a slant bar **and** a junction
8	$0.7\,(0.2\,(0.8\,c_2\;\underline{\vee}\;0.2\,c_3)\;\bar{\nabla}\;0.8\,c_0)\;\bar{\wedge}\;0.3\,c_1$	**must** have upper rounded arc **and** a junction
9	$0.27\,(0.28\,(0.62\,c_1\;\bar{\wedge}\;0.38\,c_4)\;\bar{\wedge}\;0.72\,c_3)\;\bar{\wedge}\;0.73\,c_0$	**must** have all: junction, upper rounded arc, slant bar and junction

Repository, the Diabetes Health Indicators dataset, and the ACTG 175 HIV clinical trial dataset (University of California, 2025; Teboul, 2022; aid, 1999). Across the ten medical prediction tasks, OCBM achieved the best average AUPRC rank (1.8) while pruning 57.7% of input features on average (Appendix A). When greater expressiveness is required, the retained concepts can subsequently be used with the fully connected ontology.

4.4 Emergence of Intuitive Concepts

Several studies have explored concept bottleneck models without human-annotated concepts. UCBM, for instance, discovers concepts from latent representations using unsupervised concept discovery, producing concepts that are predictive of the target task. OCBM takes a fundamentally different approach. By using graded-logical ontologies as both the decision mechanism and a structural regularizer, OCBM encourages concepts to emerge according to their role in a coherent reasoning process rather than solely their predictive utility. As a result, concept discovery and reasoning are learned jointly, promoting human-interpretable abstractions that support both accurate prediction and explainable decision making.

To investigate this property, we conducted experiments on the MNIST dataset, aiming not only to learn digit classification rules but also to uncover the underlying concepts used in the reasoning process. In this experiment, we place a 5-concept ontology with ten left-associative binary graded-logic trees, one per digit, on top of a ResNet18 backbone. The backbone is forced to emit 5 concepts shared across all ten trees; each tree combines these concepts through trainable graded-logic aggregators and input weights, with an optional per-concept negation. Crucially, the model is trained on digit labels alone, with no concept supervision. The model reaches 99.23% test accuracy with only 5 concepts, and every digit tree can be read directly as a formula over them, as summarized in Table 4.

The concept labels in Table 4 are assigned through the generalized semantic-grounding process illustrated in Figure 3, which integrates complementary validation strategies inspired by prior work such as UF^2-CBM, HIL-

CBM, Z-CBM, and related approaches. Because grounding requirements may vary across projects, we adopt a hybrid framework in which multiple forms of grounding evidence can be selected and combined. The resulting degree of grounding satisfaction is computed by a preference graded-logic tree specified by the stakeholder to encode project-specific preferences over the composition of validation evidence.

In our experiments, the LLM proposes and refines semantic interpretations of the discovered concepts, and we calculate concept grounding (G_c), structural grounding (G_s) and functional grounding (G_f). These grounding dimensions are general, but their operationalizations are evaluation-dependent; the formulations below are the specific instantiations used in our MNIST/USPS (Hull, 1994) evaluation:

Assume the concept encoder maps an image to K sigmoid concept activations $c(x) = \sigma(f_\theta(x)) \in [0,1]^K$, and each class $d \in \{0, \ldots, 9\}$ is scored by a frozen graded-logic tree $T_d : [0,1]^K \to [0,1]$. From tree d we read the *usage polarity* $\pi_{i,d} \in \{+1, -1\}$ of concept i (identity vs. negation of its leaf), and we write $\bar{c}_{i,d}$ for the mean activation of concept i over the test images of class d. Then we define:

G_{ci}: internal activation coherence. A well-grounded concept fires *high* on the classes whose rules use it positively and *low* on those that negate it. Let $D_i^+ = \{d : \pi_{i,d} = +1\}$ and $D_i^- = \{d : \pi_{i,d} = -1\}$. We define the negation-aware coherence gap as $G_{ci}^{(i)} = max(0, \frac{1}{|D_i^+|}\sum_{d \in D_i^+} \bar{c}_{i,d} - \frac{1}{|D_i^-|}\sum_{d \in D_i^-} \bar{c}_{i,d})$. A value near 0 flags a diffuse concept that fires irrespective of how the rules use it.

G_{ce}: external alignment. We test whether each concept behaves like a distinct human primitive. Given a vocabulary of human strokes $\{h_j\}$, we form the direction-agnostic identification AUC $a_{ij} = max(AUC(C_i, h_j), 1 - AUC(C_i, h_j)) \in [\frac{1}{2}, 1]$, match concepts to distinct strokes with a one-to-one (Hungarian) assignment M maximizing $\sum_i a_{i,M(i)}$ and then rescale to $[0,1]$: $G_{ce}^{(i)} = max(0, 2a_{i,M(i)} - 1)$.

G_s: logical coherence. We test whether a concept's proposed meaning is consistent with the frozen logic. Probing each tree at the neutral point $c^0 = \frac{1}{2}$ gives the signed structural influence $S_{i,d} = T_d(c^0 | c_i = 1) - T_d(c^0 | c_i = 0), \sigma_{i,d} = sign(S_{i,d})^{\mathbb{1}}[|S_{i,d}| > \epsilon]$. Given the proposer's world-knowledge membership – class P_i where the concept should be present and A_i, where it should be absent – we define a consistency and an evidence support: $cons_i = \frac{|\{d \in P_i : \sigma_{i,d} = +1\}| + |\{d \in A_i : \sigma_{i,d} = -1\}|}{|\{d \in P_i \cup A_i : \sigma_{i,d} \neq 0\}|}, s_i = |\{d \in P_i \cup A_i : \sigma_{i,d} \neq 0\}|$, and combine them so that a perfect consistency on too few classes is not mistaken for solid grounding: $G_s^{(i)} = cos_i \cdot \frac{min(s_i, s_{max})}{s_{max}}, s_{max} = 5$.

G_f: functional grounding. As a model-level functional test we measure zero-shot transfer of the whole reasoning system to the external USPS digit corpus, applying only an unsupervised per-concept quantile calibration of the concept activations (no USPS labels). G_f is the resulting calibrated top-1 accuracy. Figure 4 illustrates the result of zero-shot cross-domain transfer study on the OCBM model trained on MNIST dataset. We directly applied the model to the USPS dataset without additional training. Despite the domain shift, the model reasoning transfers at 81.37% after unsupervised calibration, suggesting that the discovered concepts capture structural properties that extend beyond a single dataset. As an ablation study, we replaced the graded-logic head with linear and MLP heads, which exhibit substantially worse and less stable cross-domain transfer, as summarized in Table 5.

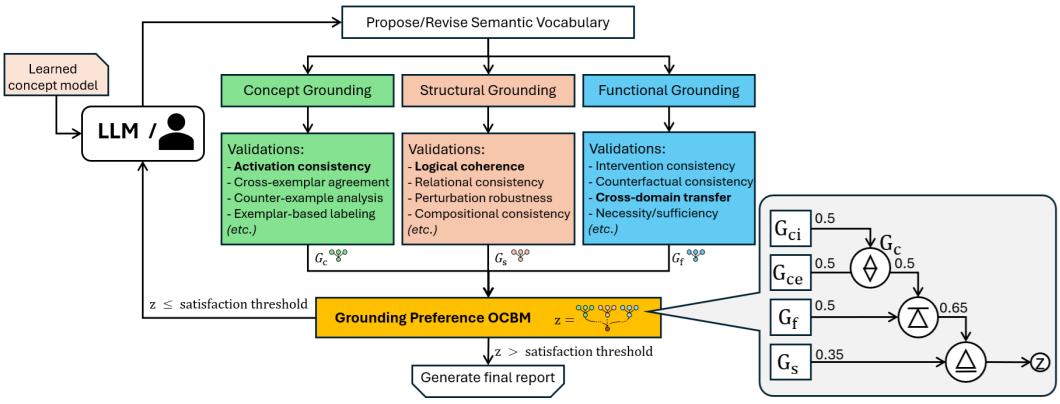

Figure 3: Concept semantic grounding workflow that supports human-AI collaboration.

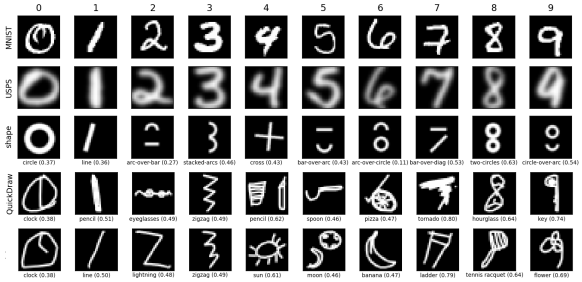

Figure 4: Concept quality evaluation. Row 1: MNIST samples; Row 2: Zero-shot transfer to USPS; Row 3: Top-activating synthetic shape probes; Row 4-5: Qualitative cross-domain concept probing on QuickDraw

Table 5: Effect of reasoning architecture on emergent concept generalization. OCBM concepts transfer better and more stable comparing to CBM with linear/MLP heads. USPS_c is result after unsupervised calibration

HEAD	**Ontology**	**MLP**	**Linear**
MNIST	$99.23_{\pm 00.43}$	$99.60_{\pm 00.04}$	$93.78_{\pm 04.78}$
USPS	$\mathbf{77.52_{\pm 11.48}}$	$50.89_{\pm 16.10}$	$49.17_{\pm 21.80}$
USPS_c	$\mathbf{81.37_{\pm 03.65}}$	$72.86_{\pm 09.28}$	$69.25_{\pm 10.47}$
G_{ci}	0.49	N/A (0)	N/A (0)
G_{ce}	0.80	0.78	0.68
G_c	0.65	0.39	0.34
G_s	0.76	N/A (0)	N/A (0)
G_f	0.81	0.73	0.69
z	**0.74**	0.23	0.20

The grounding measures are then aggregated by a preference graded-logic tree (Figure 3): G_{ci} and G_{ce} are first combined by arithmetic mean to obtain G_c. G_c and G_f are treated as mandatory criteria through hard conjunction, while G_s enters through soft conjunction. Thus, models without explicit structural grounding are penalized but are not automatically assigned zero overall grounding satisfaction. The final rows of Table 5 report these grounding measures and the resulting aggregate score z.

We further examined concept emergence in a substantially different domain using the Google QuickDraw! dataset (Ha & Eck, 2018) and observed shapes with top activations per class. The results indicate that OCBM learns intuitive geometric concepts that generalize across categories. For example, the model identified enclosed circular structures resembling the shape of the digit zero and zigzag-shaped structures resembling the digit three. These concepts emerged without explicit supervision, indicating that the ontology captures visually meaningful abstractions rather than dataset-specific artifacts.

To assess the identifiability of the learned concepts, we used the logical expressions shown in Table 4 to construct synthetic shape probes that satisfy the discovered rules. The model consistently responded to these probes in accordance with the learned concepts, demonstrating that the resulting ontology is not merely predictive but also semantically coherent. This behavior provides evidence that the learned concepts are aligned with intuitive human interpretations and can serve as meaningful building blocks for transparent reasoning systems.

Taken together, these results suggest that graded-logic ontologies can facilitate the emergence of transferable and interpretable concepts while preserving predictive performance. Such properties are essential for human-aligned reasoning, where learned concepts must not only support accurate decisions but also remain understandable, verifiable, and actionable by human users.

5 Conclusion and Future Work

We introduced Ontology Concept Bottleneck Models (OCBMs), a concept-based learning framework that replaces conventional prediction heads with class-specific graded-logic ontologies. By expressing predictions through explicit graded-logic aggregation trees, OCBMs provide transparent reasoning processes while simultaneously constraining representation learning through a shared concept space. Across diverse benchmarks, OCBMs demonstrated that graded-logic reasoning can support the emergence of transferable concepts and coherent logical rules that generalize across datasets and admit direct human inspection. These findings suggest that reasoning structures based on graded logic may offer a practical pathway toward machine reasoning that is not only interpretable, but also more closely aligned with how humans evaluate evidence, balance competing considerations, and form decisions.

Several directions remain for future work. First, we plan to investigate interactive ontology refinement, allowing domain experts to directly modify learned concepts and logical relationships. Second, we will explore larger and more complex reasoning domains, including multimodal and sequential decision-making tasks. Finally, we aim to study the role of graded-logic ontologies in autonomous agents, where transparent and auditable reasoning may improve trust, collaboration, and human oversight. We believe that OCBMs establish a foundation for further study of interpretable and auditable machine reasoning toward human alignment.

AI use statement

In this work, we used generative AI tools for spelling and grammar checks and for limited writing-style recommendations. We did not use generative AI tools for any other tasks related to the manuscript text. We also used

generative AI tools to create and edit software code used in our benchmarks. In addition, we used a coding agent for dataset management and to orchestrate long-running experiments spanning several hundred hours. We leveraged AI-assisted search tools to identify related work, including generating summaries of relevant papers to help us select candidates for benchmark comparisons. We reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, code, results, and other artifacts produced with the assistance of generative AI.

Reproducibility statement

We support reproducibility through the use of standard benchmark datasets, detailed descriptions of the experimental setup and evaluation methodology in the main paper and appendices, and publicly accessible anonymized source code at https://anonymous.4open.science/r/bacon-net. The repository contains the implementations, configurations, and scripts used to reproduce the reported experiments. Dataset preprocessing and experimental procedures are described in the corresponding experimental sections and supplementary materials. Upon acceptance, the repository will be publicly released under the MIT License.

Acknowledgments

We gratefully acknowledge Jozo Dujmović for his foundational contributions to graded logic and Logic Scoring of Preference, and for the many insightful discussions through which we deepened our understanding of these topics.

References

Hepatitis. UCI Machine Learning Repository, 1983. DOI: https://doi.org/10.24432/C5Q59J.

A randomized, double-blind phase ii/iii trial of monotherapy vs. combination therapy with nucleoside analogs in hiv-infected persons with cd4 cells of 200-500/mm^3 (clinicaltrials.gov identifier nct00000625). ClinicalTrials.gov, National Library of Medicine (U.S.), 1999. URL https://clinicaltrials.gov/study/NCT00000625. Accessed: 2025-12-30.

Cervical Cancer Behavior Risk. UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5402W.

Marzia Ahmed. Maternal Health Risk. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5DP5D.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks, 2018. URL https://arxiv.org/abs/1806.07538.

David Debot, Pietro Barbiero, Francesco Giannini, Gabriele Ciravegna, Michelangelo Diligenti, and Giuseppe Marra. Interpretable concept-based memory reasoning, 2024. URL https://arxiv.org/abs/2407.15527.

Jozo Dujmović. *Graded Logic*. Springer Cham, 2025.

Jozo J. Dujmovic. Continuous preference logic for system evaluation. *IEEE Transactions on Fuzzy Systems*, 15(6):1082–1099, 2007. doi: 10.1109/TFUZZ.2007.902041.

Jozo Dujmović. Weighted conjunctive and disjunctive means and their application in system evaluation. *Publications de la Faculté d'Electrotechnique de l'Université à Belgrade. Série Mathématiques et Physique*, 483, 01 1974.

Arslan Hilal Aktürk Selin Gülşen Mervenur Kültekin Nimet Esen, Irfan and Oğuzhan Özdemir. Gallstone. UCI Machine Learning Repository, 2024. DOI: https://doi.org/10.1097/md.0000000000037258.

Giovanni De Felice, Arianna Casanova, Francesco De Santis, Silvia Santini, Johannes Schneider, Pietro Barbiero, and Alberto Termine. Causally reliable concept bottleneck models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=UX143QGvb8.

David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23386–23397. Curran Associates, Inc., 2022. doi: 10.52202/068431-1699. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/944ecf65a46feb578a43abfd5cddd960-Paper-Conference.pdf.

Md Mohasin Hossain, Anar Amirli, Robert Leist, Md Abdul Kadir, and Daniel Sonntag. Beyond heatmaps: Unsupervised concept-graph reasoning for interpretable visual explanation, 2026. URL https://arxiv.org/abs/2607.01416.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.

Nektarios Kalampalikis, Kavya Gupta, Georgi Vitanov, and Isabel Valera. Towards reasonable concept bottlenecks, 2025. URL https://openreview.net/forum?id=IPrvnMoKHM.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks, 2018. URL https://arxiv.org/abs/1802.08665.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models, 2023. URL https://arxiv.org/abs/2304.06129.

B. Ramana and N. Venkateswarlu. Ilpd (indian liver patient dataset). UCI Machine Learning Repository, 2022.

Fawaz Sammani, Jonas Fischer, and Nikos Deligiannis. Clip-free, label free, unsupervised concept bottleneck models, 2026. URL https://arxiv.org/abs/2503.10981.

Simon Schrodi, Julian Schur, Max Argus, and Thomas Brox. Concept bottleneck models without predefined concepts, 2024. URL https://arxiv.org/abs/2407.03921.

Andrei Semenov, Vladimir Ivanov, Aleksandr Beznosikov, and Alexander Gasnikov. Sparse concept bottleneck models: Gumbel tricks in contrastive learning, 2024. URL https://arxiv.org/abs/2404.03323.

Divyansh Srivastava, Ge Yan, and Tsui-Wei Weng. Vlg-cbm: Training concept bottleneck models with vision-language guidance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 79057–79094. Curran Associates, Inc., 2024. doi: 10.52202/079017-2510. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/90043ebd68500f9efe84fedf860a64f3-Paper-Conference.pdf.

Alex Teboul. Diabetes health indicators dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset, 2022. Accessed: 2025-12-30.

Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467, 1989. doi: 10.1017/S0140525X00057046.

Athanasios Tsanas and Max Little. Parkinsons Telemonitoring. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C5ZS3N.

Irvine University of California. Uci machine learning repository medical datasets. https://archive.ics.uci.edu, 2025. Heart Disease; Breast Cancer Wisconsin (Diagnostic); Gallstone; Cervical Cancer Behavior Risk; Indian Liver Patient Dataset (ILPD); Parkinsons Telemonitoring; Hepatitis; Maternal Health Risk.

Deepika SN Vemuri, Gautham Bellamkonda, Aditya Pola, and Vineeth N Balasubramanian. Logiccbms: Logic-enhanced concept-based learning, 2025. URL https://arxiv.org/abs/2512.07383.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Mangasarian Olvi Nick Wolberg, William and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768.

Haodong Xie, Yujun Cai, Rahul Singh Maharjan, Yiwei Wang, Federico Tavella, and Angelo Cangelosi. Hierarchical, interpretable, label-free concept bottleneck model, 2026. URL https://arxiv.org/abs/2604.02468.

Haotian Xu, Tsui-Wei Weng, Lam M. Nguyen, and Tengfei Ma. Graph concept bottleneck models, 2026. URL https://arxiv.org/abs/2508.14255.

Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models, 2026. URL https://arxiv.org/abs/2502.09018.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition, 2023. URL https://arxiv.org/abs/2308.03685.

A Test result details of UCI disease datasets

This study evaluates OCBM across ten publicly available medical datasets commonly used in clinical machine learning research. The datasets span multiple diagnostic tasks and cover cardiovascular, oncological, endocrine, hepatic, neurological, and general health domains. All datasets contain tabular patient records with structured clinical features and binary outcome labels. To ensure consistency across experiments, we applied minimal and standardized preprocessing: categorical attributes were one-hot encoded, continuous values were normalized to the $[0, 1]$ interval using a sigmoid-based scaling function to reduce the influence of outliers, and missing entries were handled using dataset-provided defaults or instance removal when necessary.

Each dataset was randomly partitioned into training, validation, and test sets using a 60/20/20 split, stratified by label distribution to maintain class balance (smaller datasets were evaluated using cross-validation). All results are averaged over five independent runs to reduce variance. These tests use OCBM left-associative binary tree structure instead of a regulated full-tree in the main paper to facilitate feature selection. No dataset-specific tuning or architectural modifications were performed; the same OCBM configuration and hyperparameters were applied across all ten tasks. Table 6 provides a brief summary of datasets used in this study.

Table 6: Summary of Datasets Used in This Study

Dataset	**Domain**	**Samples**	**Feature count**	**Positive rate**	**Base accuracy**
Heart disease (Janosi & Detrano, 1989)	Cardiology	297	13	46.1%	46.7%
Breast cancer (Wolberg & Street, 1993)	Oncology	569	30	37.0%	37.3%
Diabetes (Teboul, 2022)	Endocrinology / Metabolic	253,680	21	13.9%	15.4%
Gallstone (Esen & Özdemir, 2024)	Gastroentero- logy / Hepatobiliary	319	37	49.5%	50.0%
Cervical cancer (cer, 2019)	Oncology / Gynecology	72	19	29.2%	77.8%
Liver cirrhosis (Ramana & Venkateswarlu, 2022)	Hepatology	584	10	71.4%	70.9%
Parkinsons (Tsanas & Little, 2009)	Neurology	5,875	19	25.0%	25.0%
Hepatitis (hep, 1983)	Infectious Disease / Hepatology	155	19	20.6%	77.42%
Maternal health (Ahmed, 2020)	Obstetrics	1,013	6	26.8%	26.6%
AIDS (aid, 1999)	Infectious Disease	2,139	23	24.4%	24.3%

Table 7 summarizes predictive accuracy across all ten datasets, comparing OCBM with three widely used tabular baselines: logistic regression (LR), random forest (RF), and gradient boosted trees (XGB). All baseline models were trained using the same data partitions. Their hyperparameters were selected through validation-based model selection on each dataset, while OCBM was trained using a single configuration across all tasks. No feature selection or dimensionality reduction was performed; all datasets were used in their full feature space.

To summarize performance across heterogeneous tasks in a scale-free manner, we additionally report the average AUPRC rank across datasets (Table 7, last row). OCBM attains the best average rank among all compared models, reflecting consistently strong ranking performance—particularly on imbalanced and interaction-dominated tasks—without relying on dataset-specific tuning.

Table 7: Predictive performance comparison (Accuracy, F1, AUPRC)

Dataset	LR			RF			XGB			OCBM		
	Acc	F1	AUPRC	Acc	F1	AUPRC	Acc	F1	AUPRC	Acc	F1	AUPRC
Heart disease	83.33%	0.8148	0.9194	85.00%	0.8302	0.8993	86.67%	0.8519	0.8948	**86.67%**	**0.8667**	**0.9380**
Breast cancer	92.12%	0.8989	0.9900	97.37%	0.9630	0.9938	97.37%	0.9630	0.9926	**99.12%**	**0.9880**	**0.9979**
Diabetes	70.33%	0.4222	0.3797	85.73%	0.2363	0.3535	72.28%	0.4419	0.4232	**72.72%**	**0.7081**	**0.7321**
Gallstone	78.12%	0.7667	0.8877	79.69%	0.7869	0.8824	**85.94%**	**0.8475**	**0.9213**	82.92%	0.8197	0.9036
Cervical cancer	88.89%	0.8000	0.8211	**94.44%**	**0.9091**	**1.0000**	94.44%	0.9091	0.9667	91.67%	0.8571	0.9119
Liver cirrhosis	**68.38%**	**0.8083**	**0.8951**	74.36%	0.8370	0.8888	69.23%	0.8022	0.8827	73.24%	0.8364	0.8891
Parkinson	53.26%	0.4493	0.3584	95.00%	0.9005	0.9645	**98.47%**	**0.9692**	**0.9952**	88.16%	0.8075	0.5939
Hepatitis	81.45%	0.6250	0.8310	87.10%	0.6000	0.8208	83.87%	0.6154	0.6063	**78.41%**	**0.7822**	**0.8384**
Maternal health	74.88%	0.6331	0.6664	**95.57%**	**0.9174**	**0.9472**	94.58%	0.8972	0.9437	77.79%	0.7666	0.8529
AIDS	78.97%	0.6121	0.6620	89.25%	0.7629	0.8276	88.55%	0.7656	0.8591	**84.62%**	**0.8508**	**0.9009**
Average AUPRC Rank	3.1			2.7			2.5			**1.8**		

Feature selection is done by progressing pruning the left-associative binary tree while observing performance changes. As shown in figure 5, many of the features (in the identified redundancy plateau zone) can be pruned without affecting task accuracy.

B OCBM LLM VERBALIZATION DETAILS

We feed learned graded-logic rules to a LLM model to generate verbalization of the rules. The following is a sample prompt we used:

```
System
=======
You are a report generator that produces a precise, human-readable, and
↪  insightful explanation of a Graded Logic (GL / LSP) aggregation tree.
↪  Use your knowledge of graded logic and ornithology to generate a
↪  clear, logical report.

Context
=======
This aggregation tree is a fine-grained bird-species classifier for the
↪  CUB-200 dataset. It scores how well an image matches the species
↪  **Indigo Bunting**. The leaf "features" are fuzzy attribute-truths in
↪  [0, 1] (a value near 1 means the attribute is present); "NOT" denotes
↪  a negated attribute. Each internal node is a graded-logic aggregator
↪  whose andness (its GCD code) sets how strictly its weighted children
↪  must be satisfied. Weights are the normalized relevances used inside
↪  the power mean.

Background
==========
GL aggregators are named by their andness per this table:

| Code | Name | Andness a | Verbalization |
|------|------|-----------|---------------|
| CC  | Drastic conjunction     | 2        | "must all be completely
↪  satisfied" |
| HHC | High hyper-conjunction  | [5/4, 2) | "below the lowest" |
| CP  | Product t-norm          | 5/4      | "below the lowest" |
| LHC | Low hyper-conjunction   | [1, 5/4) | "below the lowest" |
| C   | Pure conjunction        | 1        | "decided by the lowest" |
| HC+ | High hard conjunction   | 13/14    | "must have all" |
| HC  | Medium hard conjunction | 12/14    | "must have all" |
| HC- | Low hard conjunction    | 11/14    | "must have all" |
| SC+ | High soft conjunction   | 10/14    | "nice to have most" |
```

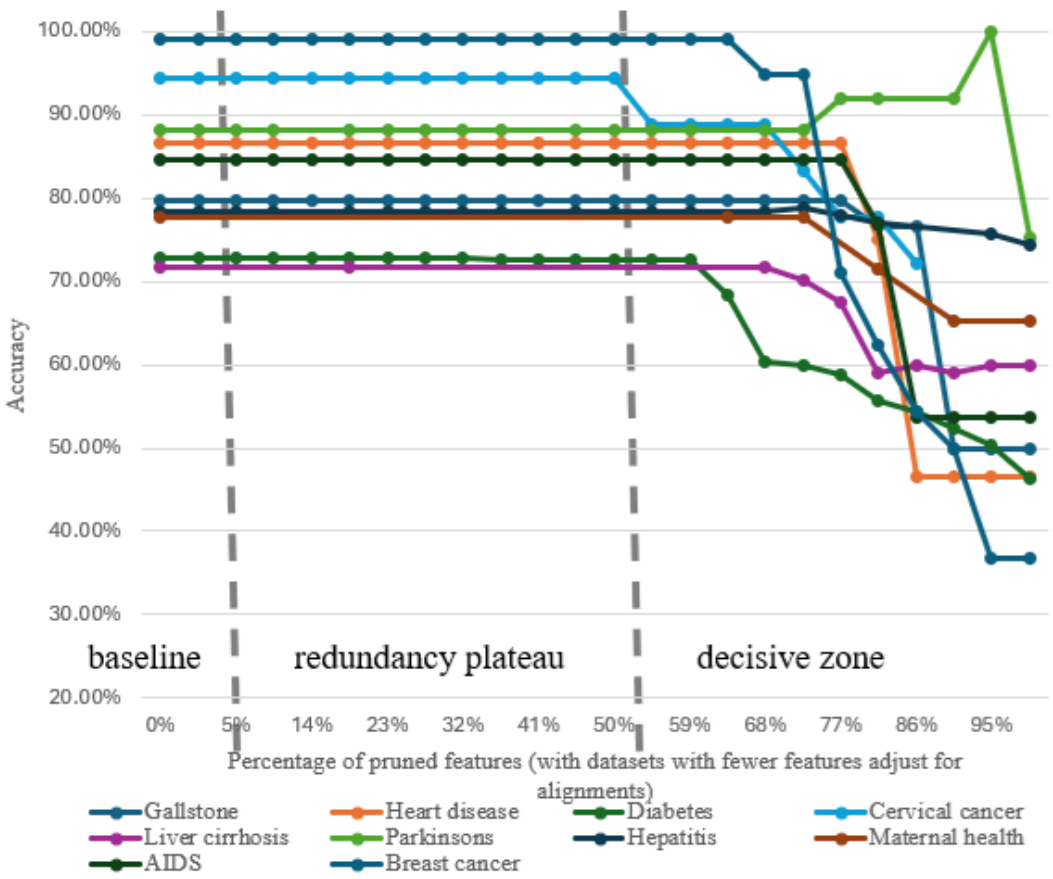

Figure 5: Accuracy as a function of retained aggregation depth across ten medical datasets. A consistent three-zone structure emerges: an early baseline region encoding a learned class prior, a redundancy plateau where correlated features contribute stably, and a decisive zone in which a small number of late-stage features govern diagnostic outcomes. Vertical dashed lines indicate typical transitions between regions.

```
| SC  | Medium soft conjunction | 9/14      | "nice to have most" |
| SC- | Low soft conjunction    | 8/14      | "nice to have most" |
| A   | Arithmetic mean         | 7/14      | "nice to have" |
| SD- | Low soft disjunction    | 6/14      | "nice to have some" |
| SD  | Medium soft disjunction | 5/14      | "nice to have some" |
| SD+ | High soft disjunction   | 4/14      | "nice to have some" |
| HD- | Low hard disjunction    | 3/14      | "enough to have any" |
| HD  | Medium hard disjunction | 2/14      | "enough to have any" |
| HD+ | High hard disjunction   | 1/14      | "enough to have any" |
| D   | Pure disjunction        | 0         | "decided by the highest" |
| LHD | Low hyper-disjunction   | (-1/4, 0] | "above the highest" |
| DP  | Product t-conorm        | -1/4      | "above the highest" |
| HHD | High hyper-disjunction  | [-1, -1/4)| "above the highest" |
| DD  | Drastic disjunction     | -1        | "true unless all are zeros"
↪  |


Instructions
============
1. Organize the report into:
   - Overview: the overall decision logic in 2-3 sentences.
```

```
   - Decision Logic Walkthrough: step by step from the root to the
    ↪  leaves. For each node, name the operator by its code AND
    ↪  verbalization (e.g. "HC+ - high hard conjunction, a strict 'must
    ↪  have all'"), and explain how it combines its children given their
    ↪  weights.
   - Ornithological Interpretation: relate the logic to the real field
    ↪  marks of Indigo Bunting; note whether the rule matches known
    ↪  diagnostic features, and flag anything surprising, redundant, or
    ↪  spurious.
2. Use plain, human-friendly language.
3. Emphasize the highest-weight features and how they drive the decision.
4. Treat low-weight children as "compensatory / nice-to-have" evidence.

Input
======
The aggregation tree (pruned to its most influential children):
```json
{
  "operator": "SC+",
  "name": "high soft conjunction",
  "andness": 0.75,
  "verbalization": "nice to have most",
  "children": [
    {
      "weight": 0.221,
      "operator": "CP",
      "name": "product t-norm",
      "andness": 1.274,
      "verbalization": "below the lowest",
      "children": [
        {
          "weight": 0.136,
          "operator": "HHC",
          "name": "high hyper-conjunction",
          "andness": 1.369,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "back pattern: spotted"
            }
          ]
        },
        {
          "weight": 0.121,
          "operator": "CP",
          "name": "product t-norm",
          "andness": 1.262,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "head pattern: masked"
            }
          ]
        },
        {
          "weight": 0.105,
          "operator": "HHC",
          "name": "high hyper-conjunction",
          "andness": 1.319,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 0.524,
              "feature": "upper tail color: blue"
```
```

```
            },
            {
              "weight": 0.307,
              "feature": "underparts color: blue"
            },
            {
              "weight": 0.169,
              "feature": "wing color: blue"
            }
          ]
        }
      ]
    },
    {
      "weight": 0.221,
      "operator": "CP",
      "name": "product t-norm",
      "andness": 1.226,
      "verbalization": "below the lowest",
      "children": [
        {
          "weight": 0.136,
          "operator": "CP",
          "name": "product t-norm",
          "andness": 1.305,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "back pattern: spotted"
            }
          ]
        },
        {
          "weight": 0.132,
          "operator": "HHC",
          "name": "high hyper-conjunction",
          "andness": 1.369,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "back pattern: spotted"
            }
          ]
        },
        {
          "weight": 0.128,
          "operator": "HHC",
          "name": "high hyper-conjunction",
          "andness": 1.338,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "nape color: white"
            }
          ]
        }
      ]
    },
    {
      "weight": 0.214,
      "operator": "C",
      "name": "pure conjunction",
      "andness": 1.004,
```

```
      "verbalization": "decided by lowest",
      "children": [
        {
          "weight": 0.467,
          "operator": "CP",
          "name": "product t-norm",
          "andness": 1.305,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "back pattern: spotted"
            }
          ]
        },
        {
          "weight": 0.333,
          "operator": "C",
          "name": "pure conjunction",
          "andness": 1.017,
          "verbalization": "decided by lowest",
          "children": [
            {
              "weight": 0.503,
              "feature": "crown color: brown"
            },
            {
              "weight": 0.497,
              "feature": "size: large (16 - 32 in)"
            }
          ]
        },
        {
          "weight": 0.201,
          "operator": "CP",
          "name": "product t-norm",
          "andness": 1.212,
          "verbalization": "below the lowest",
          "children": [
            {
              "weight": 1.0,
              "feature": "nape color: white"
            }
          ]
        }
      ]
    }
  ]
}
```


And the following is a sample generated report on Indigo Bunting:

```
# Indigo Bunting | Graded-Logic Rule Verbalization

**Model:** recttree OCBM (W64, hardness-anneal, seed 2), CUB-200
**Species 13 - Indigo Bunting** test recall 86.7% faithful rule: 13
↪  literals, 22 ops
**Rule source:** `species013_rule.json` (pruned to top-3 children / node,
↪  depth 3)

## Plain-Language Statement
> It calls a bird an **Indigo Bunting** when it sees a **mostly-blue
↪  songbird with a dark masked face and a spotted back** - falling back
↪  on a brown-crowned, large profile for females and immatures.
```

```
## Overview
The classifier scores a bird as an Indigo Bunting with a **high soft
↪  conjunction**
(SC+, andness 0.75 - "nice to have most") over three near-equally
↪  weighted
evidence bundles (each about 0.22). Because the root is only a *soft*
↪  conjunction, a
bird does not have to satisfy all three bundles perfectly - strong
↪  bundles can
compensate for a weaker one | but each individual bundle is a strict
product / hyper-conjunction, so within a bundle the traits must nearly
↪  all hold.

## Decision-Logic Walkthrough
- **Root - SC+ (high soft conjunction, a "nice to have most"):** blends
↪  the three
  branches below. Soft andness means the overall match degrades
  ↪  gracefully rather
  than collapsing if one branch is imperfect.
- **Branch 1 - CP (product t-norm, "below the lowest"):** a strict
↪  product of
  **spotted back** AND **masked head** AND a blue block
  (**blue upper-tail** 0.52 > **blue underparts** 0.31 > **blue wing**
  ↪  0.17).
  This is the core "overall-blue songbird with a dark masked face"
  ↪  signature and
  is the single most diagnostic branch.
- **Branch 2 - CP (product t-norm):** re-uses **spotted back** together
↪  with a
  **white nape**, a corroborating plumage/structure check. (The *spotted
  ↪  back*
  leaf is the **same shared node** as in Branch 1 - the DAG reuses it, it
  ↪  is not a
  duplicate.)
- **Branch 3 - C (pure conjunction, "decided by the lowest"):** **brown
↪  crown** /
  **large size** / **white nape** - a compensatory profile the soft root
  ↪  can fall
  back on.

## Ornithological Interpretation
The dominant field marks the rule keys on - an overall **blue** bird
↪  (tail,
underparts, wing) with a **masked** face - are exactly the diagnostic
↪  marks of a
breeding male Indigo Bunting, so the rule is well grounded in real avian
↪  biology.

- **Well-matched:** the blue-plumage block + masked head is textbook male
↪  Indigo
  Bunting.
- **Sensible compensation:** the **brown crown** branch plausibly
↪  captures
  females / non-breeding / immature birds (which are brown rather than
  ↪  blue) - a
  reasonable alternative route for a soft conjunction to include.
- **Shared substructure:** **spotted back** appears in multiple branches
↪  but is a
  single reused node (fan-out less or euqal to 2), i.e. the model treats
  ↪  it as a shared
  sub-condition rather than re-learning it.
- **Flag:** **white nape** is comparatively weak and less diagnostic for
↪  this
  species; it may be incidental co-occurrence rather than a true field
  ↪  mark, and
  is worth checking if you audit the rule.
```

```
**Bottom line:** a faithful, concise (~13-literal) rule whose
↪  highest-weight
evidence aligns with the recognized diagnostic features of the Indigo
↪  Bunting.
```

C MNIST GRADE-LOGIC EXPRESSION DETAILS

Table 8 shows the full graded-logic expressions for the MNIST dataset

Table 8: Emergent digit rules (K=5, full structure accuracy 99.23%). Concepts: c_0 = junction; c_1 = upper rounded arc / hook; c_2 = U-turn (varied direction); c_3 = arc; c_4 = slant bar. $\bar{\wedge}$=Hyper conjunction (below lowest), $\wedge$=Pure conjunction (deciced by lowest), $\bar{\Delta}$=Hard conjunction (must have all), $\underline{\nabla}$=Soft disjunction(nice to have some), $\bar{\nabla}$=Hard disjunction (Enough to have any) $\vee$=Pure disjunction (decided by highest), $\bar{\vee}$=Hyper disjunction (above the highest)

,

DIGIT	GRADED-LOGIC EXPRESSION
0	$0.8(0.5(0.5(0.6(c_3)\bar{\wedge}0.4\,\overline{c_0})\bar{\wedge}0.5\,\overline{c_2})\bar{\wedge}0.5\,\overline{c_4}) \vee 0.2\,c_1$
1	$0.4(0.9(0.9(0.9(\overline{c_1})\underline{\nabla}0.1\,\overline{c_2})\underline{\nabla}0.1\,\overline{c_3})\bar{\nabla}0.1\,c_4)\bar{\wedge}0.6\,\overline{c_0}$
2	$0.8(0.8(0.7(0.2(\overline{c_0})\bar{\wedge}0.8\,c_4)\bar{\wedge}0.3\,c_2)\bar{\wedge}0.2\,c_1)\underline{\nabla}0.2\,c_3$
3	$0.7(0.9(0.8(0.5(\overline{c_1})\bar{\wedge}0.5\,\overline{c_4})\bar{\Delta}0.2\,c_3)\underline{\nabla}0.1\,\overline{c_0})\underline{\nabla}0.3\,\overline{c_2}$
4	$0.4(0.1(0.3(0.1(c_1)\bar{\wedge}0.9\,c_2)\bar{\vee}0.7\,c_3)\bar{\wedge}0.9\,c_0)\bar{\wedge}0.6\,c_4$
5	$0.7(0.6(0.3(0.5(\overline{c_1})\bar{\nabla}0.5\,c_3)\underline{\nabla}0.7\,\overline{c_4})\bar{\wedge}0.4\,c_2)\underline{\nabla}0.3\,c_0$
6	$0.6(0.7(0.6(0.9(\overline{c_2})\underline{\nabla}0.1\,c_4)\bar{\wedge}0.4\,\overline{c_1})\bar{\wedge}0.3\,\overline{c_0})\underline{\nabla}0.4\,\overline{c_3}$
7	$0.4(0.2(0.8(0.9(c_1)\bar{\wedge}0.1\,c_2)\bar{\Delta}0.2\,c_3)\bar{\Delta}0.8\,c_4)\bar{\wedge}0.6\,c_0$
8	$0.7(0.2(0.9(0.8(c_2)\underline{\nabla}0.2\,c_3)\underline{\nabla}0.1\,\overline{c_4})\bar{\nabla}0.8\,c_0)\bar{\wedge}0.3\,c_1$
9	$0.3(0.3(0.6(0.1(c_2)\bar{\vee}0.9\,c_1)\bar{\wedge}0.4\,c_4)\bar{\wedge}0.7\,c_3)\bar{\wedge}0.7\,c_0$

D QUANTITATIVE CONCEPT GROUNDING

We complement the qualitative grounding workflow (Fig. 3) with three quantitative *grounding-satisfaction* scores, each in $[0, 1]$, computed per concept on the hard-frozen K=5 MNIST OCBM (99.61% test accuracy). The scores measure three distinct and non-overlapping axes of grounding—activation coherence (G_c, itself split into an internal self-consistency G_{ci} and an external alignment G_{ce}), compositional structure (G_s), and cross-domain function (G_f)—and are fused into a per-concept satisfaction score z with the same soft graded-logic aggregator used by the model. We also report a *falsification control* that verifies the structural score rejects incorrect groundings rather than accepting any assignment.

Setup and notation. The encoder maps an image to K sigmoid concept activations $\mathbf{c}(x) = \sigma(f_\theta(x)) \in [0, 1]^K$, and each class $d \in \{0, \ldots, 9\}$ is scored by a frozen graded-logic tree $T_d : [0, 1]^K \to [0, 1]$. From tree d we read the *usage polarity* $\pi_{i,d} \in \{+1, -1\}$ of concept i (identity vs. negation of its leaf), and we write $\bar{c}_{i,d}$ for the mean activation of concept i over the test images of class d.

G_c: concept coherence (activation grounding). Coherence has two complementary sub-measurements: an *internal* score G_{ci}, which asks whether a concept's activation is consistent with how the model's own rules use it, and an *external* score G_{ce}, which asks whether the concept aligns with a distinct human primitive. The reported coherence is their arithmetic mean, $G_c^{(i)} = \frac{1}{2}\big(G_{ci}^{(i)} + G_{ce}^{(i)}\big)$.

Internal self-consistency (G_{ci}). A well-grounded concept fires *high* on the classes whose rules use it positively and *low* on those that negate it. Let $D_i^+ = \{d : \pi_{i,d}{=}{+1}\}$ and $D_i^- = \{d : \pi_{i,d}{=}{-1}\}$. We define the negation-aware coherence gap

$$G_{ci}^{(i)} \;=\; \max\Big(0,\; \frac{1}{|D_i^+|}\sum_{d\in D_i^+} \bar{c}_{i,d} \;-\; \frac{1}{|D_i^-|}\sum_{d\in D_i^-} \bar{c}_{i,d}\Big). \tag{2}$$

Because it reads the model's own logic tree (the leaf polarity π), it is defined *only* for the ontology head; a value near 0 flags a concept that fires irrespective of how the rules use it.

External alignment (G_{ce}). Independently of the model, we test whether each concept behaves like a distinct human primitive. Given a vocabulary of human strokes $\{h_j\}$ (the per-digit stroke truth tables), we form the direction-agnostic identification AUC $a_{ij} = \max\big(\mathrm{AUC}(c_i, h_j),\, 1 - \mathrm{AUC}(c_i, h_j)\big) \in [\frac{1}{2}, 1]$, match concepts to *distinct* strokes with a one-to-one (Hungarian) assignment M maximising $\sum_i a_{i,M(i)}$ (no stroke reused), and rescale the matched AUC to $[0, 1]$,

$$G_{ce}^{(i)} = \max\big(0,\; 2\,a_{i,M(i)} - 1\big). \tag{3}$$

This uses only concept activations and stroke labels, so it is *head-agnostic*: it is defined for black-box (MLP, linear) heads too, which is what makes a cross-head grounding comparison possible. It is the bijective alignment (bijAUC) used in the head comparison below.

G_s: structural grounding. We test whether a concept's proposed meaning is consistent with the frozen logic. Probing each tree at the neutral point $\mathbf{c}^0 = \frac{1}{2}\mathbf{1}$ gives the signed structural influence

$$S_{i,d} = T_d(\mathbf{c}^0 \mid c_i{=}1) - T_d(\mathbf{c}^0 \mid c_i{=}0), \qquad \sigma_{i,d} = \mathrm{sign}(S_{i,d})\mathbb{1}[\,|S_{i,d}| > \epsilon\,]. \tag{4}$$

Given the proposer's world-knowledge membership—classes P_i where the concept should be present and A_i where it should be absent—we define a consistency and an evidence support,

$$\mathrm{cons}_i = \frac{|\{d \in P_i : \sigma_{i,d}{=}{+}1\}| + |\{d \in A_i : \sigma_{i,d}{=}{-}1\}|}{|\{d \in P_i \cup A_i : \sigma_{i,d} \neq 0\}|}, \qquad s_i = |\{d \in P_i \cup A_i : \sigma_{i,d} \neq 0\}|, \tag{5}$$

and combine them so that a perfect consistency on too few classes is not mistaken for solid grounding:

$$G_s^{(i)} = \mathrm{cons}_i \cdot \frac{\min(s_i,\, s_{\max})}{s_{\max}}, \qquad s_{\max} = 5. \tag{6}$$

G_f: functional grounding. As a model-level functional test we measure zero-shot transfer of the whole reasoning system to the external USPS digit corpus, applying only an unsupervised per-concept quantile calibration of the concept activations (no USPS labels). G_f is the resulting calibrated top-1 accuracy.

Fusion and acceptance. Grounding is itself a graded-logic aggregation, so we combine the pillars with the same GCD operators the model uses, respecting their roles. *Per concept* (within a single model) coherence and structure are both *mandatory*: a concept must be both activation-coherent and structurally consistent, so we combine them with a *hard conjunction* ("must have all"), realized as the weighted geometric mean,

$$z_i = \bar{\Delta}\big(G_c^{(i)}, G_s^{(i)}\big) = \sqrt{G_c^{(i)}\, G_s^{(i)}}. \tag{7}$$

At the model level we score grounding so that it is also comparable across head types (Sec. "Cross-head comparison"). The two *head-agnostic* axes—coherence G_c and function G_f—form the *mandatory* core (a grounded model must be both activation-coherent and functionally transferable), again a hard conjunction. The structural score G_s requires the logic tree, so it is available only to the ontology; we therefore fold it in as a *non-annihilating bonus* via a soft conjunction (andness ≈ 0.75, the power mean with exponent $r{=}\frac{1}{2}$), so that a head lacking structure ($G_s{=}0$) still scores its coherence-and-function core while the ontology is rewarded for the structure it uniquely earns:

$$Z = \Delta_{0.65,0.35}\Big(\bar{\Delta}\big(\overline{G_c},\, G_f\big),\, \overline{G_s}\Big), \qquad \Delta_{w_1,w_2}(a,b) = \big(w_1\, a^{1/2} + w_2\, b^{1/2}\big)^2. \tag{8}$$

The outer weights $(0.65, 0.35)$ correct for the hierarchical position of G_s: because it enters one level above the $G_c \wedge G_f$ core, equal weights would over-weight it, whereas the $0.65/0.35$ split gives the three criteria near-equal final influence ($\approx 0.325, 0.325, 0.35$). Unlike a pure conjunction, this never sends Z to zero when the bonus is absent. A concept (or model) is accepted when its score exceeds a threshold τ. We use $\epsilon = 0.05$.

Results. Table 9 reports the per-concept scores. The slant-bar (c_4) and junction (c_0) concepts are the best grounded on every measure. The diffuse c_2 ("U-turn") is the weakest overall ($z{=}0.41$), and the two coherence sub-measurements pinpoint *why*: its external alignment is respectable ($G_{ce}{=}0.74$—it does resemble a nameable stroke), but its internal self-consistency is near zero ($G_{ci}{=}0.10$) and its structural score is low ($G_s{=}0.40$). In other words c_2 *looks* like a primitive but the model does not *use* it coherently—a distinction the split coherence makes visible. The model-level functional score is $G_f = 0.74$ (USPS-calibrated; in-domain MNIST 99.6%), giving an overall grounding score $Z = \Delta_{0.65,0.35}\big(\bar{\Delta}(0.65, 0.74), 0.76\big) = \Delta_{0.65,0.35}(0.69, 0.76) = 0.72$. Collecting the three axes at the model level,

$$\overline{G_c} = 0.65\ \big(\overline{G_{ci}} = 0.49,\ \overline{G_{ce}} = 0.80\big), \qquad G_f = 0.74, \qquad \overline{G_s} = 0.76 \quad\Longrightarrow\quad Z = 0.72. \tag{9}$$

Table 9: Per-concept grounding scores for the $K{=}5$ MNIST OCBM. G_{ci}: internal self-consistency (polarity-aware gap, ontology-only); G_{ce}: external alignment (bijective concept–stroke AUC, head-agnostic); $G_c = \frac{1}{2}(G_{ci}{+}G_{ce})$; G_s: structural consistency×support; $z = \sqrt{G_c\, G_s}$. c_2 aligns externally but is weakest internally and structurally. Model-level $G_f = 0.74$ (USPS-calibrated transfer, in-domain MNIST 99.6%).

	concept	G_{ci}	G_{ce}	G_c	G_s	z
c_0	junction	0.60	0.98	0.79	1.00	0.89
c_1	upper rounded arc/hook	0.59	0.74	0.67	0.80	0.73
c_2	U-turn (varied dir.)	0.10	0.74	0.42	0.40	0.41
c_3	arc	0.35	0.54	0.45	0.60	0.52
c_4	slant bar	0.80	1.00	0.90	1.00	0.95
	mean	0.49	0.80	0.65	0.76	0.70

Table 10: Cross-head grounding. G_{ci} (internal) and G_s (structural) require the logic tree, so they are 0 / "–" for black-box heads—the ontology earns a certificate they cannot receive. G_{ce}: external alignment (bijAUC, rescaled); G_c: reported coherence; G_f: USPS-calibrated transfer, 5-seed mean ± std (the ontology's own single $K{=}5$ audit model scores 0.74, Table 9). The black-box G_{ce} (and hence G_c) are the 5-seed bijAUC; the ontology's coherence is from the $K{=}5$ model (its mean G_{ce} equals the 5-seed bijAUC). $Z = \Delta_{0.65,0.35}(\bar{\Delta}(G_c, G_f), G_s)$ is the model-level grounding score (black-box $G_s{=}0$).

head	G_{ci}	G_{ce}	G_c	G_s	G_f	Z
ontology	0.49	0.80	0.65	0.76	$0.81{\pm}0.04$	**0.74**
MLP	0.00	0.78	0.39	–	$0.73{\pm}0.09$	0.23
linear	0.00	0.68	0.34	–	$0.69{\pm}0.10$	0.20

Falsification control. A grounding score is only meaningful if it *rejects* wrong meanings. We re-score G_s under all $K!{-}1$ non-identity permutations of the concept→membership assignment (i.e. giving each concept a different concept's world-knowledge membership). Consistency drops from 1.00 to 0.85 and the mean structural score from $\overline{G_s} = 0.76$ to 0.50, confirming that the frozen structure discriminates correct groundings from permuted ones rather than accepting any assignment.

Cross-head comparison. Because G_{ce} and G_f need only concept activations—no logic tree—they are defined for any head, which lets us compare the ontology against black-box MLP and linear heads on the same axes. The internal self-consistency G_{ci} and the structural score G_s are intrinsically model-specific (they read the logic tree), so a black-box head scores $G_{ci}{=}0$ and admits no G_s: the ontology earns a grounding certificate the black-box heads structurally cannot receive. Table 10 reports the head comparison. On the coherence axes the ontology is clearly best—its external alignment G_{ce} is highest and, with the internal pillar added, its G_c nearly doubles the black-box heads'—and it uniquely earns the internal (G_{ci}) and structural (G_s) pillars. For functional transfer we report the authoritative 5-seed calibrated-USPS accuracy (transfer is high-variance, so a single run is unreliable): the ontology is both highest and most stable ($0.81{\pm}0.04$ vs. MLP $0.73{\pm}0.09$ vs. linear $0.69{\pm}0.10$). The ontology therefore leads on every comparable axis. The composite model-level score is $Z = 0.74$ for the ontology versus 0.23 (MLP) and 0.20 (linear): the black-box heads earn a nonzero coherence-and-function core but forfeit the structural bonus that only the logic head can provide.

On a fourth (evidence) pillar. We also investigated a pixel-attribution "evidence" pillar (Grad-CAM insertion–deletion, template prediction, and within-digit firing consistency). None proved a reliable grounding signal on MNIST: attribution measures spatial *localization* (compact features such as an arc score high, distributed features such as a junction score low), which is orthogonal to grounding quality; and metrics computed on top activators or within a single digit are structurally blind to c_2's failure mode, which is that it fires across *too many* classes—a cross-class property already captured by G_c and G_s. We therefore report attribution only as a descriptive property and restrict the grounding score to the three pillars above.