OPERATORS GATE GRADIENTS: ROUTING, SURVIVAL, AND REASONING-SHORTCUT CONTROL IN DIFFERENTIABLE LOGIC

Anonymous authors
Paper under double-blind review

ABSTRACT

A logical constraint can make a network learn less than it would without it: on MNIST digit-sum with the Gödel or Goguen implication, concept accuracy stays near 20% (chance 10%) through $30{,}000$ labels, below the 33% supervision alone reaches with 10. We trace this to the operator that combines sub-predictions, usually chosen by convention. It is credit assignment on two axes: gradient must be *routed* to a parameter, as the computation graph dictates, and *survive*, staying nonzero where training operates, as the objective and grounding dictate. No operator is universally best, but the best is predictable. Fuzzy description logic makes this exact: one implication gates every axiom through closed-form partials, so the chain rule predicts where each operator acts. Across $212{,}805$ runs, the implication explains 73–81% of operator-induced variance within structure, against 5–6% for the t-norm, and the effect persists under fixed-evaluator scoring; the Łukasiewicz implication is the safe default for pure satisfaction until, as predicted, the objective or grounding changes. Beyond logic, these axes predict seven of eight aggregation tasks, the one miss being prospective. The paradox is a survival failure. Gödel and Goguen *saturate*, going flat on satisfied constraints, so a wrong-concept solution satisfying them, a *reasoning shortcut*, is a zero-gradient optimum, walled off from supervision by the violated-constraint loss. Swapping in a non-saturated implication lifts the barrier without extra labels, matching supervision by 300 to $1{,}200$: the shortcut is a property of the loss, not a shortage of labels.

1 INTRODUCTION

van Krieken et al. (2022) observed that on semi-supervised MNIST the Gödel and Goguen implications train below the supervised baseline. We find that labels do not close this gap. On the digit-sum task of the same benchmark (their $\mathrm{sum}{=}9$ problem), these implications hold concept accuracy near 20% through $30{,}000$ labeled digits, while a non-saturated implication trained on the same data catches up with supervision (§ 5). Since only the implication differs between these runs, the difference lies not in the constraint but in the operator that turns it into a gradient.

Many systems train through differentiable relaxations of discrete or compositional structure, and in all of them the forward relaxation is also a backward credit-assignment mechanism, deciding where gradient flows and where it vanishes, the compositional analog of the vanishing-gradient problem (Bengio et al., 1994). Probabilistic methods propagate exact gradients through weighted model counts (Manhaeve et al., 2018; Xu et al., 2018; Winters et al., 2022), at the cost of #P-hard inference (Maene et al., 2024) that motivates approximate variants (van Krieken et al., 2023); optimization-based methods embed satisfiability into differentiable solvers (Wang et al., 2019), straight-through estimators (Yang et al., 2022), or hinge-loss random fields (Bach et al., 2017); fuzzy relaxations ground each connective through a continuous operator from many-valued logic (Badreddine et al., 2022; Diligenti et al., 2017; van Krieken et al., 2022; Marra et al., 2019; Badreddine et al., 2023). Fuzzy relaxations make the choice explicit: a conjunction may take any t-norm (Klement et al., 2000), an implication a t-norm's residuum or an S-implication. Practitioners choose once and move on.

Fuzzy description logic shows why, because its computation graph is written axiom by axiom. In fuzzy $\mathcal{ALC}$ (Straccia, 2001), every general concept inclusion $C \sqsubseteq D$ is grounded as $I(\|C\|, \|D\|)$,

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.

where $\|\cdot\| \in [0,1]$ is the truth degree a network assigns to a concept, so the implication I is the only differentiable path between the two sides and its partials $\partial_1 I$, $\partial_2 I$ act as gates. A gate can fail in two ways. Structure can close it: the Gödel implication (Göd-R) has $\partial_1 I = 0$ identically, so the antecedent receives no gradient even on violated axioms, whereas Łukasiewicz (Łuk-R) keeps $\partial_1 I = -1$ wherever the axiom is violated. No learning rate schedule can route gradient through a path whose derivative is identically zero (for hard operators; our smoothed surrogates keep this contrast on most structures, App. I.3). Training can also close it: an implication *saturates* when it is constant once an axiom is satisfied, so the logic loss has zero gradient at every satisfying assignment, including a wrong one. The gap above begins this second way; § 5 shows what then holds it.

The literature has reached opposite verdicts on these gates. van Krieken et al. (2022) first analyzed fuzzy-operator gradients and, because Łuk-R and Göd-R differ only in the antecedent derivative yet rank among the better and the worst, suggested that this derivative is required. He et al. (2024) treat the same antecedent gradient, in the Reichenbach implication, as an *implication bias* that can lead to shortcuts and should be reduced. Daniele & van Krieken (2025) show that gradient descent on Gödel logic is a discrete local search for Boolean satisfiability and is competitive on visual Sudoku. Others find that theory and practice disagree on t-norms (Grespan et al., 2021), that the ranking depends on the task (Flinkow et al., 2025), and that fuzzy logics under-perform logics whose connectives are strictly increasing (Flinkow et al., 2026). Inside fuzzy $\mathcal{ALC}$ the choice is fixed globally: Logic Tensor Networks (LTN) adopt product with Reichenbach as their best-performing configuration (Badreddine et al., 2022), FALCON treats the operators as hyperparameters yet uses the product t-norm throughout (Tang et al., 2022), and Em$\mathcal{ALC}$ fixes the product t-norm (Wu & Zhao, 2025). Each work asks which operator is best; none asks what the answer depends on.

A single principle: routing and survival. We show that these verdicts are facets of one mechanism: credit assignment along two axes, *routing*, which structure fixes, and *survival*, which the objective and the grounding fix. The single-operator gradient properties that van Krieken et al. (2022) cataloged are, in this view, not intrinsic but conditional. The routing half turns their suggestion into theorems with a quantity and a scope: along the antecedent-gradient direction, the first-order descent of Łuk-R exceeds that of Göd-R by the squared norm of that gradient where structure routes roles through the antecedent (left-quantified axioms such as $\exists R.C \sqsubseteq D$), and the two gradients coincide where it does not (Thms. 2–3). It also explains the conflict: the antecedent gradient that He et al. (2024) reduce is the source of Łuk-R's advantage wherever structure routes roles through it, and saturation, harmless when any satisfying assignment is a solution, holds a network at a wrong one when concepts must be learned. Optimality is thus a function of structure $\times$ objective $\times$ grounding.

Across $212{,}805$ training runs on 405 real-world ontologies (BioPortal and the OWL Reasoner Evaluation corpus ORE 2015) plus synthetic controls, the implication accounts for 73–81% of operator-induced variance once structure is held fixed (η^2), an order of magnitude above the t-norm, and the top-ranked implication is unchanged in 27 of 28 structure-by-loss comparisons that include the reduced implication-bias (RILL) losses of He et al. (2024) (App. F.2). Łuk-R is a safe default only for pure satisfaction on corpora of mostly plain inclusions (Prop. 5); under a learning objective two of the three saturated implications degrade, as predicted.

What the principle yields. Survival also controls *reasoning shortcuts*, where a model satisfies its constraints while learning the wrong concepts (§ 5). Saturation is necessary but not sufficient for a shortcut: Łuk-R also saturates, yet from $1{,}200$ labels on, 4 to 7 of 10 seeds escape, because what holds a shortcut is the loss on violated constraints, which in an idealized model is a steep wall for Goguen and a discontinuous jump for Gödel. The same analysis predicts behavior on pooling, fusion, and attention in seven of eight aggregation tasks (§ 4.5); the one miss was a prospective prediction.

Our contributions are: (i) the routing $\times$ survival principle, with theorems that quantify the antecedent gradient's advantage exactly where structure routes roles through it and prove it inert elsewhere, plus saturation and grounding results; (ii) operator-level control of reasoning shortcuts: saturation predicts which implications admit a shortcut, a zero-extra-label swap removes the barrier, and an idealized model orders five implications by resistance to supervision, with a pre-registered intervention confirming Gödel's hard branch as the cause; and (iii) a $212{,}805$-run validation with fixed-evaluator controls and geometric, LLM, and LTN baselines.

We do not claim a universal law of ML aggregation, only one principle: proved where routing is mathematically inspectable, and tested where it is not.

2 PRELIMINARIES

2.1 AXIOM NORMALIZATION

We study fuzzy relaxations of $\mathcal{ALC}$ (Schmidt-Schauß & Smolka, 1991), whose TBoxes are finite sets of general concept inclusions (GCIs) $C \sqsubseteq D$ (formal syntax in App. A). We adopt the seven normal forms of Wu & Zhao (2025), whose procedure rewrites any $\mathcal{ALC}$ TBox in polynomial time and preserves models on the original signature (their Lem. 1). Negation needs no separate form: normalization puts concepts in negation normal form and a negated atom is grounded as $1 - \|C\|_i$ (App. A.2). Model preservation is what makes the propositions of § 3, stated over normalized axioms, statements about the original ontology. Every axiom is then grounded as $I(u_i, v_i)$ for each individual coordinate i (Table 1). Two structural facts follow from the table and carry the analysis of § 3: the implication I is the only operator every axiom passes through, and the t-norm T and t-conorm S appear only inside the antecedent u_i and consequent v_i.

Table 1: **Normalized axiom forms and their fuzzy grounding.** Each form grounds $C \sqsubseteq D$ as $I(u_i, v_i)$, where I is a fuzzy implication, T a t-norm (fuzzy conjunction), S a t-conorm (fuzzy disjunction), u_i and v_i the antecedent and consequent truth values at individual i, and $\max_j$, $\min_j$ range over all d individuals; I always sits at the top level, while T and S build the bodies.

Form	Axiom form	(u_i, v_i)
1	$C \sqsubseteq D$	$(\mathbf{c}_i,\ \mathbf{d}_i)$
2	$C_1 \sqcap C_2 \sqsubseteq D$	$(T(\mathbf{c}_{1,i}, \mathbf{c}_{2,i}),\ \mathbf{d}_i)$
3	$C \sqsubseteq D_1 \sqcup D_2$	$(\mathbf{c}_i,\ S(\mathbf{d}_{1,i}, \mathbf{d}_{2,i}))$
4	$C \sqsubseteq \exists R.D$	$(\mathbf{c}_i,\ \max_j T(\mathbf{R}_{ij}, \mathbf{d}_j))$
5	$C \sqsubseteq \forall R.D$	$(\mathbf{c}_i,\ \min_j S(1{-}\mathbf{R}_{ij}, \mathbf{d}_j))$
6	$\exists R.C \sqsubseteq D$	$(\max_j T(\mathbf{R}_{ij}, \mathbf{c}_j),\ \mathbf{d}_i)$
7	$\forall R.C \sqsubseteq D$	$(\min_j S(1{-}\mathbf{R}_{ij}, \mathbf{c}_j),\ \mathbf{d}_i)$

2.2 OPERATORS AND THEIR DERIVATIVES

We use three standard t-norms (Gödel, product, Łukasiewicz) and the three matching t-conorms, crossed independently with each other and with five implications spanning R- and S-implication families (Hájek, 1998), giving a $3 \times 3 \times 5$ operator grid. The suffix -R marks a *residuated* implication, built from a t-norm; Reichenbach and Kleene–Dienes are S-implications, built from a t-conorm. Writing u for the antecedent value and v for the consequent value, every R-implication equals 1 once $v \geq u$, which drives the survival results of § 3.2. The only operator facts needed in the main text are the implication partials on the *active region* $\{u > v\}$ (Table 2). On that region Łuk-R is the only implication with constant nonzero partials on both branches (antecedent and consequent), while Göd-R blocks the antecedent identically; KD routes to exactly one branch at a time. These partials are not new: van Krieken et al. (2022) tabulate all five (their §8.3–8.5) and name Göd-R's blocked antecedent and KD's one-branch routing *vanishing* and *single-passing* (their §5). We take the table as given and ask when each property binds (full operator catalog, including SP and Sig of § 4.3: App. A.3).

Table 2: **Implication partials for $u > v$.**

Name	$I(u, v)$	$\partial_1 I$	$\partial_2 I$
Łukasiewicz (Łuk-R)	$\min(1,\ 1{-}u{+}v)$	-1	$+1$
Reichenbach (Rei)	$1{-}u{+}uv$	$-(1{-}v)$	u
Goguen (Gog-R)	v/u if $u > v$; 1 otherwise	$-v/u^2$	$1/u$
Gödel (Göd-R)	v if $u > v$; 1 otherwise	0	$+1$
Kleene–Dienes (KD)	$\max(1{-}u,\ v)$	-1 or 0	0 or $+1$

2.3 GROUNDING AND LOSS

The differentiable fuzzy $\mathcal{ALC}$ (DF$\mathcal{ALC}$) pipeline assigns each concept a vector $\mathbf{c} \in [0,1]^d$ and each role a matrix $\mathbf{R} \in [0,1]^{d \times d}$, with entries projected after each optimizer step (App. A.2). Here d is the number of individuals in the interpretation domain: coordinate $\mathbf{c}_i$ is the degree to which individual i belongs to the concept, and $\mathbf{R}_{ij}$ the degree to which i is R-related to j. For a concept name the embedding is its truth vector, $\mathbf{c}_i = \|C\|_i$; complex concepts inherit their truth values from T, S, and the quantifier aggregations of Table 1 (App. A.2). For axiom α, Table 1 defines antecedent and consequent truth values $u_{\alpha,i}(\theta)$ and $v_{\alpha,i}(\theta)$, where θ collects all concept and role embeddings. The *subsumption loss* for axiom α (superscript H for hierarchy) is

$$\ell_\alpha^H(\theta) \;=\; \sum_{i=1}^{d} \big[1 - I\big(u_{\alpha,i}(\theta),\ v_{\alpha,i}(\theta)\big)\big], \tag{1}$$

and $L(\theta) = \sum_\alpha \ell_\alpha^H(\theta)$ up to the minibatch averaging used in implementation; we write L^I when the implication I needs naming, as in $L^{\text{Łuk}}$ and $L^{\text{Göd}}$, and abbreviate $u_{\alpha,i}, v_{\alpha,i}$ to u_i, v_i when α is clear. § 4.4 repeats this default objective under four alternatives: a *hybrid* loss that adds a margin term $\max(0, u_i - v_i)$ to ℓ_α^H, and three RILL variants in the style of He et al. (2024) (App. F.2).

3 How Operators Gate Gradients

Our analysis rests on a single principle:

Operator o helps a parameter when two conditions hold together: the parameter lies on a loss path through o with a nonzero partial (routing), and that partial is nonzero where the task drives the groundings (survival).

Routing is fixed by ontology *structure* (which operator sits on the path to a parameter, set by the axiom form of Table 1); **survival** by the training *objective* and the *grounding*. The principle is restated for each of the three ways an operator can pass gradient (smooth, subgradient at kinks, straight-through surrogates) and is corroborated on ML pooling, fusion, and attention (App. C); fuzzy $\mathcal{ALC}$ is the controlled instance where routing is explicit and the theory sharpest. All results are for hard (non-smoothed) operators; training uses softmax/sigmoid surrogates at smoothing temperature τ=0.1, and all reported satisfaction rates (SR) and rankings rescore the learned embeddings with the hard operators at threshold 0.5. App. I.3 checks this over $\tau \in [0.01, 1]$: routing and hard-rescored rankings are stable, soft-loss rankings are not. Full proofs appear in App. B. We call a parameter vector θ *regular* if no max/min selector in the truth-value computation has a tie, and the implication I is differentiable at every grounded pair $(u_i(\theta), v_i(\theta))$ (a generic condition, violated on a measure-zero set), and say that a grounded pair (α, i) is *active violated* at θ if $I(u_i, v_i) < 1$.

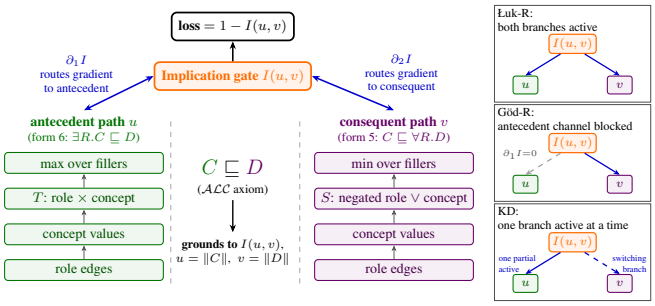


Figure 1: **Operator choice is gradient routing.** Schematic: the antecedent body of a left-existential form (left, form 6) and the consequent body of a right-universal form (right, form 5) are drawn feeding one top-level implication $I(u, v)$; in every normalized form, all gradient reaching a body passes through $\partial_1 I$ or $\partial_2 I$. Insets (violated pairs): Łuk-R keeps both branches active, Göd-R blocks the antecedent channel, KD routes one branch at a time.

3.1 Routing: the structure axis

Because T and S appear only inside u and v while I sits at the top level, implication is the only operator that directly gates the loss gradient (Fig. 1), a chain-rule identity:

Proposition 1 (Implication-gated gradient decomposition)**.** *For every normalized axiom α and every regular parameter vector θ,*

$$\nabla_\theta \ell_\alpha^H \;=\; -\sum_i \Big[\partial_1 I(u_i, v_i)\, \nabla_\theta u_i \;+\; \partial_2 I(u_i, v_i)\, \nabla_\theta v_i\Big]. \tag{2}$$

For a form-1 axiom $C \sqsubseteq D$, $u_i = \mathbf{c}_i$ and $v_i = \mathbf{d}_i$, so $\ell_\alpha^H = \sum_i [1 - I(\mathbf{c}_i, \mathbf{d}_i)]$ and both the loss and (2) are independent of T, S, and all role embeddings.

Thus $\partial_1 I, \partial_2 I$ gate every antecedent/consequent path (Fig. 1), and which paths stay open is implication-specific: the residuated family (Łuk-R, Göd-R, Gog-R) is constant on the satisfied region $v \geq u$, so both partials vanish there, whereas the S-implications (Rei, KD) keep a nonzero partial there. Which parameters those paths reach is fixed by the axiom form. For role parameters the Łuk-R/Göd-R gap is therefore structure-conditioned:

Theorem 2 (Łuk-R/Göd-R role-gradient equivalence under right quantification)**.** *On any TBox containing only forms $C \sqsubseteq \exists R.D$ and $C \sqsubseteq \forall R.D$, at regular θ the gradients with respect to the role embeddings coincide:*

$$\nabla_{\mathbf{R}} L^{\text{Łuk}}(\theta) \;=\; \nabla_{\mathbf{R}} L^{\text{Göd}}(\theta).$$

Right-quantified forms place roles only in v_i, eliminating the $\partial_1 I$ channel where Łuk-R and Göd-R differ. Left-quantified forms place roles in u_i, so Łuk-R opens a channel that Göd-R blocks:

Theorem 3 (Exact descent gap under left quantification)**.** *Let $\mathcal{A}$ be the set of active violated left-quantified pairs (forms $\exists R.C \sqsubseteq D$ and $\forall R.C \sqsubseteq D$) and let $g^L = \sum_{i \in \mathcal{A}} w_i \nabla_{\mathbf{R}} u_i$ be the weighted*

aggregate of their antecedent gradients with respect to the role embeddings $\mathbf{R}$*, with per-pair loss weights* $w_i > 0$ *(*w_i*=1 for the unweighted loss). At every regular* θ *with* $g^L \neq 0$*, the directional derivatives* $D_\Delta L := \langle \nabla_{\mathbf{R}} L, \Delta \rangle$ *along* $\Delta = -g^L$ *satisfy* $D_\Delta L^{\text{Łuk}}(\theta) = D_\Delta L^{\text{Göd}}(\theta) - \|g^L\|^2$*, so Łuk-R descends strictly faster than Göd-R. (App. B.4 gives the proof and the general form for any role subspace with its projection.)*

The comparison is pointwise and first-order: it concerns descent at a fixed θ, not final loss or convergence. The SR comparisons of § 4.4 are therefore descriptive only; the theorem is tested on gradients at initialization (App. I.2). Under the positive-derivative bounds of Cor. 11, a single active violated left-quantified pair suffices to make g^L nonzero, so Thm. 3's hypothesis is structural, and its lower bound grows linearly in their number. The algebra here is immediate from Prop. 1 once the partials are fixed; the content is in the structural side conditions.

From routing to testable predictions. Thms. 2–3 partition the landscape by structure; the remaining implications are separated by their partials alone. The sharpest case is KD, which activates exactly one top-level branch away from the switching surface $1-u = v$ (Cor. 12). KD therefore starves one branch per pair; the hybrid loss's margin term changes magnitudes, not which branch is reachable. Prop. 1 turns van Krieken et al. (2022)'s *single-passing* label from an intrinsic operator property into a statement about which parameters an axiom can reach. Two further derivative-level predictions follow the same way. Because Rei's antecedent derivative shrinks as the consequent is satisfied, Rei should overtake Łuk-R on chains of right-existential axioms, which App. I.1 confirms; and Gog-R should be unstable near $u \to 0$. The KD prediction and the Gog-R instability are confirmed by direct gradient hooks (App. L.1, with a worked two-axiom trace in App. A.4).

3.2 Survival: the objective and grounding axes

Routing decides which parameters a gradient can reach; survival decides whether that gradient is nonzero where training operates. Two regularities govern survival: the objective can drive examples into the saturated satisfied region, and the grounding can pin inputs inside an operator's interior dead zone. We call an implication *saturated* if $I(u, v) = 1$ whenever $v \geq u$; every residuated implication of a left-continuous t-norm qualifies, including hard Łuk-R, and a saturated implication has zero gradient at every regular pair with $v_i(\theta) > u_i(\theta)$ (van Krieken et al., 2022).

Proposition 4 (Vanishing logic gradient on satisfied pairs)**.** *Let* I *be saturated. At every regular* θ*, each pair with* $v_i(\theta) > u_i(\theta)$ *contributes zero gradient to* ℓ^H_α*: such pairs therefore drive no update under a satisfaction objective and supply no logic gradient to any downstream learning objective. A non-saturated implication such as Rei, with* $\partial_2 I = u$*, retains a nonzero gradient on the same pairs whenever* $u > 0$*; so does KD, on one branch at a time.*

The cost of saturation is application-dependent (van Krieken et al., 2022) and has motivated logics engineered to avoid it (Flinkow et al., 2026); here it is not uniformly a defect: under a *satisfaction* objective it is exactly what makes a saturated implication a safe default (Prop. 5), while under a *learning* objective it is the mechanism behind reasoning shortcuts (Prop. 7).

Proposition 5 (Break-even precision threshold)**.** *For a nontrivial policy that switches ontologies off the Łuk-R default (*$\beta, r, g > 0$*), with base rate* β *of profitable switches, recall* r*, mean gain* g*, mean harm* h*, and precision* p*, the expected satisfaction-rate improvement over the Łuk-R default is* $\beta\, r\, g \left(1 - \frac{h}{g} \cdot \frac{1-p}{p}\right)$*, positive iff* $p > p^* := k/(k{+}1)$ *where* $k = h/g$*.*

The Łukasiewicz t-norm on two truth values a, b, $T_L(a, b) = \max(0, a + b - 1)$, has $\nabla T_L = 0$ on half its domain, $\{a + b < 1\}$ (Evans & Grefenstette, 2018; van Krieken et al., 2022), and a randomly initialized network's failure to learn has been traced to that region (Grespan et al., 2021). What the grounding decides is whether training starts inside it.

Proposition 6 (Grounding dependence of the Łukasiewicz t-norm dead zone)**.** *Under a softmax grounding over* $K \geq 3$ *classes with class probabilities* π_i*, the uniform point has* $\pi_i + \pi_j = 2/K < 1$*, so body t-norms formed from two class coordinates start in the dead zone* $\{a + b < 1\}$*. Under free coordinatewise groundings in* $[0, 1]$*, the escape set* $\{a + b \geq 1\}$ *has positive measure (Lebesgue measure* $1/2$ *in* $[0, 1]^2$*), so the dead zone is not geometrically forced.*

That condition already has an unexplained instance in the record: van Krieken et al.'s $\mathrm{sum}{=}9$ task grounds a body t-norm over two softmax class coordinates with $K{=}10$, exactly the regime the proposition pins, and they report Łukasiewicz derivatives that vanish for the whole run.

Proposition 7 (Reasoning shortcuts as zero-gradient global minimizers)**.** *Let I be saturated (Prop. 4) and consider the logic loss $\sum_{\alpha,i}[1 - I(u_{\alpha,i}, v_{\alpha,i})]$. Every θ at which all grounded instances are satisfied with slack ($v_{\alpha,i} > u_{\alpha,i}$) is a zero-gradient global minimizer of this loss; in particular so is every constraint-satisfying but concept-incorrect assignment, i.e. a reasoning shortcut. A non-saturated I retains a nonzero gradient at such θ.*

The logic loss therefore supplies no gradient to prefer the true concepts; only supervision does. Saturation is necessary for a shortcut to be a trap, not sufficient (§ 5); the proof is in App. O.

Survival beyond logic: a sparsity threshold. On a sparse-signal aggregation model the detection-optimal readout crosses from averaging to max at a sparsity boundary, and the fuzzy dead zone (Prop. 6) is the extreme of that same axis (Prop. 14; confirmed on real CIFAR scores, App. C).

Predicting when operator choice flips. The two halves compose into a falsifiable prediction (the regime map, Fig. 25): a large operator reversal requires all three of (a) a body structure that routes the operator (§ 3.1), (b) a learning rather than satisfaction objective (Prop. 4), and (c) a grounding that pins inputs low (Prop. 6); Prop. 5 adds the satisfaction-regime default. § 4.3 confirms the reversal where all three hold (classifier-grounded MNIST digit-sum) and its near-absence where the grounding is free (real ontologies), as predicted. In DF$\mathcal{ALC}$ the implication inside $\forall R.D$ is fixed to $\neg R \sqcup D$ and only the top-level inclusion uses the swept implication; the survival experiments instead write $\forall$ as a product of implication values over pairs, so the swept implication sits inside the quantifier (App. K).

4 Experiments

4.1 Setup

We implement the DF$\mathcal{ALC}$ pipeline of § 2.3 over the $3 \times 3 \times 5 = 45$ operator grid, which the BioPortal corpus uses as is and the synthetic and ORE 2015 corpora extend (below); together they form the main 212,805-run archive; optimizer, dimension, seed, learning-rate and epoch budgets are in App. E. Two downstream tasks probe the learned embeddings: D2 holds out 20% of each ontology's axioms and asks the trained embeddings to rank each held-out axiom above 10 copies whose consequent concept is replaced at random (mean reciprocal rank, MRR, and AUC-ROC); D4 does the same for role-bearing axioms but replaces the role. D2 uses its own BioPortal slice: all 308 parseable ontologies, with no axiom cap (the baselines of App. J use the 291 with at least 20 axioms; Apps. G.1, H.1). The synthetic corpus varies 10 structure dimensions one-factor-at-a-time (the columns of Fig. 2), giving 35 settings, crossed with a 162-configuration extended grid and five repetitions, and 28,350 deduplicated runs (the paper-core 45-cell subset is 7,875 runs); the real corpora are 175 BioPortal ontologies (Whetzel et al., 2011) (10–1000 axioms, 39,375 runs) and 230 ORE 2015 ontologies (Parsia et al., 2017) (145,080 runs, smaller but richer in role structure; App. G.1); that count includes operator controls beyond the 45-cell core, whose deduplicated size is $51{,}750$. A third corpus, $\mathcal{EL}$-real (260), is used only for the fragment-restriction checks: the 149 $\mathcal{EL}$ ontologies among the 175 plus 111 further ones outside the 10–1000-axiom window, so it is disjoint from ORE and not independent of BioPortal. The primary metric is satisfaction rate (SR: fraction of TBox axioms whose implication values $I(u_i, v_i)$, aggregated over individuals by the t-norm, reach 0.5); variance is decomposed by operator role via η^2 computed separately for implication, t-norm, and t-conorm. Mean truth degree, the secondary metric, averages over TBox axioms the same aggregated truth value that SR thresholds at 0.5.

4.2 Routing: implication dominance and when the t-norm matters

Fig. 2 gives the variance decomposition: implication dominates on every structure dimension. Conditioning on the structural setting (removing the between-structure confound), the implication's *within-structure* η^2 is 0.73–0.81 across all three corpora (synthetic, BioPortal, ORE 2015), an order of magnitude above the t-norm (0.05–0.06) and t-conorm (<0.01; per-corpus values and cluster-bootstrap CIs in Table 7); merging corpora re-mixes the between-structure variance, so merged effect sizes are smaller. This follows from Prop. 1: $\partial_1 I$ and $\partial_2 I$ gate every top-level path, so changing I rewires the routing topology of the whole loss, whereas changing T/S only rescales subpaths already attenuated by the gate. The pure-subsumption regime (form 1) is the cleanest confirmation: t-norm and t-conorm η^2 are exactly zero, as the form-1 case of Prop. 1 predicts.

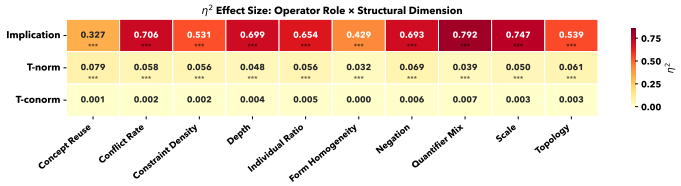


Figure 2: **Implication dominates operator-induced variance across structures.** Per-dimension η^2 of satisfaction rate in the one-factor-at-a-time synthetic sweep, decomposed by operator role: the implication row (0.33–0.79) is 4 to 20 times the t-norm row (<0.08) and far above the t-conorm row (<0.01). Stars: one-way F-test $p < 0.001$ and $\eta^2 > 0.01$; every implication and t-norm cell has $p < 10^{-17}$ and survives Bonferroni correction over the 30 cells, and no t-conorm cell reaches $\eta^2 = 0.01$. Conditioning on the full structural setting raises the implication's share to 0.73–0.81 (§ 4.2).

No universal ranking. Nine of the ten structure dimensions interact significantly with operator choice under both SR and mean truth degree (type-II ANOVA, level × implication, on the 45-operator core; all nine survive Bonferroni correction over the ten dimensions), the exception being quantifier mix ($p = 0.09$ and 0.10); the strongest is form homogeneity (one axiom form versus a mix): Göd-R beats Łuk-R on subsumption-only ontologies (its zero $\partial_1 I$ harmless with no role gradient) but falls behind as ∃/∀ axioms increase (Łuk-R's constant $\partial_1 I = -1$ keeps the role channel open). A broad ordering Łuk-R > Gog-R > Göd-R > Rei > KD persists, but local reversals make it a distributional pattern, not a law; the framework even predicts where the dominant operator fails (on chains of form-4 axioms $C \sqsubseteq \exists R.D$, Rei drives the training loss to exactly zero while Łuk-R stops at a residue of about 10^{-4}; App. I.1). The t-norm is not inert: stratified by conjunction density its effect is structure-gated and survives fixed, operator-independent evaluators (App. K.2).

4.3 Survival: Objective and Grounding

Objective: satisfaction versus learning. Prop. 4 predicts saturated operators are matched to satisfaction but lose signal under a learning objective. A satisfaction objective only asks that the constraints become true; a learning objective asks the logic loss to train something else, here either a classifier whose label passes through the implication (the discrimination objective) or digit concepts (MNIST). A controlled sweep over six implications, the five of § 2.2 plus the sigmoid implication $\sigma(10(v-u))$ (Sig, not the sigmoidal implication SP of the MNIST experiments), confirms this (Fig. 26): under satisfaction four of the six tie at ≈ 1.0, including all three saturated ones; under a discrimination objective two of those three degrade while the non-saturated hold ($\rho = -0.44$ across objectives, a one-directional collapse, not a clean inversion; Göd-R is the saturated exception that stays high). A real learning task shows the same mechanism: on classifier-grounded MNIST digit-sum (600 labels; target sums 5, 9, 13), Łuk-R reaches only 35% best-epoch concept (digit) accuracy versus 91% (SP, a sigmoid transform of Rei, van Krieken et al., 2022) and 74% (Rei), with the sum=9 collapse analyzed as a reasoning shortcut in § 5.

Grounding: softmax dead zones versus free embeddings. Prop. 6 predicts that the Łukasiewicz dead zone is entered only when groundings are pinned low. We isolate the grounding axis two ways. (i) A per-form learning scan under softmax groundings finds operator choice inert on the non-quantifier forms but decisive on quantifier ones: on existential forms, where the body t-norm aggregates softmax class probabilities through a max, the Łukasiewicz t-norm collapses to chance while t-norms with nonzero gradient there learn, exactly the dead-zone condition $a + b = 2/K < 1$ of Prop. 6. (ii) The same sweep on real ontologies (free sigmoid groundings) shows the predicted escape: the t-norm is inert for learning and the Łukasiewicz t-norm does not collapse (nominally best; App. K). On the softmax simplex this is output geometry, as Prop. 6 states, not ontology-specific (App. K.1). The existential collapse is therefore a property of the (classifier-grounding × existential-structure) regime, common in classifier-grounded neuro-symbolic systems but not in free-grounded fuzzy ALC, rather than a universal law. The ∀ analog is likewise scoped: a controlled MNIST-relational probe shows the collapse is *relation-dependent* (Göd-R collapses to chance only on bijective relations, unrescued by supervision; on non-bijective ones a small label anchor suffices; App. K.3).

4.4 Real Ontologies: Validation, Safe Default, and Robustness

A measurement-caliber control. Because SR is itself operator-computed, self-scoring is confounded even on form 1, where the t-norm is off the loss path, so we recheck every effect under fixed, operator-independent evaluators: the headline η^2 is self-scored, and on a 20-ontology train×eval

grid the implication effect persists under each fixed evaluator (η^2=0.10–0.70 vs. 0.21 self-scored, App. H.1). Under that control the spurious "t-norm governs satisfaction" reading collapses, while the genuine conjunction-dense effect and the implication effect persist under every fixed evaluator (App. K.2). Routing also reshapes the learned representations, with a measurable downstream cost: on real subsumption prediction the 0/1-clustered embeddings of Göd-R and Rei that win AUC on concept-corruption (D2) lose on role-corruption (D4), a within-run rank reversal of $\rho = -0.80$ over five implications, while the continuous Łuk-R/Gog-R embeddings stay stable across both, so no single training operator is best across the two corruption targets we test (App. G.2, H.2, K.2).

Real-world validation of the routing theorems. We test Thm. 3's prediction on three real corpora: 175 BioPortal ontologies; 260 $\mathcal{EL}$-real (§ 4.1); and the 227/230 ORE 2015 ontologies with a left-quantified axiom (the theorem is silent on the other 3). At initialization the antecedent-gradient norm tracks the theorem's structural bound (ρ=0.53/0.45/0.44; App. I.2). The complementary prediction of Thm. 2 is visible in the same corpora: right-existential share is the strongest single correlate of Łuk-R's advantage shrinking (ρ=+0.343), which is where the theorem says the left-quantified mechanism is unavailable (App. G.4).

The safe default and its precision barrier. We call a corpus inclusion-heavy when most of its normalized axioms are plain inclusions $C \sqsubseteq D$ (form 1); the median share is 93% on BioPortal and 99% on ORE 2015. Within the satisfaction regime (inclusion-heavy real ontologies, where routing dominates and saturation is a feature), Łuk-R is a strong safe default: only 8 of 405 real ontologies admit a higher self-scored SR under any alternative, a count that favors Łuk-R, the pointwise largest implication (App. G.3). A break-even analysis (Prop. 5; Table 8) sharpens this into a *precision barrier*: profitable switching needs near-perfect precision ($p^* \geq 0.994$, because false switches cost two orders of magnitude more than correct ones gain). It holds in the satisfaction regime; the objective and grounding axes are exactly where it ceases to hold.

Robustness and a cross-framework probe. The implication ranking that Prop. 1 accounts for survives every check in our robustness battery, including the $\mathcal{EL}$ fragment, five loss variants, softened and straight-through quantifier surrogates, and per-operator best learning rates (Apps. F, F.6). An LTN replication agrees descriptively ($\eta^2_{\text{impl}} = 0.77$, $\rho = 0.90$ over five implications, exact two-sided permutation $p = 0.083$; App. F.6). Every theoretical claim, including the predictions our own tests rejected, is mapped to its evidence with an evidential-status label (Table 6).

4.5 Beyond logic: pooling, fusion, and attention

Routing and survival concern an operator's gradient profile, so the same predictions extend to the *fixed analytic* aggregation operators (pooling, fusion, attention) mainstream systems share; on seven real ML aggregation benchmarks (each a dataset paired with the operator slot it exercises) and a synthetic pilot, seven of the eight predictions hold (Table 3): the Łukasiewicz t-conorm collapses instance-level AUC on CIFAR-10 multiple-instance learning (0.47 vs. 0.91). On aspect-based sentiment analysis (ABSA; SemEval-2014 Task 4 restaurant reviews), our prospective prediction that attention would beat mean pooling failed (mean 0.70 vs. attention 0.58–0.63); we then narrowed the routing scope to signals that can reach the output only through a routing operator.

5 Controlling Reasoning Shortcuts: Predict, Mitigate, Derive

The *survival* half, the mechanism behind the MNIST collapse of § 4.3, also yields a practical capability: operator-level control of reasoning shortcuts.

The standard remedy for a reasoning shortcut is dense concept supervision (Marconato et al., 2023; 2024), and here it fails: at the benchmark's logic-loss weight (10), Göd-R and Gog-R stay at $\approx 20\%$ concept accuracy (chance 10%) through 30,000 labeled digits, below supervision alone at 10 (33%). The concept loss is a batch mean, so labels add coverage, not gradient magnitude, and the barrier depends on the logic-loss weight: at weight 1 Gog-R escapes (93% at 300 labels) while Göd-R stays trapped (32%), and at 0.1 both escape, so Göd-R needs the lower weight (App. O). A non-saturated implication trails supervision alone when labels are scarce but catches up (KD by 300 labels, 94% vs. 93%; SP by 600; Rei by 1,200). This does not contradict Marconato et al. (2023): their Prop. 5 counts optima of the concept cross-entropy under assumption A1, whereas our claim concerns the *gradient* of a fuzzy logic loss, a fuzzy setting they flag (their Remark 8) but do not study.

The MNIST digit-sum collapse of § 4.3 is thus a textbook *reasoning shortcut* (Marconato et al., 2023) (Fig. 3), and the survival mechanism gives the field's open problems of identifying and mitigating it (Bortolotti et al., 2025) an operator-level handle. Concurrent work (Takemura & Inoue, 2026) relates t-norm gradient flow to shortcuts on MNIST tasks; our handle is the implication.

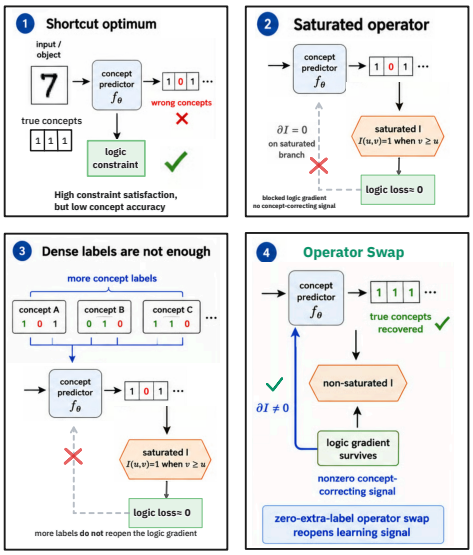


Figure 3: **Shortcuts and gradient survival.** Schematic: (1–2) a saturated implication gives a wrong-concept shortcut zero gradient; (3) labels do not restore it; (4) a non-saturated one does. Concepts are drawn as abstract binary vectors.

Prediction: an implication that is *saturated* ($I{=}1$ on $\{v \geq u\}$, so $\partial I{=}0$) makes every constraint-satisfying assignment a zero-gradient optimum of the logic loss, which then cannot prefer the ground-truth concepts (Prop. 7); so which operators induce a shortcut is predictable a priori from their derivative. The standard shortcut signature is constraint satisfaction well above concept accuracy (Marconato et al., 2023). On the MNIST digit-sum task (10 seeds, 300 labels) the saturated Gog-R and Łuk-R show it, constraint satisfied at 72–76% with concepts at 20–22% (for Gog-R at 3,000 labels, constraint 90% vs. concept 19%); whereas a swap to KD or SP attains both (concept 94/79% at constraint 89/83%): genuine learning, not a shortcut. RSBench (Bortolotti et al., 2024) corroborates this on its out-of-distribution split: without concept labels, operator choice tells a shortcut from a collapse in which neither is learned, but does not mitigate (App. O).

Mitigation: the implication swap above is that handle (Fig. 4); it removes the barrier, adds a small gain over supervision alone (KD: 0.2–1.7 points at 300–3,000 labels), and requires the swapped implication to lie on the constraint's gradient path.

Derivation: saturation alone makes a shortcut stationary, not stable: Łuk-R is saturated as well, yet 4 to 7 of 10 seeds escape at every budget from 1,200 labels. What holds the shortcut is the loss on the violated side of the constraint: in an idealized model, a $\Theta(K)$ wall for Gog-R against a $\Theta(1)$ slope for Łuk-R (Prop. 25); for Göd-R, the wall is its hard branch indicator $\mathbf{1}[u \leq v]$. A pre-registered τ-intervention confirms that this indicator causes the trap: smoothing only it unlocks escape at 3,000 labels while the hard operator stays trapped at 30,000 (Fig. 37).

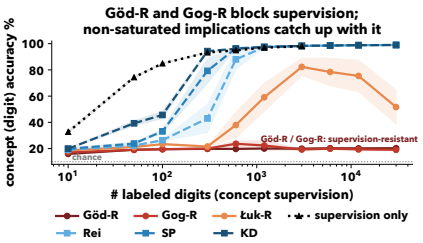


Figure 4: **Best-epoch concept accuracy on MNIST digit-sum** (logic-loss weight 10, 10 seeds, SE bands; dotted: supervision only, 5 seeds). Łuk-R is bimodal: escaped seeds end at 97–99%, the rest near 20% after transient peaks that lift the mean; escapes are not monotone in labels, cause unknown.

6 Discussion and Conclusion

The aggregation operator in a differentiable logic is a *derivable, structure-aware design variable*, not a per-domain convention. The ontology contribution is a *mechanism for choosing operators* plus a default rule, not a SOTA ontology-prediction result: on pure subsumption (form 1), an order embedding (Vendrov et al., 2016) and a box embedding both outscore every DF$\mathcal{ALC}$ implication, and DF$\mathcal{ALC}$ ties four zero-shot LLM probes, but neither baseline scores the role-quantified axioms the mechanism targets (Apps. J.1, J.2). Future work includes operator-aware remedies (exploration noise), learned structure-conditioned routing, and SOTA-scale ML benchmarks. Operator comparisons that omit structure, objective, and grounding rank a regime, not the operators; and for Göd-R and Gog-R a reasoning shortcut is not a shortage of labels but a property of the loss geometry, which an operator swap chosen a priori from its derivative, or a lower logic-loss weight, removes.

Reproducibility Statement

Code and data. We release anonymized code at `https://anonymous.4open.science/r/differentiable-logic-routing/`, covering the DF$\mathcal{ALC}$ pipeline (§ 2.3), experiment grid and runners, analysis scripts, and figure code. All corpora are public; the operator grid and corpus construction are specified in § 4.1 and App. G.1. **Training settings.** All settings are fixed and reported in App. E: d=50, Adam (PyTorch defaults), batch size 64, and 5 fixed seeds. The three main archives use learning rate 2×10^{-4}, a 25,000-epoch maximum, and early stopping with patience 10 at 100-epoch checkpoints; every other experiment states its own budget. **Theory and evidence.** All proofs appear in App. B. Every theoretical claim, including the predictions our own tests rejected, is paired with its empirical evidence and an explicit evidential-status label in the theory–evidence audit map (Table 6). **Results.** Raw result files are append-only with per-run seeds; the general-ML runs additionally carry host/commit/seed provenance manifests. The appendices document every statistic needed to recompute the reported numbers, including the headline η^2, the operator rankings, and the shortcut curves, end-to-end from the released artifacts. **Compute.** The experiments use commodity CPUs and GPUs and no large-scale training; the LLM baselines run on an A40 GPU. The operator rankings are budget-robust: the broad tier structure is preserved at learning rates of at least 2×10^{-4} (App. F.6) and at $d \geq 16$ (App. F.1), and is recovered with half the epoch budget (App. L.3), so the main conclusions can be recovered at a small fraction of the full 212,805-run archive.

AI Use Statement

We used LLM-based assistants for implementation, experiment management, figure generation, and editing, including consistency checks between the main text and the appendices. The research ideas, theory, experimental design, and conclusions are the authors' own; every reported number is computed from logged results by the released scripts, and the authors verified all claims and take full responsibility for the accuracy, originality, integrity, and final form of this paper.

Ethics Statement

This work analyzes how aggregation operators route and preserve gradients. It collects no data from people and uses only public resources under their original licenses: public ontologies (BioPortal, ORE 2015) and standard ML benchmarks, namely MNIST, CIFAR-10, the multiple-instance learning sets MUSK1 and Elephant, bAbI, ERASER movie rationales, SemEval-2014 Task 4 aspect-based sentiment analysis (restaurant reviews), CelebA, and CREMA-D. CelebA (face images) and CREMA-D (audio-visual recordings of consenting actors) involve people; we use them only in the attribute-conjunction and audio-visual fusion analyses, respectively, and train no person-identifying or attribute-inference system intended for deployment. Four open-weight instruction-tuned language models serve only as zero-shot subsumption baselines. The contribution is a mechanism and a set of diagnostic design rules for choosing operators in differentiable logic, not a deployable system, so we foresee no harmful societal impact specific to this analysis.

References

Dariusz Max Adamski and Jedrzej Potoniec. Reason-able embeddings: Learning concept embeddings with a transferable neural reasoner. *Semantic Web*, 15(4):1333–1365, 2024.

Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

Franz Baader, Ian Horrocks, Carsten Lutz, and Ulrike Sattler. *An Introduction to Description Logic*. Cambridge University Press, 2017.

Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *J. Mach. Learn. Res.*, 18:109:1–109:67, 2017.

Samy Badreddine, Artur S. d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic Tensor Networks. *Artif. Intell.*, 303:103649, 2022.

Samy Badreddine, Luciano Serafini, and Michael Spranger. logLTN: Differentiable Fuzzy Logic in the Logarithm Space. *CoRR*, abs/2306.14546, 2023.

Yoshua Bengio, Patrice Y. Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Networks*, 5(2):157–166, 1994.

Tarek R. Besold, Artur S. d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro M. Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luís C. Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, Frontiers in Artificial Intelligence and Applications, pp. 1–51. IOS Press, 2021.

Samuele Bortolotti, Emanuele Marconato, Tommaso Carraro, Paolo Morettin, Emile van Krieken, Antonio Vergari, Stefano Teso, and Andrea Passerini. A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts. In *Advances in Neural Information Processing Systems 37 (NeurIPS'24)*, 2024.

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and Identifiability in Concept-based Models from a Neuro-Symbolic Lens. In *Advances in Neural Information Processing Systems 38 (NeurIPS'25)*, 2025.

Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. OWL2Vec*: embedding of OWL ontologies. *Mach. Learn.*, 110(7):1813–1845, 2021.

Alessandro Daniele and Emile van Krieken. Noise to the Rescue: Escaping Local Minima in Neurosymbolic Local Search. *CoRR*, abs/2503.01817, 2025.

Michiel de Jong and Fei Sha. Neural Theorem Provers Do Not Learn Rules Without Exploration. *CoRR*, abs/1906.06805, 2019.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proc. ACL'20*, pp. 4443–4458. Association for Computational Linguistics, 2020.

Michelangelo Diligenti, Marco Gori, and Claudio Saccà. Semantic-based regularization for learning and inference. *Artif. Intell.*, 244:143–165, 2017.

Ivan Donadello, Luciano Serafini, and Artur S. d'Avila Garcez. Logic Tensor Networks for Semantic Image Interpretation. In *Proc. IJCAI'17*, pp. 1596–1602. ijcai.org, 2017.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Richard Evans and Edward Grefenstette. Learning Explanatory Rules from Noisy Data. *J. Artif. Intell. Res.*, 61:1–64, 2018.

Thomas Flinkow, Barak A. Pearlmutter, and Rosemary Monahan. Comparing differentiable logics for learning with logical constraints. *Sci. Comput. Program.*, 244:103280, 2025.

Thomas Flinkow, Ekaterina Komendantskaya, Matteo Capucci, and Rosemary Monahan. Quantitative Linear Logic for Neuro-Symbolic Learning and Verification. *CoRR*, abs/2605.13845, 2026.

Mattia Medina Grespan, Ashim Gupta, and Vivek Srikumar. Evaluating Relaxations of Logic for Neural Networks: A Comprehensive Study. In *Proc. IJCAI'21*, pp. 2812–2818. ijcai.org, 2021.

Petr Hájek. *Metamathematics of Fuzzy Logic*, volume 4 of *Trends in Logic*. Kluwer, 1998.

Hao-Yuan He, Wang-Zhou Dai, and Ming Li. Reduced implication-bias logic loss for neuro-symbolic learning. *Mach. Learn.*, 113(6):3357–3377, 2024.

Patrick Hohenecker and Thomas Lukasiewicz. Ontology Reasoning with Deep Neural Networks. *J. Artif. Intell. Res.*, 68:503–540, 2020.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *Proc. ICML'18*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2132–2141. PMLR, 2018.

Yuri I. Ingster. On some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

Mathias Jackermeier, Jiaoyan Chen, and Ian Horrocks. Dual Box Embeddings for the Description Logic $\mathcal{EL}^{++}$. In *Proc. WWW'24*, pp. 2250–2258. ACM, 2024.

Erich-Peter Klement, Radko Mesiar, and Endre Pap. *Triangular Norms*, volume 8 of *Trends in Logic*. Springer, 2000.

Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, and Robert Hoehndorf. EL Embeddings: Geometric Construction of Models for the Description Logic $\mathcal{EL}^{++}$. In *Proc. IJCAI'19*, pp. 6103–6109. ijcai.org, 2019.

Chen-Yu Lee, Patrick W. Gallagher, and Zhuowen Tu. Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree. In *Proc. AISTATS'16*, volume 51 of *JMLR Workshop and Conference Proceedings*, pp. 464–472. JMLR.org, 2016.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proc. ICML'19*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Binghui Li, Zhixuan Pan, Kaifeng Lyu, and Jian Li. Feature Averaging: An Implicit Bias of Gradient Descent Leading to Non-Robustness in Neural Networks. In *Proc. ICLR'25*. OpenReview.net, 2025.

Jaron Maene, Vincent Derkinderen, and Luc De Raedt. On the Hardness of Probabilistic Neurosymbolic Learning. In *Proc. ICML'24*, Proceedings of Machine Learning Research, pp. 34203–34218. PMLR / OpenReview.net, 2024.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural Probabilistic Logic Programming. In *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, pp. 3753–3763, 2018.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts. In *Advances in Neural Information Processing Systems 36 (NeurIPS'23)*, 2023.

Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Antonio Vergari, Andrea Passerini, and Stefano Teso. BEARS Make Neuro-Symbolic Models Aware of their Reasoning Shortcuts. In *Proc. UAI'24*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2399–2433. PMLR, 2024.

Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, and Marco Gori. Integrating Learning and Reasoning with Deep Logic Models. In *Proc. ECML-PKDD'19*, LNCS, pp. 517–532. Springer, 2019.

Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From Statistical Relational to Neurosymbolic Artificial Intelligence: A Survey. *Artif. Intell.*, 328:104062, 2024.

André F. T. Martins and Ramón Fernandez Astudillo. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In *Proc. ICML'16*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1614–1623. JMLR.org, 2016.

Olga Mashkova, Fernando Zhapa-Camacho, and Robert Hoehndorf. Enhancing Geometric Ontology Embeddings for $\mathcal{EL}^{++}$ with Negative Sampling and Deductive Closure Filtering. In *Proc. NeSy'24*, LNCS, pp. 331–354. Springer, 2024.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. Learning Reasoning Strategies in End-to-End Differentiable Proving. In *Proc. ICML'20*, Proceedings of Machine Learning Research, pp. 6938–6949. PMLR, 2020.

Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak A. Rao, and Bruno Ribeiro. Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs. In *Proc. ICLR'19*. OpenReview.net, 2019.

Simon Odense and Artur d'Avila Garcez. A semantic framework for neurosymbolic computation. *Artif. Intell.*, 340:104273, 2025.

Özgür Lütfü Özçep, Mena Leemhuis, and Diedrich Wolter. Embedding Ontologies in the Description Logic ALC by Axis-Aligned Cones. *J. Artif. Intell. Res.*, 78:217–267, 2023.

Bijan Parsia, Nicolas Matentzoglu, Rafael S. Gonçalves, Birte Glimm, and Andreas Steigmiller. The OWL Reasoner Evaluation (ORE) 2015 Competition Report. *J. Autom. Reason.*, 59(4):455–482, 2017.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *Proc. CVPR'22*, pp. 8228–8237. IEEE, 2022.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse Sequence-to-Sequence Models. In *Proc. ACL'19*, pp. 1504–1519. Association for Computational Linguistics, 2019.

Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, and Giuseppe Marra. From Statistical Relational to Neuro-Symbolic Artificial Intelligence. In *Proc. IJCAI'20*, pp. 4943–4950. ijcai.org, 2020.

Ryan Riegel, Alexander G. Gray, Francois P. S. Luus, Naweed Khan, Ndivhuwo Makondo, Ismail Yunus Akhalwaya, Haifeng Qian, Ronald Fagin, Francisco Barahona, Udit Sharma, Shajith Ikbal, Hima Karanam, Sumit Neelam, Ankita Likhyani, and Santosh K. Srivastava. Logical Neural Networks. *CoRR*, abs/2006.13155, 2020.

Tim Rocktäschel and Sebastian Riedel. End-to-end Differentiable Proving. In *Advances in Neural Information Processing Systems 30 (NeurIPS'17)*, pp. 3788–3800, 2017.

Manfred Schmidt-Schauß and Gert Smolka. Attributive Concept Descriptions with Complements. *Artif. Intell.*, 48(1):1–26, 1991.

Prithviraj Sen, Breno W. S. R. de Carvalho, Ryan Riegel, and Alexander G. Gray. Neuro-Symbolic Inductive Logic Programming with Logical Neural Networks. In *Proc. AAAI'22*, pp. 8212–8219. AAAI Press, 2022.

Luciano Serafini, Ivan Donadello, and Artur S. d'Avila Garcez. Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proc. SAC'17*, pp. 125–130. ACM, 2017.

Iman Sharifi, Peng Wei, and Saber Fallah. ANDRE: An Attention-based Neuro-symbolic Differentiable Rule Extractor for Inductive Logic Programming. *CoRR*, abs/2605.04193, 2026.

Natalia Slusarz, Ekaterina Komendantskaya, Matthew L. Daggitt, Robert J. Stewart, and Kathrin Stark. Logic of Differentiable Logics: Towards a Uniform Semantics of DL. In *Proc. LPAR'23*, EPiC Series in Computing, pp. 473–493. EasyChair, 2023.

Umberto Straccia. Reasoning within Fuzzy Description Logics. *J. Artif. Intell. Res.*, 14:137–166, 2001.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28 (NeurIPS'15)*, pp. 2440–2448, 2015.

Akihiro Takemura and Katsumi Inoue. Differentiable Logic Programming to Mitigate Reasoning Shortcuts in Neurosymbolic Systems. *CoRR*, abs/2607.21185, 2026.

Zhenwei Tang, Tilman Hinnerichs, Xi Peng, Xiangliang Zhang, and Robert Hoehndorf. F$\mathcal{ALC}$ON: Sound and Complete Neural Semantic Entailment over $\mathcal{ALC}$ Ontologies. *CoRR*, abs/2208.07628, 2022.

Emile van Krieken, Erman Acar, and Frank van Harmelen. Analyzing Differentiable Fuzzy Logic Operators. *Artif. Intell.*, 302:103602, 2022.

Emile van Krieken, Thiviyan Thanapalasingam, Jakub M. Tomczak, Frank van Harmelen, and Annette ten Teije. A-NeSI: A Scalable Approximate Method for Probabilistic Neurosymbolic Inference. In *Advances in Neural Information Processing Systems 36 (NeurIPS'23)*, 2023.

Emile van Krieken, Pasquale Minervini, Edoardo M. Ponti, and Antonio Vergari. On the Independence Assumption in Neurosymbolic Learning. In *Proc. ICML'24*, Proceedings of Machine Learning Research, pp. 49078–49097. PMLR / OpenReview.net, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS'17)*, pp. 5998–6008, 2017.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-Embeddings of Images and Language. In *Proc. ICLR'16*, 2016.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the Limitations of Representing Functions on Sets. In *Proc. ICML'19*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6487–6494. PMLR, 2019.

Po-Wei Wang, Priya L. Donti, Bryan Wilder, and J. Zico Kolter. SATNet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. In *Proc. ICML'19*, Proceedings of Machine Learning Research, pp. 6545–6554. PMLR, 2019.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *Proc. ICLR'16*, 2016.

Patricia L. Whetzel, Natalya Fridman Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, 39(Web-Server-Issue):541–545, 2011.

Thomas Winters, Giuseppe Marra, Robin Manhaeve, and Luc De Raedt. DeepStochLog: Neural Stochastic Logic Programming. In *Proc. AAAI'22*, pp. 10090–10100. AAAI Press, 2022.

Xuan Wu and Yizheng Zhao. A Neuro-Symbolic Approach to Symbol Grounding for $\mathcal{ALC}$-Ontologies. In *Proc. KDD'25*, pp. 3240–3249. ACM, 2025.

Bo Xiong, Nico Potyka, Trung-Kien Tran, Mojtaba Nayyeri, and Steffen Staab. Faithful Embeddings for $\mathcal{EL}^{++}$ Knowledge Bases. In *Proc. ISWC'22*, LNCS, pp. 22–38. Springer, 2022.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In *Proc. ICML'18*, Proceedings of Machine Learning Research, pp. 5498–5507. PMLR, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *Proc. ICLR'19*. OpenReview.net, 2019.

Weidi Xu, Jingwei Wang, Lele Xie, Jianshan He, Hongting Zhou, Taifeng Wang, Xiaopei Wan, Jingdong Chen, Chao Qu, and Wei Chu. LogicMP: A Neuro-symbolic Approach for Encoding First-order Logic Constraints. In *Proc. ICLR'24*. OpenReview.net, 2024.

Fan Yang, Zhilin Yang, and William W. Cohen. Differentiable Learning of Logical Rules for Knowledge Base Reasoning. In *Advances in Neural Information Processing Systems 30 (NeurIPS'17)*, pp. 2319–2328, 2017.

Zhun Yang, Joohyung Lee, and Chiyoun Park. Injecting Logical Constraints into Neural Networks via Straight-Through Estimators. In *Proc. ICML'22*, Proceedings of Machine Learning Research, pp. 25096–25122. PMLR, 2022.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems 30 (NeurIPS'17)*, pp. 3391–3401, 2017.

Fernando Zhapa-Camacho and Robert Hoehndorf. Lattice-preserving $\mathcal{ALC}$ ontology embeddings. In *Proc. NeSy'24*, LNCS, pp. 355–369. Springer, 2024.

Matthieu Zimmer, Xuening Feng, Claire Glanois, Zhaohui Jiang, Jianyi Zhang, Paul Weng, Dong Li, Jianye Hao, and Wulong Liu. Differentiable logic machines. *Trans. Mach. Learn. Res.*, 2023, 2023.

Supplementary Material: Table of Contents

Notation. Symbols keep their main-text meaning throughout; a symbol introduced inside an appendix section is local to that section unless stated otherwise. *Reading guide.* Appendices A–C provide definitions, proofs, and the general-ML theory. Appendices D–E cover prior-work scope and implementation details. Appendices F, I, and L are the main claim-defense appendices, collecting the robustness, structural-crossover, and training-dynamics evidence. Appendices G, H, J, K, and O document practical consequences and scope checks: real-corpus switching, downstream geometry, external (geometric and LLM) baselines, survival-axis probes, and reasoning-shortcut analyses. Appendices M–N state limitations and broader related work.

Appendix A Extended Preliminaries 18
A.1 Formal $\mathcal{ALC}$ Syntax and Semantics 18
A.2 Fuzzy Relaxation: Syntax, Semantics, and Loss 18
A.3 Operator Catalog: T-Norms, T-Conorms, and Implications 19
A.4 A Worked Micro-Example: Gradient Routing Through Two Axioms 20

Appendix B Full Proofs 20
B.1 Proof of Proposition 1 (Implication-Gated Gradient Decomposition) 20
B.2 Role-Free Antecedent and Consequent Lemmas 21
B.3 Proof of Theorem 2 (Role-Gradient Equivalence) 21
B.4 Proof of Theorem 3 (Exact Descent Gap) 21
B.5 Proof of Corollary 11 (Linear Lower Bound) 22
B.6 Proof of Proposition 5 (Break-Even Precision Threshold) 23
B.7 Proof of Proposition 4 (Vanishing Logic Gradient) 23
B.8 Proof of Proposition 6 (Grounding Dependence of the Dead Zone) 24

Appendix C General-ML Validation: Theory, Scorecard, and Statistics 24
C.1 Quantitative Survival Thresholds 25

Appendix D Prior-Work Comparison and Scope Boundaries 32

Appendix E Implementation Details 34

Appendix F Robustness Battery 35
F.1 Dimension Sensitivity 36
F.2 Loss-Family Robustness: Expanded Sweep 36
F.3 Expanded Parametric-Family Sweep 37
F.4 Targeted Extra-Seed Robustness 38
F.5 ML-Friendly Rank Summaries 38
F.6 Claim-Survival Summary 39
F.7 Threshold Robustness and Additional Sigmoid Baseline 41
F.8 Softened Quantifier Ablation 41

Appendix G Real-World Corpus Analysis 42
G.1 Dataset Selection Bias and Corpus Complementarity 42

G.2 Embedding Pathology 42
G.3 Switch-Case Detail 42
G.4 Online-Switching Ablation and Structural-Classifier Feasibility 43

Appendix H Embedding Geometry and Downstream Impact 46
H.1 D2 Task-Type Metric Split: Detail 46
H.2 D4 Cross-Task Role-Corruption: Detail 47

Appendix I Structural Crossover Details 48
I.1 Chain $\times$ Right-Existential Structural Law: Detail 48
I.2 Multi-Corpus Validation of Theorem 3 and Corollary 11 49
I.3 Soft-Surrogate τ-Sweep Audit 50
I.4 Soft-vs-Hard Endpoint Re-check on the Chain Cell 51

Appendix J Extended Scope Checks 51
J.1 Geometric Baseline on the Shared D2 Slice 51
J.2 Scope Check Against Axiom-Only LLM Probes 52

Appendix K Cross-Task Survival Experiments: Details 53
K.1 Softmax Simplex Dead-Zone View 53
K.2 The Measurement-Caliber Confound, Across Both T-Norm Sweeps 54
K.3 The Universal-Quantifier Collapse Is Relation-Dependent 54
K.4 Survival in a Standard Regularizer: Softmax Consistency, and When the Dead Zone Bites 55

Appendix L Training Dynamics 57
L.1 Direct Gradient Routing Observation 57
L.2 Convergence Speed and Training Stability 57
L.3 Early-Stopping Learning Curve 58

Appendix M Extended Limitations and Discussion 60
M.1 Theoretical Scope: Łuk-R/Göd-R Only 60
M.2 Logic Fragment: $\mathcal{ALC}$ Only 60
M.3 Mixed-Operator Strategies 60
M.4 Synthetic-to-Real Transfer 61

Appendix N Extended Related Work 61
N.1 Fuzzy Neuro-Symbolic Systems and Operator Conventions 61
N.2 Ontology Embedding Methods 61
N.3 Differentiable Logic Beyond Fuzzy Relaxation 62
N.4 Neuro-Symbolic Surveys and Taxonomies 62

Appendix O Reasoning-Shortcut Theory: Predict, Mitigate, Derive 62

A Extended Preliminaries

Roadmap. The formal background the main text compresses: full $\mathcal{ALC}$ syntax and fuzzy-relaxation semantics, the operator catalog with residuum derivations and the derivative atlas (Fig. 5), and a worked two-axiom gradient trace.

A.1 Formal $\mathcal{ALC}$ Syntax and Semantics

Description logics are a family of knowledge representation formalisms that describe a domain in terms of concepts (unary predicates over individuals), roles (binary relations between individuals), and their logical combinations (Baader et al., 2017). A set of such logical statements about a domain is called an *ontology*; in this paper, an ontology is synonymous with a *TBox* (terminological box), a finite set of general concept inclusions (GCIs) of the form $C \sqsubseteq D$.

The concept language of $\mathcal{ALC}$ is defined by the grammar

$$C \ ::= \ A \ \mid \ \neg C \ \mid \ C \sqcap C \ \mid \ C \sqcup C \ \mid \ \exists R.C \ \mid \ \forall R.C,$$

where $A \in N_C$ is a concept name and $R \in N_R$ is a role name. An *interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ consists of a non-empty domain $\Delta^{\mathcal{I}}$ and a mapping $\cdot^{\mathcal{I}}$ that assigns each concept name A a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and each role name R a relation $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. Complex concepts are interpreted compositionally as follows:

$$\begin{aligned}
(\neg C)^{\mathcal{I}} &= \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}, \\
(C \sqcap D)^{\mathcal{I}} &= C^{\mathcal{I}} \cap D^{\mathcal{I}}, \\
(C \sqcup D)^{\mathcal{I}} &= C^{\mathcal{I}} \cup D^{\mathcal{I}}, \\
(\exists R.C)^{\mathcal{I}} &= \{x \in \Delta^{\mathcal{I}} \mid \exists y.\, (x, y) \in R^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}, \\
(\forall R.C)^{\mathcal{I}} &= \{x \in \Delta^{\mathcal{I}} \mid \forall y.\, (x, y) \in R^{\mathcal{I}} \rightarrow y \in C^{\mathcal{I}}\}.
\end{aligned}$$

In words: $\neg C$ contains everything not in C; $C \sqcap D$ and $C \sqcup D$ are the intersection and union of two concepts; $\exists R.C$ contains every individual that has at least one R-related individual in C; and $\forall R.C$ contains every individual whose R-related individuals all belong to C. A GCI $C \sqsubseteq D$ is *satisfied* by $\mathcal{I}$ (written $\mathcal{I} \models C \sqsubseteq D$) iff $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$. An interpretation $\mathcal{I}$ is a *model* of a TBox $\mathcal{T}$ iff it satisfies every GCI in $\mathcal{T}$.

A.2 Fuzzy Relaxation: Syntax, Semantics, and Loss

The DF$\mathcal{ALC}$ pipeline replaces crisp set-theoretic semantics with continuous truth values in $[0, 1]$. Each concept name A is assigned a *concept embedding* $\mathbf{a} \in [0, 1]^d$, where entry $\mathbf{a}_i$ represents the degree to which individual i belongs to concept A. Each role name R is assigned a *role embedding* $\mathbf{R} \in [0, 1]^{d \times d}$, where entry $\mathbf{R}_{ij}$ represents the degree to which individual i is R-related to individual j. Complex concepts are interpreted by replacing the set-theoretic operations with their fuzzy counterparts as follows:

$$\begin{aligned}
\|\neg C\|_i &= 1 - \|C\|_i, \\
\|C \sqcap D\|_i &= T(\|C\|_i,\, \|D\|_i), \\
\|C \sqcup D\|_i &= S(\|C\|_i,\, \|D\|_i), \\
\|\exists R.C\|_i &= \max_j T(\mathbf{R}_{ij},\, \|C\|_j), \\
\|\forall R.C\|_i &= \min_j S(1 - \mathbf{R}_{ij},\, \|C\|_j),
\end{aligned}$$

where T is a t-norm (fuzzy conjunction), S is a t-conorm (fuzzy disjunction), $\|C\|_i \in [0, 1]$ denotes the truth degree of individual i in concept C, and $\max_j$/$\min_j$ aggregate over all individuals j (fuzzy existential/universal quantification).

A GCI $C \sqsubseteq D$ is grounded for each individual i as $I(\|C\|_i, \|D\|_i)$, where I is a fuzzy implication. Normalization rewrites arbitrary nested GCIs into a fixed set of shallow patterns so that every axiom's truth-value computation follows the same two-step structure: first compute the antecedent and consequent truth values u_i and v_i using T and S, then apply the implication I at the top level. Under the normalization of § 2.1, every GCI is rewritten into one of the seven *axiom forms* (Table 1 in the main text), each of which decomposes the truth-value computation into an antecedent u_i and a consequent v_i such that the per-axiom loss takes the form $\ell_\alpha = \sum_i [1 - I(u_i, v_i)]$. The total loss is $L(\theta) = \sum_\alpha \ell_\alpha(\theta)$, where θ denotes all embedding parameters. A single choice of t-norm T,

t-conorm S, and implication I is called a *configuration*; the paper's *operator grid* consists of all $3 \times 3 \times 5 = 45$ configurations formed from the three t-norms, three t-conorms, and five implications defined in App. A.3. As described in § 2.3, the implication I always appears at the top level, while the t-norm T and t-conorm S appear only inside u_i or v_i. This architectural asymmetry is the structural basis of the gradient-routing analysis in § 3.

A.3 Operator Catalog: T-Norms, T-Conorms, and Implications

This subsection provides the complete definitions of the t-norms, t-conorms, and fuzzy implications used in the paper. The main text (§ 2.2) lists the formulas and derivatives; here we explain where each operator comes from.

T-Norms. A t-norm $T\colon [0,1]^2 \to [0,1]$ is commutative, associative, monotone, and satisfies $T(1,x) = x$. The three t-norms used in this paper are:

$$\begin{aligned} \text{Gödel:} \quad & T_G(x,y) = \min(x,y), \\ \text{Product:} \quad & T_P(x,y) = xy, \\ \text{Łukasiewicz:} \quad & T_L(x,y) = \max(0,\, x+y-1). \end{aligned}$$

T-Conorms. Each t-norm T induces a dual t-conorm $S(x,y) = 1 - T(1-x, 1-y)$; the operator grid uses the three resulting t-conorms, crossed independently with the three t-norms:

$$\begin{aligned} \text{Gödel:} \quad & S_G(x,y) = \max(x,y), \\ \text{Product (probabilistic sum):} \quad & S_P(x,y) = x + y - xy, \\ \text{Łukasiewicz (bounded sum):} \quad & S_L(x,y) = \min(1,\, x+y). \end{aligned}$$

Fuzzy Implications. A fuzzy implication $I\colon [0,1]^2 \to [0,1]$ is decreasing in the first argument, increasing in the second, and satisfies $I(0,0) = I(1,1) = 1$ and $I(1,0) = 0$. We use five implications from two families.

R-implications (residua) are derived from a t-norm via the *residuum* construction: $I_T(u,v) = \sup\{z \in [0,1] : T(u,z) \le v\}$. Intuitively, $I_T(u,v)$ is the largest truth degree z such that the conjunction $T(u,z)$ does not exceed v. Three of our five implications are R-implications:

$$\begin{aligned} \text{Łukasiewicz (R):} \quad & I_{\text{Łuk}}(u,v) = \min(1,\, 1-u+v) && \text{[residuum of } T_L\text{]}, \\ \text{Gödel (R):} \quad & I_{\text{Göd}}(u,v) = \begin{cases} 1 & \text{if } u \le v, \\ v & \text{otherwise} \end{cases} && \text{[residuum of } T_G\text{]}, \\ \text{Goguen (R):} \quad & I_{\text{Gog}}(u,v) = \begin{cases} 1 & \text{if } u \le v, \\ v/u & \text{otherwise} \end{cases} && \text{[residuum of } T_P\text{]}. \end{aligned}$$

S-implications are constructed from a t-conorm: $I_S(u,v) = S(1-u, v)$, corresponding to the classical tautology $A \to B \equiv \neg A \vee B$. Two of our five implications are S-implications:

$$\begin{aligned} \text{Reichenbach:} \quad & I_{\text{Rei}}(u,v) = 1 - u + uv && \text{[from } S_P\text{: } S_P(1-u,v) = (1-u) + v - (1-u)v\text{]}, \\ \text{Kleene–Dienes:} \quad & I_{\text{KD}}(u,v) = \max(1-u,\, v) && \text{[from } S_G\text{: } S_G(1-u,v) = \max(1-u,v)\text{]}. \end{aligned}$$

The key distinction for gradient routing is that R-implications and S-implications produce qualitatively different partial derivatives $\partial_1 I$ and $\partial_2 I$. R-implications can have $\partial_1 I = 0$ in the satisfied region ($u \le v$), creating gradient-blocking behavior (Gödel, Goguen). Łukasiewicz is the unique R-implication (among the three studied) with constant nonzero derivatives on both branches. S-implications inherit structure from the t-conorm: Reichenbach's $\partial_1 = -(1-v)$ is input-scaled (self-damping), while KD's $\max$-based structure produces branch-exclusive routing. The derivative table in § 2.2 summarizes these properties, and Fig. 5 maps them over the entire (u,v) plane.

Implications outside the operator grid. Two further operators appear only in § 4.3. Sig is $I_{\text{Sig}}(u,v) = \sigma(10(v-u))$, with $\sigma(x) = 1/(1+e^{-x})$ the logistic function; it is a smooth surrogate rather than a fuzzy implication in the strict sense above, since $I_{\text{Sig}}(0,0) = \frac{1}{2}$. SP is the Reichenbach-sigmoidal implication of van Krieken et al. (2022) (their Def. 9 with $b_0 = -\frac{1}{2}$, Eq. 49), used with $s = 9$:

$$I_{\text{SP}}(u,v) = \frac{1}{e^{s/2}-1}\Big(\big(1+e^{s/2}\big)\,\sigma\big(s\,(I_{\text{Rei}}(u,v) - \tfrac{1}{2})\big) - 1\Big),$$

which rescales $\sigma(s(I_{\text{Rei}} - \frac{1}{2}))$ onto $[0,1]$ so that I_{SP} keeps the boundary values of I_{Rei}.

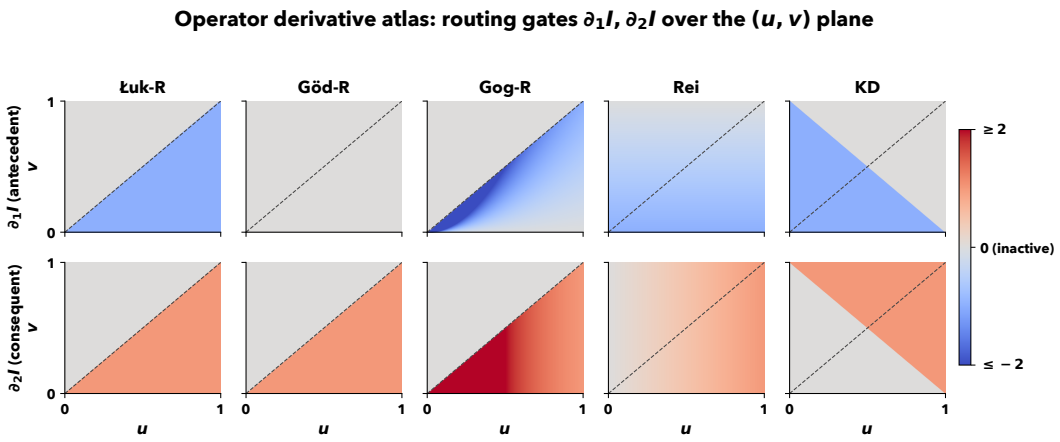


Figure 5: **Operator derivative atlas.** The routing gates $\partial_1 I(u, v)$ (antecedent, top row) and $\partial_2 I(u, v)$ (consequent, bottom row) for the five implications over the truth-value plane; the dashed line is the satisfied boundary $v{=}u$. A colored cell is where gradient can reach that branch (blue = antecedent-pulling $\partial_1{<}0$, red = consequent-pulling $\partial_2{>}0$); white/gray is an inactive gate. The atlas summarizes the routing picture: Łuk-R activates both branches on the violated region; Göd-R's $\partial_1 I$ is *identically zero* (antecedent channel blocked); Gog-R's partials diverge as $u \to 0$ (instability); Rei is smooth and self-damped everywhere (no dead zone); KD is branch-exclusive across the switching surface $1{-}u{=}v$. The three insets of Fig. 1 are one slice of this atlas.

A.4 A Worked Micro-Example: Gradient Routing Through Two Axioms

Consider the two-axiom ontology $\{A \sqsubseteq B \text{ (form 1)};\ \exists R.A \sqsubseteq C \text{ (form 6)}\}$ and one individual x with groundings $a = \|A\|(x) = 0.7$, $b = \|B\|(x) = 0.4$ (axiom 1 violated). **Axiom 1** contributes $1 - I(a, b)$: under Łuk-R, $I = 1 - a + b = 0.7$ with $\partial_1 I = -1,\ \partial_2 I = 1$—both branches receive gradient (a is pushed down, b up); under Göd-R, $I = b = 0.4$ with $\partial_1 I = 0,\ \partial_2 I = 1$—the antecedent branch is *invisible*: no learning-rate schedule can move A's embedding through this axiom; under KD, $I = \max(1 - a, b) = \max(0.3, 0.4) = 0.4$—the max selects the consequent branch, so only b receives gradient (branch starvation, Cor. 12); had a been 0.5, the antecedent branch would be selected instead. **Axiom 2** places the *role* inside the antecedent: $u = \max_y T(\|R\|(x, y), \|A\|(y))$, so the role embedding $\mathbf{R}$ is reachable only through $\partial_1 I$. Under Łuk-R ($\partial_1 I = -1$ on the active region) $\mathbf{R}$ trains; under Göd-R ($\partial_1 I = 0$) $\mathbf{R}$ is *frozen*—the role-freezing of Thm. 3, visible already in this two-axiom trace. Every body result is this arithmetic at scale: which partials are nonzero (routing), and on which input regions (survival).

B Full Proofs

Roadmap. Complete proofs of every proposition, theorem, corollary, and lemma stated in the main text, together with the supporting lemmas of App. B.2.

All proofs use the notation and axiom-form definitions from § 2. The gradient results hold at every regular θ as defined in § 3.

B.1 Proof of Proposition 1

Define $\ell_{\alpha,i}(\theta) = 1 - I(u_i(\theta), v_i(\theta))$. By the chain rule,

$$\nabla_\theta \ell_{\alpha,i} = -\partial_1 I(u_i, v_i)\ \nabla_\theta u_i - \partial_2 I(u_i, v_i)\ \nabla_\theta v_i.$$

Summing over individuals gives (2). In the pure form-1 case $u_i = \mathbf{c}_i$, $v_i = \mathbf{d}_i$, so neither T, S, nor any role embedding enters the computation graph and $\nabla_{\mathbf{R}} \ell_\alpha = 0$, giving the stated independence. The max/min operations inside u_i and v_i act as piecewise-differentiable selectors. At non-tie points, the gradient passes through exactly one branch of each selector, confirming almost-everywhere validity (ties form a measure-zero set).

Quantifier branch-selection lemma. For existential restrictions, $\text{exists}_i = \max_j T(\mathbf{R}_{ij}, \mathbf{d}_j)$; away from ties, gradient flows only through the argmax filler. For universal restrictions, $\text{forall}_i =$

$\min_j S(1-\mathbf{R}_{ij}, \mathbf{d}_j)$; away from ties, gradient flows only through the argmin filler. This bridges the general chain-rule proposition to the structure-dependent path-survival interpretation. □

B.2 Role-Free Antecedent and Consequent Lemmas

The proofs of Thms. 2 and 3 use two structural lemmas stating that roles are absent from the consequent or the antecedent in pure-form regimes. Both are immediate from the axiom-form definitions in § 2; they hold for the implementation's hard $\max/\min$ selectors away from selector ties.

Lemma 8 (Role-free consequents under left quantification)**.** *Let O contain only forms 6 ($\exists R.C \sqsubseteq D$) and 7 ($\forall R.C \sqsubseteq D$). Then for every axiom of O, every individual i, and every role coordinate, $\partial v_i/\partial \mathbf{R} = 0$, since the consequents of forms 6 and 7 contain no role.*

Proof. In forms 6 and 7, the consequent v_i is the concept embedding $\mathbf{d}_i$ (modulated by negation flags). No role coordinate enters the consequent computation. Hence $\partial v_i/\partial \mathbf{R} = 0$. □

Lemma 9 (Role-free antecedents under right quantification)**.** *Let O contain only forms 4 ($C \sqsubseteq \exists R.D$) and 5 ($C \sqsubseteq \forall R.D$). Then for every axiom of O, every individual i, and every role coordinate, $\partial u_i/\partial \mathbf{R} = 0$, since the antecedents of forms 4 and 5 contain no role.*

Proof. In forms 4 and 5, the antecedent u_i is the concept embedding $\mathbf{c}_i$ (modulated by negation flags). No role coordinate enters the antecedent computation. Hence $\partial u_i/\partial \mathbf{R} = 0$. □

B.3 Proof of Theorem 2

Fix O pure right-quantified and θ regular (§ 3). Let i be an active violated axiom (so $u_i(\theta) > v_i(\theta)$ on the loss-relevant branch).

By Prop. 1,

$$\nabla_{\mathbf{R}}\, \ell_i^H(\theta) \;=\; -\,\partial_1 I(u_i, v_i)\, \nabla_{\mathbf{R}} u_i(\theta) \;-\; \partial_2 I(u_i, v_i)\, \nabla_{\mathbf{R}} v_i(\theta).$$

By Lem. 9, $\nabla_{\mathbf{R}} u_i(\theta) = 0$. So the first term vanishes for any implication, and

$$\nabla_{\mathbf{R}}\, \ell_i^H(\theta) \;=\; -\,\partial_2 I(u_i, v_i)\, \nabla_{\mathbf{R}} v_i(\theta).$$

For the violated branch of hard Łuk-R outside the saturation corner, $I_{\text{Łuk}}(u, v) = 1 - u + v$, so $\partial_2 I_{\text{Łuk}} = 1$. For the violated branch of hard Göd-R, $I_{\text{Göd}}(u, v) = v$, so $\partial_2 I_{\text{Göd}} = 1$. Both implications therefore satisfy

$$\nabla_{\mathbf{R}}\, \ell_i^{\text{Łuk}}(\theta) \;=\; -\nabla_{\mathbf{R}} v_i(\theta) \;=\; \nabla_{\mathbf{R}}\, \ell_i^{\text{Göd}}(\theta).$$

Summing over all active violated axioms (with weights w_i) gives

$$\nabla_{\mathbf{R}} L_O^{\text{Łuk}}(\theta) \;=\; \nabla_{\mathbf{R}} L_O^{\text{Göd}}(\theta).$$

□

B.4 Proof of Theorem 3

We use a directional decomposition that holds for arbitrary mixed ontologies on any role subspace; this is Prop. 10 below, which we state and prove first.

Proposition 10 (Role-subset decomposition of the descent difference between Łuk-R and Göd-R)**.** *Let O be a DF$\mathcal{ALC}$ ontology, θ regular, and let $U \subseteq \mathrm{Roles}(O)$ be a role subset. Let $\Theta_{R,U}$ be the corresponding role-coordinate subspace and fix $\Delta \in \Theta_{R,U}$. Define*

$$A_L^U(\theta) = \{\, i \in \mathcal{A}(\theta) : \mathrm{form}(i) \in \{6, 7\},\ \mathrm{role}(i) \in U \,\},$$
$$A_R^U(\theta) = \{\, i \in \mathcal{A}(\theta) : \mathrm{form}(i) \in \{4, 5\},\ \mathrm{role}(i) \in U \,\}.$$

Then

$$DL_O^{\text{Łuk}}(\theta)[\Delta] - DL_O^{\text{Göd}}(\theta)[\Delta] \;=\; \sum_{i \in A_L^U(\theta)} w_i\, Du_i(\theta)[\Delta].$$

Proof. The directional derivative decomposes axiom-wise. We classify active violated axioms by form: *(a) Forms $6/7$ with role in U.* The direction Δ acts only on the antecedent (Lem. 8 gives $Dv_i[\Delta] = 0$). Hard Łuk-R contributes $D\ell_i^L[\Delta] = Du_i[\Delta] - 0 = Du_i[\Delta]$; hard Göd-R contributes $D\ell_i^G[\Delta] = -Dv_i[\Delta] = 0$. The Łuk–Göd difference per axiom is $Du_i[\Delta]$. *(b) Forms $4/5$ with role in U.* The direction Δ acts only on the consequent (Lem. 9). On the violated branch, $\partial_2 I = 1$ for both implications, so both contribute $-Dv_i[\Delta]$, yielding zero per-axiom difference. *(c) Roles*

outside U, or forms $1/2/3$. The direction $\Delta \in \Theta_{R,U}$ does not act on any coordinate touched by these axioms. They contribute zero on each side. Summing the per-axiom Łuk–Göd differences leaves only contributions from (a), proving the claim. □

We can now prove Thm. 3. Choose $\Delta_U := -g_U^L(\theta; S) \in \Theta_{R,U}$. By Prop. 10 applied with $S = A_L^U(\theta)$,

$$DL_S^{\text{Łuk}}(\theta)[\Delta_U] - DL_S^{\text{Göd}}(\theta)[\Delta_U] \;=\; \sum_{i \in \mathcal{A}} w_i \,\langle \nabla u_i(\theta),\, \Delta_U \rangle.$$

Because $\Delta_U \in \Theta_{R,U}$, the inner product is unchanged by projecting the gradient onto $\Theta_{R,U}$:

$$\sum_{i \in \mathcal{A}} w_i \,\langle \nabla u_i, \Delta_U \rangle \;=\; \sum_{i \in \mathcal{A}} w_i \,\langle P_{R,U} \nabla u_i, \Delta_U \rangle \;=\; \langle g_U^L(\theta; S),\, \Delta_U \rangle.$$

Substituting $\Delta_U = -g_U^L$ gives $\langle g_U^L, -g_U^L \rangle = -\|g_U^L\|^2$. Since $g_U^L(\theta; S) \neq 0$ by hypothesis, $\|g_U^L\|^2 > 0$, completing the proof. □

Corollary 11 (Linear lower bound in the number of active left-quantified pairs)**.** *Under form-disjointness (no active violated form-6 entry writes to the same (role, output coordinate, argmax) triple as a form-7 entry, so their contributions cannot cancel) and positive derivative lower bounds,* $\|g^L\|^2 \geq \kappa \cdot \#\{$*active violated left-quantified pairs*$\}$ *for a constant* $\kappa > 0$ *determined by those bounds, linearly in the active left-existential count in pure* $\mathcal{EL}$*. In particular one active violated left-quantified pair already certifies* $g^L \neq 0$*, the hypothesis of Thm. 3. (Prop. 13 gives the general form-disjoint form; this is its uniform-constant specialization.)*

Corollary 12 (Single-branch routing of Kleene–Dienes)**.** *For Kleene–Dienes implication* $I_{KD}(u, v) = \max(1-u,\, v)$*, the gradient activates exactly one top-level branch away from the switching surface* $1-u = v$*: if* $1-u > v$*, then* $\partial_1 I = -1$ *and* $\partial_2 I = 0$*; if* $1-u < v$*, then* $\partial_1 I = 0$ *and* $\partial_2 I = 1$*.*

B.5 Proof of Corollary 11

We first state the structural lower bound on $\|g_U^L\|^2$ in the form-disjoint case; Cor. 11 is its uniform-constant specialization.

Proposition 13 (Structural lower bound on the antecedent gradient)**.** *Let* θ *be regular and satisfy the form-disjoint hypothesis: no form-6 active-violated scalar entry shares its* $(\text{role}, \text{output}, \text{argmax})$ *triple with any form-7 active-violated entry. Define* $g_6^{\text{Roles}(O)}$ *and* $g_7^{\text{Roles}(O)}$ *as the form-6 and form-7 components of* $g^L_{\text{Roles}(O)}(\theta)$ *respectively. Then*

$$\|g^L_{\text{Roles}(O)}(\theta)\|^2 \;\geq\; \beta_{\text{disj},6}(\theta)\, N_6^{\text{act}}(\theta)\, w_{\min}^2\, (\alpha_T^{\min}(\theta))^2 \;+\; \beta_{\text{disj},7}(\theta)\, N_7^{\text{act}}(\theta)\, w_{\min}^2\, (\alpha_S^{\min}(\theta))^2,$$

where $\alpha_T^{\min}(\theta)$ *and* $\alpha_S^{\min}(\theta)$ *are minimum t-norm/t-conorm derivative magnitudes on the active region,* $\beta_{\text{disj},6}$ *and* $\beta_{\text{disj},7}$ *are form-disjointness ratios in* $[0, 1]$*, and* $N_6^{\text{act}}, N_7^{\text{act}}$ *count active violated axioms in forms* 6 *and* 7 *respectively.*

Proof. Write $g^L_{\text{Roles}(O)}(\theta) = \sum_{i \in \mathcal{A}} w_i\, P_{\text{Roles}(O)} \nabla u_i(\theta)$ over the active-violated family $\mathcal{A}$, where $P_{\text{Roles}(O)}$ projects onto the role-output coordinate blocks. On the violated branch $\partial_1 I_{\text{Łuk}} = -1$ (§ 2.2), so each axiom's contribution to g^L is carried entirely by its antecedent gradient ∇u_i (Prop. 1). By the DF$\mathcal{ALC}$ side map, a form-6 axiom routes its role through the body t-norm and writes to a single $(\text{role}, \text{output}, \text{argmax})$ scalar coordinate with magnitude $|w_i\, \partial_{\text{role}} T| \geq w_{\min}\, \alpha_T^{\min}(\theta)$; a form-7 axiom routes through the body t-conorm and writes with magnitude $\geq w_{\min}\, \alpha_S^{\min}(\theta)$, where $\alpha_T^{\min}, \alpha_S^{\min}$ are the minimum t-norm/t-conorm derivative magnitudes on the active region.

Index the scalar entries of $g^L_{\text{Roles}(O)}$ by their $(\text{role}, \text{output}, \text{argmax})$ triple t, so that $\|g^L_{\text{Roles}(O)}\|^2 = \sum_t \big(\sum_{i \text{ writing to } t} w_i\, \partial_i\big)^2$ is a sum of squared per-triple aggregates; cancellation can therefore occur only within a triple. The form-disjoint hypothesis states that no form-6 active entry shares a triple with any form-7 active entry, so form-6 and form-7 contributions occupy disjoint triples and cannot cancel one another. Let $\beta_{\text{disj},6}(\theta) \in [0, 1]$ (resp. $\beta_{\text{disj},7}(\theta)$) be the fraction of the N_6^{act} form-6 (resp. N_7^{act} form-7) active entries whose triple carries no opposite-sign same-form entry, i.e. that survive intra-form cancellation. Restricting the sum over t to these surviving, mutually disjoint triples and lower-bounding each surviving triple by its per-axiom magnitude squared gives

$$\|g^L_{\text{Roles}(O)}\|^2 \;\geq\; \beta_{\text{disj},6}\, N_6^{\text{act}}\, (w_{\min}\alpha_T^{\min})^2 \;+\; \beta_{\text{disj},7}\, N_7^{\text{act}}\, (w_{\min}\alpha_S^{\min})^2,$$

the stated bound. (The exact consequent-silent specialization—$\beta_{\text{disj}} = 1$ on an owned subspace—recovers $\|g^L\|^2$ as the squared aggregate antecedent field of Prop. 1.) □

Cor. 11 now follows by setting $\beta_{\text{disj},6} = \beta_{\text{disj},7} = \beta_0$, $w_{\min} = 1$, and $\alpha_T^{\min} = \alpha_T^*$, $\alpha_S^{\min} = \alpha_S^*$, and absorbing $w_{\min}^2$ into β_0. □

B.6 Proof of Proposition 5

Proof of Prop. 5. Let n be the number of ontologies, n_b the number of genuinely profitable switches, and $\beta = n_b/n$. Among the profitable cases, a policy with recall r switches on rn_b ontologies in expectation. Since precision is the fraction of selected ontologies that are genuinely profitable, the expected total number of switched ontologies is $(rn_b)/p$. Hence the expected number of harmful switches is

$$\frac{rn_b}{p} - rn_b \;=\; rn_b \frac{1-p}{p}.$$

Each correct switch contributes mean gain g, while each harmful switch incurs mean loss h relative to the Łuk-R default. Therefore the expected total SR improvement is

$$rn_b g - rn_b \frac{1-p}{p} h \;=\; rn_b g\Big(1 - \frac{h}{g}\frac{1-p}{p}\Big).$$

Dividing by n gives

$$\mathbb{E}[\Delta] \;=\; \beta r g\Big(1 - k\frac{1-p}{p}\Big),$$

where $k = h/g$. Because $\beta, r, g > 0$ for any nontrivial switching policy, the sign is determined entirely by the bracketed term:

$$1 - k\frac{1-p}{p} > 0 \iff p > \frac{k}{k+1} =: p^*.$$

Recall appears only as the multiplicative factor $\beta r g$, so lowering recall rescales the magnitude of the improvement but cannot change its sign. When $k \gg 1$, we have $p^* = 1 - \frac{1}{k+1} \approx 1 - \frac{1}{k}$, yielding the near-perfect precision requirement claimed in the proposition. □

B.7 Proof of Proposition 4

Fix an axiom-individual pair and write $u = u_i(\theta)$, $v = v_i(\theta)$. Assume the point is regular and away from the saturation boundary $v = u$; the boundary is a codimension-one set and is handled by one-sided derivatives or subgradients. On the satisfied branch $v > u$, hard Łuk-R is the constant function

$$I_{\text{Łuk}}(u, v) = \min(1, 1 - u + v) = 1.$$

Therefore

$$\partial_1 I_{\text{Łuk}}(u, v) = 0, \qquad \partial_2 I_{\text{Łuk}}(u, v) = 0 \qquad (v > u).$$

The pair loss is $\ell_i(\theta) = 1 - I_{\text{Łuk}}(u_i(\theta), v_i(\theta))$. For any parameter block ϕ that enters u_i or v_i, the chain rule gives

$$\nabla_\phi \ell_i = -\partial_1 I_{\text{Łuk}}(u_i, v_i)\, \nabla_\phi u_i - \partial_2 I_{\text{Łuk}}(u_i, v_i)\, \nabla_\phi v_i = 0.$$

Hence every satisfied pair on the constant branch contributes no gradient to the logic loss.

For the pure satisfaction objective, each pair loss is nonnegative because fuzzy implications take values in $[0, 1]$. If all pairs satisfy $v_i \geq u_i$ under the saturated implication, then every pair has $I_{\text{Łuk}}(u_i, v_i) = 1$ and the total satisfaction loss is $L^H = 0$, a global minimum. The claim is only this global-minimum and gradient-support statement: it does not assert that every stationary point is satisfying, since nonsmooth boundaries, projection to $[0, 1]$, and cancellations across unsatisfied pairs can also create stationary behavior.

Now consider a composite objective $L = L^H + \lambda L^{\text{task}}$ whose task term depends on the same grounding parameters. On any satisfied pair from the constant branch, the contribution of L^H to the shared-gradient path is zero by the calculation above, so the composite gradient through that pair is supplied only by L^{task}. This is the sense in which a saturated implication is well matched to pure satisfaction but can stop providing auxiliary signal once the logic constraint is already satisfied.

Reichenbach provides the contrast. For

$$I_{\text{Rei}}(u, v) = 1 - u + uv,$$

we have

$$\partial_1 I_{\text{Rei}}(u,v) = -(1-v), \qquad \partial_2 I_{\text{Rei}}(u,v) = u.$$

Thus the Reichenbach pair loss $1 - I_{\text{Rei}}(u,v) = u(1-v)$ has consequent derivative $-u$, which is nonzero whenever $u > 0$, including on fuzzy-order-satisfied points $v \geq u$ unless the antecedent is zero. It can therefore retain gradient support in regions where a saturated implication has already become constant. This proves the proposition.

Consistency note. The proposition is a gradient-support result, not a convergence-rate theorem. It matches the objective-sweep reading in § 4: saturation is compatible with exact satisfaction, while discrimination objectives can require continued signal after such constraints go inactive. □

B.8 Proof of Proposition 6

The Łukasiewicz t-norm is

$$T_L(a,b) = \max(0, a+b-1).$$

On the open half-plane $a + b < 1$, the maximum selects the constant branch 0. Therefore

$$\nabla T_L(a,b) = (0,0) \qquad \text{whenever } a+b<1.$$

On the open half-plane $a + b > 1$, the selected branch is $a + b - 1$ and the gradient is $(1, 1)$. The kink $a + b = 1$ is the only nondifferentiable boundary; it has Lebesgue measure zero in $[0, 1]^2$ and can be treated with the convex subdifferential $\text{co}\{(0, 0), (1, 1)\}$ if needed.

For the softmax claim, let $p \in \Delta^{K-1}$ be a softmax output over $K \geq 3$ classes. At the uniform point $p_k = 1/K$,

$$\pi_i + \pi_j = \frac{2}{K} \leq \frac{2}{3} < 1$$

for every pair of coordinates i, j. Thus every body t-norm $T_L(\pi_i, \pi_j)$ lies on the constant-zero branch and emits zero gradient to both selected coordinates. More generally, the same conclusion holds for any pair satisfying $\pi_i + \pi_j < 1$. In particular, the conclusion is stable in a neighborhood of the uniform point: if

$$\|p - \tfrac{1}{K}\mathbf{1}\|_\infty < \frac{1 - 2/K}{2},$$

then $\pi_i + \pi_j < 1$ for all coordinate pairs. This formalizes the low-pinned classifier-grounding regime without claiming that every point in the simplex has all coordinate pairs in the dead zone.

For free coordinatewise grounding, the pair (a, b) ranges over the unit square rather than the simplex. The escape set is

$$E = \{(a,b) \in [0,1]^2 : a + b \geq 1\}.$$

It is the upper-right triangle of the unit square, with area

$$\int_0^1 \int_{1-a}^1 db\, da = \int_0^1 a\, da = \frac{1}{2}.$$

Thus free grounding has a positive-measure region on which the Łukasiewicz t-norm is not on the zero branch. The dead zone can still be entered under free grounding, but it is not forced by the geometry of the grounding itself.

Consistency note. The proposition separates routing from dead-zone severity. Form-2 conjunction can make the t-norm a real path on free-grounded ontologies, as in the fixed-evaluator conjunction sweep, while the dramatic Łukasiewicz collapse in classifier-grounded existential tasks requires the additional low-pinned condition established above. □

C General-ML Validation: Theory, Scorecard, and Statistics

Survival: the clipped-sum Łukasiewicz t-conorm collapses CIFAR-10 multiple-instance learning (MIL) localization to chance (instance AUC 0.47 vs. 0.91 for noisy-OR, 10 seeds) and CelebA conjunction likewise, the dead zone living in the aggregation, not the encoder. **A controlled ablation** isolates the operator: one CIFAR instance scorer is trained then frozen (AUC 0.92) and only the readout is swept across bag size K: mean dilutes as the predicted $K^{-1/2}$ (Fig. 10), `max` leads or ties at every K, and the Łukasiewicz t-conorm and noisy-OR saturate to chance; classic MIL and ERASER give the same-encoder controls for the other two task types. **And the negatives:** on distributed-signal text (ERASER, which formed the routing scope, and aspect-based sentiment analysis, ABSA, on SemEval-2014 Task 4 restaurant reviews) averaging suffices. On ABSA this

contradicts the call committed before the run, which predicted that attention would win; we report it as a failed prospective prediction.

Roadmap. The general-ML validation in full: the substrate-independent routing/survival theory ladder, the eight-task predictive scorecard (including the negatives and the one failed prospective call), the end-to-end retrieval-augmented generation (RAG) pooling prescription, and per-benchmark setups, seeds, and statistics.

We lift the survival/routing statements to an arbitrary aggregation operator $A : [0,1]^K \to [0,1]$ combining per-instance/per-branch scores $s_i = \sigma(f_{\theta_i}(x_i))$, with loss $L = \ell(A(s), y)$. *We deliberately do not claim one theorem for "arbitrary differentiable A"*: the operators of interest are not all smooth, so we split into three regimes, each with its own statement and proof.

C.1 Quantitative Survival Thresholds

Proposition 14 (Operator-dependent detection thresholds under sparse signals)**.** *On the Gaussian sparse-witness bag (K items per bag, of which $m{=}K^{1-\beta}$ carry the signal) at shift μ, the minimal signal $\mu^*(A,\beta)$ at which a readout A makes the witnesses detectable (hence learnable) is operator-dependent: $\mu^*_{\text{mean}} \asymp K^{\beta-1/2}$ (the $K^{-1/2}$ dilution), $\mu^*_{\max} \asymp \sqrt{2\log K}\,(1-\sqrt{1-\beta})$ (logarithmic), the detection-optimal pooling operator crosses from mean to max at $\beta^* \to 1/2$, and a witness-vanishing (saturated / dead-zone) operator has $\mu^*{=}\infty$.*

Proof. On the Gaussian sparse-witness bag (positive bag: $m{=}\lceil K^{1-\beta}\rceil$ witness coordinates $s_i \sim \mathcal{N}(\mu,1)$, the rest $\mathcal{N}(0,1)$; negative bag: all $\mathcal{N}(0,1)$), let $\mu^*(A,\beta)$ be the smallest signal *order* at which the likelihood-free test $\{A(s) > t\}$ separates the two hypotheses with both error probabilities $\to 0$ as $K \to \infty$. *Mean (exact).* $A_{\text{mean}} \mid H_0 \sim \mathcal{N}(0, 1/K)$ and $\mid H_1 \sim \mathcal{N}(\mu m/K, 1/K)$; both errors vanish iff the standardized gap $\mu m/\sqrt{K} \to \infty$, so $\mu^*_{\text{mean}} \asymp \sqrt{K}/m = K^{\beta-1/2}$. *Max (extreme value).* Write u_n for the Gaussian-maximum centering ($u_n = \sqrt{2\log n} + o(1)$). Then $\max_j s_j \mid H_0$ concentrates at u_K with Gumbel width $O(1/u_K)$, while $\max_j s_j \mid H_1 \geq \mu + u_m$. A threshold placed in the quantile gap separates the two iff $\mu - (u_K - u_m) \gg 1/u_K + 1/u_m$, so $\mu^*_{\max} \asymp u_K - u_m = \sqrt{2\log K}\,(1-\sqrt{1-\beta})$ (using $b_m = \sqrt{1-\beta}\,b_K$ for $m = K^{1-\beta}$): a *difference of square roots*, $\Theta(\sqrt{\log K})$ for fixed β and logarithmic in K throughout (top-q, the mean of the q largest scores, and sparsemax are conjectured to share this order; not proved here). *Crossover.* $K^{\beta-1/2} \asymp \sqrt{2\log K}\,(1-\sqrt{1-\beta})$ at $\beta^* = \frac{1}{2} + \frac{\log\log K}{2\log K} \to \frac{1}{2}$: mean is detection-optimal below the boundary, max above it. *Saturation ($\mu^* = \infty$).* If $\partial A/\partial s_i = 0$ for $i \in W$ on a region that contains the positive witnesses with probability $\to 1$ (the Łuk-R satisfied branch; the softmax $K{\geq}3$ dead zone $\pi_i{+}\pi_j{=}2/K{<}1$, Prop. 6), then $A \mid H_1$ and $A \mid H_0$ have the same law in the witness signal and no test based on A detects: $\mu^* = \infty$ for every β — the survival failure of Props. 4–6 placed on the μ^* axis. *Scope:* leading-order in the Gaussian model with iid coordinate noise (the noise term is $\sqrt{\nabla A^\top \Sigma \nabla A}$ in general); a *survival* (detectability) statement, not a separation/optimality one (the Higher-Criticism separation remains a negative scope result, App. C.1). □

End-to-end prescription (RAG pooling). On real retrieval-augmented text classification (Fig. 7), a concentrated or sharp-τ readout (top-q or tuned attention) improves depth-512 accuracy over the mean-pooling default in every dataset×encoder cell, with the largest gains in the sparse 20-class setting: mean accuracy drops $0.090/0.081$ at 20 classes vs. $0.022/0.014$ at 4, and the *best* concentrated/sharp-τ readout's gap over mean at $K{=}512$ is $+0.111/+0.097$ vs. $+0.023/+0.016$ (MiniLM/MPNet; all 95% bootstrap CIs over seed × target excluding 0). The mechanism: such classifiers aggregate the votes of a query's top-K neighbors, and as K grows the on-class fraction shrinks (sparse witnesses), so Prop. 14 predicts the mean dilutes while a concentrated readout survives—a usable rule beyond the frozen-scorer crossover (Fig. 6). Here top-1/max is the clean sparse extreme—strongest on the 20-class data but, on the denser 4-class AG News, depth-robust yet not above mean, where the tuned attention carries the gap. The degradation magnitude is set by the dataset's class-sparsity and is invariant to the encoder, in the direction Prop. 14 predicts (an asymptotic sparse-witness statement, not an exact natural-retrieval magnitude law). This supports the survival-threshold prescription—concentrate rather than mean-pool deep retrieval—and is not a new RAG algorithm over known kNN/MIL dilution.

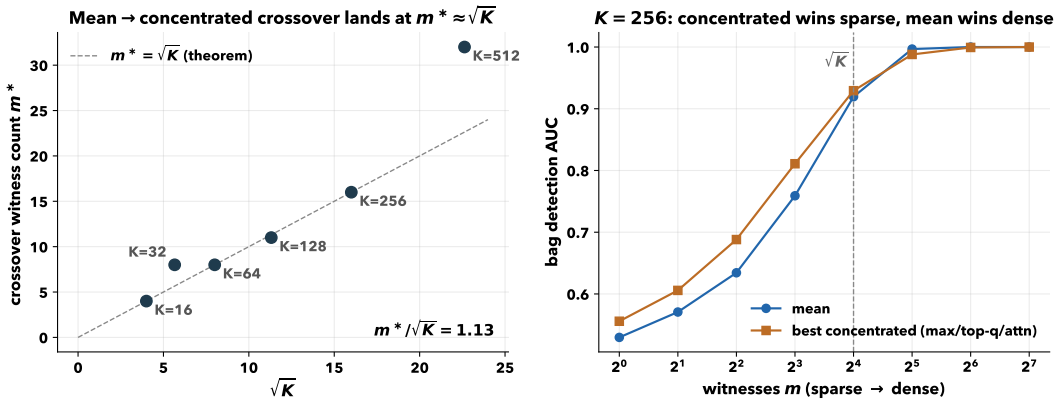


Figure 6: **Prop. 14 confirmed on real CIFAR-derived scores** (one frozen instance scorer, AUC 0.92; only the readout and the bag sparsity vary; 5 seeds). **Left:** the witness count m^* at which the detection-optimal readout flips from concentrated (max/top-q/attention) to mean tracks the predicted boundary $m^* \approx \sqrt{K}$ ($m^*/\sqrt{K} = 1.13$ over $K \in \{16, \ldots, 512\}$). **Right:** at K=256, concentrated readouts win in the sparse regime and mean wins in the dense regime, crossing at $m \approx \sqrt{K}$: the toy-model survival threshold realized on real-image aggregation.

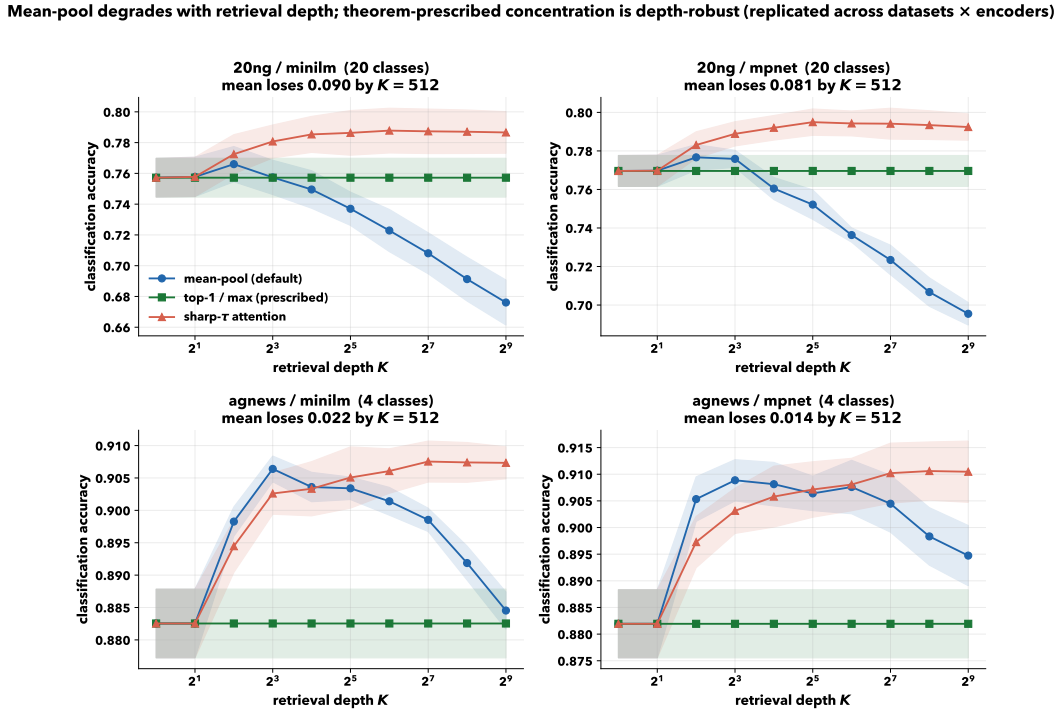


Figure 7: **The survival threshold as a usable RAG-pooling rule** (end-to-end retrieval-augmented classification; real 20 Newsgroups/AG News, frozen MiniLM/MPNet rerankers; natural top-K relevance, no planted evidence; 5 seeds). Mean-pooling the retrieved neighbors' votes (blue) degrades as retrieval depth K grows; the prescribed concentrated top-1/max (green) and sharp-τ attention (red) are depth-robust (attention strongest; top-1/max clearly above mean only in the sparse 20-class regime). The degradation magnitude is set by class-sparsity (20-class strong, 4-class mild) and is invariant to the encoder, in the direction the survival threshold predicts. Bands are ± 1 s.d. over seeds.

Regime I: smooth / piecewise-smooth (genuine gradient).

Proposition 15 (A1: dead zones from local constancy)**.** *Let A be locally constant on an ambient-open neighborhood $R \ni s(\theta_0)$ (for boundary points of $[0,1]^K$, on a local C^1 extension). Then $\nabla_\theta L(\theta_0) = 0$, so the loss term routes no gradient to the encoder while $s(\theta) \in R$. For $K \geq 2$ the*

Łukasiewicz t-norm $A_\wedge = \max(0, \sum_i s_i - (K-1))$ is flat ($A_\wedge \equiv 0$) on the open set $\{\sum_i s_i < K-1\}$ and the Łukasiewicz t-conorm $A_\vee = \min(1, \sum_i s_i)$ is flat ($\equiv 1$) on $\{\sum_i s_i > 1\}$.

Proof. A constant on the open set R gives $DA(s(\theta_0)) = 0$; the chain rule $\nabla_\theta L = \ell'(A, y)\, DA\, D_\theta s$ then vanishes. The two regions are open and nonempty for $K \geq 2$ (e.g. $s = \mathbf{0}$ and $s = \mathbf{1}$ respectively), and A equals a constant on each, so $DA = 0$ there. □

Proposition 16 (A2: gradient shares of smooth aggregators)**.** *If $A \in C^1$ and parameters are branch-disjoint ($\nabla_{\theta_i} s_j = 0$ for $j \neq i$), then $\nabla_{\theta_i} L = \ell'(A, y)\, (\partial A / \partial s_i)\, \nabla_{\theta_i} s_i$, so $\partial A / \partial s_i$ is branch i's gradient share. In particular $\partial_i A_{\text{noisyOR}} = \prod_{j \neq i}(1 - s_j)$ (self-damping: vanishes once another branch is confident), $\partial_i A_{\text{mean}} = 1/K$ (uniform), and $\partial_i A_{\text{prod}} = \prod_{j \neq i} s_j$ (boundary starvation when a co-factor $\to 0$, but no flat open region: product is not a dead zone).*

Proof. Branch-disjointness kills cross terms in the chain rule, leaving the single product. The three partials are direct derivatives of $1 - \prod_j (1 - s_j)$, $\frac{1}{K}\sum_j s_j$, and $\prod_j s_j$. For product, the zero set $\{s_j = 0\}$ has empty interior, so A_{prod} is constant on no open set. □

A2 says *where* gradient goes; the next lemma says *what each readout can recover*, formalizing the qualitative routing scope of App. C (attention necessary for localized signal, mean-pooling best for distributed signal) as a quantitative crossover in the set size K under an idealized score model. It is a statement about identifiability of the readout rather than about gradients; it is a formal scope lemma connecting standard aggregation statistics to the routing/survival framing, making the scope a consequence of the score model rather than an empirical coincidence.

Proposition 17 (A3: routing crossover between localized and distributed signals)**.** *Let a set of K items carry scores $s_i \in [0, 1]$ that are* independent within a set *with centered σ-sub-Gaussian noise of variance σ^2, read out by $A_{\text{mean}} = \frac{1}{K}\sum_i s_i$ or $A_{\max} = \max_i s_i$. Write the signal-to-noise ratio $\text{SNR} = |\mathbb{E}[A \mid \text{pos}] - \mathbb{E}[A \mid \text{neg}]| / \sqrt{\text{Var}[A \mid \text{pos}] + \text{Var}[A \mid \text{neg}]}$, the d' separability index (the Bayes threshold separation for Gaussian readouts; a moment proxy otherwise). **(a) Localized signal.** If positive sets have one* key *item of mean μ^+ and $K-1$ background items of mean $\mu^- < \mu^+$, negative sets have K background items, and all item noises share variance σ^2, then*

$$\text{SNR}_{\text{mean}} = \frac{\mu^+ - \mu^-}{\sigma\sqrt{2K}} = \Theta\big(K^{-1/2}\big), \qquad \text{SNR}_{\max} = \Omega(1),$$

*the latter in the margin-dominated regime $\mu^+ - \mu^- \geq 2\sigma\sqrt{2 \log K}$. Within that regime $\text{SNR}_{\max}/\text{SNR}_{\text{mean}} = \Omega(\sqrt{K})$: the mean readout's class separation vanishes as K grows while the max readout's persists (routing necessary). **(b) Distributed signal.** If the target is the population mean $y = \frac{1}{K}\sum_i z_i$ of item labels $z_i \in [0, 1]$, then A_{mean} is unbiased for y (with fixed noise $s_i = z_i + \varepsilon_i$, $\mathbb{E}[(A_{\text{mean}} - y)^2] = \sigma^2/K \to 0$: consistent), while the noiseless max readout $A_{\max} = \max_i z_i$ has bias $\mathbb{E}[\max_i z_i] - \mathbb{E}[y] \to \operatorname{ess\,sup} z - \mathbb{E}z > 0$ for non-degenerate i.i.d. z_i (inconsistent: its bias does not vanish as K grows; routing is harmful).*

Proof. **(a)** The positive-set mean is $\mu^- + (\mu^+ - \mu^-)/K$ and the negative-set mean is μ^-, so the means separate by $(\mu^+ - \mu^-)/K$; each readout averages K independent scores, so $\text{Var} = \sigma^2/K$ and the combined noise is $\sigma\sqrt{2/K}$, giving $\text{SNR}_{\text{mean}} = (\mu^+ - \mu^-)/(\sigma\sqrt{2K})$. For the max, centered σ-sub-Gaussian scores give $\mathbb{E}[\max \text{ of } m] \leq \mu + \sigma\sqrt{2 \log m}$, so $\mathbb{E}[A_{\max} \mid \text{neg}] \leq \mu^- + \sigma\sqrt{2 \log K}$ while $\mathbb{E}[A_{\max} \mid \text{pos}] \geq \mu^+$; under $\mu^+ - \mu^- \geq 2\sigma\sqrt{2 \log K}$ the separation is $\geq (\mu^+ - \mu^-)/2 = \Theta(\mu^+ - \mu^-)$ and the max fluctuation is $\Theta(\sigma)$, so $\text{SNR}_{\max} = \Omega(1)$ and the ratio is $\Omega(\sqrt{K})$. **(b)** $A_{\text{mean}} - y = \frac{1}{K}\sum_i \varepsilon_i$ has mean 0 and variance σ^2/K. For the noiseless max, $A_{\max} = \max_i z_i$; by the law of large numbers $y \to \mathbb{E}z$, and for non-degenerate i.i.d. z_i the maximum increases to $\operatorname{ess\,sup} z > \mathbb{E}z$, so the bias is bounded below by a positive constant independent of K. □

Both rates are confirmed numerically: the localized $\text{SNR}_{\text{mean}}/\text{SNR}_{\max}$ ratio tracks $K^{-1/2}$, and the distributed $\mathbb{E}|A_{\max} - y|$ rises to $\operatorname{ess\,sup} z - \mathbb{E}z$ while $\mathbb{E}|A_{\text{mean}} - y| \to 0$. This is exactly the bAbI/MIL (localized ⇒ routing necessary) versus ERASER/ABSA (distributed ⇒ mean necessary) scope pair, here placed on a formal footing under the idealized score model.

The mean/max dichotomy is the boundary of a one-parameter family. Softmax-attention pooling $A_\tau(s) = \sum_i w_i s_i$ with $w = \text{softmax}(s/\tau)$ has $A_\tau \to A_{\max}$ as $\tau \to \infty$ (uniform weights) and

$A_\tau \to A_{\max}$ as $\tau \to 0^+$ (weights concentrate on the argmax set), so temperature is a continuous *routing-strength* knob. This lets us state attention's value precisely: the temperature family *contains* a temperature that avoids the penalty a fixed wrong pool incurs—an expressivity (oracle-τ) statement; we do not claim gradient descent finds the optimal τ.

Proposition 18 (A4: temperature-interpolated attention and the fixed-pool penalty)**.** *Let A_τ be softmax-attention pooling, $\tau \in (0, \infty]$, with $A_\infty = A_{\text{mean}}$ and $A_{0^+} = A_{\max}$. **(a)** Under the localized model of Prop. 17(a), $\sup_\tau \mathrm{SNR}(A_\tau) \geq \mathrm{SNR}(A_{\max}) = \Omega(1)$ while $\mathrm{SNR}(A_{\text{mean}}) = \Theta(K^{-1/2})$, so the family separates the classes $\Omega(\sqrt{K})$ better than mean-pooling. **(b)** Under the distributed model of Prop. 17(b), $\inf_\tau \mathbb{E}[(A_\tau - y)^2] \leq \mathbb{E}[(A_{\text{mean}} - y)^2] = \sigma^2/K \to 0$, while the max endpoint carries bias $\to \operatorname{ess\,sup} z - \mathbb{E}z > 0$. Thus the penalty for a fixed wrong choice is $\Omega(\sqrt{K})$ separation (mean on localized signal) or $\Omega(1)$ bias (max on distributed signal).*

Proof. $A_\infty = A_{\text{mean}}$ and $A_{0^+} = A_{\max}$ are the standard softmax limits (weights tend to uniform, and to a one-hot indicator of the argmax when it is unique, else to the uniform distribution on the argmax set). $\tau{=}\infty$ lies in the family, giving $\inf_\tau \mathbb{E}[(A_\tau - y)^2] \leq \mathbb{E}[(A_{\text{mean}} - y)^2]$ immediately; for the localized case, $\mathrm{SNR}(A_\tau)$ is continuous in τ with a nonzero denominator and $A_\tau \to A_{\max}$, so $\sup_\tau \mathrm{SNR}(A_\tau) \geq \lim_{\tau \to 0^+} \mathrm{SNR}(A_\tau) = \mathrm{SNR}(A_{\max})$. The endpoint rates are Prop. 17(a,b). □

The bound only brackets the family by its endpoints; a temperature sweep confirms the interior is well-behaved—under localized signal $\mathrm{SNR}(A_\tau)$ increases as τ falls toward the routing endpoint and the small-τ advantage over mean grows with K, while under distributed signal the MSE is minimized at $\tau \to \infty$. This is why the family *brackets* the fixed-pool penalty: it contains the signal-appropriate routing strength, of which mean and max (the ERASER/ABSA and bAbI endpoints) are the two extremes. We do not claim a closed form for the optimal τ, nor that gradient descent attains it; the contribution is the bracketing and its tie to the routing/survival framing.

An information-theoretic gap: a sparse-detection view of pooling choice. A3/A4 compare specific readouts; the next result removes the readout from the picture and asks what any method can do, exposing the operator-choice gap as an information-theoretic suboptimality of linear pooling rather than an artifact of a chosen pool. Cast localized-signal aggregation as single-spike detection: observe $s \in \mathbb{R}^K$ and test $H_0 : s \sim \mathcal{N}(0, \sigma^2 I_K)$ against $H_1 : s \sim \mathcal{N}(\delta e_J, \sigma^2 I_K)$ for an unknown coordinate J (one item carries the signal).

Proposition 19 (A5: detection boundary and rate optimality of max routing)**.** *For the single-spike detection problem above, write $\delta = \sigma\sqrt{2r \log K}$. (i) The only permutation-invariant linear pooling statistic (Zaheer et al., 2017) is $T_{\text{lin}} \propto \sum_i s_i$, and it has power bounded away from its level only if $\delta = \Omega(\sigma\sqrt{K})$ (i.e. $r = \Omega(K/\log K)$). (ii) The extreme-sparse detection problem has a phase transition at $r = 1$: no test of any form has nontrivial asymptotic power for $r < 1$, while the routing statistic $T_{\max} = \max_i s_i$ (Bonferroni) has power for $r > 1$ and is first-order minimax optimal in this regime (Arias-Castro et al., 2011; Ingster, 1997; Donoho & Jin, 2004). (iii) Hence linear pooling is suboptimal by the factor $\delta_{\text{lin}}/\delta^* = \Theta(\sqrt{K/\log K}) \to \infty$, a gap intrinsic to the pooling nonlinearity, not to a particular architecture. (iv) The soft-pool $\mathrm{LSE}_\tau(s) = \tau \log \sum_i e^{s_i/\tau}$ (the log-sum-exp analog of the temperature-τ attention pool of Prop. 18, sharing its mean/max endpoints) satisfies $\max_i s_i \leq \mathrm{LSE}_\tau \leq \max_i s_i + \tau \log K$, so with $\tau_K \log K \to 0$ it inherits the max test's boundary attainment, while $\tau \to \infty$ recovers linear pooling; the routing strength that closes the gap is the attention temperature of Prop. 18.*

Proof. (i) A linear functional $\sum_i c_i s_i$ invariant to relabeling the exchangeable coordinates must have all c_i equal, hence $\propto \sum_i s_i$; under H_0 this is $\mathcal{N}(0, K\sigma^2)$ and under H_1 it is $\mathcal{N}(\delta, K\sigma^2)$, with standardized separation $\delta/(\sigma\sqrt{K})$, which is $o(1)$ unless $\delta = \Omega(\sigma\sqrt{K})$. (ii) is the extreme-sparse (1-of-K) case of the sparse-normal-means detection phase diagram: with $\delta = \sigma\sqrt{2r \log K}$ the boundary is at $r = 1$, below which all tests are asymptotically powerless and above which the maximum/Bonferroni test has power and is first-order minimax optimal; we cite these as sparse-detection phase-diagram references (Higher Criticism is the adaptive optimum at *intermediate* sparsity and is not needed for the single-spike max regime). (iii) is the ratio of (i) and (ii). (iv) The sandwich $\max_i s_i \leq \mathrm{LSE}_\tau \leq \max_i s_i + \tau \log K$ is immediate, so for $\tau_K \log K \to 0$, $\mathrm{LSE}_\tau \to \max_i s_i$ uniformly and inherits its boundary; for fixed K, $\mathrm{LSE}_\tau(s) = \tau \log K + \bar{s} + O(\max_i s_i^2/\tau)$ as $\tau \to \infty$, i.e. linear pooling up to an additive constant that does not affect the test. □

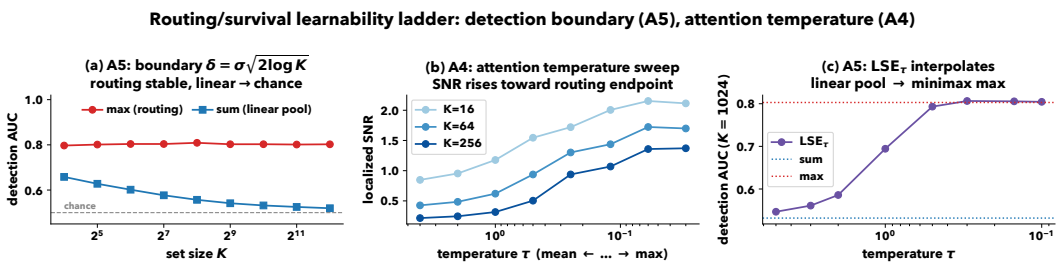


Figure 8: **The routing/survival ladder (Props. 17–19), verified on the idealized score model. (a)** A5: at the detection boundary $\delta = \sigma\sqrt{2\log K}$, the max (routing) test holds AUC $\approx$0.8 as K grows while the sum (linear pool) test decays to chance, the $\sqrt{K/\log K}$ separation. **(b)** A4: localized-signal $\mathrm{SNR}(A_\tau)$ versus attention temperature for three set sizes; SNR rises toward the routing endpoint (small τ), and the gain over mean grows with K. **(c)** A5: the soft-pool LSE_τ interpolates the detection AUC from the linear-pool level (large τ) up to the minimax max level (small τ), rising from 0.56 at τ=3 to 0.81 at τ=0.1.

Numerically (Fig. 8): at $\delta = \sigma\sqrt{2\log K}$ (i.e. $r = 1$) the max test holds AUC $\approx$0.80 for K from 16 to 4096 while the sum test decays to 0.5 (chance); LSE_τ at that δ rises from 0.56 (τ=3, sum-like) to 0.81 (τ=0.1, max-like). This connects fixed linear pooling to the classical sparse-detection boundary (Arias-Castro et al., 2011; Ingster, 1997; Donoho & Jin, 2004): for localized signal, tail-sensitive routing (max or low-temperature LSE) attains the single-spike detection rate, whereas linear permutation-invariant pooling is information-theoretically *suboptimal*, requiring a factor $\Theta(\sqrt{K/\log K})$ more signal. The boundary is classical; our role is to *identify* operator choice (and its attention-temperature interpolation) with it. The oracle-family negative in App. C.1 shows this is a *fixed-linear*-vs-routing-strength gap, not a claimed separation between tuned softmax and order-statistic families. This complements the discriminative-power view of Xu et al. (2019) (sum-pooling separates multisets that mean/max cannot): for the orthogonal task of *localized detection* the ordering reverses—max/low-temperature attention is rate-optimal and the linear sum is not—so neither aggregator dominates across tasks, exactly the no-universal-best thesis.

Scope of the detection-boundary connection (a pre-registered negative). A natural *stronger* claim would be an operator-*family* separation: that the value-weighted softmax family $\mathcal{F}_{\mathrm{sm}}$ (temperature-τ attention, LSE) is suboptimal on an intermediate-sparsity band where order-statistic pools $\mathcal{F}_{\mathrm{os}}$ (top-q, Higher Criticism) are strictly better. We tested this directly, and report that it does not hold—a deliberate scope check so the A5 connection is not over-read. On the canonical sparse-normal-means model (where the Donoho–Jin phase diagram is known), with $K \in \{256, 1024, 4096\}$, sparsity exponents $\beta \in \{0.5, \ldots, 1.0\}$, signal at 1–2$\times$ the detection boundary, and *each family given its oracle knob* (best τ for $\mathcal{F}_{\mathrm{sm}}$, best q for top-q, parameter-free Higher Criticism on the known null), the separation $\Delta = \mathrm{AUC}_{\mathcal{F}_{\mathrm{os}}} - \mathrm{AUC}_{\mathcal{F}_{\mathrm{sm}}}$ never reaches our pre-registered 0.02 threshold in any β-band; in the moderately-sparse band (β=0.6) it is significantly *negative* and the softmax advantage *grows* with K ($\Delta = -0.011, -0.013, -0.021$ at $K = 256, 1024, 4096$; 20 seeds, 10k-bootstrap 95% CIs all excluding 0)—i.e. a temperature-tuned softmax *beats* Higher Criticism, by a margin that increases with K (AUC 0.82 vs. 0.78 at K=4096), the opposite of the hypothesized order-statistic advantage. The reason is that softmax-with-temperature is already an adaptive, near-likelihood-ratio detector for this Gaussian-location model, so it is not pinned to the mean/max endpoints. We therefore claim only the *fixed-linear*-vs-routing gap of A5 and read the boundary as a connection, not a separation we own; the operative axis is fixed-vs-learnable routing *strength*, consistent with the attention-temperature result (Prop. 18).

Regime II: subgradient / kink (a.e. differentiable).

Proposition 20 (B1: single-branch routing of max and min)**.** *Where the argmax is unique, $A_{\max} = \max_i s_i$ is differentiable with $\nabla A_{\max} = e_m$, $m = \arg\max_i s_i$ (gradient routes to one branch; the rest are starved). The tie set $\bigcup_{i<j}\{s_i = s_j\}$ has Lebesgue measure zero, and there the Clarke subdifferential is $\partial_C A_{\max} = \mathrm{conv}\{e_i : i \in \arg\max\}$; min is symmetric (concave) with $\partial_C A_{\min} = \mathrm{conv}\{e_i : i \in \arg\min\}$.*

Proof. A unique maximizer with margin $\delta > 0$ persists on a ball of radius $\delta/2$, where $A_{\max}(s+h) = s_m + h_m$, giving $\nabla A_{\max} = e_m$. $A_{\max}$ is the pointwise max of the linear forms $s \mapsto s_i$; the subdifferential of a finite max of affine functions at a point is the convex hull of the gradients of the active pieces, i.e. $\mathrm{conv}\{e_i : i \in \arg\max\}$. Each tie set is the intersection of a hyperplane with $[0,1]^K$, and so has Lebesgue measure zero. □

Proposition 21 (B2: Clarke subdifferential at the Łukasiewicz clamp)**.** *At the boundary* $\{\sum_i s_i = K-1\}$ *of* $A_\wedge$ *(resp.* $\{\sum_i s_i = 1\}$ *of* $A_\vee$*) the Clarke subdifferential is* $\mathrm{conv}\{\mathbf{0}, \mathbf{1}_K\} = \{\alpha \mathbf{1}_K : \alpha \in [0,1]\}$.

Proof. $A_\wedge = \max(0,\, g(s))$ with $g(s) = \sum_i s_i - (K-1)$ affine, $\nabla g = \mathbf{1}_K$. On the kink the two active pieces are 0 (gradient $\mathbf{0}$) and g (gradient $\mathbf{1}_K$); the Clarke subdifferential of $\max$ of two affine pieces is the convex hull of their gradients. $A_\vee$ is analogous. □

Regime III: straight-through / surrogate (biased update, not ∂A).

Proposition 22 (C1: self-reinforcing dead components under hard top-1 routing)**.** *Let hard top-1 routing select component* $m = \arg\max_j q_j$ *with surrogate update* $\tilde{g}_{\theta_j} = h_j\, G_j$ *(*h_j *the selection weight,* $h_j = 0$ *if* $j \neq m$*). Suppose over a training interval: (i) the data are fixed; (ii) no shared parameter affects any* q_j*; (iii) no surrogate gradient reaches gate parameters in a way that changes any gap* $q_j - q_m$*; (iv) no momentum/weight-decay/noise acts on the relevant variables; (v) ties are broken deterministically. Then an unselected component* j *receives* $\tilde{g}_{\theta_j} = 0$ *at every step and never becomes selected: collapse changes* utilization*, not the accuracy of the selected path.*

Proof. Under (i)–(v) the selection map is constant over the interval, so for $j \neq m$, $h_j \equiv 0$ and $\tilde{g}_{\theta_j} \equiv 0$; θ_j is unchanged, q_j is unchanged (ii,iii), and the ordering $q_j < q_m$ is preserved, so j stays unselected. The selected path's parameters update normally, so task accuracy is governed by the live components only. □

The fuzzy-$\mathcal{ALC}$ results of § 3 are exactly the instances: Prop. 4 and Prop. 6 are Regime-I (A1) statements for the implication's satisfied/dead regions; Cor. 12 (KD branch starvation) and the $\max/\min$ role-gradient gaps (Thm. 3) are Regime-II (B1) selections; B2 characterizes the kink at the *boundary* of the A1 dead zone that the CIFAR/CelebA Łukasiewicz collapses sit against (training drives $\sum_i s_i$ across it without crossing back). The statements above were verified numerically against the released implementation.

Predictive scorecard and scope checks. The scorecard (Table 3) matches on seven of eight aggregation tasks, with evidential status marked per row. Survival predictions are a priori (Prop. 15); the routing scope was formed on ERASER (DeYoung et al., 2020) and prospectively confirmed on bAbI (predicted attention-necessary), while the prospective ABSA call, which predicted that attention would beat mean pooling, failed (mean 0.70 vs. attention 0.58–0.63), after which we narrowed the routing scope to signals that cannot be injected without routing; the softmax attention these benchmarks use follows Vaswani et al. (2017). Per-benchmark setups, seeds, and provenance manifests are listed below, and Fig. 9 shows every benchmark's operator comparison at a glance (including the secondary classic-MIL replication and the ERASER/ABSA negatives, where no operator collapses).

Statistics. Bootstrap 95% CIs (10k resamples over seeds; classic-MIL over seed$\times$fold) exclude 0 for every headline contrast. With 10 seeds (paired sign-test $p \leq 0.004$): CIFAR Łuk$-$noisy-OR $\Delta = -0.433\,[-0.498, -0.362]$; CelebA Łuk$-$product $\Delta = -0.218\,[-0.257, -0.178]$; bAbI softmax$-$mean $\Delta = +0.554\,[+0.550, +0.557]$; CREMA-D max$-$mean (visual) $\Delta = -0.092\,[-0.110, -0.074]$ (negative on all 10; joint $\Delta = -0.035\,[-0.051, -0.021]$). With 5 seeds, ERASER softmax$-$mean $\Delta = -0.056\,[-0.085, -0.029]$ (the predicted distributed-signal scope). All real-data runs are written append-only with host/commit/seed provenance manifests; the synthetic MoE check is an exploratory pilot.

Controlled operator-only ablations (one per task type). Where the question is operator choice at a *fixed* representation, three benchmarks are run as controlled ablations—the encoder/scorer is shared and frozen (or cross-validated) and only the aggregation operator is swapped, so any gap is attributable to the operator alone (same encoder, same compute). (i) *MIL over images*: one CIFAR instance

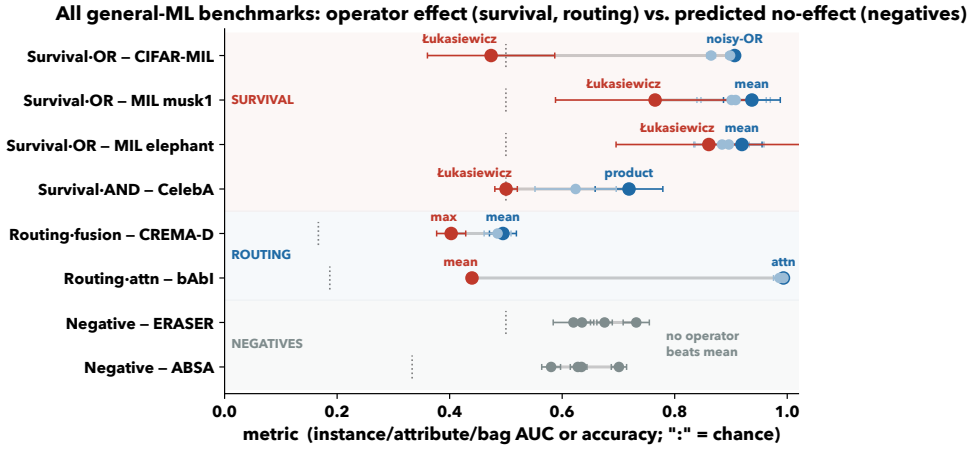


Figure 9: **Operator choice matters on the localized-signal benchmarks and not on the distributed-signal ones.** All general-ML benchmarks as a dumbbell: each benchmark's operators on its metric axis (mean ± SD over seeds; dotted tick = chance), so the winner→regime-wrong-operator span *is* the operator effect. **Survival:** the Łukasiewicz operator (red) collapses toward chance on the disjunctive MIL benchmarks (dramatically on CIFAR-10, modestly but consistently on classic MUSK1/Elephant) and on conjunctive CelebA. **Routing:** `max` starves the weak modality on CREMA-D and uniform mean pooling fails the bAbI retrieval task. **Negatives:** on the two distributed-signal benchmarks (ERASER, ABSA) no routing operator beats mean pooling, which an "operators always matter" account cannot produce; on ABSA this contradicts the prospective call (Table 3).

scorer is trained then frozen (instance AUC 0.92) and the readout operator is swept across bag size K (Fig. 10); mean dilutes as $K^{-0.466}$ (R^2=0.999), `max` leads or ties at every K, the Łukasiewicz t-conorm saturates to chance by K=16, and noisy-OR by K=256. (ii) *MIL over feature vectors*: on MUSK1 and Elephant a shared MLP scorer is cross-validated (5 seeds × 5 folds) with only the pooling operator varied; the Łukasiewicz t-conorm is the worst aggregator on both (0.77/0.86 vs. mean 0.94/0.92). (iii) *Text attention*: on ERASER (movies) a shared embedding and query are fixed and only the attention operator is swapped; mean wins (0.73 vs. softmax 0.68, sparsemax 0.62)—the predicted distributed-signal negative. Each isolates the operator at a fixed encoder, the cleanest form of the external-validity check.

Table 3: **The geometry predicts the operator outcome on seven of eight aggregation tasks; one prospective call failed.** Predictive scorecard: the geometry's prediction vs. the observed outcome across eight aggregation tasks spanning images, molecules, faces, audio, and text, one of them a synthetic pilot. Seven of eight match; the prospective ABSA call, which predicted a routing effect, failed. Two rows are *negatives* where operator choice does not matter (ERASER, which formed the scope, and the mixture-of-experts (MoE) pilot). The **Sign** column separates the two predicted outcomes; the **Status** column is orthogonal and states each row's evidential status, not just a match: **prospective** = the call was committed to version control before the run; **confirmatory** = an a-priori theory prediction verified in-infrastructure; **scope** = the benchmark that *formed* the routing scope hypothesis (not an independent test); **exploratory** = a synthetic pilot.

Benchmark	Prediction (regime)	Sign	Observed	Status
CIFAR-10 MIL (OR)	Łuk t-conorm collapses (A1)	matters	inst-AUC 0.47 vs 0.91	✓confirmatory
Classic MIL (OR)	Łuk worst (A1)	matters	worst on both sets	✓confirmatory
CelebA (AND)	Łuk t-norm collapses (A1)	matters	attr-AUC 0.50 vs 0.72	✓prospective
CREMA-D fusion	max starves weak modality (B1)	matters	vis 0.40 vs 0.50, 10/10	✓confirmatory
bAbI task1	routing necessary (retrieval)	matters	attn 0.99 vs mean 0.44	✓prospective
ERASER (distributed)	routing unnecessary	**negative**	mean 0.73 > attn	✓scope
ABSA (distributed)	routing helps (attn > mean)	matters	mean 0.70 > attn	× prospective
MoE (slack capacity)†	utilization only (C1)	**negative**	acc unchanged	✓exploratory

†Synthetic pilot (Regime-III check with a straight-through estimator, STE); the other seven are provenance-tracked real-data benchmarks.

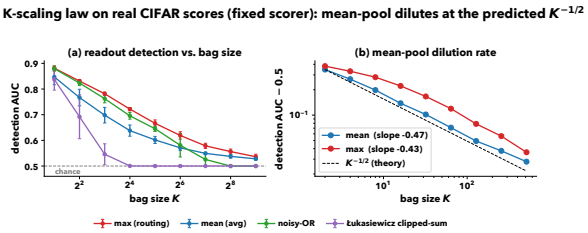


Figure 10: **Controlled MIL operator ablation (one frozen instance scorer, AUC 0.92; only the readout operator swapped).** Detection AUC vs. bag size K on real CIFAR instance scores: mean dilutes as $K^{-0.466}$ (R^2=0.999, predicted $K^{-1/2}$), `max` leads or ties at every K, while the Łukasiewicz t-conorm and noisy-OR saturate (every bag scores 1.0), so detection falls to chance (AUC 0.5): a same-encoder, operator-only test of the survival prediction.

D Prior-Work Comparison and Scope Boundaries

Roadmap. Point-by-point positioning against prior work, and the explicit scope boundary of each claim (what we do and do not assert).

Positioning of the general-ML validation. A note on learned aggregation modules (learned OR/AND blocks, learned pooling): fixed analytic operators expose predictable routing/survival geometry precisely because their partials are known a priori; learned modules absorb the routing knob into trainable parameters, trading predictability for adaptivity. Our two pre-registered negatives quantify this: a learnable scorer absorbs the attention temperature, and a temperature-tuned softmax already attains the order-statistic detection optimum (App. C.1)—so the analysis here applies at design time to the fixed-operator layer, and predicts when a learned aggregator will simply re-learn the regime-appropriate fixed operator. Each general-ML benchmark (App. C) touches a known phenomenon; the contribution is the single predictive geometry that unifies them, not their discovery. van Krieken et al. (2022) established that fuzzy-operator gradients differ for learning (single operators); permutation-invariant pooling and the discriminative power of sum/mean/max aggregators are studied by Zaheer et al. (2017); Murphy et al. (2019); Xu et al. (2019); Wagstaff et al. (2019), attention-based MIL pooling by Ilse et al. (2018) and set attention by Lee et al. (2019); modality-competition work (OGM-GE (Peng et al., 2022)) established multimodal under-training and fixes; sparse attention (sparsemax/entmax (Martins & Astudillo, 2016; Peters et al., 2019)) and memory networks (Sukhbaatar et al., 2015; Weston et al., 2016) established, respectively, that attention sparsity and attention-based retrieval matter. We do not claim any of these. What is new is (i) one substrate-independent routing $\times$ survival law that predicts all of them—and predicts when an operator choice does not matter (out-of-sample scorecard, App. C, 7/8, with one failed prospective call); (ii) the survival dead zone as one mechanism across the conjunctive and disjunctive sides on real images/faces (the AND-side and real-image collapses not in prior single-operator analyses); (iii) recasting modality competition and attention-necessity as two faces of one allocation law with a predicted, out-of-sample-confirmed scope; and (iv) the three-regime placement (smooth / subgradient / surrogate) that says which operators admit a true-gradient statement. Three threads share our gradient-allocation lens. Learned/gated pooling (Lee et al., 2016) showed that the pooling *choice* reshapes gradient flow; we supply the structure/objective/grounding law behind it and the regime where a learned gate cannot help (the fixed-scale survival collapse). Winner-take-all credit-assignment pathologies in neural theorem provers (de Jong & Sha, 2019) are an instance of routing structurally starving a sub-module, here generalized beyond proof search to operator-level routing. And the implicit bias of gradient descent toward feature *averaging* (Li et al., 2025) is, on the aggregation side, exactly the survival cost our mean-pool dilution ($K^{-1/2}$, Prop. 17) and reasoning-shortcut analyses make operator-explicit and predictable.

Operator Space. The paper core uses a standard comparable space of $3\times3\times5 = 45$ operator combinations. An extended stress-test space adds 6 additional t-norms (Yager, Hamacher, nilpotent, Frank, drastic, Einstein) and 1 additional implication (Sig, $\sigma(10(v-u))$), yielding $9\times3\times6 = 162$ combinations. Claims are about this evaluated family, not about all possible fuzzy operators. App. F.3 reports an 8-structure, 960-run Frank/Yager/Hamacher sweep; it yields 0 strict wins over the best discrete anchor, with only two saturated ties, and the drastic t-norm fails on bottleneck paths (Fig. 11).

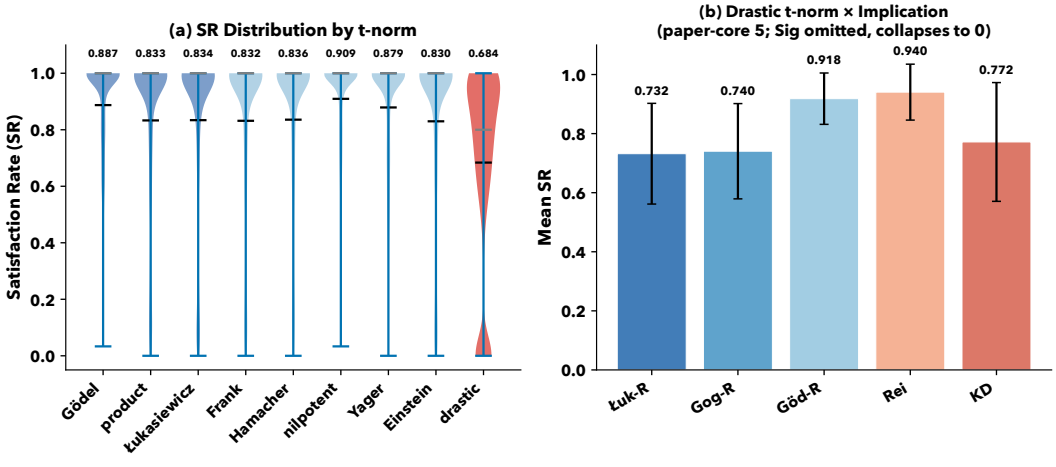


Figure 11: **Internal operator roles matter when they sit on a bottleneck path.** Structure-linked failure under the drastic t-norm. Gradients collapse when the t-norm occupies a bottleneck internal path, confirming that internal operator roles matter when they sit on critical routes (Prop. 1).

Mitigating an operator collapse: annealing and learned gates. A natural question is whether the survival collapses—e.g. the Łukasiewicz/clipped-sum t-conorm sitting at chance on localized-signal MIL—are an *expressivity* limit or a *gradient-survival* one, and whether learned gates or annealing curricula repair them. On a synthetic localized-signal MIL (instance scores initialized so $\sum_i s_i > 1$, saturating the clamp; 5 seeds) the clipped-sum t-conorm collapses (instance-localization AUC 0.53 ± 0.22) while noisy-OR survives (0.98). Two interventions recover it, each consistent with the prediction (Fig. 12): (i) *operator annealing* from noisy-OR to clipped-sum reaches 0.98 *while ending at the collapsing operator* (β=1)—the early surviving-operator phase supplies the gradient that separates the instances, after which the clamp is no longer saturated; so the collapse is survival/optimization, not expressivity, and the principle *prescribes* the fix (anneal from a non-saturated operator). (ii) A *learned convex gate* $\beta = \sigma(\theta)$ over the two operators reaches 0.98 by routing *away* from the collapsing operator ($\beta \to 0.09$): it replaces rather than rescues it. Neither repairs clipped-sum in place—both confirm the diagnosis and answer when a learned gate helps (it routes to a surviving operator; it cannot resurrect a dead-zone one). The same survival signature appears in the third operator family, the *quantifier aggregator*: on a synthetic task with a spurious label-correlated instance, max-like p-mean aggregators lock onto the spurious instance (localization AUC 0.79–0.85, high variance—some seeds fully shortcut) while mean-pooling resists (0.91), as the sparse-vs-dense-gradient survival argument predicts. Operator-induced shortcuts thus span all three families on the constraint path—implication, t-norm, and aggregator—one mechanism.

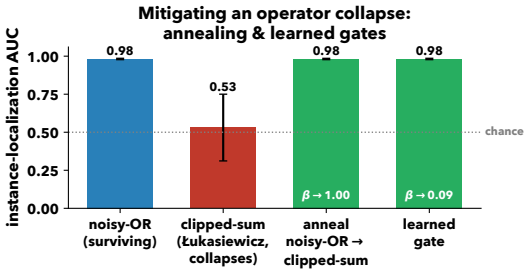


Figure 12: **An operator collapse is a survival/optimization failure, not an expressivity one** (synthetic localized-signal MIL, 5 seeds). The clipped-sum (Łukasiewicz) t-conorm collapses to chance; *annealing* noisy-OR→clipped-sum recovers while ending *at* the collapsing operator (β=1); a *learned gate* recovers by routing *away* ($\beta \to 0.09$).

Difference Table. Table 4 summarizes how this work differs from closely related prior work.

Novelty positioning. We claim: the first *unified* routing × survival account under which fuzzy operator optimality is a function of (structure × objective × grounding); the accompanying structure-

Table 4: **Point-by-point differentiation from related work.**

Reference	Structure-aware?	Object of study	Our difference
van Krieken et al. (2022)	No (aggregate)	Operator derivatives as intrinsic properties (vanishing, single-passing; the five-implication table)	We take their table and both terms as given, and make each conditional: single-passing becomes structure-relative parameter reachability (routing), vanishing becomes objective- and grounding-relative (survival). We add the descent-dominance theorems and the shortcut consequence.
He et al. (2024)	Loss-level	Implication bias in loss	We study operator choice, not loss design; we test a robustness matrix (§ 4.4)
Wu & Zhao (2025)	No	Framework + grounding	We study operator choice within an existing framework
Badreddine et al. (2023)	No	Operator configuration	We study which operator to choose, not how to stabilize a chosen one
Flinkow et al. (2025)	No (aggregate)	Logic comparison	We condition on structure; same operator ranks differently under different structures

conditioned ranking law with an implication-gated role decomposition; the survival propositions linking saturation to the objective and the Łukasiewicz dead zone to the grounding; and the satisfaction-regime safe-default barrier. We do not claim to be first to show that operators matter for learning—that is van Krieken et al. (2022), whose derivative table and *vanishing*/*single-passing* vocabulary we take as given (Table 4)—nor first to compare operators across frameworks (Flinkow et al., 2025) or to study fuzzy downstream tasks. Our advance is the unifying mechanism and the precise conditions under which each operator wins or fails.

E Implementation Details

Roadmap. Implementation details: the τ-surrogates and numerical stabilizers, corpus construction, the fixed training settings used throughout, and pointers from appendix results to scripts and data.

Several operators in the paper-core family involve hard $\max$/$\min$ or piecewise definitions that are not differentiable at switching boundaries. Our implementation uses standard smooth approximations with a fixed temperature parameter $\tau = 0.1$:

- **Softmax/softmin:** $\mathrm{smax}_\tau(x_1, x_2) = \frac{x_1 e^{x_1/\tau} + x_2 e^{x_2/\tau}}{e^{x_1/\tau} + e^{x_2/\tau}}$. Used for Gödel min/max, Łukasiewicz $\max(\cdot, 0)$/$\min(\cdot, 1)$, and KD $\max(1-u, v)$.
- **Sigmoid gating:** $\sigma((v-u)/\tau)$ replaces the hard indicator $\mathbf{1}[u \le v]$. Used for the Göd-R and Gog-R implications.
- **Division stabilizer:** $\varepsilon_{\mathrm{div}} = 10^{-6}$ prevents zero-division in Gog-R (v/u).

Several operators require no approximation because they are already smooth: the product t-norm ($T(a, b) = ab$), the Reichenbach implication ($I(u, v) = 1-u+uv$), and the Łuk-R implication uses a hard clamp $\min(1, 1-u+v)$ that is piecewise-linear and differentiable almost everywhere.

Training Parameterization. Concept and role embeddings are initialized with independent uniform draws from $[0.25, 0.75]$. The optimizer is `torch.optim.Adam` with only the learning rate overridden; $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ are PyTorch defaults. Training iterates over the seven normalized axiom forms; within each non-empty form, axioms are shuffled and processed in

batches of size 64, per-batch losses are accumulated, and one full-epoch optimizer step is taken. After every optimizer step, concept and role embeddings are clamped to $[0, 1]$; the special $\top$ and $\bot$ concept rows are fixed at 1 and 0 respectively. Evaluation is full-batch over all axioms in each form.

The training budgets are experiment-family specific. The 28,350-run synthetic archive, the 39,375-run BioPortal archive, and the 145,080-run ORE 2015 archive use learning rate 2×10^{-4}, batch size 64, a 25,000-epoch maximum, and early-stopping patience 10 evaluated at 100-epoch checkpoints (an improvement must lower the loss by at least 10^{-4} in absolute terms, so patience 10 allows 1,000 epochs without such an improvement). The D2 held-out subsumption prediction benchmark (§ 4.4) uses learning rate 10^{-3}, 500 maximum epochs, and patience 30; it is reported separately because it evaluates learned embeddings on a downstream task rather than contributing to the main SR variance decomposition. The older standard 45-grid synthetic run used the same learning rate and batch size with 5,000 maximum epochs and patience 3; it is retained as a historical robustness/provenance slice, not as the 28,350-run paper-facing synthetic archive.

Synthetic Level Sets and Deduplication. The one-factor synthetic corpus varies one structure dimension at a time around the default setting: scale $\{10, 30, 100, 200\}$, quantifier mix $\{\exists$-only, $\forall$-only, balanced$\}$, nesting depth $\{0, 1, 2, 3\}$, topology $\{$chain, tree, dag, with disjoint$\}$, conflict rate $\{0, 0.1, 0.3\}$, concept-pool fraction $\{0.3, 0.6, 1.0\}$, constraint density $\{0.5, 2.0, 5.0\}$, form homogeneity $\{$subsumption-only, existential-only, universal-only, diverse$\}$, individual ratio $\{0.5, 2.0, 5.0\}$, and negation density $\{0, 0.3, 0.6, 0.9\}$. These 35 levels are crossed with the 162-configuration extended grid and five repetitions, yielding the 28,350-run synthetic archive; the paper-core 45-grid subset has 7,875 runs. Synthetic runs are deduplicated by $(\text{dimension}, \text{level}, T, S, I, \text{repetition})$; real-corpus runs are deduplicated by $(\text{ontology}, T, S, I, \text{repetition})$.

Routing Topology Preservation. The gradient routing topology predicted by Prop. 1 holds exactly for the exact operators (product, Reichenbach, Łuk-R) and asymptotically for the soft-approximated ones: at $\tau = 0.1$, the losing branch of a softmax receives a gradient weight that decays exponentially with the branch gap, preserving the qualitative routing character described in the operator table in § 2.2. Quantifier aggregation ($\exists$: $\max$ over fillers, $\forall$: $\min$ over fillers) uses hard $\max/\min$ without softening, so the quantifier branch-selection lemma (App. B.1) holds exactly at non-tie points. App. F.8 reports a matched softened/stabilized aggregation ablation; it refines some local top-1 outcomes in quantifier-heavy settings, while a more aggressive `logmeanexp` variant fails on three representative structures rather than overturning the implication-dominant ordering.

Reproduction pointers. The released code maps each appendix result to its script or data file: the numerical checks of Props. 17–19 and Fig. 8 (App. C.1) to `verify_routing_crossover.py`; the sparsity phase diagram (App. C.1) to `scripts/general_ml/sparsity_phase_synth.py` and `results/general_ml/sparsity_phase_synth.jsonl`; the aggregator shortcut (App. D) to `results/general_ml/aggregator_shortcut.jsonl`; the per-operator learning rates (App. F.6) to `analyze_lr_per_operator.py`; the structure-stratified effect sizes (App. G) to `analyze_mode0_stratification.py`; the fixed-evaluator disentangling (App. H.1) to `analyze_disentangle_fixed_eval.py`; and, in App. O, the FashionMNIST replication to `b3cross_fashion.jsonl`, the trapping-model thresholds to `verify_trapping_model.py`, the τ-intervention to `run_godel_tau_sweep.py` and `analyze_godel_tau.py`, the basin map to `basin_map_b0.jsonl`, the spectrum check to `basin_b2_jacobian.py`, and the noise-escape, K-sweep, race, and ratio probes to `b3noise_K6_n10.jsonl`, `ksweep_sustained.jsonl`, `b3p4_race_K6.jsonl`, and `ratio_probe_K6.jsonl`.

F Robustness Battery

Roadmap. Each subsection perturbs one axis of the base setup—loss family, latent dimension, seeds, quantifier surrogate, and parametric operator family—and checks the implication ranking survives (HermiT-filtered negatives are checked in App. H.1).

F.1 Dimension Sensitivity

We tested a 4 structure × 45 operator × 3 dimension × 3 repetition matrix with $d \in \{50, 100, 200\}$. Exact-combo rank correlations remain high in $7/8$ pairwise comparisons ($5/8 \geq 0.95$, $7/8 \geq 0.9$), with the weakest case in the saturated *universal-only* regime ($\rho = 0.768$ for $50 \to 200$). In the hardest tested structure (*with disjoint, depth 2*), the best exact t-norm/t-conorm pairing changes with d, but the winning implication remains Łuk-R throughout. Under the fixed training budget, absolute optimization quality degrades monotonically as d increases (overall mean SR: $0.625 \to 0.599 \to 0.558$; mean KL increases accordingly), so the correct interpretation is ranking robustness without dimension-neutral optimization cost (Fig. 13).

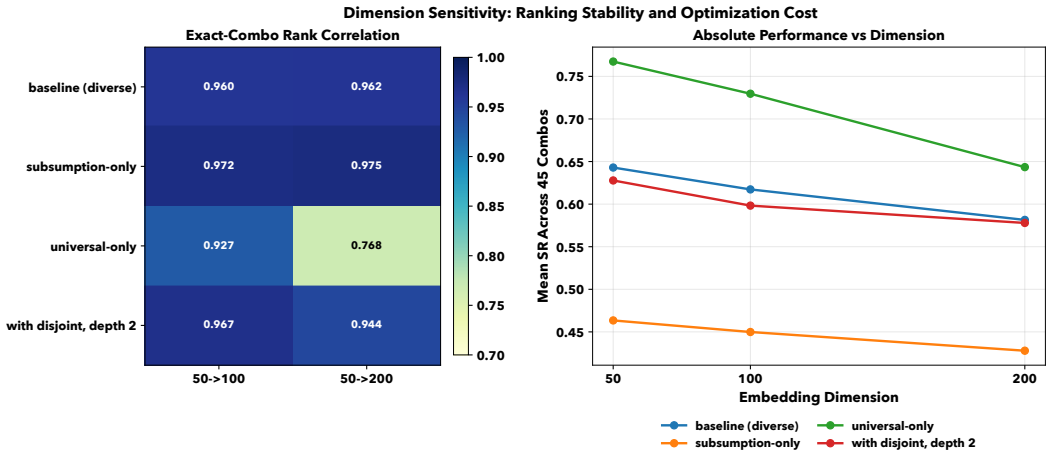


Figure 13: **The implication ranking is robust to latent dimension.** Dimension sensitivity. Left: exact-combo rank correlations remain high across most $d \in \{50, 100, 200\}$ comparisons. Right: absolute mean SR decreases as d increases under the fixed training budget, separating ranking robustness from optimization-cost robustness.

To decouple latent representation size from the number of individuals more directly, we also ran a factorized true-d control over 4 representative structures, 15 matched operator triples, $d \in \{16, 32, 50, 100, 200\}$, and 3 repetitions (900 runs, 0 failures). Using the smallest tested dimension ($d = 16$) as the baseline reference, the winning operator is preserved in 19/20 structure-by-dimension cells. Only *baseline (diverse)* at $d = 100$ changes winner (from Gog-R to Göd-R), while the other three structures are fully stable: *subsumption-only* stays Göd-R throughout, *universal-only* stays Göd-R throughout, and *with disjoint, depth 2* stays Łuk-R throughout. Spearman rank correlations remain high in the informative structures ($\rho = 0.71$–0.97 in *baseline (diverse)*/*subsumption-only*, 0.45–0.97 in *with disjoint, depth 2*), and the global mean-SR implication order remains

$$\text{Łuk-R} > \text{Gog-R} > \text{Göd-R} > \text{Rei} > \text{KD}.$$

The paper-safe reading is therefore stronger than the main-text proxy sweep alone: the structure-conditioned implication pattern survives a clean decoupling of latent dimension from the ontology-size axis used in the native DF$\mathcal{ALC}$ implementation.

F.2 Loss-Family Robustness: Expanded Sweep

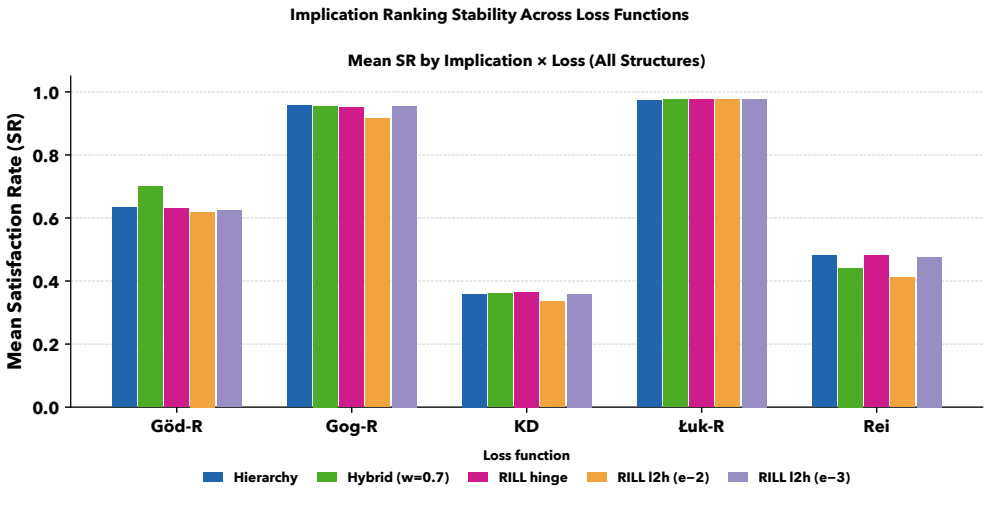
Core Matrix. We first extended the 4-structure baseline with a RILL-style fixed-threshold variant (He et al., 2024) that down-weights already-satisfied examples (hinge form, $\varepsilon = 10^{-3}$). Across the same 4 structures and 5 implications, the top-1 implication is unchanged in all tested cells and rank correlations with the Hierarchy baseline (the default loss ℓ^H of § 2.3) remain at 0.9–1.0. The effect is margin modulation and lower-rank reshuffling, not winner replacement.

Expanded 7-structure × 5-loss Sweep. We further ran a full 2,625-run sweep (7 structures, 5 losses, 5 implications, 3 t-norm/t-conorm pairs, 5 reps, 0 failures) covering *Hierarchy*, *HierarchyHybrid* ($w = 0.7$, the hybrid loss of § 2.3), and three RILL variants (hinge $\varepsilon = 10^{-3}$; L2+hinge $\varepsilon \in \{10^{-3}, 10^{-2}\}$). Results: top-1 implication preserved in 27/28 structure×loss comparisons; Spearman ρ mean 0.993, median 1.000, min 0.900, all ≥ 0.9. The crossover law replicates exactly: Łuk-R

wins outright in 5/7 structures under all 5 losses; *universal-only* is a saturated top-tier tie (Łuk-R and Gog-R both reach SR $= 1.000$; Table 5 marks the tie). The single deviant cell (*low individual ratio* $\times$ RILL hinge) is likewise a numerical tie, at SR $= 0.9978$ for both Łuk-R and Gog-R (ΔSR $= 0.000$, $\rho = 1.000$). The L2+hinge variant at large $\varepsilon = 10^{-2}$ reduces mean SR by ≈ 0.030 globally (excess soft-region width dilutes gradient signal) but does not alter any ranking. Fig. 14 summarizes the ρ matrix and per-implication SR profiles, and Table 5 gives the per-structure winners.

Table 5: **Implication dominance is not a loss artifact.** Hierarchy (H) and HierarchyHybrid (HH, w=0.7) induce identical implication rankings in 5 of 7 structures (ρ=1.0) and ρ=0.9 in the other two (*existential-only*, *subsumption-only*: a lower-rank Göd-R/Rei swap). Łuk-R is top-1 under both losses in every structure, outright in six and in an exact tie with Gog-R at SR$= 1.000$ in *universal-only* (†). SR is the mean over 3 t-norm pairs $\times$ 5 repetitions.

Structure	ρ	Winner (H)	SR (H)	Winner (HH)	SR (HH)
Baseline (diverse)	1.0	Łuk-R	0.984	Łuk-R	0.991
Existential-only	0.9	Łuk-R	0.993	Łuk-R	0.989
High density	1.0	Łuk-R	0.958	Łuk-R	0.976
Low individual ratio	1.0	Łuk-R	1.000	Łuk-R	1.000
Subsumption-only	0.9	Łuk-R	0.971	Łuk-R	0.973
Universal-only	1.0	Łuk-R=Gog-R†	1.000	Łuk-R=Gog-R†	1.000
With disjoint, depth 2	1.0	Łuk-R	0.910	Łuk-R	0.907



All values averaged over 3 t-norm pairs × 5 repetitions. ρ: Spearman rank correlation vs. Hierarchy baseline. Per-cell rank stability (28 cells = 7 structures × 4 alternative losses): 27/28 strict OK (same top-1 as Hierarchy); 1 tied exception (low individual ratio × RILL hinge: top-1 tied, ρ=1.000); 2 cells with ρ=0.900 (existential-only and subsumption-only on Hybrid).

Figure 14: **The implication tier survives every loss family.** Loss-family robustness: mean satisfaction rate per implication across 5 loss designs (averaged over 7 structures and 3 t-norm pairs). The tier ordering Łuk-R>Gog-R>Göd-R>Rei>KD is preserved under every loss; $26/28$ cells are at Spearman $\rho = 1.0$ versus the Hierarchy baseline (the two exceptions, $\rho = 0.9$, are lower-rank Göd-R/Rei swaps under HierarchyHybrid in *existential-only* and *subsumption-only*); top-1 is preserved in $27/28$ cells, the exception being the saturated *low individual ratio*$\times$RILL-hinge Łuk-R/Gog-R tie at SR 0.998 (App. F.2).

F.3 Expanded Parametric-Family Sweep

We expanded the earlier family check to an 8-structure sweep over Frank, Yager, and Hamacher families with the same three discrete anchors (the Gödel, product, and Łukasiewicz t-norms), totaling 24 configurations and 960 runs. Following Klement et al. (2000), this should still be read as bounded family coverage rather than exhaustive operator-space enumeration, but it is materially broader than the earlier 4-structure probe. Across all 8 structures, no continuous setting strictly beats the best discrete anchor. Two regimes (*existential-only*, *universal-only*) tie at SR $= 1.0$ and are saturated; the remaining 6 informative structures all stay below the best anchor. The structure-wise best continuous

gap has mean -0.201 and median -0.126, with the hardest losses in *subsumption-only* (-0.853) and *high density* (-0.293). By family, Frank, Hamacher, and Yager each remain below zero on every informative structure. The resulting conclusion is stronger than the earlier pilot claim: within the tested parametric families, interpolation does not introduce new winners beyond the paper-core discrete anchors (Fig. 15).

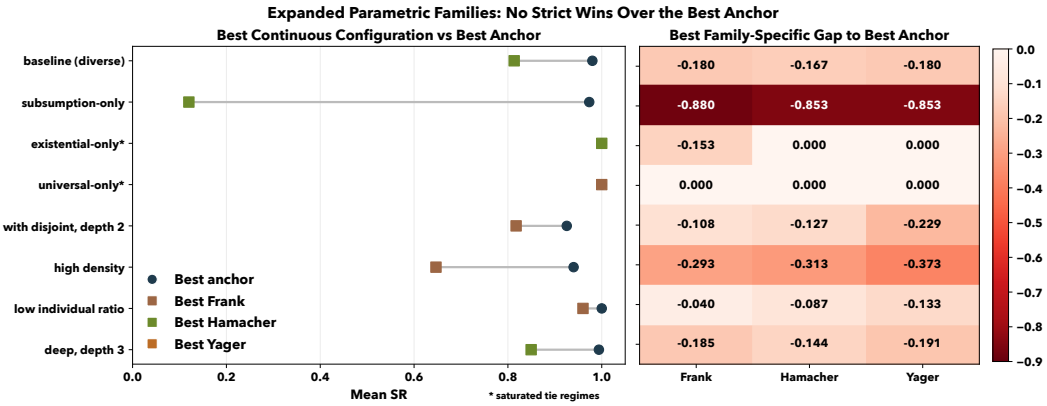


Figure 15: **Continuous operator families introduce no new winners beyond the discrete anchors.** Expanded parametric-family sweep across 8 structures. Left: best continuous configuration versus the best discrete anchor. Right: best family-specific gap to the best anchor. No strict wins appear; only the two saturated regimes tie at 0.000.

F.4 Targeted Extra-Seed Robustness

We reran 12 key exact configurations across the same 4 representative structures with 20 seeds each (240 runs total). In the easier structures, the strongest two implications remained tied at SR $= 1.0$ while the weaker comparator stayed clearly below them. In the hardest exact-combo case (*with disjoint, depth 2*), the original winner (the Łukasiewicz t-norm, product t-conorm, Łuk-R configuration) remained top-1 with mean SR 0.949 against 0.919 and 0.917 for the two nearest competitors (Fig. 16), so the exact-combo winner does not appear to be a small-sample artifact of the original 5-seed setup.

F.5 ML-Friendly Rank Summaries

To complement the ANOVA-style effect-size presentation, we also computed ontology-level pairwise win-rate matrices and bootstrap confidence intervals on mean implication ranks for BioPortal and ORE 2015 after averaging over t-norm, t-conorm, and repetitions within each ontology. Fig. 17 visualizes the pairwise win-rate matrices. On BioPortal, Łuk-R has mean rank 1.85 (95% bootstrap CI $[1.75, 1.95]$) and beats Gog-R on 45.7% of ontologies while losing on only 0.6%; ties account for the rest because many strong implications saturate at SR $= 1.0$. On ORE 2015, the same broad ordering persists in rank form (Łuk-R best, KD worst), again showing that the main ranking conclusions do not depend on accepting η^2 as the only summary language.

Corpus-Level Statistical Audit. For the corpus-level η^2 decomposition, we report bootstrap CIs for the observed effect sizes and permutation tests for role-label exchangeability. The bootstrap unit is the structural cell for synthetic data and the ontology for real data. The aggregate SR implication effects are 0.5908 $[0.5796, 0.6020]$ on synthetic, 0.3782 $[0.3660, 0.3909]$ on BioPortal, and 0.3107 $[0.3056, 0.3159]$ on ORE 2015; all corresponding permutation tests give $p < 10^{-3}$ (no permuted statistic reached the observed one in $5{,}000$ permutations). As a nonparametric check on the synthetic one-factor design, Kruskal–Wallis tests agree with the ANOVA significance decisions across all 10 structure dimensions. Nine of the ten structure dimensions interact significantly with operator choice under both SR and mean truth degree (type-II ANOVA, level $\times$ implication, on the 45-operator core; all nine survive Bonferroni correction over the ten dimensions); the exception is quantifier mix ($p = 0.09$ and 0.10). The Kruskal–Wallis routine reports significance only, not an effect size; the per-dimension implication η^2 quoted in the main text is the parametric between-/total-SS estimator read from the same sweep.

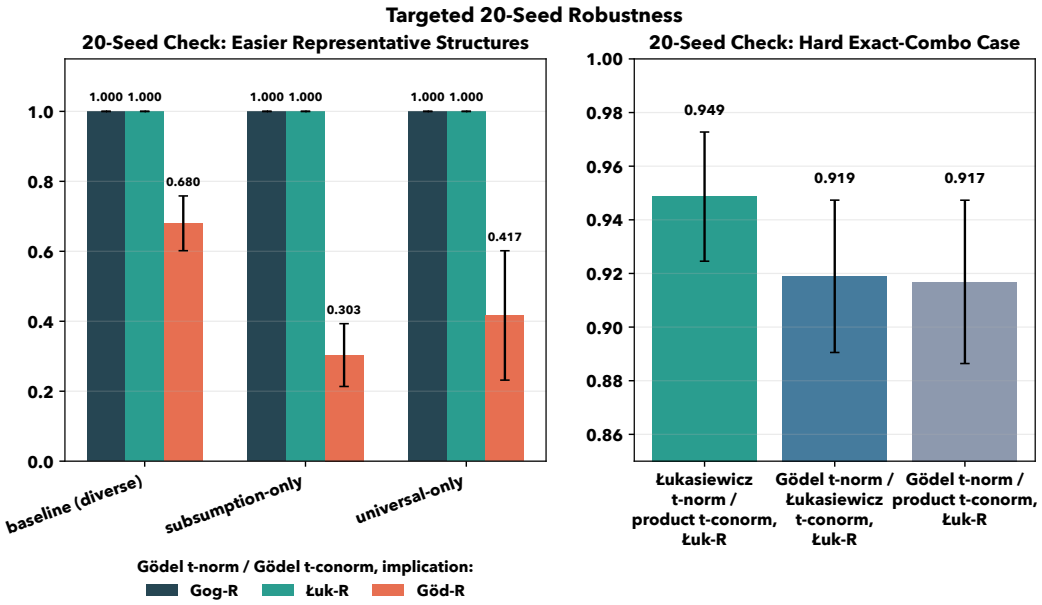


Figure 16: **The ranking holds under 20-seed averaging.** Targeted 20-seed robustness checks. Left: the stronger implications remain saturated in the easier representative structures. Right: in the hardest exact-combo case, the original winner remains top-1 with a nontrivial gap under 20 seeds.

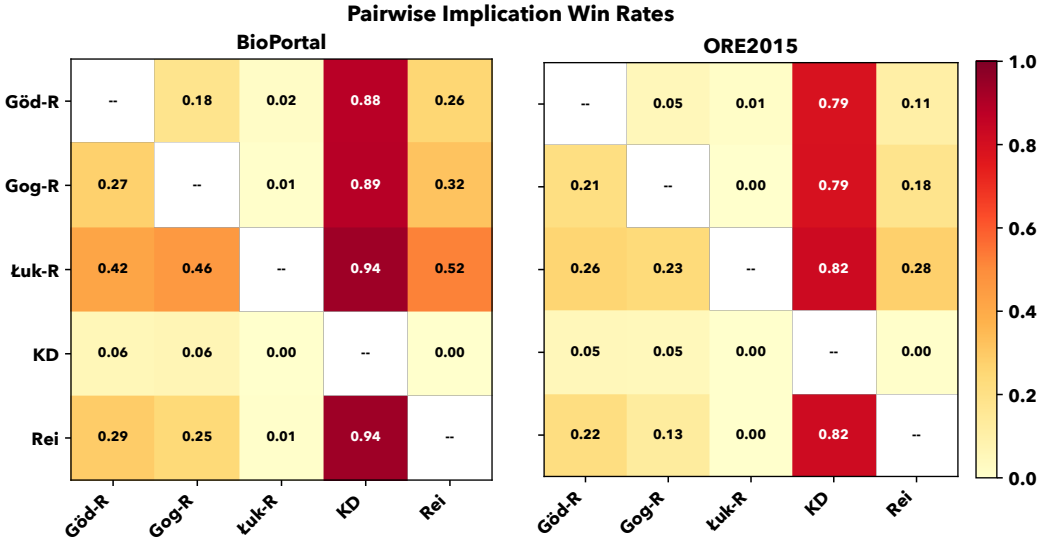


Figure 17: **Pairwise win-rates confirm the ranking on both real corpora.** Pairwise implication win-rate matrices for BioPortal and ORE 2015. The main ranking conclusions remain visible in a reviewer-friendly format that does not rely on η^2 alone.

F.6 Claim-Survival Summary

Per-operator learning-rate re-ranking. Re-ranking each implication at its own best learning rate from the five-point learning-rate grid (5×10^{-5}–10^{-3}) of our learning-rate sensitivity sweep (the most generous per-operator tuning the sweep supports) leaves every ordering unchanged: Spearman $\rho = 1.0$ against the shared-schedule ranking, both globally and within each of the 6 structures; only Gog-R even selects a different best rate (5×10^{-4} vs 10^{-3}, ΔSR 0.001). The shared-schedule caveat is thus a control, not an open risk.

Table 6: **Every theoretical claim is paired with its evidence and an evidential-status label.** Theory–evidence map with evidential status: **exact** theorem **/ model** (exact within an idealized model) **/ causal** (pre-registered intervention) **/ emp.** (empirical) **/ failed** (rejected prediction, reported). It maps the main text's full evidence chain.

Claim	Prediction	Evidence	Status
Prop. 1: implication gates every top-level path	implication dominates operator η^2; form-1 t-norm/t-conorm $\eta^2 = 0$	within-structure η^2 0.73–0.81 vs. 0.05–0.06 (t-norm); form-1 exactly 0 (§ 4.2)	exact + emp.
Thms. 2, 3: structure routes the Łuk-R/Göd-R gap	Łuk-R $\succeq$ Göd-R where left-quantified, no gap where right-quantified	init-gradient ρ=0.53/0.45/0.44; right-existential share is the top correlate of the Łuk-R advantage vanishing (ρ=+0.343); fit counts descriptive (§ 4.4)	exact + emp.
Props. 4, 6: survival × objective/grounding	satisfaction→learning flip; softmax collapse, free escape	0.69 vs 0.82–0.84 (Göd-R deviates); 0.11 vs. no collapse under free groundings (§ 4.3, App. K)	exact + emp.
Prop. 5: switch break-even	precision barrier	$p^* \geq 0.994$ on real corpora (§ 4.4)	exact + emp.
Prop. 7 + Lem. 24: saturated shortcuts	predict/mitigate/derive resistance	Figs. 4, 37: swap recovers at 300; τ-cliff intervention	exact/model; causal
fine $\sim 1/\tau$ scaling; model K-ratio scaling	smooth τ-threshold; Łuk-R/Rei gap widens in K	sharp transition; non-monotone ratio at K=6/8/10 (App. O)	failed
Basin-selection law (Thms. 27–28):	collapse rate $\propto \partial_2 I/K^3$; race law; $1/T$ floor	exact spectrum at K=6/8/10; double-T differential; ordering+immunity alphabet-robust (480-run K-sweep)	model; causal
Props. 15–22: substrate-independent forms	out-of-sample scorecard incl. negatives	7/8 aggregation tasks; the prospective ABSA call failed (App. C)	emp.

Each paper-core claim and the main perturbation that stresses it (numbers in the cited subsections; "preserved" = the qualitative reading survives, "narrowed" = scope refined, not overturned):

- **Implication dominates operator effects** — *preserved* under $\mathcal{EL}$, $\mathcal{EL}$-real, LTN, and fixed-evaluator downstream transfer: role asymmetry survives beyond base fuzzy $\mathcal{ALC}$.
- **Structure-conditioned ranking law** — *preserved but narrowed* (real corpora, synthetic crossovers, the tree-topology control for the chain-existential case; App. I): the surviving non-default positive is topology-qualified, not a competing default.
- **Łuk-R is a strong safe default on real corpora** — *preserved* (BioPortal, ORE 2015, $\mathcal{EL}$-real, regret/1%-safe audits; App. G.3): profitable deviations stay sparse and low-margin.
- **Loss formulation does not drive the ranking** — *preserved* (Hierarchy vs. HierarchyHybrid, 7-structure/5-loss sweep; Table 5), modulo one exact tie in a saturated cell.
- **Not a threshold or smoothing artifact** — *preserved but narrowed* (threshold sweep, SR-vs-mean truth degree, sigmoid baseline, softened quantifiers; Apps. F.7, F.8): the global order holds; the aggressive `logmeanexp` variant fails on three structures rather than winning.
- **Not a small-seed or single-d accident** — *preserved* (20-seed reruns, $d \in \{50, 100, 200\}$ proxy plus a factorized true-d control): 19/20 structure-by-dimension cells stable.
- **Not a learning-rate artifact** — *preserved* on two distinct tests: (i) re-ranking each operator at its *own* best rate is *exact* (ρ=1.0 globally and per structure, above); (ii) *rate-by-rate* across the pilot 5×10^{-5}–10^{-3} the Łuk-R/Gog-R top tier and KD weakness persist, with local saturated-cell swaps only at the most aggressive rates (top-tier robustness, not exact invariance at every individual rate).

Two statistical caveats qualify the results above. LTN agreement uses only five implications ($\eta^2_{\text{impl}} = 0.77$, $\rho = 0.90$, exact two-sided permutation $p = 0.083$), so we read it as descriptive cross-framework agreement rather than a confirmatory test; the learning-rate pilot covers 5×10^{-5}–10^{-3} and supports top-tier robustness, not exact rank invariance at aggressive rates. Within the tested range, the broad tier structure is stable for learning rates at or above 2×10^{-4}; the weakest cases at 5×10^{-5} are under-converged rather than evidence for a different operator ranking.

F.7 Threshold Robustness and Additional Sigmoid Baseline

Explicit Stored Threshold Sweep. The main archives store SR at threshold 0.5 together with mean truth degree, but not full per-axiom truth vectors. To provide an explicit multi-threshold check without rerunning the whole corpus, we therefore use the stored gradient-branch stress archive (3 structures $\times$ 3 t-norm backbones $\times$ 5 implications $\times$ 3 reps), which records SR at thresholds 0.3, 0.5, and 0.7 for every run. After aggregation, the overall implication order is identical at all three thresholds:

$$\text{Łuk-R} > \text{Gog-R} > \text{Göd-R} > \text{Rei} > \text{KD},$$

with Spearman $\rho = 1.0$ between every pair of thresholded rankings. Łuk-R remains top-1 throughout; at the cell level, safe-set and winner-set overlap stays high but is not literally identical in every stress cell, so the correct reading is threshold robustness rather than exact threshold invariance.

Broader Main-Corpus Metric Check. On the main synthetic and real archives, SR and mean-truth-degree rankings are exactly aligned on BioPortal and ORE 2015 and highly aligned on synthetic. We therefore use mean truth degree as the broad threshold-free proxy on the full corpora and the explicit $0.3/0.5/0.7$ archive as the direct threshold audit.

Additional Sigmoid Implication Baseline. The extended stress-test space also includes a sigmoid implication $I(u, v) = \sigma(10(v - u))$ (Sig), which differs from the sigmoidal implication SP used in the MNIST experiments. It remains a clear negative baseline rather than a new structure-sensitive winner. On the extended synthetic matrix, Sig is last by mean SR (0.407), with mean rank $6.0/6$, zero top ties, and zero appearances in the 1%-safe set. On the extended ORE 2015 slice, Sig is again last by mean SR (0.742), while Łuk-R remains top by mean SR (0.9989), mean rank (1.013), and 1%-safe-set coverage ($230/230$). Including this smoothed implication baseline therefore does not alter the paper's structure-conditioned ranking story.

F.8 Softened Quantifier Ablation

To test whether the main ranking pattern is an artifact of hard existential/universal routing, we ran a matched DF$\mathcal{ALC}$ ablation over 4 representative structures (*baseline (diverse)*, *subsumption-only*, *universal-only*, *with disjoint, depth 2*), 5 implications, and 4 quantifier aggregators (`hard`: exact max/min; `soft_tau`: their τ-smoothed surrogate; `pmean`: power mean; `logmeanexp`: log-mean-exp), with 3 repetitions per cell (240 planned cells). Training-time and evaluation-time quantifier aggregation are kept matched in every cell; the softened settings use a fixed $\tau = 0.1$ throughout, the power-mean variant uses fixed exponent $p = 2.0$, and the `logmeanexp` variant is evaluated without per-structure retuning, so the whole ablation is a no-retuning stress test rather than a per-structure hyperparameter search.

A fifth arm adds a *straight-through* surrogate (`ste`: forward = the hard max/min value, backward = the `soft_tau` gradient; 60 runs, same protocol): the per-structure implication ranking is preserved against both endpoints (Spearman ρ vs. `hard` $= 0.9/0.9/1.0/1.0$ and vs. `soft_tau` $= 1.0/0.9/1.0/0.9$ across the four structures, with only the familiar top-tier Łuk-R/Gog-R local swaps)—so the routing reading is robust across the surrogate family, including the biased-estimator regime (Regime III of App. C).

The correct reading is refinement, not exact invariance. Under `hard`, `soft_tau`, and `pmean`, the paper-core conclusion is unchanged at the aggregate level:

$$\text{Łuk-R} > \text{Gog-R} > \text{Göd-R} > \text{Rei} > \text{KD}.$$

On *subsumption-only*, hard-vs-soft rankings are identical ($\rho = 1.0$ for `soft_tau`, `pmean`, and `logmeanexp`) and Łuk-R remains top-1. On *universal-only*, the top tier stays fully saturated under `hard`, `soft_tau`, and `pmean`: Göd-R, Gog-R, and Łuk-R all achieve SR $= 1.0$, while Rei and KD remain far below. The quantifier-heavier cells show only local top-tier shifts under `soft_tau` and `pmean`: in *baseline (diverse)*, the hard top implication Gog-R gives way to a strict Łuk-R win under both softened variants ($\rho = 0.975$ and 0.947 vs. hard); in *with disjoint, depth 2*, the hard Łuk-R lead reverses to Gog-R under both softened variants ($\rho = 0.900$ for both). These are real local changes, but they stay inside the Łuk-R / Gog-R top tier, and KD remains last throughout the successfully completed structures.

The added `logmeanexp` variant sharpens the boundary rather than producing a stronger alternative baseline. It runs cleanly only on *subsumption-only*, where Łuk-R remains top-1 and the full impli-

Table 7: **Within-structure η^2 by operator role and corpus.** Point estimates with cluster-bootstrap CIs (2,000 resamples).

Corpus	Implication	t-norm	t-conorm
Synthetic (one-factor-at-a-time sweep)	0.738 [0.680, 0.787]	0.060 [0.047, 0.076]	0.004
BioPortal (175)	0.811 [0.743, 0.879]	0.061 [0.034, 0.090]	0.004
ORE 2015 (230)	0.732 [0.711, 0.752]	0.046 [0.042, 0.051]	0.003

cation ranking is unchanged. On *baseline (diverse)*, *universal-only*, and *with disjoint, depth 2*, the matched `logmeanexp` cells fail systematically across all five implications and all three repetitions. The paper-safe conclusion is therefore narrower than hard-routing invariance but stronger than a pure artifact dismissal: softened quantifier aggregation can perturb local top-1 outcomes on quantifier-heavy structures, while more aggressive log-space-style stabilization is itself structure-sensitive in this matched DF$\mathcal{ALC}$ setting and does not overturn the implication-dominant story.

G Real-World Corpus Analysis

Roadmap. Real-corpus analysis: dataset selection and complementarity, embedding pathology, the safe-default break-even map, the structure-risk map of profitable switches, and the online-switching ablation and a feasibility test of its structural classifier.

Within-structure effect sizes by corpus. Table 7 gives the within-structure η^2 behind the 0.73–0.81 range of § 4.2, with CIs from $2{,}000$ cluster-bootstrap resamples.

Structure-stratified effect sizes. Stratifying the 175 BioPortal ontologies by form-1 fraction, the implication effect is substantial in every quartile ($\eta^2_{\text{impl}} = 0.21/0.64/0.52/0.51$ from mixed to pure form-1; ontology-bootstrap 95% CIs lie above 0.14 in all four)—in contrast to the t-norm, which is off the loss path on form-1 and whose self-scored effect vanishes under a fixed evaluator (App. K.2). The routing-relevant contrast sharpens accordingly: the Łuk-R-vs-Göd-R mean-SR gap doubles on ontologies with left-quantified axioms ($+0.064$ $[0.039, 0.093]$, n=29) relative to those without ($+0.033$ $[0.023, 0.045]$, n=146)—a *differential* that the pointwise loss ordering (App. I.2) cannot produce, since it applies equally to both strata.

G.1 Dataset Selection Bias and Corpus Complementarity

The BioPortal main corpus is a size-bounded subset of the parseable local BioPortal sample. Of 329 local files, 308 are parseable and non-empty; the paper corpus keeps the 175 ontologies with 10–1000 supported axioms, excluding 10 smaller and 123 larger ontologies. Of the 175 BioPortal ontologies, 149 fall in the $\mathcal{EL}$ bucket (134 fully $\mathcal{EL}$ and 15 near-$\mathcal{EL}$, i.e. at least 95% $\mathcal{EL}$ axioms and no negation); together with 111 further $\mathcal{EL}$ ontologies outside the 10–1000 axiom window they form the 260-ontology $\mathcal{EL}$-real corpus. Relative to the excluded large-ontology tail, the selected BioPortal set is much smaller (median scale 237 vs. 2645 axioms; Mann–Whitney $p \approx 2.4 \times 10^{-37}$), less quantifier-heavy, and less constraint-dense. ORE 2015 remains useful not as a claim of recency, but as a complementary second corpus: compared with BioPortal it is smaller in scale (median 73 vs. 237 axioms) but richer in roles (median 24 vs. 6) and individuals-per-concept (median 0.75 vs. 0.31), so agreement across the two corpora is complementary rather than duplicate evidence (Fig. 18).

G.2 Embedding Pathology

Across the SR<0.3 failed-run subset (n=5 configurations), the aggregate bimodality index is 0.85 and the effective rank is 14.2, confirming degeneration rather than collapse. The extreme case is Göd-R, which drives embeddings into 0/1 bimodal clusters with effective rank 1.9, consistent with its antecedent-blind derivative profile (§ 2.2).

G.3 Switch-Case Detail

On BioPortal, the 5 profitable switch targets are heterogeneous: 3 go to Göd-R, 1 to Rei, 1 to Gog-R. On ORE 2015, all 3 profitable switches go to Göd-R. This asymmetry explains why a single fallback rule does not transfer cleanly across corpora (Fig. 20).

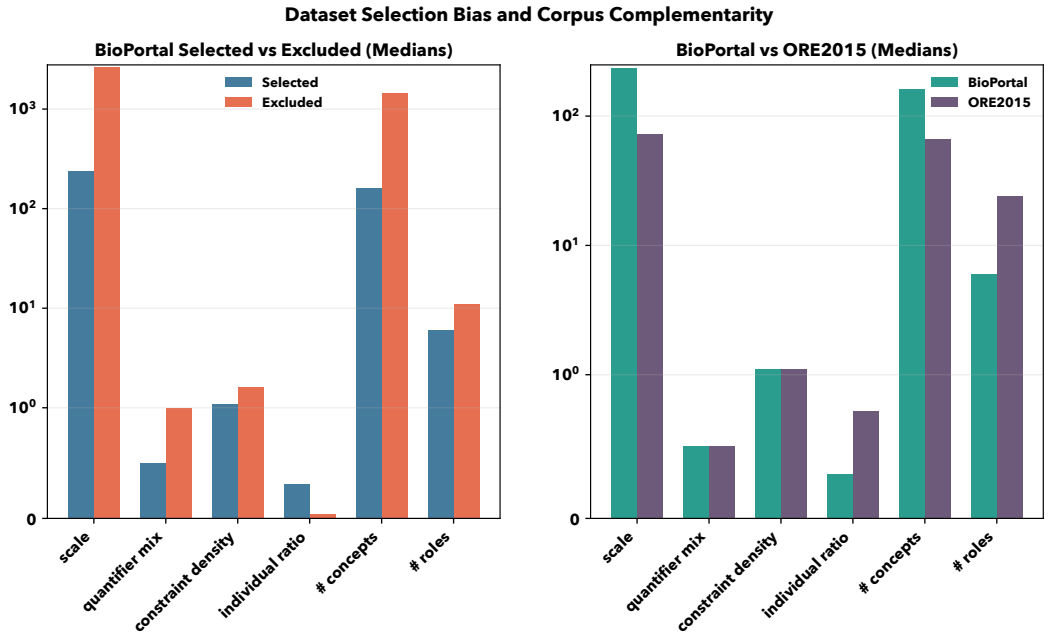


Figure 18: **The two real corpora are complementary, not redundant.** Dataset selection and corpus complementarity. Left: the selected BioPortal subset is primarily distinguished from the excluded tail by scale and related feasibility-linked structural properties. Right: ORE 2015 is smaller but more role-rich and higher in individual ratio than BioPortal, making it complementary rather than redundant as a second real corpus.

Mixed-operator / adaptive selection (a negative). The precision barrier also bounds the obvious next move—selecting the operator *per axiom* or *per ontology* by structure rather than fixing one. We probed this directly: a retrospective per-structure oracle and a random-forest gate trained on the structural features (17 features, AUROC 0.73) both fail to beat the fixed Łuk-R default under paired evaluation on the real corpora, and the learned HGB selectors of Table 9 do no better—profitable switching opportunities are too sparse ($8/405$ ontologies) and too corpus-specific (the structure→gap mapping does not transfer out-of-corpus) to clear the $p^* \geq 0.994$ bar. Adaptive selection therefore does not overturn the safe default on real ontologies; it is the structure-conditioned *ranking* that is the portable object, not a structure-conditioned *switcher*.

The feasibility bounds quantify the difficulty (Table 8): Fig. 19 visualizes the precision barrier. The best single fallback (Göd-R) needs $\sim$99.49% precision on BioPortal and $\sim$99.40% on ORE 2015 to break even with the Łuk-R default.

Table 8: **Mapping the break-even threshold to data.** Each variable in Prop. 5 is paired with its empirical estimator and measured value.

Variable	Estimator	BioPortal	ORE 2015
β (base rate)	$n_{\text{beneficial}}/n$	5/175 (2.9%)	3/230 (1.3%)
g (mean gain)	$\overline{\text{SR}_{\text{alt}} - \text{SR}_{\text{Łuk-R}}}$ over profitable	$\sim 10^{-4}$	$\sim 10^{-4}$
h (mean harm)	$\overline{\text{SR}_{\text{Łuk-R}} - \text{SR}_{\text{alt}}}$ over harmful	$\sim 10^{-1}$	$\sim 10^{-1}$
$k = h/g$	direct ratio	195	165
$p^* = k/(k{+}1)$	derived	≥ 0.994	≥ 0.994

G.4 Online-Switching Ablation and Structural-Classifier Feasibility

All gains in this section are in held-out D2 MRR. The online implication-switching policy, a test of the precision barrier of § 4.4, contrasts a pre-selected 4-ontology opportunity slice ($+0.02140$ vs. the fixed Łuk-R baseline `always_lukas`; method names follow Tables 10 and 11) with the unselected real corpus ($+2.6 \times 10^{-6}$ on the 270 benchmark ontologies with at most 5,000 axioms, $+5.3 \times 10^{-3}$

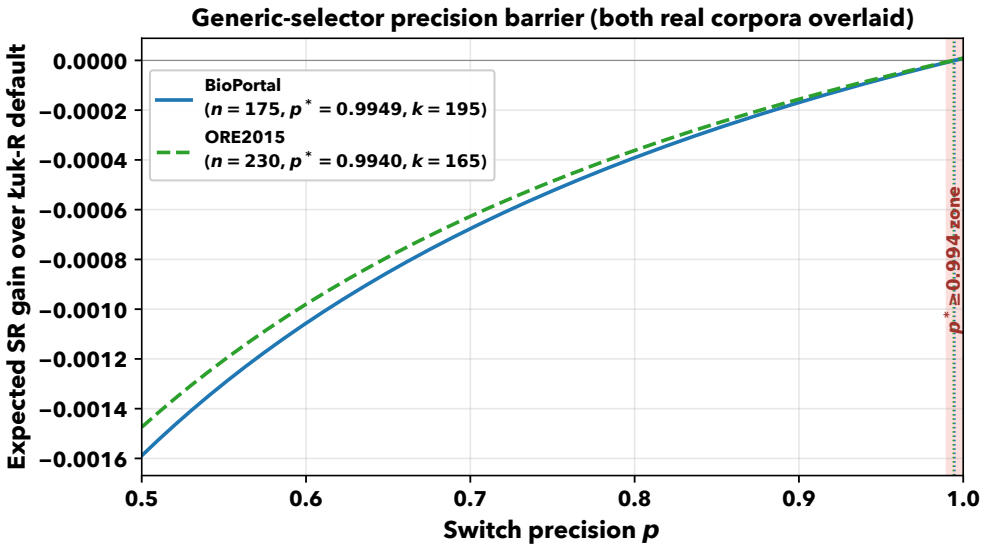


Figure 19: **Departing from Łuk-R requires near-perfect precision ($p^* \geq 0.994$).** Break-even precision barrier (Prop. 5). $p^* \geq 0.994$ on both corpora because false switches cost 165–195× more than correct switches gain.

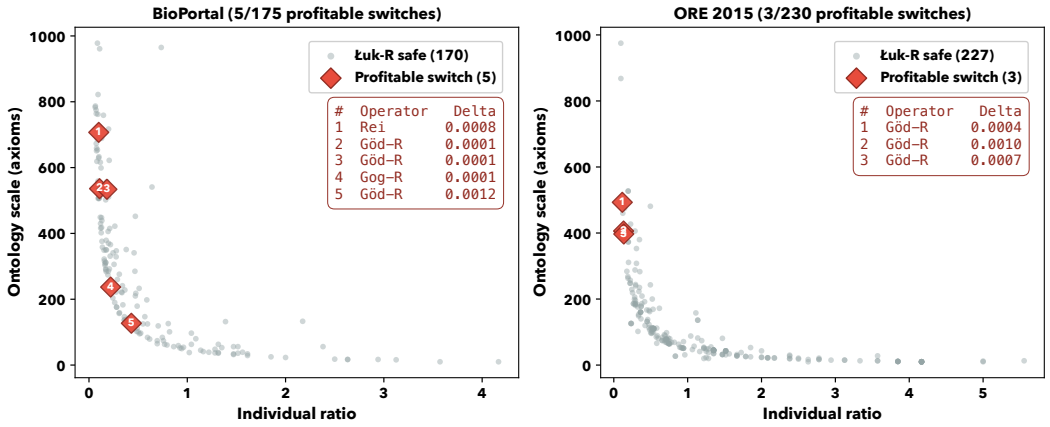


Figure 20: **Profitable switches are sparse and confined to a low-individual-ratio region.** Structure-risk map of profitable switches. Gray dots are ontologies where Łuk-R is optimal; red diamonds are the 8 profitable switch cases. All exceptions lie in the low-individual-ratio region, but BioPortal switches are heterogeneous across operators while ORE 2015 switches uniformly favor Göd-R. Regret deltas (Δ) are annotated.

on the 12 ontologies with clean records in the large-ontology extension (the 20 benchmark ontologies with 5,000–16,656 axioms), totalling 282/291 ontologies of the shared-benchmark target), and attributes the bulk of the opportunity-slice gain to checkpoint preservation. Tables 10 and 11 give the per-component and full-corpus numbers.

Component Decomposition (Leave-One-Out). The finalized policy stacks four components on top of a baseline that always retains Łuk-R: *(i)* `outer_restore_best` (heldout-best checkpoint restoration on the outer eval), *(ii)* candidate-positive warm phase (early Gog-R warm-start on ontologies whose offline classifier predicts $q_{\text{Gog-R}} \geq 0.02$ and ≤ 0.2), *(iii)* offline confidence threshold $q_{\text{Gog-R}} \geq 0.7$ for a switch label, and *(iv)* exact Łuk-R fallback on offline-abstained ontologies. Dropping each in turn (5 seeds, 500 epochs, GPU) gives the per-component contribution in Table 10. Restore-best is 83% of the gain on its own; the warm phase, abstention threshold, and hard fallback account for the rest; their leave-one-out drops (9%, 10%, 2%) are not additive. Per-ontology, dropping the

Table 9: **Implication-selection results on BioPortal (175 ontologies) and ORE 2015 (230 ontologies).** "Regret" is mean SR shortfall vs. oracle. "W1%" is the within-1%-of-oracle rate. HGB = HistGradient Boosting classifier; LOO = leave-one-out cross-validation. The Łuk-R default achieves 100% W1% on both corpora; no learned selector improves on this.

		BioPortal		ORE 2015	
Row type	Policy	Regret	W1%	Regret	W1%
Safe default	Łuk-R always	1.4×10^{-5}	100%	9.0×10^{-6}	100%
Best zero-shot	HGB-regret-gated	1.4×10^{-5}	100%	9.0×10^{-6}	100%
Best real-LOO	HGB-regret-gated	1.7×10^{-5}	100%	6.5×10^{-4}	98.3%
Oracle switch	Göd-R fallback	1.8×10^{-4}	99.4%	0	100%
Oracle best	Per-ontology best	0	100%	0	100%

warm phase *helps* `mcbcc` ($+0.04212 \to +0.04722$) and *hurts* `onlira` ($+0.04302 \to +0.03052$); the entire warm-phase contribution is concentrated on a single ontology.

Table 10: **Leave-one-out ablation of the online-switching policy on the 4-ontology opportunity slice (`esfo, mcbcc, onlira, icps`; 5 seeds, 500 epochs).** *Drop* is the loss in `warm_online` mean improvement (held-out D2 MRR) vs. the FULL stack.

Variant	Mean Δ	Drop	Share
FULL stack: restore-best + warm + $q_{\text{Gog-R}}\geq0.7$ + hard fallback	$+0.02140$		100%
(i) drop `outer_restore_best`	$+0.00358$	-0.01781	**83**%
(ii) drop candidate-positive warm phase	$+0.01955$	-0.00185	9%
(iii) drop $q_{\text{Gog-R}}\geq0.7$ threshold	$+0.01926$	-0.00214	10%
(iv) soft abstention (no hard Łuk-R fallback)	$+0.02090$	-0.00050	2%

Full-Corpus Reading. On the unselected real corpus, the same finalized policy is a statistical tie (Table 11). On these 270 ontologies, `warm_online` switches on 47% of ontologies, but, of its 127 switches, 34 are beneficial, 76 harmful, and 17 neutral, an empirical precision of $34/110 = 0.309$, below the Prop. 5 break-even for this metric: across these 270 ontologies, switching to Gog-R gains 0.00930 held-out MRR on average where it helps (101 ontologies) and loses 0.00988 where it hurts (145), so $k = h/g = 1.062$ and $p^* = 0.515$ (the SR-based $p^* \geq 0.994$ of Table 8 does not apply to MRR). The largest positive mean is `neutral_online` at $+0.00346$ (the other positive point estimates, `warm_online` at $+2.6\times10^{-6}$ and `offline_proxy` at $+4.0\times10^{-4}$, are effectively ties), with a 3-seed validation slice on a 12-ontology subsample reproducing the same direction ($+0.00264$ on `neutral_online`, $+0.00990$ on `warm_online`) within bootstrap noise. Even a per-ontology oracle on these 270 ontologies is bounded above by $\sum_O \max(0, \text{MRR}_{\text{Gog-R}} - \text{MRR}_{\text{Łuk-R}})/n = +0.00348$, and `warm_online` recovers none of it ($+2.6 \times 10^{-6}$). `neutral_online` ($+0.00346$) is not a selector but a near-uniform Łuk-R/Gog-R blend (gate weight 0.49–0.52 on every ontology) whose per-ontology gains and losses ($+0.015/-0.012$) largely cancel.

Table 11: **Full-corpus online-switching results on the 5,000-axiom slice (270 real ontologies, $n_{\text{ax}}\leq5{,}000$, 1 seed; gains in held-out D2 MRR).** `warm_online` is the final online-switching policy; `neutral_online` uses a neutral warm phase; `frozen_proxy` takes the offline classifier's hard call.

Method	Mean Δ vs. Łuk-R	Switch	Ben./harm	Prec.
`warm_online` (FULL stack)	$+2.6\times10^{-6}$	47.0%	34/76	30.9%
`neutral_online`	$+3.46\times10^{-3}$	94.8%	96/137	41.2%
`offline_proxy`	$+4.0\times10^{-4}$	6.3%	13/4	76.5%
`always_goguen`	-1.82×10^{-3}	100%	101/145	41.1%
`frozen_proxy`	-7.57×10^{-3}	28.9%	49/27	64.5%

Structural-Classifier Feasibility. A 5-fold stratified classifier of $\text{sign}(\text{MRR}_{\text{Gog-R}} - \text{MRR}_{\text{Łuk-R}})$, the sign of the held-out D2 MRR gap between training under Gog-R and under Łuk-R (fixed product/product/Łuk-R evaluator), over the 270 benchmark ontologies with at most 5,000 axioms using 17 pure structural features (form-fraction histogram, individual ratio, role/concept reuse, scale,

form entropy; no D2-leak) reaches AUC-ROC 0.722 (logistic) / 0.736 (random forest) and precision 0.671 (LR) / 0.674 (RF) at recall $=0.30$. The strongest single Spearman correlate of the oracle gap is `mode_3_frac` (feature names keep the released code's `mode` prefix, with `mode_`k denoting form $k{+}1$ of Table 1) (right-existential share, $\rho = +0.343$, $p = 7.2\times10^{-9}$), exactly the regime in which Thm. 2 forbids the left-quantified mechanism that gives Łukasiewicz its safe-default advantage; `mode_homogeneity` (-0.287) and `mode_0_frac` (-0.211) are the next two strongest, both negative.

Selected-slice policy. On the top-20 RF-picked ontologies under multi-seed evaluation (5 seeds), the same final online-switching policy yields `warm_online` $= +0.01370$ with $20/20$ beneficial switches and 0 harmful, consistent with the feasibility estimate that operating at recall $=0.30$ and precision $=0.67$ would deliver an aggregate full-corpus gain of approximately $+0.0022$.

Deployment caveat. This is not a deployable system: the operating point is selected in-sample, so its precision of 0.67, although above the MRR break-even $p^* = 0.515$, is not an out-of-sample estimate. It does, however, close the loop between the structural laws of § 4.2 and the decision-theoretic barrier of § 4.4: the structural features that the theory identifies (form-4 share above all) are exactly the ones a simple classifier needs to clear the empirical precision barrier on a structurally selected slice.

H Embedding Geometry and Downstream Impact

Roadmap. How routing reshapes the learned representations and what it costs downstream: the D2 task-type metric split and the D4 cross-task reversal (embedding pathology is in App. G.2).

H.1 D2 Task-Type Metric Split: Detail

Protocol. D2 trains DF$\mathcal{ALC}$ under each of the 5 paper-core implications on the held-out subsumption benchmark (308 BioPortal ontologies, 80/20 stratified split, fixed evaluator with product t-norm and t-conorm and Łuk-R implication, 5 seeds). For each held-out axiom (all seven forms are held out), 10 negatives are generated by replacing its consequent concept D with a random non-superclass concept; both MRR (rank of the positive among 11 candidates) and AUC-ROC (over all positive–negative pairs) are computed on identical embedding pairs.

Threshold robustness. The split is stable across $\min n_{\text{ax}} \in \{0, 5, 10, 20, 50\}$ (308, 304, 298, 291 and 275 ontologies): at every threshold MRR favors Łuk-R and Gog-R, while AUC-ROC favors Göd-R and Rei. All D2 numbers in this section and in Fig. 21 use the unfiltered 308 ontologies; the 291-ontology subset with at least 20 axioms is the shared benchmark of the geometric and LLM baseline comparisons (App. J).

Uncertainty. All headline D2 intervals use ontology-cluster bootstrap over the 308 ontologies. On MRR, Łuk-R is 0.542 $[0.522,\ 0.562]$ and Gog-R is 0.541 $[0.521,\ 0.561]$; on AUC, Göd-R is 0.725 $[0.708,\ 0.742]$ and Rei is 0.721 $[0.704,\ 0.739]$. Paired bootstrap differences preserve the leading-pair orderings: Łuk-R $-$ Gog-R MRR $= 0.0015$ $[0.0005,\ 0.0025]$ and Göd-R $-$ Rei AUC $= 0.0036$ $[0.0006,\ 0.0067]$. On a separate 5×5 train$\times$eval grid (20 BioPortal ontologies, 3 seeds), the train$\times$eval interaction is $\eta^2_{\text{train}\times\text{eval}} = 0.265$ with ontology-bootstrap 95% CI $[0.176,\ 0.351]$ (main effects on the same grid: train 0.179, eval 0.230). Applying the t-norm caliber control of App. K.2 to the implication side on this grid: under *each* fixed evaluator the train-implication effect persists on the evaluator's SR—Göd-R $\eta^2{=}0.12$ $[0.08, 0.27]$, Gog-R 0.10 $[0.05, 0.44]$, Łuk-R 0.43 $[0.29, 0.59]$, KD 0.62 $[0.45, 0.83]$, Rei 0.70 $[0.51, 0.89]$ (ontology-bootstrap CIs; fixed-mean-truth-degree values agree within 0.02)—versus 0.21 self-scored. Unlike the form-1 t-norm artifact, which vanishes under this control, the implication effect is not a scoring artifact; self-scoring if anything *understates* it under most evaluators. Filtering the negatives through a DL reasoner (HermiT, with a transitive-closure fallback used for 44% of cells) to remove the $\approx 1.3\%$ of corruptions that are actually entailed leaves the mean implication ranking over the 308 ontologies unchanged (Spearman $\rho = 1.0$ for AUC and MRR) and raises mean AUC by $+0.007$ $[+0.005, +0.008]$ for every implication.

Formal inversion check. The MRR/AUC split is not only a difference in global means. After averaging seeds within each ontology, Göd-R beats Łuk-R on AUC by $+0.0106$ (ontology-bootstrap 95% CI $[+0.0046,\ +0.0168]$; one-sided Wilcoxon $p = 3.3\times10^{-5}$), while Łuk-R beats Göd-R on MRR by $+0.0484$ ($[+0.0407,\ +0.0564]$; one-sided Wilcoxon $p = 1.4\times10^{-31}$). The per-ontology winner agrees between AUC and MRR on only $100/308$ ontologies (32.5%), confirming that the inversion reflects metric-dependent embedding utility rather than a plotting artifact.

Per-ontology beat-counts vs Łuk-R. On AUC-ROC, Göd-R wins on 59% of ontologies, Rei on 47%, Gog-R on 30%, KD on 17%. On MRR, Göd-R wins on only 17%, Rei on 10%, Gog-R on 34%, KD on 3%. Fig. 21 visualizes the global means (panel a) and the per-ontology distributions (panel b).

Caveats. (i) The MRR vs AUC split is reported on a single benchmark; the cross-task replication on D4 role-corruption (App. H.2) is the second corruption-target check. (ii) The mechanism interpretation (bimodal vs continuous embedding geometry) is supported by the embedding-pathology analysis (App. G.2) but a full geometric characterization across all 5 implications is left to future work.

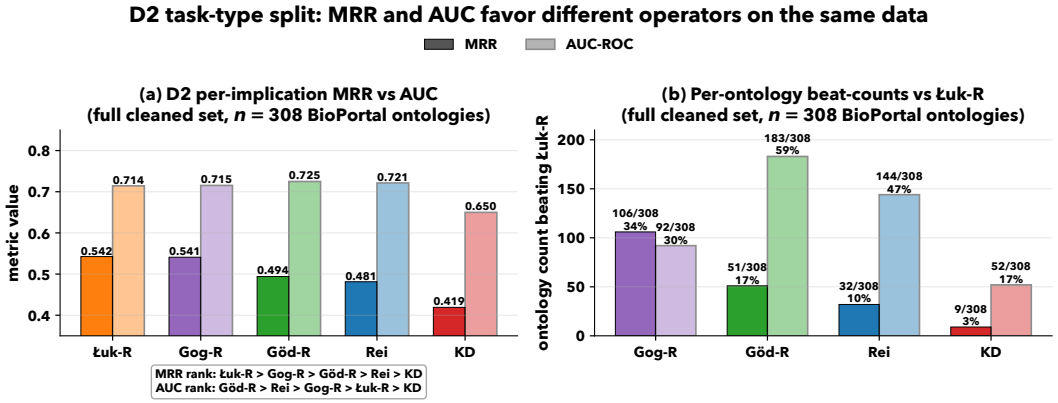


Figure 21: **MRR and AUC favor different operators on the same data.** D2 task-type split. (a) Per-implication mean MRR and AUC-ROC over all 308 D2 ontologies, with rankings annotated. (b) Per-ontology beat-counts against Łuk-R over the same 308 ontologies: Göd-R beats Łuk-R on 59% of ontologies under AUC but on only 17% under MRR, and Gog-R beats Łuk-R on 34% under MRR and on 30% under AUC. The split is mechanistically tied to bimodal Göd-R embedding geometry (case-study effective rank 1.9, bimodality 0.89) vs. continuous Łuk-R geometry (case-study effective rank 5.2, bimodality 0.11).

H.2 D4 Cross-Task Role-Corruption: Detail

Protocol. D4 reuses the D2 anti-tautology pipeline (DF$\mathcal{ALC}$ training under each of the 5 paper-core implications, fixed evaluator with product t-norm and t-conorm and Łuk-R implication), except that training uses learning rate 2×10^{-4} and at most 25,000 epochs, and replaces the 80/20 stratified subsumption holdout with a *role-corruption* negative set: for each held-out role-quantified axiom (forms 4–7), 10 negatives are generated by replacing the role index (right[0] for forms 4/5, left[0] for forms 6/7) with a random alternative role, keeping concept indices intact. Concept-corruption negatives (the standard D2 ones) are computed on the same held-out set in the same training pass, so each $(\text{ontology}, \text{implication}, \text{seed})$ produces *paired* concept-AUC and role-AUC measurements on identical trained embeddings.

Selection. 38 BioPortal ontologies sampled evenly from the shared D2 manifest, restricted to those with ≥ 10 form-4–7 axioms and $\in [20, 2000]$ total axioms. On CPU, 5 implications $\times$ 5 seeds = 25 runs per ontology, 901 ok rows after keeping only runs whose held-out split contains at least one role-bearing axiom (forms 4–7) with a role-corrupted negative.

Findings.

- Role-AUC: Łuk-R (0.620) $\approx$ KD (0.617) $\approx$ Gog-R (0.611) $>$ Göd-R (0.578) $\approx$ Rei (0.571); spread 0.049.
- Same-runs concept-AUC reproduces D2 (within sample): Göd-R (0.668) $>$ Rei (0.657) $>$ Gog-R (0.629) $\approx$ Łuk-R (0.627) $>$ KD (0.563).
- Within-run rank Spearman: AUC $\rho = -0.80$, MRR $\rho = +0.50$.
- Per-ontology beat-counts vs Łuk-R on role-AUC: Göd-R wins $12/36$, Gog-R $9/37$, Rei $10/36$, KD $17/36$. The bimodal Göd-R geometry that helps D2 cEmb discrimination is below Łuk-R on 67% of ontologies for D4 rEmb discrimination (Fig. 22).

The interpretation is mechanistically aligned with § 4.4: bimodal Göd-R embeddings produce sharp axis-aligned cEmb partitions that are well-suited for concept-pair pairwise ordering (D2 AUC), but role embeddings (rEmb $\in \mathbb{R}^{|R|\times d\times d}$) do not inherit the same bimodality, so the same operator that wins on cEmb discrimination loses on rEmb discrimination. Łuk-R's continuous embeddings perform consistently across both corruption targets.

Caveats. (i) The $\rho = -0.80$ inversion is computed on 5 implications; it is a strong qualitative signal but the small space of operators limits the statistical power of the rank correlation itself. (ii) D4 is a single second corruption target; the manuscript's "metric×corruption-target structural conditioning" framing claims two corruption targets, not the full task space. A third behavioral readout (e.g., concept-membership prediction with ABox individuals, or hits@1 on multi-class concept retrieval) would further constrain the geometry-mechanism interpretation.

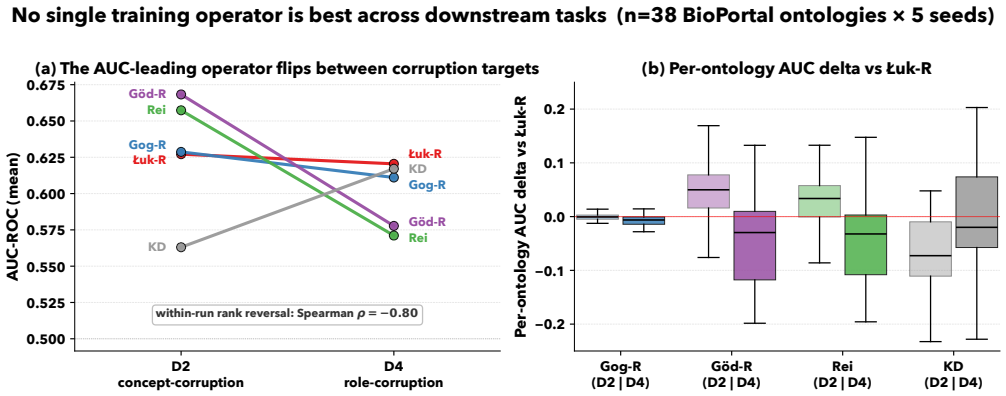


Figure 22: **No single training operator is best across downstream tasks.** (a) Slopegraph of per-implication mean AUC-ROC from D2 concept-corruption to D4 role-corruption (38 BioPortal ontologies × 5 seeds): the crossing lines are the reversal: the bimodal Göd-R/Rei cEmb geometry *leads* on D2 but *plunges* on D4, KD climbs the opposite way, while the continuous Łuk-R/Gog-R embeddings stay flat across both; the within-run rank Spearman between concept-AUC and role-AUC is $\rho = -0.80$. (b) Per-ontology AUC delta vs. Łuk-R: Göd-R and Rei boxes are positive on D2 (winning) but negative on D4 (losing), so the inversion holds per ontology, not just in the mean.

I Structural Crossover Details

Roadmap. The structure-conditioned crossovers in detail, including the chain × pure-existential law—the one regime where the safe default is itself beaten (Reichenbach exact-zero).

I.1 Chain × Right-Existential Structural Law: Detail

Setup. Synthetic chain-topology ontologies with form 4 ($C \sqsubseteq \exists R.D$) only, $n_{\text{axioms}} \in \{60, 120, 240\}$, depth 1 (the first chain grid also at depth 2), $T{=}S{=}$ product, Adam at lr$=0.05$, 10 matched seeds (first grid: 60 runs per implication over both depths; the second grid adds Gog-R, Göd-R, and KD for the five-way ranking, total 90 runs). Training runs for at most 1500 epochs and stops early when the loss fails to improve by an *absolute* 10^{-4} over 10 checkpoints; the reported loss is therefore the best 100-epoch checkpoint, not a converged value. All operators are the $\tau{=}0.1$ surrogates of App. E.

Per-cell ranking (two paired chain grids). Reichenbach reaches exactly zero training loss on $60/60$ paired runs (no ties) at every n_{axioms} cell; the order of the other four, all within 10^{-3} of zero, is not stable, moving with the learning rate, the stopping rule, and whether the endpoint is scored with the soft surrogate or the hard operator, and even second place alternates between Göd-R and Gog-R (App. I.4), so we claim only Reichenbach's exact zero here. KD's deficit in this cell is an artifact of its $\tau{=}0.1$ softmax surrogate: rescoring the same trained embeddings with the hard $\max(1-u, v)$ puts KD at exactly zero (App. I.4), so this cell does not adjudicate an R-vs-S family split, which we leave open. The causal evidence for the derivative shape rather than the family is the 2×2 derivative-flattening ablation below (Fig. 23).

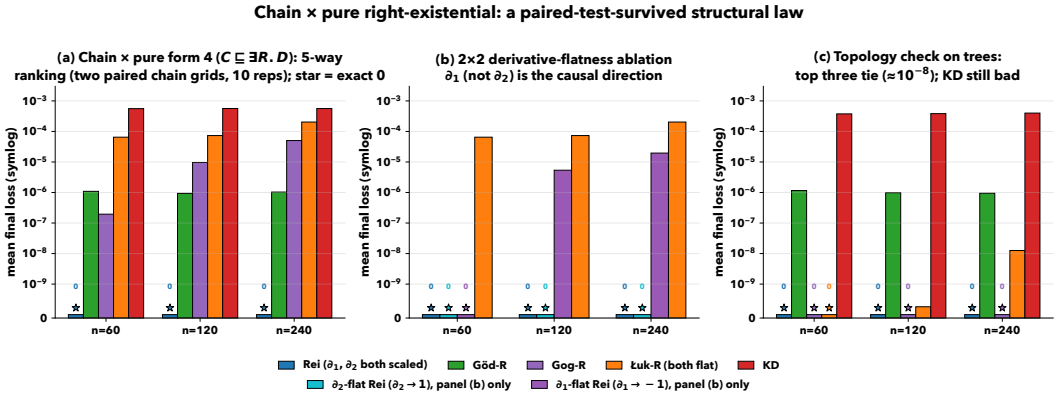


Figure 23: **On chain × pure-existential structure the safe default is itself beaten (Rei reaches exactly zero loss).** Chain × pure right-existential structural law. (a) Five-way ranking on the two paired chain grids (chain topology): Rei reaches exactly zero loss on every paired run; KD's position here reflects its τ=0.1 softmax surrogate; under the hard operator it is exactly zero, so this panel does not adjudicate a family split. (b) The ∂_2-flat Rei ablation (flattening $\partial_2 I$ to 1) leaves exact-zero intact across all 30 runs; the complementary ∂_1-flat Rei ablation (flattening ∂_1 to -1) loses exact-zero and degrades to a residue $\approx 10\times$ smaller than Łuk-R's, localizing the dominant cause to input-scaled $\partial_1 = -(1-v)$ self-damping on consequent satisfaction. (c) Topology check on trees: the top three (Rei, Gog-R, Łuk-R) tie at $\approx 10^{-8}$ loss while Göd-R stays near 10^{-6} and KD remains poor, confirming the law is chain-specific.

Mechanism ablation (2×2 **derivative-flatness,** 30 **runs each cell).** Two custom-backward Reichenbach variants isolate the role of ∂_1 vs ∂_2:

- ∂_2-*flat Rei* (input-scaled $\partial_1 = -(1-v)$, flat ∂_2=1): exact 0 loss on all 30 runs.
- ∂_1-*flat Rei* (flat $\partial_1 = -1$, input-scaled ∂_2=u): exact 0 only at n_{ax}=60 (10/10); residue 5.4×10^{-6} at n_{ax}=120, 2.0×10^{-5} at n_{ax}=240 ($\approx 10\times$ smaller than Łuk-R).

Conclusion: input-scaled ∂_1 self-damping is the dominant cause; the Reichenbach forward shape contributes a secondary $10\times$ residue reduction.

Topology check on trees. On tree topology (depth 1, form 4, otherwise identical), Rei and Gog-R reach exactly zero loss and Łuk-R 4.3×10^{-9} (essentially tied), while Göd-R stays at 1.0×10^{-6} and KD at 3.9×10^{-4}. The chain-existential law is therefore topology-conditioned: the mechanism activates on chain because antecedent satisfaction depends multiplicatively on consequent satisfaction along the chain, while tree decouples the per-axiom updates.

I.2 Multi-Corpus Validation of Theorem 3 and Corollary 11

Init-state quantitative validation (Cor. 11). For each of 175 BioPortal main, 260 $\mathcal{EL}$-real, and 227 ORE 2015 ontologies, we compute the structural lower bound LB and the projected antecedent gradient norm $\|g_U^L(\theta_0)\|^2$ at training initialization under hard Łuk-R. Pooled rank correlations are ρ=0.53 (BioPortal, $p < 10^{-13}$), ρ=0.45 ($\mathcal{EL}$-real, $p < 10^{-13}$), and ρ=0.44 (ORE 2015, $p < 10^{-11}$). On a controlled synthetic form-6/7 corpus, Cor. 11 holds as a loose lower bound on 91.7% of cases. The pooled rank correlation is driven by the LB-vacuous/LB-active dichotomy: 29/175 (16.6%) BioPortal, 48/260 (18.5%) $\mathcal{EL}$-real, and 22/227 (9.7%) ORE 2015 ontologies have $N_6^{\text{act}}+N_7^{\text{act}} > 0$ at θ_0 (LB-active); the remaining LB-vacuous majority have LB = 0 but small positive $\|g_U^L\|^2$ from concept-pathway contributions. Within-LB-active subsamples, rank correlations are weaker (ρ=0.04 on BioPortal, $\rho = -0.08$ on $\mathcal{EL}$-real), confirming Cor. 11 is a loose lower bound rather than a tight predictor (Fig. 24).

Qualitative dominance (Thm. 3). Per-ontology training comparison of hard Łuk-R vs hard Göd-R (5 seeds each, mean over seeds): Łuk-R achieves strictly lower training loss on 172/175 (98.3%) BioPortal main, 248/260 (95.4%) $\mathcal{EL}$-real, and 219/227 (96.5%) ORE 2015 ontologies. One-sided binomial tests against $p_0 = 0.5$ give $p < 10^{-46}$, $p < 10^{-58}$, and $p < 10^{-54}$, respectively. The

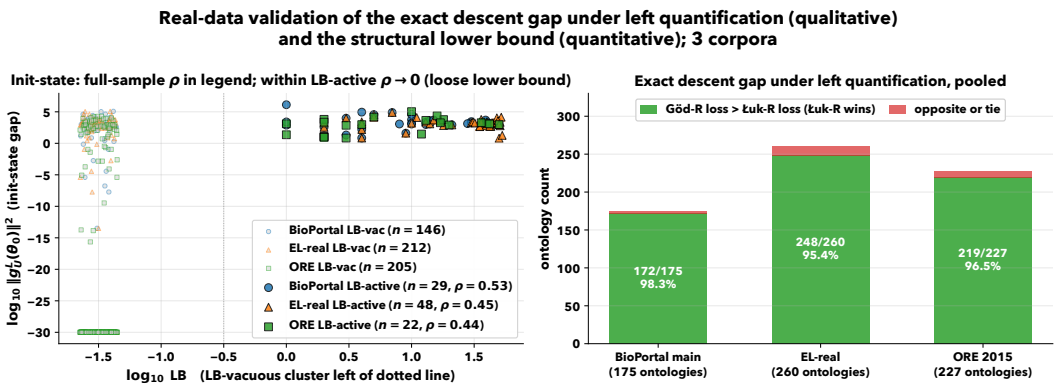


Figure 24: **The structural lower bound activates the predicted init-gradient on real corpora.** Real-data validation of Thm. 3 and Cor. 11. Left: log-log scatter of structural lower bound LB vs. init-state directional gap $\|g_U^L(\theta_0)\|^2$ on the LB-active subsamples of BioPortal main (n=29), $\mathcal{EL}$-real (n=48), and ORE 2015 (n=22); within-subsample rank correlation is weak, consistent with Cor. 11 as a loose lower bound rather than a tight prediction. Right: qualitative dominance counts (Thm. 3 via concept-pathway), $172/175$ (98.3%), $248/260$ (95.4%), and $219/227$ (96.5%) Łuk-R wins on the three corpora (descriptive; see the pointwise-ordering caveat in the text).

remaining $3 + 12 + 8 = 23$ comprise 13 exact ties (both losses 0, all LB-vacuous) and 10 reversals with $|\text{gap}| \leq 5 \times 10^{-4}$, below noise.

Caliber caveat on the dominance counts (pointwise ordering). The counts above compare each arm's *own* training loss, and the two pair losses are pointwise ordered at every θ: on the violated branch $1 - I_{\text{Łuk}} = \max(0, u-v) \leq 1 - v = 1 - I_{\text{Göd}}$. The comparison is therefore partially forced by the loss scale, and we treat it as descriptive corroboration only; the theorem-level evidence is the init-state gradient measurement above, which compares gradients, not losses. A fixed-currency cross-check on the 5×5 train$\times$eval satisfaction grid (App. H.1) makes the caliber point explicit: the direction of a satisfaction comparison follows the evaluator—Łuk-R-trained models beat Göd-R-trained ones under the Łuk-R evaluator (10 wins/10 ties/0 losses over 20 ontologies) and lose under the other four evaluators (e.g. $0/3/17$ under Göd-R's)—while on the held-out learning currency the comparison splits by metric (Łuk-R best MRR, Göd-R best AUC; App. H.1). The implication-η^2 headline is unaffected: it is caliber-checked separately (Apps. K.2, H.1).

Multi-epoch trajectory analog (null on BioPortal). Replacing θ_0 with the per-epoch parameter trajectory and measuring the mean directional gap along $\Delta_U(\theta_t)$ produces ρ=0.07 on BioPortal (no significant signal; the θ_t trajectory is dominated by concept-pathway descent). On $\mathcal{EL}$-real the multi-epoch correlation is weakly positive (ρ=0.24) but not statistically established.

I.3 Soft-Surrogate τ-Sweep Audit

The implementation replaces the hard $\max/\min$ selectors with softmax/sigmoid surrogates at τ=0.1. We audit that choice over $\tau \in \{1.0, 0.5, 0.2, 0.1, 0.05, 0.01\}$ on four structures from the generator family of App. F (*subsumption-only*, *existential-only*, *baseline (diverse)*, *with disjoint, depth 2*; 30 axioms, 50 individuals), the 5 paper-core implications, and 5 seeds (600 runs). Each run's endpoint is scored twice: under the soft operator it was trained with, and under the hard operator of the theorems.

Routing topology is stable. On the elements where the gate is still sensitive, Łuk-R splits its sensitivity evenly between the two branches at every τ (0.500), as $\partial_1 I = -1$, $\partial_2 I = +1$ requires. Göd-R is consequent-skewed on three of the four structures (1.000 at τ=0.1) but only 0.625 on *with disjoint, depth 2*, where the endpoint is not settled and the soft gate stays open near $u \approx v$: the skew is typical, not uniform.

The ranking is stable only under hard rescoring. The hard-rescored ordering matches the τ=0.1 ordering in all 20 non-reference comparisons (tie-aware ρ=1; several operators tie at exactly zero). The soft-surrogate ordering does not: it breaks on three of the four structures, most sharply on *existential-only* (ρ=0.40 at τ=0.01). This is a statement about rescored endpoints, not about hard-operator training: every model is trained under its own surrogate.

KD approaches branch starvation as $\tau \to 0$. Its per-element dominant-branch share rises from 0.67 (τ=1.0) to 0.85 (τ=0.1) and 0.91 (τ=0.01), while its aggregate consequent share stays near 0.5: different elements take different branches. At τ=0.1 the operator is a softmax-weighted mean, not a hard max, so Cor. 12 is approached but not reached.

All values are fixed-budget endpoints, not converged solutions: 354 of 600 runs meet our settling criterion, and only 37 of 150 on *with disjoint, depth 2*.

I.4 Soft-vs-Hard Endpoint Re-check on the Chain Cell

The endpoint losses of the chain × pure-form-4 cell come from τ=0.1 surrogates, an absolute early-stopping threshold of 10^{-4}, and lr=0.05, each of the same order as the tail spread they rank, or larger. We therefore re-ran the cell in 900 runs, a 600-run grid of 5 implications × $\{\text{lr}=0.05, 3\times10^{-3}\}$ × {early stop on, off} × $n_{\text{axioms}} \in \{60, 120, 240\}$ × 10 matched seeds (the seeds of the two chain grids), plus a τ sweep at n=60. Each run's final parameters are scored twice under identical batching and normalization, with the optimized soft operator and with the hard operator of the theorems.

The tail ranking is a surrogate effect, not an operator effect. Under soft scoring KD is last in 12/12 cells (5.0×10^{-4} to 8.9×10^{-4}); under hard scoring it is exactly zero in all 120 runs. Neither switching off early stopping nor lowering the learning rate 17× moves it: KD's soft residual varies only ∼1.75× across the grid. The cause is the surrogate itself: our KD is a softmax-weighted mean, which falls below $\max(1-u, v)$ by $\delta/(1+e^{\delta/\tau})$ whenever the two branches differ by δ. The τ sweep agrees: KD's soft residual falls from 4.2×10^{-2} at τ=1.0 to 0 at τ=0.01, hard scoring stays at zero throughout, and KD is no longer last once $\tau \le 0.05$.

What survives. Reichenbach is exactly zero in every cell under both scorings. Göd-R's antecedent partial under the soft surrogate is numerically zero at the endpoint on this cell ($|\partial_1 I| \le 1.3\times10^{-4}$ in every run of the grid and the τ sweep, because the consequent saturates, $v \to 1$), consistent with the hard operator's $\partial_1 I = 0$. What does not survive is the order of the other four operators, whose second place alternates between Göd-R and Gog-R, and with it the use of KD's last place as a negative control against a family explanation. Two caveats bound this re-check: only 8 of the 120 KD runs meet our pre-registered settling criterion, so these are endpoints, not certified converged values; and the endpoints are rescored, not retrained with hard operators, so a genuine effect of the surrogate on the optimization path is not excluded.

How the exact zeros are reached. On a further 200-run grid at n=60 (depth 1, early stopping off, τ=0.1, lr $\in \{0.05, 10^{-2}, 3\times10^{-3}, 10^{-3}\}$, 10 matched seeds), we attribute every axiom–individual pair at the endpoint to one of four cases: satisfied with antecedent <0.01, with consequent >0.99, with both, or with both endpoints interior. For Reichenbach and KD an exact hard zero requires u=0 or v=1 analytically, so a boundary solution is forced, not observed; the attribution adds which boundary. At lr=3×10^{-3} Reichenbach reaches zero almost entirely by saturating the consequent (0.96 of pairs), not by emptying the antecedent (0.02), and its antecedent sparsity falls from 0.80 to 0.04 of coordinates as the learning rate drops: the exact zero does not depend on antecedent collapse, though collapse is severe at lr=0.05. Łuk-R and Gog-R, which can satisfy exactly with both endpoints interior, do so (0.998 and 0.996 of pairs at lr=3×10^{-3}, 1.000 at 10^{-3}). Endpoint loss alone therefore does not order these solutions: the operator that reaches exact zero does so on the boundary, and the operators with a small residue converge to interior ones. This attribution covers the n=60, depth-1 cell only.

J Extended Scope Checks

Roadmap. External-baseline scope checks: two geometric ontology embeddings and four zero-shot LLM probes on the matched slices, scoping the contribution as mechanism rather than SOTA.

This section positions the operator-routing findings against alternative paradigms that do not expose the same operator triple.

J.1 Geometric Baseline on the Shared D2 Slice

We benchmark two geometric baselines, Order Embedding (Vendrov et al., 2016) and a Box Embedding implementation, on a shared form-1 slice of the D2 benchmark (the 244 of the benchmark's

291 BioPortal ontologies whose held-out split has at least 5 training and 3 test form-1 subsumptions; form-1 $C \sqsubseteq D$ axioms only, since neither baseline scores role-quantified subsumptions; the DF$\mathcal{ALC}$ rows keep the all-form D2 metric). Both geometric baselines and DF$\mathcal{ALC}$ use 5-seed training under the same fixed evaluator semantics. Order embedding scores subsumption by negative reverse-product distance; box embedding represents each concept as an axis-aligned box $[\ell_C, u_C] \in \mathbb{R}^d$ and scores $C \sqsubseteq D$ by negated containment violation $-\sum_k(\mathrm{ReLU}(\ell_D - \ell_C) + \mathrm{ReLU}(u_C - u_D))$; DF$\mathcal{ALC}$ uses the implication value $I(u_i, v_i)$. Table 12 reports the results.

Table 12: **On the pure-subsumption slice two geometric embeddings outscore DF$\mathcal{ALC}$.** Two geometric baselines vs. DF$\mathcal{ALC}$ on the shared form-1 D2 slice (244 ontologies, 5 seeds; the geometric rows use the form-1 subsumption metric and the DF$\mathcal{ALC}$ rows the all-form D2 metric, both under the same fixed evaluator). Both order- and box-embedding outperform every DF$\mathcal{ALC}$ implication on this slice; box-embedding (which can represent existential restrictions in principle but is here scored only on form-1) further outperforms order-embedding by $+0.020$ MRR / $+0.034$ AUC. The geometric advantage on this slice is reproducible across two distinct geometric baselines, not an order-embedding-specific artifact. The comparison is fair only within this slice because neither geometric baseline scores role-quantified subsumptions (forms 4–7).

Method	MRR	AUC-ROC	Hits@10	Form-1 only?
Box Embedding	**0.6863**	**0.8284**	0.9894	yes (closed-form)
Order Embedding	0.6664	0.7944	**0.9906**	yes (closed-form)
DF$\mathcal{ALC}$ Łuk-R	0.5812	0.7449	0.9202	no (handles forms 1–7)
DF$\mathcal{ALC}$ Gog-R	0.5791	0.7444	0.9201	no
DF$\mathcal{ALC}$ Göd-R	0.5321	0.7523	0.9171	no
DF$\mathcal{ALC}$ Rei	0.5165	0.7483	0.8821	no
DF$\mathcal{ALC}$ KD	0.4453	0.6701	0.7749	no

Interpretation. Both geometric baselines' closed-form constructions are targeted at $\mathcal{EL}$-style $C \sqsubseteq D$ subsumption and outperform every DF$\mathcal{ALC}$ implication on this slice; box-embedding further outperforms order-embedding, demonstrating that the geometric advantage on $\mathcal{EL}$ is reproducible across two distinct baselines and not a single-method artifact. This positions fuzzy $\mathcal{ALC}$ as the relevant framework for role-quantified content (forms 4–7) where neither geometric baseline has a representation, not as a competitor on $\mathcal{EL}$. The result also clarifies that the safe-default reading inside fuzzy $\mathcal{ALC}$ (Łuk-R top among DF$\mathcal{ALC}$ implications) is internally valid even when $\mathcal{EL}$-targeted geometric baselines outperform the entire family on this form-1 slice.

J.2 Scope Check Against Axiom-Only LLM Probes

This scope check compares axiom-only LLM probes against the matched DF$\mathcal{ALC}$ slice; the paper's main claims concern operator-induced gradient routing inside differentiable fuzzy $\mathcal{ALC}$.

We benchmark 4 instruct-tuned LLMs (Qwen3-4B, Qwen3-8B, Qwen3-14B, Llama-3.1-8B) on the shared D2 subsumption benchmark under a pairwise natural-language-inference prompt, scoring $\mathrm{score}{=}p_{\mathrm{yes}}{=}\mathrm{softmax}([\ell_{\mathrm{yes}}, \ell_{\mathrm{no}}])_0$ on the model's own $\{\mathrm{yes}, \mathrm{no}\}$ token IDs. Setup: 30 structurally diverse BioPortal ontologies, 240 candidate groups, 5 seeds. Inference is run in `bfloat16` on an A40 (Ampere) with flash attention, which permits the *full* train-fold context with no truncation; the same evaluator is shared with DF$\mathcal{ALC}$ for the slice-matched paired comparison.

Precision audit (fp16 → bf16). An initial fp16 sweep produced $46{-}73\%$ fully-tied candidate groups on every Qwen3 size and a $3\times$ apparent Llama-vs-Qwen gap in strict-top-1. We isolated this to `float16` underflow at the final 2-element softmax over $\{\mathrm{yes}, \mathrm{no}\}$ logits: with $\ell_{\mathrm{no}} - \ell_{\mathrm{yes}} \geq 11$ on confident negatives, $e^{\ell_{\mathrm{yes}} - \ell_{\mathrm{no}}} \approx 1.7 \times 10^{-5}$ rounds to identical fp16 values across distinct prompts. A controlled fp16-logits-with-fp32-softmax run on Qwen3-14B reproduces the bf16 result (fully-tied $0/240$ vs $176/240$ under fp16) while leaving the raw fp16 logit means bit-identical, isolating the artifact to the softmax stage alone. All numbers below use `bfloat16`; the fp16 archives are superseded and excluded from every table below.

Context fairness (head-16 → full). A `head−16` truncated context underestimates LLM performance relative to the matched DF$\mathcal{ALC}$ slice trained on the full fold, most severely for the largest model. Restoring the full context raises LLM AUC across the board (Table 13), most for Qwen3-14B ($0.600 \to 0.836$, $\Delta{=}{+}0.236$) and Qwen3-8B ($0.662 \to 0.818$); the apparent "inverse scaling" seen

under `head−16` is itself a truncation artifact and disappears under full context (the ordering becomes the normal $14\text{B} > \text{Llama} > 8\text{B} > 4\text{B}$).

Table 13: **Under fair full-context prompting DF$\mathcal{ALC}$ is statistically tied with all four LLMs.** LLM subsumption AUC under `head−16` vs. full train-fold context (bf16, A40; per-group AUC, cluster bootstrap over 30 ontologies, 4000 resamples). Full context helps every model; the strongest model was handicapped most by truncation.

Model	head-16	full	Δ (full−head) [95% CI]
Qwen3-4B	0.782	0.796	$+0.014\ [-0.043, +0.071]$
Qwen3-8B	0.662	0.818	$+0.156\ [+0.097, +0.216]$
Qwen3-14B	0.600	0.836	$+0.236\ [+0.160, +0.319]$
Llama-3.1-8B	0.801	0.824	$+0.023\ [-0.005, +0.052]$

Slice-matched DF$\mathcal{ALC}$ paired head-to-head (full context). DF$\mathcal{ALC}$ (Łuk-R, trained on the same full folds; mean train-SR 0.998) scored the same 240 groups with the fixed anti-tautology evaluator (Łuk-R/product/product); per-group AUC $0.781\ [0.712, 0.845]$. Paired cluster bootstrap (DF$\mathcal{ALC}$ − LLM, 4,000 resamples over ontologies):

- DF$\mathcal{ALC}$ vs Qwen3-4B: $\Delta\text{AUC}=-0.015\ [-0.091, +0.063]$.
- DF$\mathcal{ALC}$ vs Qwen3-8B: $\Delta\text{AUC}=-0.036\ [-0.114, +0.047]$.
- DF$\mathcal{ALC}$ vs Qwen3-14B: $\Delta\text{AUC}=-0.055\ [-0.121, +0.008]$.
- DF$\mathcal{ALC}$ vs Llama-3.1-8B: $\Delta\text{AUC}=-0.043\ [-0.114, +0.029]$.

Every CI includes 0, so DF$\mathcal{ALC}$ is *statistically tied* with each model; but every point estimate favors the LLM (DF$\mathcal{ALC}$ wins only 25–30% of groups), and the gap peaks against the strongest model in our set, Qwen3-14B.

Take-away. Under full-context prompting (bf16, no truncation), DF$\mathcal{ALC}$ is statistically tied with all four zero-shot LLM probes, with every point estimate slightly favoring the LLM (furthest below Qwen3-14B); a `head−16` comparison understates the LLMs, most severely for the largest model. The paper's claim is therefore the operator-routing *mechanism*, not state-of-the-art subsumption prediction. The $\mathcal{EL}$-slice comparison in App. J.1 similarly shows two geometric baselines ahead in the form-1 regime where they have a representation. (Scope: 240 groups / 30 ontologies; wide cluster CIs; DF$\mathcal{ALC}$ = Łuk-R anchor.)

K Cross-Task Survival Experiments: Details

Roadmap. The survival-axis experiments in detail: the objective sweep, the softmax dead-zone geometry, the ∀-relational probe, the measurement-caliber confound, and the softmax-consistency probe on a standard classifier.

Objective sweep, summary statistic. Across the six swept implications, the Spearman rank correlation between the operator ranking under the satisfaction objective and the ranking under the discrimination objective is $\rho = -0.44$ (six implications; four tie at 1.0 under satisfaction, exact permutation $p = 0.40$, so we read it descriptively): a one-directional degradation (Łuk-R and Gog-R collapse, Göd-R stays high, the non-saturated implications hold), not a clean inversion. The MNIST numbers are means over target sums 5, 9 and 13 (three seeds each) of the best test-set digit classification accuracy during training, with 600 labeled digits: 35.4 (Łuk-R), 90.6 (SP) and 73.8 (Rei).

K.1 Softmax Simplex Dead-Zone View

Fig. 27 visualizes Prop. 6 without any ontology-specific structure. For a $K = 3$ softmax grounding, a pair of class coordinates $(a, b) = (\pi_i, \pi_j)$ lies in the triangle $a \geq 0, b \geq 0, a+b \leq 1$. The Łukasiewicz body t-norm is active only on the boundary $a + b = 1$ and has zero gradient throughout the simplex interior, including the uniform point $(1/3, 1/3)$. By contrast, free coordinatewise groundings range over the square $[0, 1]^2$, where the escape set $a + b \geq 1$ has area $1/2$. This is the output-geometry version of the grounding-axis claim: the same operator can be harmless under free embeddings and pathological under low-pinned softmax outputs.

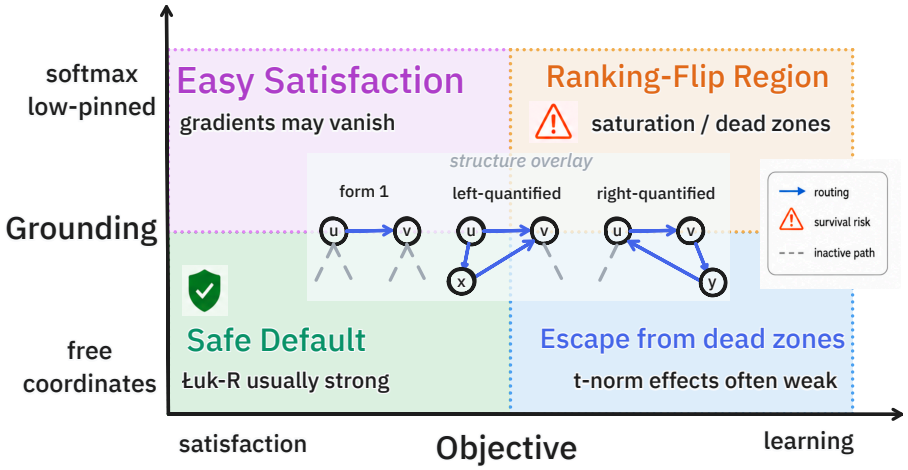


Figure 25: **No universal best operator: the prediction space** (referenced from § 3.2). The best operator depends on where the task sits along structure × objective × grounding. Satisfaction with free coordinatewise groundings is a safe-default region (Łuk-R strong, routing dominates); a learning objective or low-pinned softmax grounding enters a survival-risk region where saturation and dead zones flip the ranking. The structure overlay sets which branch receives gradient before objective/grounding decide whether it survives.

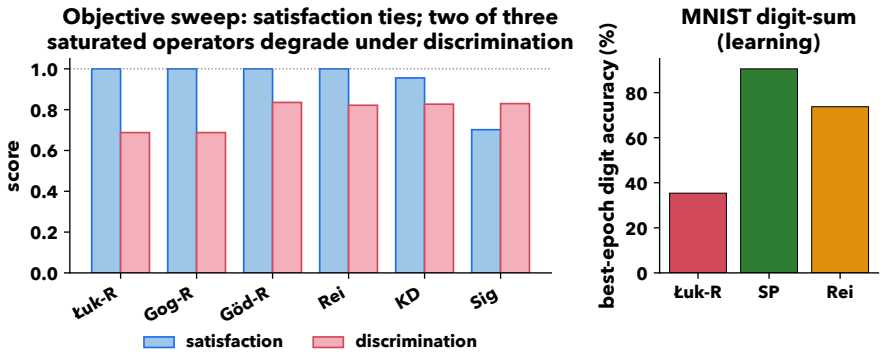


Figure 26: **Saturated operators are matched to satisfaction but lose signal under a learning objective.** Objective axis (§ 4.3). **Left:** on the same data/rule, the satisfaction objective leaves the top operators tied while the discrimination objective collapses two of the three saturated ones (Łuk-R, Gog-R; Göd-R stays high), a one-directional degradation. **Right:** MNIST digit-sum (classifier-grounded learning; 600 labels, target sums $\{5, 9, 13\}$, best-epoch concept (digit) accuracy): Łuk-R starves (35%) relative to Sigmoidal (SP, 91%) and Reichenbach (Rei, 74%).

K.2 The Measurement-Caliber Confound, Across Both T-Norm Sweeps

Fig. 28 makes the fixed-evaluator argument of § 4.4 concrete across both real-ontology t-norm sweeps (conjunction-axis and existential-axis, each high- vs low-density). Self-scored caliber-SR inflates the t-norm η^2 everywhere—including bands where the t-norm is provably off the loss path—because Gödel's min-aggregation saturates the SR metric by construction. Under a fixed evaluator only the *conjunction-on-path* effect survives ($\eta^2 = 0.71$ high-conjunction); the existential effect falls to 0.002 (high density) and 0.11 (low density), and all off-path controls to ≈ 0. The existential self-scored "satisfaction" effect ($\eta^2 = 0.41/0.78$) is thus revealed as caliber, not signal.

K.3 The Universal-Quantifier Collapse Is Relation-Dependent

The survival experiments operationalize $\forall$ in the LTN form, as a product of implication values over the constrained pairs, sweeping the implication (distinct from DF$\mathcal{ALC}$'s t-conorm $\forall$, see § 3.2). A controlled MNIST-relational probe, a CNN trained through a universal constraint over a fixed digit relation π with a fraction of labels supervised, shows that the collapse of Göd-R under a universal

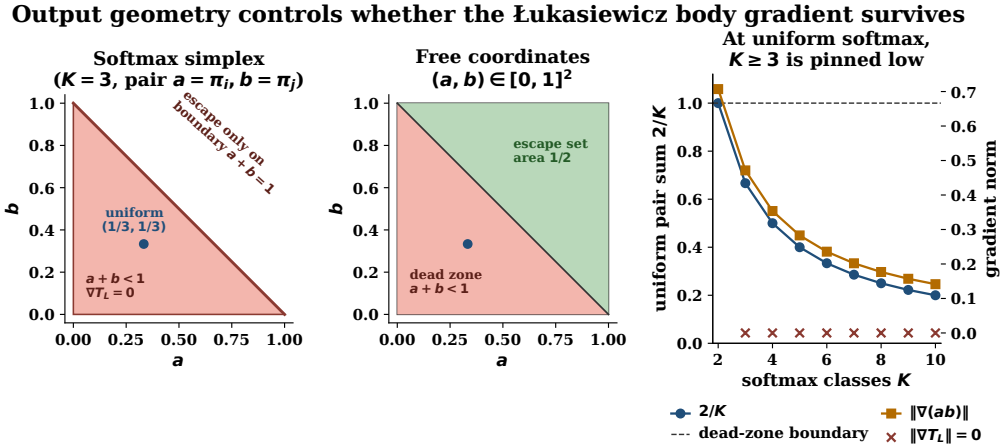


Figure 27: **The Łukasiewicz dead zone is forced by softmax geometry, not the ontology.** Softmax simplex view of the Łukasiewicz dead zone. **Left**: for a $K = 3$ softmax pair, the simplex interior satisfies $a+b < 1$, so the Łukasiewicz body t-norm has zero gradient except on the boundary. **Middle**: free coordinates have a positive-measure escape set ($a + b \geq 1$). **Right**: at the uniform softmax point, $2/K < 1$ for all $K \geq 3$, while the product t-norm retains a nonzero gradient.

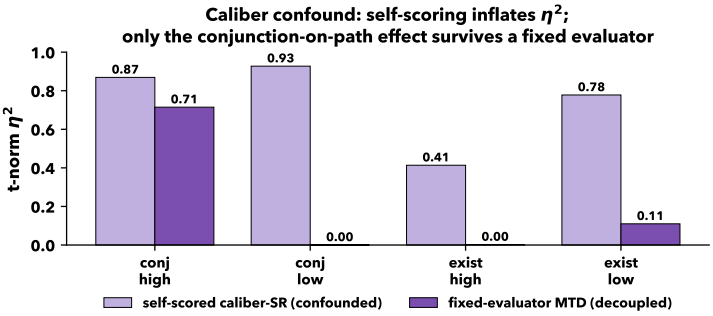


Figure 28: **The t-norm "satisfaction" effect is measurement caliber, not signal.** T-norm η^2 under self-scored caliber-SR vs. the fixed-evaluator mean truth degree (MTD), for the conjunction- and existential-axis real-ontology sweeps (high/low density). Self-scoring inflates every bar; only the conjunction-on-path effect survives the fixed evaluator, isolating genuine operator effects from measurement caliber.

constraint is not a general law but a relation-dependent diagnostic (Fig. 29); the probe folds $\forall$ with a product and does not test $\min$ aggregation. Göd-R collapses to chance only on *bijective* (permutation) relations π, and increasing supervision does not rescue it (it remains at chance even at a 0.5 label fraction); on *non-bijective* relations Göd-R recovers fully once a small label anchor is present. A constraint-off (CE-only) control confirms every operator recovers digits without the constraint, so the collapse is constraint-induced. The mechanism is that bijective relations admit a cyclic-relabeling degeneracy into which Göd-R settles: its gradient is identically zero on satisfied pairs and, on violated pairs, reaches only the consequent; non-bijective relations break that symmetry. We therefore report this only as a scoped diagnostic.

K.4 Survival in a Standard Regularizer: Softmax Consistency, and When the Dead Zone Bites

To probe the survival mechanism outside logical/ontology settings, we instantiate it in a *standard* tool: consistency regularization for semi-supervised classification. The rule "two augmented views of an image share a class" is the existential $\bigvee_c T(p_1[c], p_2[c])$ over a softmax classifier; the consistency loss is $1 - \bigvee_c T(p_1[c], p_2[c])$. This is exactly the dead-zone-prone form of Prop. 6, now over an ordinary classifier rather than an ontology.

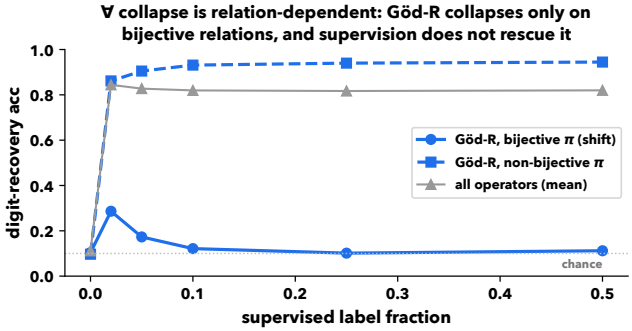


Figure 29: **The ∀ collapse is relation-dependent, not universal.** ∀ isolation (semi-supervised MNIST-relational probe, LTN-form ∀). Göd-R collapses to chance only on bijective relations and is not rescued by supervision; on non-bijective relations it recovers with a small label anchor. The collapse is relation- and grounding-dependent, not a general ∀ law.

The mechanism transfers (Fig. 30). At initialization the softmax is near-uniform, so every class-coordinate pair has $p_1[c] + p_2[c] \approx 2/K < 1$: the Łukasiewicz consistency loss has *identically zero* gradient to the network, while product and Gödel do not. A net-independent sweep over the agreement level m (both views placing probability m on the shared class) shows Łukasiewicz's consistency gradient is exactly 0 for $2m < 1$ and switches on only once $2m > 1$ — a cold-start trap: escaping the dead zone requires the two views to *already* agree, but reaching agreement requires the gradient the dead zone removes. This is geometry, not tuning: scaling the (random) logits does not escape it, because the two views peak on different classes.

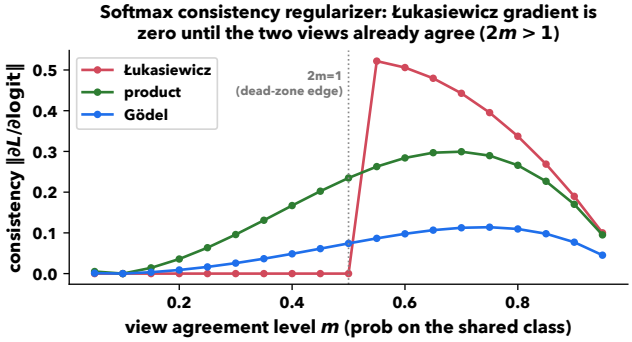


Figure 30: **The dead zone appears under a standard softmax-consistency objective too.** The grounding dead zone in a standard softmax consistency regularizer. Łukasiewicz's consistency gradient is exactly zero until the two augmented views already agree on the shared class ($2m > 1$); product and Gödel retain gradient throughout. This is Prop. 6 in an ordinary classifier, not an ontology.

Accuracy impact under supervision. The dead zone itself is label-independent: across a supervision sweep ($n_{\text{labeled}} \in \{10, 30, 100, 250\}$, 5 seeds) Łukasiewicz's dead-zone occupancy is 1.000 and its encoder consistency-gradient is exactly 0 at every label count, while product's is nonzero throughout. Yet *having* the gradient never converts to an accuracy advantage: product does not beat Łukasiewicz at any label count. At few labels consistency *hurts* every operator equally (pulling two views together before the network has learned reinforces a poor model: both $\approx -3.6\%$ at n=10); at 250 labels both *help* equally ($+2.2\%$ vs. $+2.2\%$; confidence-masked $+1.4\%$ vs. $+1.7\%$), because the supervised cross-entropy concentrates predictions enough that Łukasiewicz escapes the dead zone during joint training. Dead-zone gradient-absence is thus *necessary but not sufficient* for an accuracy penalty: it costs accuracy only when the dead-zoned gradient is the main, irreplaceable productive signal — as in the softmax-grounded per-form scan of § 4.3, where perception is trained only through the operator and the existential form aggregates class probabilities through the body t-norm: there the Łukasiewicz t-norm stays at chance (0.11 class-recovery accuracy, chance 0.125) while product reaches 0.37. In an *auxiliary* regularizer the missing gradient is simply substituted by the labeled signal. This sharpens, rather than weakens, the survival claim: the dead zone is a real, predictable

property of the (operator × softmax-grounding) pair, and we delimit exactly when it does and does not reach the loss.

L TRAINING DYNAMICS

Roadmap. Direct training-dynamics evidence for the routing prediction: per-branch gradient-routing hooks, convergence and stability, and the early-stopping learning curve.

L.1 DIRECT GRADIENT ROUTING OBSERVATION

We instrument the training loop with per-branch gradient hooks on the antecedent and consequent tensors passed to the implication function. Each backward pass records the gradient norm, mean absolute gradient, and active-element ratio for both branches. The experiment covers 3 structures (*baseline (diverse)*, *subsumption-only*, *universal-only*) × 5 implications × 3 t-norm pairs × 3 repetitions = 135 runs at 1,500 epochs each. Aggregate consequent-routing ratios are Łuk-R 0.500 (95% CI $[0.500,\ 0.500]$) and Göd-R 0.600 ($[0.595,\ 0.605]$; per-structure means 0.590–0.617). The hooks record the loss-gradient norm on the antecedent tensor u over the whole trajectory, where the τ=0.1 gate still passes $-\sigma'(1-v)/\tau$; the hard operator's $\partial_1 I = 0$ is its $\tau \to 0$ limit.

Fig. 31 shows the routing ratio summary; Fig. 32 shows the full per-epoch gradient traces.

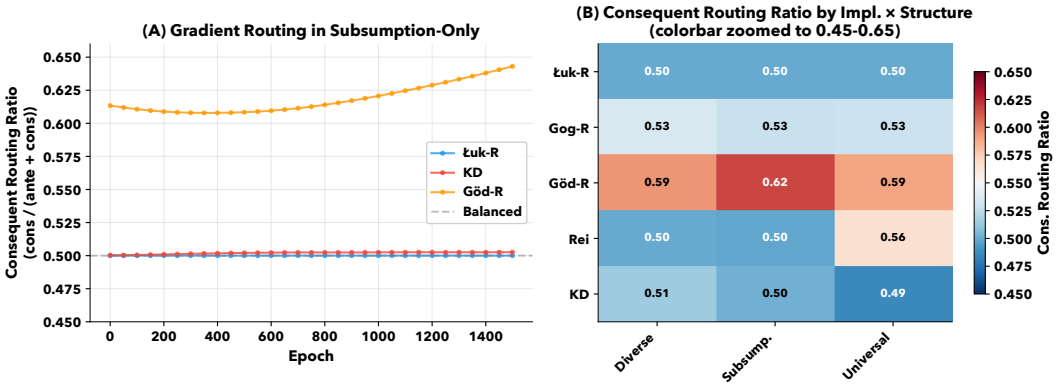


Figure 31: **Direct gradient hooks confirm the predicted branch routing.** **(A)** Consequent routing ratio over training epochs in subsumption-only structure: Łuk-R maintains balanced routing near 0.5 while Göd-R is consequent-skewed and KD, which routes to one branch per element, averages near 0.5. **(B)** Mean routing ratio heatmap across all structures and implications; values near 0.5 indicate balanced gradient support on both branches.

The traces confirm the derivative-structural predictions: Łuk-R maintains near-equal gradient norms on both branches across all structures and throughout training, consistent with its constant partials $\partial_1 I = -1$, $\partial_2 I = +1$; KD routes each grounded pair to a single branch, confirming Cor. 12, but which branch varies across pairs, so its aggregate ratio stays near 0.5; Göd-R is consistently consequent-skewed under the smooth surrogate, matching the hard operator's $\partial_1 I = 0$ limit; and Gog-R and Rei route intermediately, varying with input and structure.

L.2 CONVERGENCE SPEED AND TRAINING STABILITY

We analyze convergence speed and training stability across all 184,231 converged runs from the main corpora. Łuk-R's headline SR / mean-truth-degree advantage in the main text is not an artifact of silently dropping non-converged runs: the main ranking summaries are computed from final checkpoints of all completed runs, while this appendix isolates the converged subset only for convergence-epoch statistics and reports convergence rates separately. Łuk-R converges fastest (median epoch 8–13 across corpora) with near-perfect convergence rate ($> 99.8\%$). KD has the lowest convergence rate (Fig. 33) (12.4% on synthetic, 50.6% on BioPortal, 62.2% on ORE 2015) and the slowest median convergence when it does converge. Mann–Whitney U tests confirm that Łuk-R converges significantly faster than Göd-R, Rei, and KD on all corpora ($p < 10^{-100}$, rank-biserial

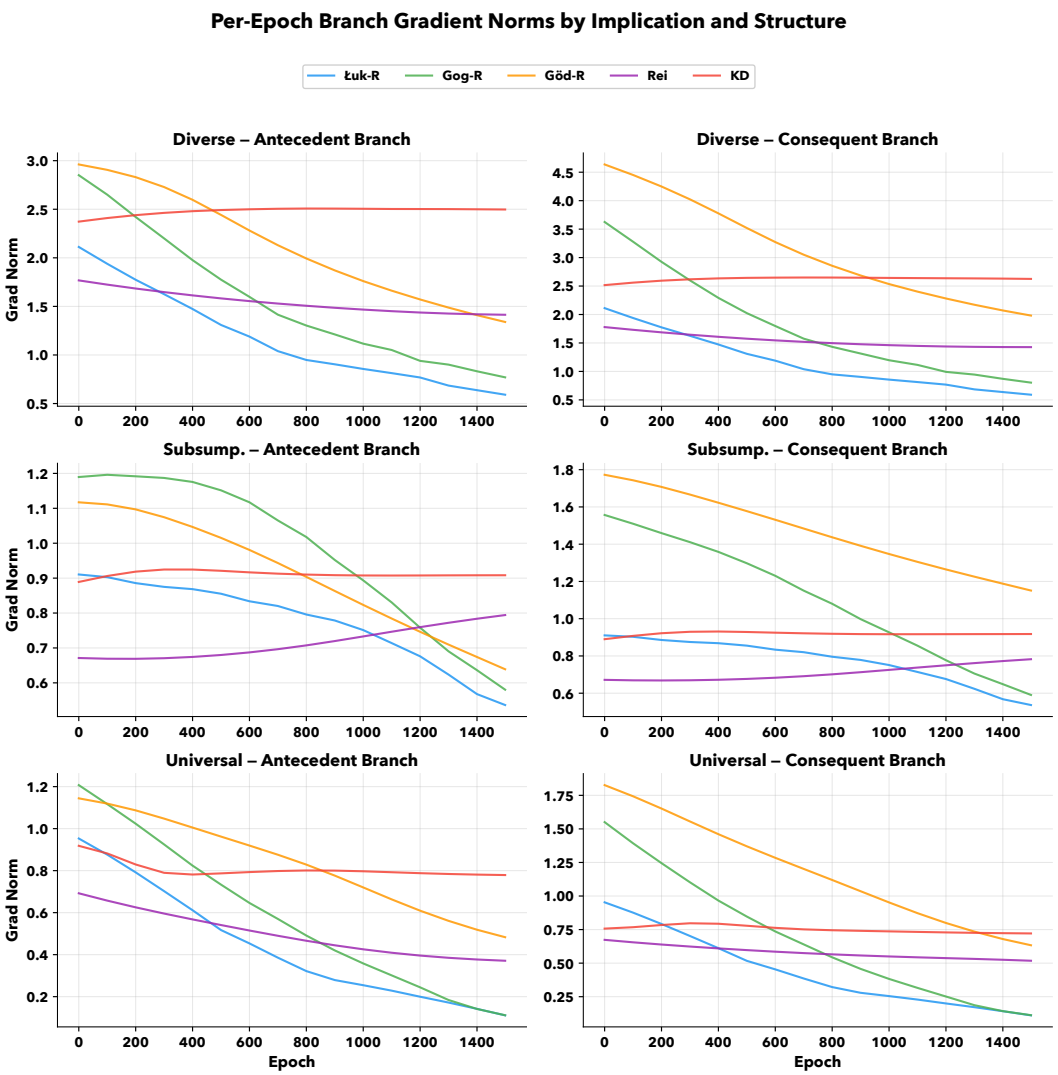


Figure 32: **Per-epoch gradient norms match the predicted per-operator routing.** Per-epoch gradient norms on antecedent (left column) and consequent (right column) branches across three representative structures. Łuk-R (blue) maintains balanced support; KD (red) routes to one branch per element but balances in aggregate; Göd-R (orange) is consequent-skewed.

$r > 0.7$); only Gog-R is comparable in speed. No NaN or Inf events were observed in any run; the stability differences manifest as convergence failure rather than numerical instability. Gog-R exhibits extremely high gradient norm outliers (Fig. 34) (mean cEmb gradient norm $> 500{,}000$ on BioPortal), indicating occasional gradient explosion that does not trigger NaN but may explain its lower convergence rate relative to Łuk-R.

L.3 Early-Stopping Learning Curve

We test whether the implication ranking is preserved under reduced epoch budgets (Table 14). The experiment sweeps four budgets of 375, 750, 1125, and 1500 epochs (25%, 50%, 75%, 100% of the full budget) across the same 3 structures, 5 implications, and 3 t-norm pairs as the per-branch gradient-routing experiment (App. L.1), with 3 repetitions each (540 runs total).

Key observations:

- The learning curves (Fig. 35) show that the top-2 order Łuk-R $>$ Gog-R is preserved at every budget, including the 25% checkpoint.

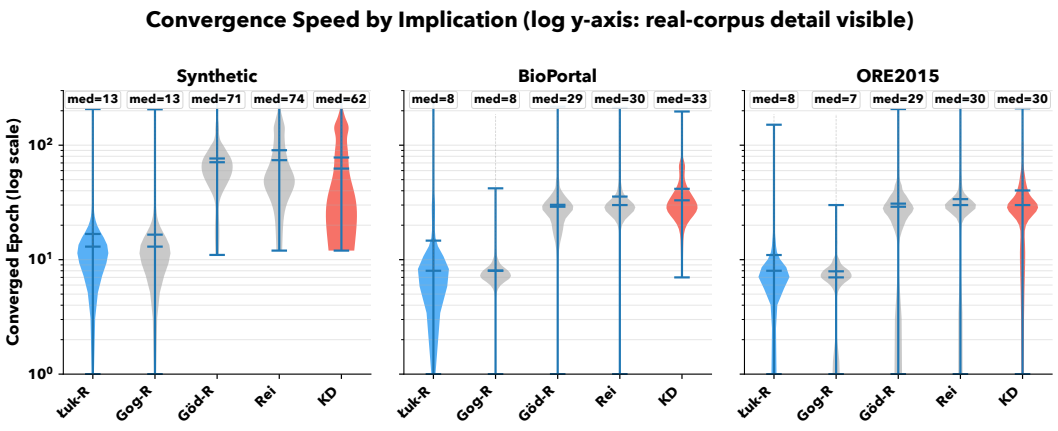


Figure 33: **Łuk-R converges fastest and most reliably; KD often fails to converge.** Convergence epoch distributions by implication across three corpora. Łuk-R (blue) converges fastest and most reliably; KD (red) frequently fails to converge within the training budget.

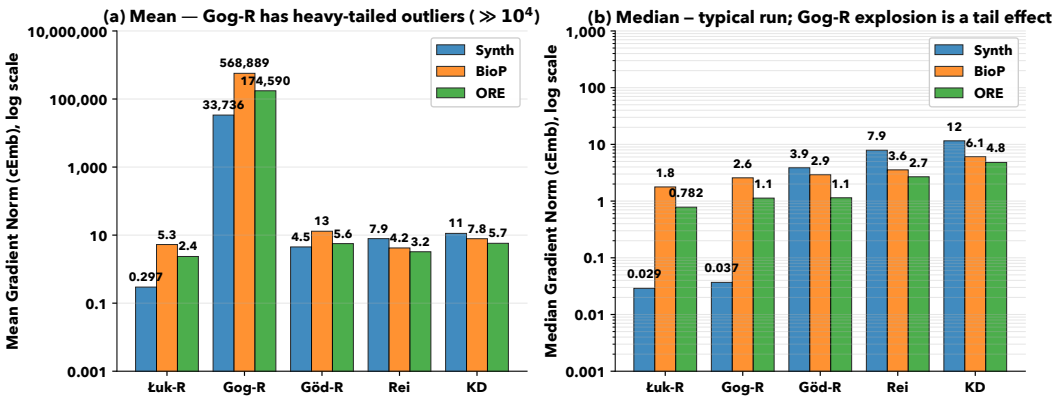


Figure 34: **No NaN or training-failure events; Gog-R shows occasional gradient explosion.** Training gradient norms by implication and corpus. NaN rate and training failure rate are zero throughout and are therefore not shown. Gog-R exhibits high gradient norm outliers on BioPortal, indicating occasional gradient explosion that does not trigger NaN but may explain its lower convergence rate relative to Łuk-R.

Table 14: **Mean SR(0.5) by implication and epoch budget, pooled over the 3 structures and 3 t-norm pairs (3 repetitions each, 540 runs).**

Implication	375 ep (25%)	750 ep (50%)	1125 ep (75%)	1500 ep (100%)
Łuk-R	0.686	0.832	0.980	0.988
Gog-R	0.575	0.746	0.946	0.984
Göd-R	0.199	0.359	0.469	0.552
Rei	0.283	0.291	0.280	0.409
KD	0.032	0.063	0.173	0.303

- The canonical five-way ranking (Łuk-R > Gog-R > Göd-R > Rei > KD) is fully recovered by 50% of the budget; only the Rei/Göd-R pair is reversed at 25% (0.283 vs. 0.199).
- Łuk-R and Gog-R reach high SR faster than slower-converging implications, consistent with the convergence speed analysis (App. L.2).
- The universal-only structure shows Gog-R > Łuk-R at the full budget, a structure-conditioned crossover that is already detectable at the 75% checkpoint.

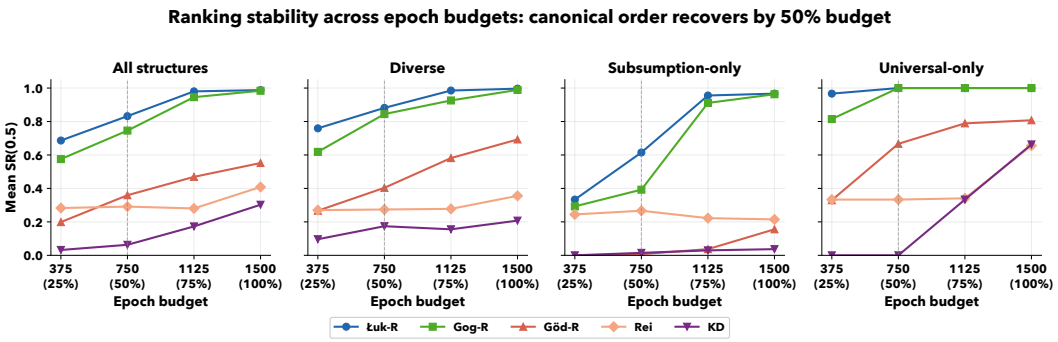


Figure 35: **The canonical ranking is recovered by half the epoch budget.** SR(0.5) vs. epoch budget for each implication. **(A)** Aggregated across all structures. **(B–D)** Per-structure breakdown. Dashed vertical line marks the 50% budget (750 epochs), at which point the canonical ranking is fully recovered. Łuk-R (blue) leads at all checkpoints overall.

M Extended Limitations and Discussion

Roadmap. Extended limitations: description-logic fragment coverage, mixed-operator strategies, and the bounds of synthetic-to-real transfer.

We discuss four limitations of the present work, covering the technical obstacles, the extent to which the current evidence addresses them, and the research directions they suggest.

M.1 Theoretical Scope: Łuk-R/Göd-R Only

The formal descent bounds (Thms. 2 and 3) compare Łuk-R and Göd-R because these two implications have the maximal qualitative contrast in $\partial_1 I$: constant nonzero (-1) versus identically zero. Extending formal bounds to Reichenbach ($\partial_1 = -(1-v)$) or Goguen ($\partial_1 = -v/u^2$) would require bounding the input-scaled derivatives, which depend on the joint distribution of (u, v) during training. Specifically, Reichenbach's directional derivative involves a weighted sum $\sum_i w_i(1-v_i)\langle \nabla u_i, \Delta \rangle$, where the $(1-v_i)$ factors are not constant across axioms and change as training progresses. Bounding this sum would require distributional assumptions on how v_i concentrates during optimization, which is itself operator-dependent. Despite this theoretical gap, the partial-derivative structure in § 2.2 generates accurate qualitative predictions for all five implications (§ 3, "From routing to testable predictions"), and the 135-run direct gradient measurement in App. L.1 confirms these predictions quantitatively. A natural next step is to formalize these predictions under mild distributional assumptions on the inputs.

M.2 Logic Fragment: $\mathcal{ALC}$ Only

The seven normalized forms in Table 1 cover the standard $\mathcal{ALC}$ constructors. More expressive description logics introduce additional constructors: number restrictions ($\geq n\, R.C$, $\leq n\, R.C$) in $\mathcal{SHIQ}$ would require a counting aggregation inside the truth-value computation, adding a new internal operator whose routing effects are currently unexplored; transitive role axioms ($R^+ \sqsubseteq R$) in $\mathcal{SROIQ}$ would chain role embeddings, lengthening the internal gradient path between the implication gate and the role parameters. In both cases, the core observation that implication sits at the top level and gates all downstream paths remains valid. What changes is the depth and complexity of the internal paths, which could amplify or attenuate the t-norm/t-conorm effects that are currently small ($\eta^2 < 0.08$). Extending the experimental framework to $\mathcal{SHIQ}$ is a concrete next step, as several ontology embedding systems already support this fragment.

M.3 Mixed-Operator Strategies

The current design fixes one implication for the entire TBox. The structure-conditioned crossovers in § 4.2 suggest a more flexible approach: different axiom forms could use different implications. For example, left-quantified axioms (forms 6, 7) could use Łuk-R (leveraging its descent dominance from Thm. 3), while right-quantified chain-existential axioms (form 4, chain topology) could use Reichenbach (leveraging its self-damping advantage from the chain-existential analysis). The techni-

cal challenge is that a mixed-operator strategy breaks the factored structure of the operator grid. In the current framework, the $3 \times 3 \times 5$ grid treats the three operator roles as independent choices; a form-conditioned implication creates dependencies between the axiom structure and the operator, turning the grid into a structured search space. App. G.4 already probed real-corpus *whole-TBox* switching directly and found no deployable gain (the online policy's empirical switch precision, 0.309, stays below the MRR break-even $p^* = 0.515$ on the unselected corpus); the narrower direction left open here, which those whole-TBox selector probes did not test, is the *per-axiom-form* rule above: assigning implications by axiom form within one TBox rather than one operator per ontology.

M.4 Synthetic-to-Real Transfer

The synthetic-to-real transfer of the aggregate implication-dominance finding is robust: $\eta^2_{\text{impl}} \geq 0.31$ on both BioPortal and ORE 2015, and the qualitative Łuk-R-vs-Göd-R dominance pattern predicted by Thm. 3 appears on $\geq 95\%$ of real ontologies (§ 4.4), subject to the loss-scale caliber caveat of App. I (the satisfaction comparison is descriptive corroboration; the theorem-level evidence is the init-state gradient measurement). The transfer of specific crossover patterns is weaker: the aggregate five-implication rank correlation is $\rho = 0.30$ between synthetic and BioPortal ($n = 5$, $p = 0.624$) and $\rho = 0.40$ between synthetic and ORE 2015 ($n = 5$, $p = 0.505$). We treat these as descriptive distribution-matching diagnostics rather than confirmatory tests because the units are aggregate operator ranks, not exchangeable ontology-level samples. This gap reflects the distributional mismatch between the one-factor-at-a-time synthetic generator (which produces extreme structural settings like pure subsumption or pure existential) and the mixed-form structure of real ontologies. The theoretical predictions do not depend on synthetic-to-real transfer at all: Thm. 3 is stated in terms of axiom-form counts, which are directly computable from any real ontology. These weak aggregate correlations apply to distributional matching ("do real ontologies look like synthetic setting X?"), not to the structural laws themselves.

N Extended Related Work

Roadmap. Extended related work across differentiable logic, single-operator gradient analysis, ontology embeddings, and reasoning shortcuts.

The main text (§ 1) situates the paper within three families of differentiable logic. This appendix expands on the positioning against specific systems and research threads that share technical surface area with our work.

N.1 Fuzzy Neuro-Symbolic Systems and Operator Conventions

The operator-choice question studied here arises in every system that grounds logical connectives through fuzzy operators. LTN (Badreddine et al., 2022; Serafini et al., 2017; Donadello et al., 2017) ground first-order formulas as real-valued tensors, typically under a product t-norm and the Reichenbach implication, but the framework permits arbitrary choices. Logical Neural Networks (Riegel et al., 2020) take the opposite design: they learn tight bounds on fuzzy truth values using a weighted form of the Łukasiewicz t-norm, making the operator itself a parameterized component. Sen et al. (2022) extend LNN to inductive logic programming, retaining the learned-weight Łukasiewicz convention. LYRICS and Deep Logic Models (Marra et al., 2019) use product-based fuzzy semantics to inject logical constraints into neural classifiers. In each case, the operator is chosen once per system and treated as a design constant. Our gradient-routing analysis applies to all of these: the implication sits at the top level of every constraint, and its partials gate the learning signal whether the system is rule-based (LTN, LYRICS), bound-based (LNN), or subsumption-based (DF$\mathcal{ALC}$).

N.2 Ontology Embedding Methods

Ontology embedding research has progressed from geometric approaches for lightweight fragments to neural approaches for expressive logics. For $\mathcal{EL}^{++}$, Kulmanov et al. (2019) introduced EL Embeddings with order-preserving constraints, Xiong et al. (2022) proposed faithful embeddings, and Jackermeier et al. (2024) developed DualBox with lattice-preserving geometry. Zhapa-Camacho & Hoehndorf (2024) extended lattice preservation to $\mathcal{ALC}$, and Mashkova et al. (2024) studied categorical embeddings. OWL2Vec* (Chen et al., 2021) combines structural graph walks with

lexical features, bypassing explicit fuzzy operators entirely. Adamski & Potoniec (2024) learn transferable concept embeddings that generalize across ontologies. For full $\mathcal{ALC}$, Hohenecker & Lukasiewicz (2020) used deep networks for ontology reasoning, Özçep et al. (2023) studied lattice-based embeddings, and Em$\mathcal{ALC}$ (Wu & Zhao, 2025) and FALCON (Tang et al., 2022) implement the full fuzzy-$\mathcal{ALC}$ pipeline that our operator grid probes. None of these systems has studied how the choice of fuzzy operators interacts with ontology structure to shape gradient flow. Our work fills this gap within the fuzzy-$\mathcal{ALC}$ regime, while the geometric and text-augmented approaches operate in complementary regimes where the operator-choice question does not arise.

N.3 Differentiable Logic Beyond Fuzzy Relaxation

Other differentiable reasoners either avoid fuzzy connectives or place them in a different architecture, which helps locate our results. NeuralLP (Yang et al., 2017) composes TensorLog's relation operators through a soft attention controller and uses no fuzzy connective. ∂ILP (Evans & Grefenstette, 2018) forward-chains with softmax-weighted clauses generated from rule templates and chooses its fuzzy connectives for gradient flow: it takes the product t-norm over the Gödel and Łukasiewicz ones because the Gödel t-norm sends no gradient to the larger argument and the Łukasiewicz t-norm none at all when the arguments sum below 1, and it amalgamates conclusions with the probabilistic sum rather than $\max$ for the same reason, an early instance of the routing and survival effects formalized here. End-to-end differentiable proving (Rocktäschel & Riedel, 2017; Minervini et al., 2020) uses vector similarity as a soft unification mechanism. LogicMP (Xu et al., 2024) formulates logical message passing as a differentiable fixed-point iteration. Slusarz et al. (2023) proposed a meta-language, the Logic of Differentiable Logics, that provides a uniform semantics for reasoning about the properties of differentiable logics, including fuzzy ones. In each case, the "top-level gate" that controls gradient flow is architecturally determined (attention weights, similarity scores, message functions) rather than selected from a family of named operators. Our gradient-routing observation generalizes to these settings in the structural sense identified in § 6: the top-level differentiable operation, whatever it is called, gates all downstream gradient paths. The practical question of whether the specific operator choice matters depends on whether the system exposes a choice at all, and if so, whether the alternatives have qualitatively different derivative structures. Two recent differentiable-ILP systems make this axis explicit in the way our principle predicts. Differentiable Logic Machines (Zimmer et al., 2023) place learnable weights on *predicates* rather than rule templates, so the continuous relaxation's gradient flow is governed directly by the chosen connective semantics. ANDRE (Sharifi et al., 2026) replaces product-t-norm conjunction/disjunction with attention-based operators precisely because the product t-norm has vanishing gradients at the Boolean extremes ($a{=}b{=}0$ or $a{=}b{=}1$), and proves its attention operators approximate hard min/max more faithfully—an independent instance of our *survival* reading (an operator whose derivative vanishes on the active region cannot route a learning signal), with attention as the gradient-preserving router that the product t-norm is not.

N.4 Neuro-Symbolic Surveys and Taxonomies

Several surveys organize the neuro-symbolic landscape. Raedt et al. (2020) identify parallels between statistical relational and neuro-symbolic AI along seven dimensions and use them to position neuro-symbolic approaches. Marra et al. (2024) provide a comprehensive survey connecting probabilistic, fuzzy, and constraint-based approaches. Besold et al. (2021) edited a collection covering foundational and applied aspects. Odense & d'Avila Garcez (2025) proposed a semantic framework for neuro-symbolic computation that formally separates the "neural" and "symbolic" components. Under any of these taxonomies, our work analyzes how the "relaxation interface" between the symbolic and neural components shapes optimization.

O Reasoning-Shortcut Theory: Predict, Mitigate, Derive

Roadmap. The reasoning-shortcut theory in full: the identification proposition, the supervision barrier, the exact trapping-threshold hierarchy, the SGD escape-time dichotomy, the basin-selection law, and the RSBench corroboration.

A *reasoning shortcut* (Marconato et al., 2023) is a neuro-symbolic model that satisfies its logical constraint while learning concepts that disagree with the ground truth; van Krieken et al. (2024) gave a probabilistic-side analog (independence assumptions create disconnected, biasing landscapes). The

survival half of our principle yields a fuzzy-side, *operator-predictable* account: the very saturation that makes Łuk-R a satisfaction-regime safe default (Prop. 4) also turns constraint-satisfying-but-wrong concept assignments into stationary optima of the logic loss.

A scope note: van Krieken et al. (2024) observe (their App. A) that the Łukasiewicz *t-conorm* does not bias toward deterministic solutions, unlike the product and Gödel ones, and suggest this may explain its practical effectiveness. That observation concerns a different operator role (disjunction, not implication) and a different mechanism (the direction in which gradient is biased, not whether gradient exists); they separately note that Łukasiewicz logic suffers vanishing gradients. Our claim, that the Łukasiewicz *implication* is saturated and so stabilizes shortcuts, agrees with both.

The concept×constraint shortcut diagnostic (provenance for § 5). Table 15 backs the body's headline numbers: the field-standard diagnostic (constraint satisfaction vs. true-concept accuracy (Marconato et al., 2023)) on the van Krieken MNIST digit-sum (K=10, sum target 9, 10 seeds, 300 labeled digits, best-epoch). Saturated/intermediate operators satisfy the constraint while their concepts stay near chance (the shortcut); KD/SP attain both (genuine learning). Same pipeline as Fig. 4, read at the single 300-label point.

Table 15: **Saturated operators satisfy the constraint with wrong concepts.** Concept×constraint shortcut diagnostic (K=10, 300 labels, 10 seeds; chance concept = 10%). High constraint + low concept = reasoning shortcut.

	Göd-R	Gog-R	Łuk-R	Rei	SP	KD
concept acc. (%)	20	20	22	43	79	94
constraint sat. (%)	61	72	76	78	83	89
reading	collapse	**shortcut**	**shortcut**	partial	learns	learns

Prop. 7 (stated in § 3.2) makes this precise.

Proof of Prop. 7. Saturation gives $I(u_k, v_k) = 1\,\forall k$, so $L_{\text{logic}} = \sum_k (1-1) = 0$, its minimum ($I \leq 1$). By the chain rule $\nabla_\theta L_{\text{logic}} = -\sum_k [\partial_1 I\, \nabla u_k + \partial_2 I\, \nabla v_k]$, and $\partial_1 I = \partial_2 I = 0$ on $\{v > u\}$ ("with slack" means $v_k > u_k$, so no instance sits on the boundary $v = u$, where Göd-R is discontinuous), so it vanishes; a shortcut satisfies the constraint, hence lies in this region. For non-saturated I (e.g. Rei $\partial_2 = u > 0$, or KD $\max(1-u, v) < 1$ on the satisfied region), the gradient is generically nonzero. □

Logic-loss-weight control. A logic-loss-weight control (same protocol, seeds 0–2) separates the two saturated operators. At weight 1 Gog-R escapes: concept accuracy 92.7% at 300 labels and 98.0% at 3,000, level with supervision alone (93.2%, 98.0%). Göd-R stays trapped at 300 labels (31.9%) and at 3,000 escapes on one seed of three (97%; the others reach 52% and 26%). At weight 0.1 both escape on every seed (92.6% and 93.3% at 300 labels, 98.0% and 98.1% at 3,000), level with supervision alone. Göd-R thus needs a lower logic-loss weight than Gog-R to escape, consistent with its hard branch indicator being the causal core (Fig. 37). KD is unaffected by the weight (93.3% at 300 for weights 1 and 0.1).

The proposition predicts which operators admit shortcuts purely from their derivative structure: Łuk-R, Göd-R, and Gog-R are saturated; Rei, KD, and SP are not—KD= $\max(1-u, v) < 1$ on the satisfied region, so, contrary to a naive "starvation" reading, it *learns*. We confirm this on MNIST digit-sum, where a shortcut means satisfying "the digits sum to the target" with the wrong digits (constraint-satisfaction vs concept (digit) accuracy; 600 labels, target sums $\{5, 9, 13\}$, five seeds each, best-epoch concept accuracy, of which § 4.3 reports the three-seed subset; Fig. 36). Concept accuracy is sharply operator-determined: pooled across target sums the three saturated operators learn the true digits poorly (concept accuracy 20–35%: Göd-R 20, Gog-R 23, Łuk-R 35) while still reaching far higher constraint satisfaction, whereas the non-saturated operators learn them (70–91%: Rei 70, SP 86, KD 91). The shortcut signature is cleanest at sum=9, where the saturated operators show a +37 to +43 constraint-minus-concept gap and the non-saturated ones sit on the diagonal (gap ≈ -4); pooled over all three target sums the saturated gap is +19 to +47. The one nuance the sweep exposes is itself predicted: when the constraint is also hard to satisfy (Gog-R at sum=13) saturation yields outright *collapse* rather than a clean shortcut, and a borderline operator (Rei at

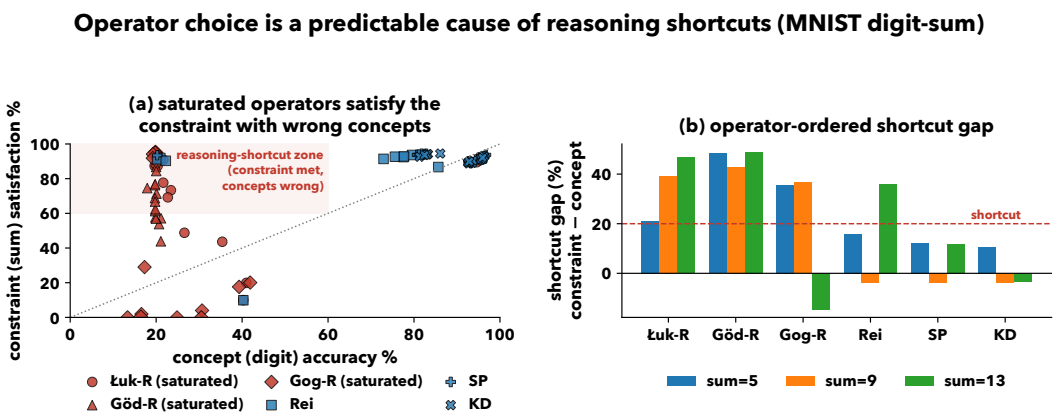


Figure 36: **Operator choice is a predictable cause of reasoning shortcuts.** (MNIST digit-sum; target sums $\{5, 9, 13\}$, five seeds). **(a)** saturated implications (Łuk-R, Göd-R, Gog-R; red) reach high constraint satisfaction with low concept (digit) accuracy (left of the diagonal; shaded shortcut zone), while non-saturated ones (Rei, KD, SP; blue) learn the true concepts (upper right); the concept axis cleanly separates the two families. **(b)** constraint-minus-concept gap per target sum: large and positive for saturated operators; small at sum=9 for learners, with the predicted nuances shown faithfully (Gog-R *collapses* at sum=13; Rei degrades at sum=13), as Prop. 7 predicts.

sum=13) drifts toward the shortcut—both are the same missing concept gradient surfacing as either a satisfying-but-wrong optimum or a degenerate one. The property that makes Łuk-R a safe *satisfaction* default is thus exactly what impairs its ground-truth concept learning: one derivative-structural cause, predicted not diagnosed. The split is modality-general, not an MNIST-pixel artifact: rerun on *FashionMNIST* (sum=9, K=10, 300 labels, five seeds) the same operator partition holds—saturated Göd-R/Gog-R/Łuk-R stay near chance (concept 19/21/23%) while non-saturated SP/KD recover the concepts (67/66%, with Rei intermediate at 28%)—so the saturation mechanism is a property of the operator's derivative, not the input distribution.

Why concept supervision cannot rescue a saturated shortcut: a logic barrier. Prop. 7 shows the logic loss is *silent* at the shortcut, so only supervision could correct it. The mitigation experiment (Fig. 4) shows that for Gog-R, whose violated-side wall is $\Theta(K)$, and Göd-R, whose wall is a jump (Prop. 25), even dense supervision does not: concept accuracy stays near chance to 30,000 labels. The following makes that mechanism precise: the shortcut sits behind a *positive logic barrier* that a bounded supervision budget cannot cross.

Proposition 23 (Supervision barrier around saturated shortcuts)**.** *Let θ^* be a shortcut at which every unsupervised grounding lies in the interior of the saturation region $R_I = \{\partial_1 I = \partial_2 I = 0\}$, so $L_{\text{logic}}(\theta^*) = 0$ and $\nabla L_{\text{logic}}(\theta^*) = 0$ (Prop. 7); let N be the connected flat satisfied component through θ^*, and train $F = L_{\text{logic}} + \lambda R$ with concept supervision $R \geq 0$ (per-label loss $\leq r_{\max}$, $n_{\sup}$ labels). Assume (i) θ^* is a local minimum of R restricted to N (the labeled constraints are also met by the shortcut, as under constraint-consistent weak supervision), and (ii) the truth-aligned set is separated from N by a positive logic mountain-pass barrier $\Delta := \inf_{\gamma:\theta^*\to\theta_{\text{true}}} \max_t L_{\text{logic}}(\gamma(t)) > 0$. Then for effective budget $\lambda\, n_{\sup}\, r_{\max} < \Delta$, every path from θ^* to the truth crosses a positive barrier of F, so descent in the basin of N does not reach the ground-truth concepts. For a* non-saturated *I, R_I has empty interior on $\{v \geq u\}$, so generically $\nabla L_{\text{logic}}(\theta^*) \neq 0$ and θ^* is not stationary for F—the flat trap, and with it this barrier, is absent. Assumption (i) holds under constraint-only supervision; with direct concept labels, as in our MNIST experiment, it need not, and the proposition then describes only the logic term's share. The benchmark's concept loss is also a batch mean, so its effective budget scales with the supervision weight rather than with $n_{\sup}$.*

Proof. On N, $L_{\text{logic}} \equiv 0$, so $F = \lambda R + \text{const}$ and, by (i), θ^* is a local minimum of $F|_N$ with $F(\theta^*) = \lambda R(\theta^*) \leq \lambda\, n_{\sup}\, r_{\max}$. Since $R \geq 0$, $F \geq L_{\text{logic}}$ pointwise, so any path γ to θ_{true} has $\max_t F(\gamma(t)) \geq \max_t L_{\text{logic}}(\gamma(t)) \geq \Delta$. Hence the escape barrier is $\max_t F(\gamma(t)) - F(\theta^*) \geq \Delta - \lambda\, n_{\sup}\, r_{\max} > 0$. For non-saturated I, $\nabla L_{\text{logic}} = -\sum_k [\partial_1 I\, \nabla u_k + \partial_2 I\, \nabla v_k]$ has a nonvanishing summand a.e. on the satisfied set, so θ^* generically fails to be stationary. □

This explains the mitigation result mechanistically: a non-saturated implication removes the flat trap (at the implementation point its ρ^* is 1.4 for Rei and 0 for KD, so a free swap recovers concepts), whereas a saturated one keeps a barrier Δ that supervision must out-budget, and dense supervision fails when Δ is large. In general the exact Δ depends on the grounding geometry; in the *canonical* geometry of the experiment, however, the threshold—and the full resistance *ordering*—can be computed exactly, which we do next.

From a budget bound to an exact threshold: deriving the resistance hierarchy. Consider the one-dimensional path model matched to the experiment's geometry (the softmax grounding forces the satisfying concept classes to be *disconnected*: one class's probability must fall before another's rises, so the constraint is violated in transit). Along a path $\theta \in \mathbb{R}$ from the shortcut point θ_B (component B) to the supervised truth target $\theta_{\text{true}} \in \operatorname{int} A$ (component A; we write the shortcut as θ_B here, freeing θ_{true} for the target, whereas Prop. 23 wrote the shortcut θ^*), the groundings are $u(\theta) \in (0, 1]$ (antecedent) and $v(\theta) \in [0, 1]$ (consequent), piecewise C^1, with $v < u$ on the gap G between B and A; supervision contributes $\lambda n R(\theta)$ with $R = \frac{1}{2}(\theta - \theta_{\text{true}})^2$ (n labels, per-label weight λ), and the logic loss is $L_I(\theta) = 1 - I(u(\theta), v(\theta))$. *Pinning hypothesis* (P_K): on G, $u = \Theta(1/K)$ with $|u'| = O(1/K)$, while $|v'| = \Theta(1)$—the precise form of "softmax-pinned grounding" (cf. Prop. 6), with K the number of classes.

Lemma 24 (Exact trapping threshold)**.** *Suppose L_I admits one-sided derivatives, and the initialization has entered the shortcut-to-wall basin ($d^+ L_I \le 0$ outside the wall $W = \{\theta \in G : d^+ L_I(\theta) > 0\}$). Define*

$$\rho^* \;=\; \frac{1}{\lambda} \sup_{\theta \in W} \frac{d^+ L_I(\theta)}{\theta_{\text{true}} - \theta}.$$

If $n > \rho^*$, $F_n = L_I + \lambda n R$ is strictly decreasing along the whole path outside A, and gradient flow from any $\theta_0 \in B$ reaches A. If $n < \rho^*$, the flow stops at the first point with $d^+ F_n \ge 0$—a trapped local minimum strictly outside A. ($n = \rho^*$ is left unclassified; kinks are read in the differential-inclusion sense.) Note the threshold is gated by the wall's slope*, not its area: strictly sharper than the budget bound of Prop. 23.*

Proof. $F_n' = L_I' + \lambda n(\theta - \theta_{\text{true}})$, and $\theta_{\text{true}} - \theta > 0$ on the path, so $F_n' < 0$ at θ iff $\lambda n(\theta_{\text{true}} - \theta) > d^+ L_I(\theta)$; taking the supremum over the wall gives both directions. For $n < \rho^*$ some wall point has $d^+ F_n > 0$; the flow, which moves right while $F_n' < 0$, halts at the first zero. □

Proposition 25 (Five-operator hierarchy of violated-side walls under (P_K))**.** *On the violated gap ($v < u$), direct computation of $L_I = 1 - I(u, v)$ gives the wall slopes, hence by Lem. 24 the thresholds, up to the constants of (P_K): (i) Gog-R: $L = 1 - v/u$, slope $\Theta(K)$; (ii) Łuk-R: $L = u - v$ on $\{v < u\}$, slope $\Theta(1)$ with wall height $\le u$; (iii) Rei: $L = u(1 - v)$, slope $\Theta(1/K)$; (iv) KD: $L = \min(u, 1 - v) = u$ constant in v on the active branch—a* flat-topped *wall, slope $0 + O(|u'|) = O(1/K)$: no v-trap under exact pinning; (v) hard Göd-R has a* jump *of height $1 - u$ at the satisfied boundary $v = u$ (as $v \uparrow u$, $I = v \to u \ne 1$): no finite n crosses it under gradient flow ($\rho^* = \infty$), an idealized statement, since autograd sees no force from the jump itself; under the implementation's τ-smoothed surrogate the cliff becomes a wall of slope $\Theta((1 - u)/\tau)$, so ρ^*(Göd-R$_\tau$) $= \Theta(1/(\lambda\tau))$—independent of K (it saturates). Hence ρ^*: Gog-R $= \Theta(K) >$ Göd-R$_\tau = \Theta(1/\tau) >$ Łuk-R $= \Theta(1) >$ Rei $= \Theta(1/K) >$ KD ≈ 0.*

At the implementation point ($\tau = 0.1$, $K = 10$) the model evaluates to $\rho^* = 124 \,/\, 34 \,/\, 12.5 \,/\, 1.4 \,/\, 0$ for Gog-R/Göd-R$_\tau$/Łuk-R/Rei/KD, matching the observed mitigation ordering on every *resolved* rank (Fig. 4, escape = final concept accuracy $\ge 90\%$: Göd-R and Gog-R unbeaten through $30{,}000$ labels, so their relative order is censored; Łuk-R escapes on only 4 to 7 of 10 seeds at any budget from $1{,}200$ labels; Rei escapes on 9 of 10 seeds at 600 labels and on all 10 from $1{,}200$, and KD on all 10 from 300). We state this as an idealized mechanism (gradient flow, one path, model constants), not a derivation of the experiment. It yields three falsifiable predictions the data did not train: (i) the Łuk-R/Rei threshold ratio grows $\sim$linearly in K (in-model: $4.4/8.9/17.8/35.6$ at $K{=}5/10/20/40$); (ii) Gog-R's threshold grows with K while Göd-R$_\tau$'s saturates, so a larger label sweep should see Göd-R escape before Gog-R; (iii) Göd-R's resistance is τ-controlled ($\rho^* \sim 1/\tau$): the most surgical causal test, since it varies neither task nor K.

Testing the predictions: the resistance comes from the discontinuity. We pre-registered two interventional tests (predictions and kill-switches committed before the runs), and report both

halves of the outcome. *τ-intervention (prediction (iii)).* We replace only Göd-R's hard branch indicator $\mathbf{1}[u \le v]$ with $\sigma((v-u)/\tau)$ ($\tau \to 0$ recovers the hard operator exactly; task, data, and hyperparameters unchanged) and sweep $\tau \in \{1.0, 0.5, 0.25, 0.1\} \times$ supervision $\{300, 1200, 3000\}$ (3 seeds, extended to 5 on the escaped cells). The dichotomous, causal core is *confirmed*: τ=0.1 is indistinguishable from the hard operator at every budget (concept accuracy $19.9/20.3/20.2\%$ vs. hard $19.4/19.6/19.7\%$), while $\tau \ge 0.25$ *unlocks* escape at 3000 labels (48–61% mean)—the resistance is located in the operator's cliff, not in the task (Fig. 37). The finer monotone $\sim 1/\tau$ scaling did not resolve: within the escaped group the ordering is non-monotone at five seeds (bimodal, stochastic escape dynamics), so the transition appears sharp rather than smooth—we report this half as failing our pre-registered monotonicity gate (a larger τ also weakens the logic pull globally, a two-sided effect the one-dimensional model does not capture). *Extended-budget race (prediction (ii)).* Hard Göd-R remains at 19.9–20.4% concept accuracy at $6000/12000/30000$ labels—supervision-immune across a $3000\times$ range ($10 \to 30{,}000$; half of MNIST labeled), exactly as the uncrossable-jump case predicts; Gog-R is similarly flat through 30,000 (satisfying the constraint at $\approx$86% with chance-level concepts—a canonical shortcut at half-dataset supervision), so the $\Theta(K)$-vs-∞ distinction remains right-censored at our largest budget (the model's absolute scale underestimates the real resistance). In summary: smoothing only the branch indicator unlocks at 3,000 labels what 30,000 labels cannot achieve for the hard operator—a $10\times$-budget causal demonstration that supervision-resistance resides in the operator's derivative structure (Fig. 37).

From the deterministic gate to stochastic escape times. Lem. 24 is a gradient-flow statement; real training is SGD with noise, where a deterministic trap is not absolute—noise can cross any finite wall eventually. The question becomes *rates*, and it admits a self-contained answer in the same path model: an escape-time dichotomy (Thm. 26).

Theorem 26 (Escape-time dichotomy under SGD)**.** *Consider $X_{t+1} = X_t - \eta\,(g(X_t) + \xi_t)$ on the path model, with g a one-sided selection of ∂F_n, noise ξ_t i.i.d. mean-zero sub-Gaussian(σ^2), and bounded steps $\eta|g + \xi_t| \le B$ a.s.* ***(a)*** *If $[\alpha, \beta)$ is any sub-interval of the gap with $g \ge \gamma > 0$ on it and width $W = \beta - \alpha > 2B$, then for every T, $\Pr\big[X$ hits $[\beta, \infty)$ within T steps $\mid X_0 \le \alpha\big] \le T\, e^{-\gamma(W-B)/(\eta\sigma^2)}$; hence $\mathbb{E}[T_{\mathrm{escape}}] \ge \frac{1}{4}\exp(\Gamma^*)$ with the sub-interval-optimized exponent $\Gamma^* = \sup_{[\alpha,\beta)\subseteq G} \inf_{[\alpha,\beta)} g \cdot (W-B)/(\eta\sigma^2)$.* ***(b)*** *If instead the drift is favorable, $-g \ge \delta > 0$, on the whole reachable region left of θ_{true}, then $\mathbb{E}[T_{\theta_{\mathrm{true}}}] \le (\theta_{\mathrm{true}} - X_0 + B)/(\eta\delta)$: linear time.*

Proof. (a) On $[\alpha, \beta)$, $\mathbb{E}[e^{\mu\Delta X} \mid X] \le e^{-\mu\eta\gamma + \mu^2\eta^2\sigma^2/2} \le 1$ for $\mu = \gamma/(\eta\sigma^2)$, so $M_t = e^{\mu X_t}$ is a supermartingale there. An excursion starts when the process first lies in $[\alpha, \beta)$ having been below α at the previous step; bounded steps put the entry point at most $\alpha + B$, and $W > 2B$ precludes jumping over the interval. Stopping the excursion at the first time $X \ge \beta$ or $X < \alpha$, M is bounded by $e^{\mu(\beta+B)}$ on the stopped interval, so optional stopping applies with no integrability caveat: $e^{\mu(\alpha+B)} \ge \mathbb{E}[M_{\mathrm{stop}}] \ge \Pr[\text{exit right}]\, e^{\mu\beta}$, giving a *uniform* conditional bound $e^{-\mu(W-B)}$ on the success of each excursion given its history. Any path reaching $[\beta, \infty)$ completes at least one successful excursion, each consuming at least one step, so a union bound over the at most T excursions started gives the display; the expectation bound follows from $\Pr[T_{\mathrm{escape}} \le \frac{1}{2}e^{\mu(W-B)}] \le \frac{1}{2}$. No assumption on the behavior left of α is used. (b) $Y_t = X_{t\wedge T} - \eta\delta(t \wedge T)$ is a submartingale; optional stopping and $X_{t\wedge T} \le \theta_{\mathrm{true}} + B$ (bounded steps) give $\mathbb{E}[t \wedge T] \le (\theta_{\mathrm{true}} + B - X_0)/(\eta\delta)$; letting $t \to \infty$ gives the same bound on $\mathbb{E}[T]$ by monotone convergence. □

The exponent Γ^* is a *wall action* (slope $\times$ width over the best sub-interval), not a peak slope: for the σ-smoothed cliff (height $H \approx (1-u) - \lambda n\,\Delta R$, layer width $\sim\tau$), the central layer has slope $\ge H/(8\tau)$ over width $\sim 2\tau$, so $\Gamma^* \ge c\,H/(\eta\sigma^2)$ with c absolute—**τ-independent**, provided the layer is resolved by the step size ($\tau > B$; the hard operator is the degenerate limit, which we state as a remark rather than a lemma since a single step can spatially cross a zero-width layer). Linear ramps (Łuk-R, Gog-R, Rei) give $\Gamma^* =$ action up to $O(B \cdot \text{slope})$; KD's flat top gives no adverse sub-interval, so (b) applies with $\delta = \lambda n P_{\min}$. The resulting *two-mechanism* picture: the *slope* gate (Lem. 24) decides whether a trap exists at all—this is what τ and n control, and what the intervention's threshold between τ=0.25 and 0.1 measures—while the *action* sets the escape time of trapped runs and is τ-independent for the cliff. Three post-hoc retrodictions follow (flagged as such, not pre-registered): trapped temperatures behave exactly like the hard operator rather than "slower" ($19.9/20.3/20.2\%$ vs $19.4/19.6/19.7\%$), because once trapped the action is the same; the

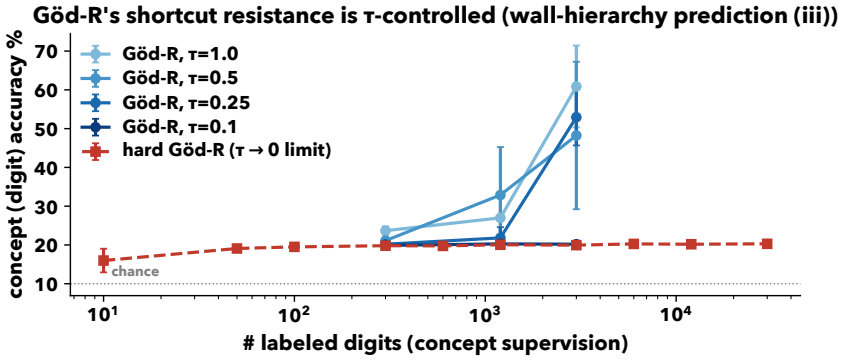


Figure 37: **The resistance comes from the operator's discontinuity: smoothing only Göd-R's branch indicator unlocks escape.** The pre-registered τ-intervention (prediction (iii) of Prop. 25): replacing only Göd-R's hard branch indicator with $\sigma((v-u)/\tau)$ controls its shortcut supervision-resistance. τ=0.1 (and the hard operator, red dashed, flat to 30,000 labels) stays supervision-immune; $\tau \geq 0.25$ unlocks concept recovery at 3000 labels. The monotone fine-structure among escaped temperatures does not resolve at five seeds (a sharp, not smooth, transition).

30,000-label immunity corresponds to an $\exp(\Gamma^*)$ escape time that is exponentially large in the wall action; and near the slope threshold escape is a rare event, producing the sharp, seed-bimodal transition that the smooth $\sim 1/\tau$ prediction missed. One qualitative fingerprint is checkable in our logs: across all 48 runs of the τ-sweep, *no run trapped at the first evaluation checkpoint ever escaped later* (every escape is already visible at the first of seven checkpoints)—consistent with exponentially suppressed late escape, and refining the bimodality reading: seeds select a basin during the early logic-dominated transient, and metastability is why trapped runs never recover.

Scope of the shortcut theory: proved, idealized, confirmed, open. **Proved (exact):** Prop. 7 (saturated implications make every slack-satisfying assignment a zero-gradient global optimum of the logic loss); Lem. 24, Prop. 25, and Thm. 26, each exactly, within the one-dimensional path model under the stated hypotheses (pinning (P_K); gradient flow, or bounded-step sub-Gaussian SGD). **Idealized:** the path model itself—one transit coordinate, pinned antecedent, model constants; none of these results is a statement about the full multi-parameter training system. **Confirmed (causally or observationally):** that the resistance is located in the operator's cliff (τ-intervention: same task, data, and hyperparameters; only the branch indicator changed); the hard operator's supervision-immunity across a 3000× label range; the absence of late escape (0/48 runs). **Open (stated, not claimed):** the Göd-R-vs-Gog-R relative order (right-censored at 30,000 labels: both remain trapped at the largest budget the 60k-image corpus admits before the unsupervised set degenerates); the magnitudes of supervision barriers in the real system; the extension of the basin-selection law below—proved in the shared-classifier model—to the full convolutional system; and logit-temperature or label-smoothing interventions as alternative dead-zone levers.

The basin-selection law: which seeds get trapped. The fingerprint above leaves one upstream question: trapped runs never recover, but which runs get trapped is decided in the early transient. We answer it within a minimal shared-classifier model faithful to the experiment's clause wiring: a linear classifier over the K class prototypes with logits $W \in \mathbb{R}^{K\times K}$, $\pi_i(c) = \mathrm{softmax}(W)_{ic}$; a per-pair predicate $u_{ab} = \mathrm{sigmoid}(\omega_{ab})$ supervised toward the true indicator $\mathbf{1}[a+b=K-1]$; and the swept implication applied as $I(u_{ab}, D_{ab})$ on all cross pairs, where $D_{ab} = \sum_{i+j=K-1} \pi_i(a)\,\pi_j(b)$ is the satisfying-assignment disjunction (this is the structure of the benchmark's sum=9 clause). Training: gradient flow on $w_{\mathrm{rl}}\,\overline{(1-I)} + w_{\mathrm{sup}}\,\mathrm{BCE}(u) + \lambda\,\mathrm{CE}(W)$ from $W \sim \mathcal{N}(0, s^2)$, $\omega = 0$. *Hypotheses*: K even (so $\sigma_f(x) = K-1-x$ is a fixed-point-free involution preserving the constraint), per-pair u (expressivity is necessary: tying u to one scalar destroys the trap layer below), small s (the linearized window). This model reproduces the empirical five-operator escape ordering and the observed collapse labelings; a 20-seed-per-cell empirical basin map on the real system gives the target probabilities (evaluated at epoch 100, escape = best-epoch concept accuracy $\geq 40\%$: Göd-R 0.00, KD 1.00, Łuk-R at 600 labels 0.20, Rei at 600 labels 0.90).

Theorem 27 (Unstable spectrum at the uniform stationary point)**.** *At the uniform stationary point (W=0, u_0=$\frac{1}{2}$, D_0=$1/K$): (i) the all-pairs mean of D depends on W only through the label*

masses $S_i = \sum_a \pi_i(a)$, namely $\overline{D} = K^{-2}\sum_i S_i S_{\sigma_f(i)}$; (ii) every implication in the family is affine in D on its active branch near the uniform point, so the logic Hessian factorizes exactly as $-\,w_{\mathrm{rl}}\,\partial_2 I(u_0, D_0)\cdot \mathrm{Hess}_W[\overline{D}]$; (iii) the unstable spectrum is exactly

$$-\,\frac{2\,w_{\mathrm{rl}}\,\partial_2 I(u_0, D_0)}{K^3} \qquad \textit{with multiplicity } \tfrac{K}{2}-1,$$

supported on the σ_f-symmetric label-mass forms orthogonal to the total mass—and empty for KD, whose active branch is $1-u$. The truth and flip directions carry no logic curvature.

Proof. (i) Averaging D_{ab} over all (a,b) collects each satisfying index pair once: $\overline{D} = K^{-2}\sum_i\big(\sum_a \pi_i(a)\big)\big(\sum_b \pi_{\sigma_f(i)}(b)\big)$. (ii) $D_0 = 1/K < u_0 = \frac{1}{2} < 1 - D_0$ places each operator strictly inside one branch: $I_{\text{Göd}} = D$, $I_{\text{Gog}} = D/u$, $I_{\text{Łuk}} = 1-u+D$, $I_{\text{Rei}} = 1-u+uD$ are affine in D with slopes $\partial_2 I \in \{1, 1/u_0, 1, u_0\}$ and $I_{\text{KD}} = 1-u$ has slope 0; affinity kills all second-order I terms. (iii) For a σ_f-symmetric, mean-zero label vector x, the form $\delta W_{ia} = x_i$ (uniform across columns) has $\delta S_i = x_i$ through the softmax Jacobian at uniform, hence $\delta^2\overline{D} = 2K^{-2}\sum_i x_i x_{\sigma_f(i)} = 2K^{-2}\|x\|^2$, while $\|\delta W\|^2 = K\|x\|^2$: eigenvalue $2/K^3$. Column-varying or σ_f-antisymmetric perturbations contribute non-positive curvature, and the total-mass direction is annihilated by the softmax Jacobian; the σ_f-symmetric subspace has dimension $K/2$, minus the mean-zero constraint. □

The spectrum is verified exactly (four decimals, including the degeneracies and KD's empty spectrum) at $K = 6, 8, 10$. The eigenmodes are label-mass concentrations on σ_f-conjugate class pairs—predicting precisely the collapse labelings the experiments show (e.g. the binary $\{4,5\}$ relabeling).

Theorem 28 (Truth-basin probability in the small-initialization window)**.** *Within the linearized window and up to absorbed constants, the collapse amplitudes are the projections of the Gaussian initialization onto the $\frac{K}{2}-1$ degenerate eigenmodes—i.i.d. $\mathcal{N}(0, s^2)$—growing at the positive rate $r_{\mathrm{op}} = 2\,w_{\mathrm{rl}}\,\partial_2 I(u_0, D_0)/K^3$, while the truth coordinate carries no logic curvature and is driven at rate λc_2 by the supervision. With Δ the window-exit scale and $\beta = \Delta'/c_2$,*

$$\Pr[\textit{truth basin}] \;=\; \Big[\,2\Phi\big(\tfrac{\Delta}{s}\,e^{-\beta r_{\mathrm{op}}/\lambda}\big) - 1\Big]^{K/2-1},$$

so the escape threshold is $\lambda^* = \max\big(\lambda_0,\ \beta r_{\mathrm{op}}/\ln(\Delta/s)\big)$ with $\lambda_0 \propto 1/T$ the finite-horizon drive floor: the race layer is linear in $\partial_2 I(u_0, D_0)$ *and horizon-independent, the floor is operator-independent and $\propto 1/T$.*

Proof sketch. In the linearized window the $\frac{K}{2}-1$ degenerate collapse forms evolve independently as $a_i(t) = a_i(0)\,e^{r_{\mathrm{op}}t}$ with $a_i(0) \sim \mathcal{N}(0, s^2)$ i.i.d. (Thm. 27 gives the eigenbasis and the shared rate r_{op}), while the truth coordinate carries no logic curvature and is driven by supervision at rate λc_2, exiting the window at t^* with $\lambda c_2 t^* = \Delta'$, i.e. $t^* = \beta/\lambda$ for $\beta = \Delta'/c_2$. The run lands in the truth basin iff no collapse form reaches the window-exit scale Δ by t^*, i.e. $|a_i(0)|\,e^{r_{\mathrm{op}}t^*} < \Delta$ for every i, equivalently $|a_i(0)/s| < (\Delta/s)\,e^{-\beta r_{\mathrm{op}}/\lambda}$. Each event has probability $2\Phi\big((\Delta/s)e^{-\beta r_{\mathrm{op}}/\lambda}\big) - 1$ for a standard normal, and independence across the $\frac{K}{2}-1$ forms gives the displayed product. Setting that argument to $\Theta(1)$ yields the threshold $\lambda^* = \beta r_{\mathrm{op}}/\ln(\Delta/s)$, and $\lambda_0 \propto 1/T$ is the separate finite-horizon condition that the supervision drive exits the window within T steps ($\lambda c_2 T \geq \Delta'$). Constants are absorbed and the bound is asserted only inside the small-init linear window. □

Both layers are verified causally: doubling the horizon halves KD's threshold (the floor layer; KD has no unstable logic form) and leaves Łuk-R's unchanged (the race layer); the race thresholds order by $\partial_2 I$ across KD/Rei/Łuk-R. We do not claim the displayed formula outside the small-init linear window; initial truth-coordinate noise is absorbed into the constants.

Proposition 29 (The antecedent partial as the unique relief channel)**.** *At the mechanism level, post-collapse escape toward the truth labeling drives D on true pairs below u, re-activating the wall ($I < 1$). The antecedent partial is the unique relief channel: if $\partial_1 I \neq 0$ (Gog-R, $\partial_1 I = -D/u^2$), the wall force transiently lowers u and self-releases ($u \leq D$ restores I=1), so the escape threshold stays at the race scale; if $\partial_1 I = 0$ exactly (Göd-R), u stays pinned by its supervision and the wall is permanent—escape must climb it against w_{rl} with only λ CE, the threshold geometry of Lem. 24. We state this at the mechanism level with verified thresholds (Göd-R never escapes at any tested λ; Gog-R escapes at the Łuk-R race scale), not as a global phase-portrait theorem.*

The three layers reduce to the operator's two partials at the uniform point: $\partial_2 I$ sets the collapse-entry rate, the finite horizon sets the supervision floor, and $\partial_1 I$ with the branch structure sets trap permanence—the basin-level counterpart of the routing (Thms. 2–3), survival (Props. 4–6), and trapping (Lem. 24) results. An operator-independent antecedent anchor (the benchmark's $\forall\exists$ clause) shifts the race constants without changing the ordering.

Alphabet robustness: a pre-registered full-system K-sweep. We varied the digit alphabet itself—$K \in \{6, 8, 10\}$ (digits $0..K{-}1$, target sum $K{-}1$, keeping the flip involution fixed-point-free), 4 operators $\times$ 4 label budgets $\times$ 10 seeds $=$ 480 runs—with three predictions and kill-switches committed before launch. **Confirmed (out of sample):** the escape ordering KD $\geq$ Rei $>$ Łuk-R $>$ Göd-R$=0$ holds at every (K, n_{sup}) cell, and hard Göd-R's supervision-immunity is K-uniform (flat at $\approx 2\times$ chance at every alphabet size—the collapse-labeling accuracy itself is a K-predictable constant). **Failed (reported):** the one-dimensional model's *quantitative* prediction that the Łuk-R-vs-Rei threshold ratio widens with K is not supported (non-monotone at our grid resolution)—the model's ordering transfers, its constants do not, consistent with its stated scope. At $K{=}6$ the trapped Łuk-R basin is metastable but *noise-escapable*—a fact that both reinterprets the supervised coverage ratio $\rho_b{=}B/n_{\text{sup}}$ and exports a fix. The escape-time bound (Thm. 26) says a trapped run does not recover on its own: minibatch supervision at any ρ_b produces only a *transient* best-epoch excursion (peak $\sim 65\%$ at epoch ~ 5) that re-collapses to the $\approx 2/K$ plateau—no run sustains even at 600 epochs $(0/8)$. The pre-registered coverage-ratio "band" is therefore a transient-peak readout under coarse checkpoints, not a dynamical escape law: with fine (5-epoch) checkpointing the high-ρ_b cells also peak $\sim$65–75%, merely later (epoch ~ 5 vs. ~ 0), so the apparent high-ρ_b "wall" is checkpoint aliasing, and every cell collapses by the horizon. The basin is, however, escapable by *injected* exploration noise: adding noise to the logic-path logits above a variance threshold flips the run to a *sustained* truth-basin solution. Logistic scale ≥ 1.0 and uniform scale ≥ 2.0 both reach $\approx 99\%$ concept accuracy that holds to the horizon ($10/10$ seeds each), while sub-threshold noise (logistic 0.5: $0/10$; uniform 1.0: $1/10$) and the noise-free baseline $(0/10)$ stay trapped; the threshold is set by the noise *variance* (logistic variance $\pi^2/3 \approx 3.3$, so logistic $1.0 \sim$ uniform 2.0), not its type—realizing the noise-escapes-local-minima principle of Daniele & van Krieken (2025) (there for discrete Gödel local search) in the gradient-descent reasoning-shortcut regime. The trap is *operator-specific* and its depth shrinks with the alphabet K: a 600-epoch K-sweep gives a Łuk-R sustained-escape fraction rising monotonically ($0/16$, $2/16$, $10/16$ at $K{=}6, 8, 10$), whereas the fast operator KD already sustains at $K{=}6$ $(8/8)$—the small alphabet is the deep trap that implicit supervision cannot escape, but a non-saturated operator, a larger alphabet, or injected noise can (cf. the operator map, Fig. 4).

Corroboration on a second benchmark and operator family. The same survival logic governs a *different* operator axis on the standard reasoning-shortcut benchmark. We ran RSBench's shortcut-MNIST (Bortolotti et al., 2024)—whose training split is built to elicit shortcuts—through its native LTN model, sweeping the conjunction t-norm and the p-mean quantifier aggregator (three seeds), and used RSBench's held-out out-of-distribution (OOD) digit-combinations as the shortcut diagnostic: a model that learned the true digits generalizes to the OOD combinations, whereas a shortcut satisfies the in-distribution constraint but fails OOD. The t-norm, with the p-mean aggregator modulating Product, sets the *failure form*.

Gödel/min and Product satisfy the digit-sum constraint in-distribution ($\approx 92\%$ and $\approx 86\%$, the latter only under max-like aggregation $p{=}6$) yet collapse to near-chance on the OOD split ($\approx 7\%$ and $\approx 2\%$; Fig. 38)—RSBench's own diagnostic confirming that the satisfaction is a reasoning shortcut, not learned concepts—while the Łukasiewicz t-norm collapses outright (never satisfies, any p). Because this split carries no concept supervision, every satisfying solution is a shortcut, so here the operator selects shortcut (Gödel/Product t-norms) versus collapse (Łukasiewicz t-norm) rather than shortcut versus *learn* (the operator-differentiated learn-vs-shortcut split is the weakly supervised differentiable fuzzy logic result above). This pattern is consistent with the survival structure: the Łukasiewicz collapse matches the t-norm dead zone of Prop. 6 (the softmax grounding over the ten digit classes pins the body t-norm T_L in $\{a + b < 1\}$, where $\nabla T_L = 0$), and the Product-versus-aggregator interaction (it needs the max-like $p{=}6$) instantiates the aggregator dichotomy of Prop. 17. The survival half of the principle is thus reproduced on a second benchmark with a second operator family (t-norm and aggregator rather than implication). On RSBench's out-of-distribution split, operator choice serves as a diagnosis (shortcut-vs-collapse); it is not a mitigation here—neither an operator swap nor added concept supervision recovers out-of-distribution accuracy, because that split

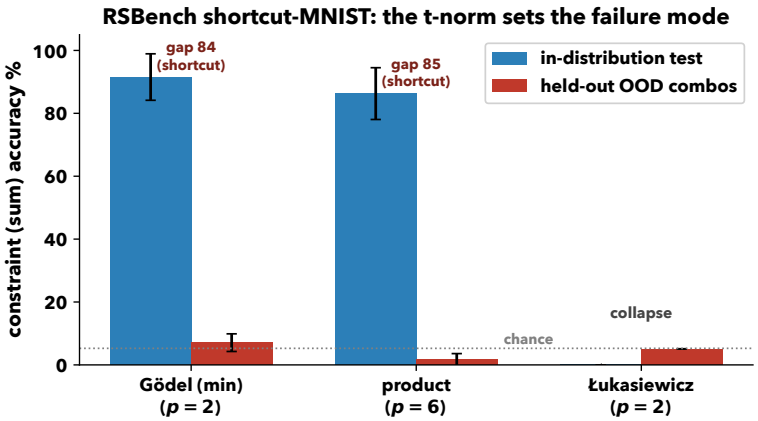


Figure 38: **RSBench corroborates the shortcut mechanism on a second benchmark and operator family.** Corroboration on RSBench's shortcut-MNIST with its native LTN model (out-of-distribution split of unseen digit combinations, no concept supervision, three seeds). On this benchmark the conjunction t-norm, with the p-mean aggregator modulating Product, sets the failure form: Gödel/min and Product satisfy the digit-sum constraint in-distribution but collapse to near-chance on the held-out OOD digit-combinations (an $\approx$ 85-point gap = a confirmed reasoning shortcut), while the Łukasiewicz t-norm collapses outright. Together these reproduce the survival-axis pattern on a second benchmark and operator family (t-norm/aggregator).

provides no concept signal for the held-out combinations. The mitigation result therefore rests on the supervised MNIST digit-sum task (Fig. 4); RSBench serves only as diagnostic corroboration.