THEME ARTICLE: Metacognitive Prediction of AI Behavior

Relevance Structures are a Necessary Prerequisite for Artificial Metacognition

Alexander M. Berenbeim, *United States Military Academy, West Point, NY, 10996, USA*

Ramneet Kaur, *SRI International, Menlo Park, CA, 94025, USA*

Adam Cobb, *SRI International, Arlington, VA, 22209, USA*

Brian Matejek, *SRI International, Arlington, VA, 22209, USA*

Anirban Roy, *SRI International, Menlo Park, CA, 94025, USA*

Susmit Jha, *SRI International, Arlington, VA, 22209, USA*

Nathaniel D. Bastian, *United States Military Academy, West Point, NY, 10996, USA*

Abstract—AI-enabled systems deployed in dynamic, open-world, safety- and mission-critical environments often lack mechanisms for reliable self-monitoring and self-regulation under shift and perturbation. Artificial Metacognition (AM) addresses this by enabling systems to assess and adapt their own inference and learning processes. We show that relevance structures, which are order-theoretic substrates for comparing generalization and specialization and for quantifying under-coverage versus overreach, are structurally unavoidable for computable self-regulation. In turn, relevance structures provide an operational control theory for metacognitive tasks. We introduce exploratory, parsimonious, and balanced relevance scores and embed them into version-space algebra to regulate hypothesis search and inductive strategy selection. The resulting framework yields competence-aware learning systems that can both correct and explain their behavior by explicitly balancing exploration and parsimony. We illustrate relevance structures and their application with a medical visual question answering study on the SLAKE dataset.

Introduction

Artificial intelligence (AI) now operates in interconnected, dynamic, and failure-prone safety-critical settings. Conventional learning machines often degrade under domain shift and lack principled mechanisms to self-regulate against perturbations. Additional regulatory considerations for accountability, transparency, and value alignment further raise deployment risk by demanding that systems both *behave robustly* and *communicate their limits* in ways that are usable by operators, auditors, and other systems. We frame these joint technical and socio-technical pressures as *metacognitive* requirements: for our purposes, an Artificial metacognition (AM) system is one that can monitor, assess, and adapt its own inference, while also producing artifacts that support intelligible self-assessment at the point of inference.

Whereas intelligence is a measure of capacity and capability, metacognition, which has been extensively studied in foundational psychological theory, is concerned with the application of intelligence. Flavell [1979] introduced the phrase *thinking about thinking* and distinguished metacognitive knowledge from metacognitive regulation: knowledge about one's own cognitive strategies versus monitoring and controlling one's cognition. We do not claim that any single formal treatment exhausts the term "metacognition" across disciplines. Our focus is on *computable* systems that implement self-monitoring and self-regulation in deployed AI.

In natural language and pragmatics, "relevance" is similarly treated as a relational notion governing what information is appropriate or useful in context. Operationally, "relevance" reflects a trade-off between informativeness and cost/effort in communication and

XXXX-XXX © 2026 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

reasoning. Our contribution is to make this broadly familiar albeit informal notion precise in a way that is compatible with computable learning and control, and argue for consciously incorporating *relevance structures* into design of AM-enabled systems, first introduced in Berenbeim and Bastian [2025].

A central result motivating this work, which we prove in this paper, is that *any computable (partial) function, let alone any computable regulation mechanism, induces a relevance structure*. Whenever a system optimizes, selects, abstains, prunes, or otherwise regulates behavior relative to a target, it imposes a preorder over candidate states or hypotheses, and therefore admits algebraic operations corresponding to shared content and exclusion. In this sense, *relevance structures are structurally unavoidable*, formalizing the minimal comparative substrate required for a system to judge "better vs. worse", "more vs. less applicable", and "what remains after removing what has been already been accounted for", all relative for a given task.

Importantly, the existence of a relevance structure does not imply that it is arbitrarily *fine-grained*, or that the one provided is the appropriate one for whatever the 'true' human needs are that pertain to the task. In the limiting case, a task may induce only a coarse (indicator-like) valuation, corresponding to a binary partition between acceptable and unacceptable hypotheses, reducing to classical binary decision settings rather than rendering the notion vacuous. Nontrivial metacognitive regulation arises when the structure is refined enough to distinguish *under-coverage* from *overreach* and to support graded self-assessment. Our framework characterizes both the minimal substrate and the richer refinements needed for control and explanation.

Our contributions in this article are primarily formal and operational. First, we identify relevance structures [Berenbeim and Bastian, 2025] as the mathematical substrate for scoring internal hypotheses against targets, and embed the exploratory δ, parsimonious ψ, and balanced β scores into version–space algebra to regulate generalization–specialization during learning. Second, by leveraging Lemmas 1–3 and Theorem 1, we prove that in the ambient settings where AI and machine learning are set (e.g. computable Polish, or lattice-based spaces) relevance structures necessarily exist and enable competence-based optimization. Finally, we provide a demonstration of relevance structures within a multimodal learning pipeline for medical visual question answering (VQA), showing that the certainty-relevance-competence (CRC) framework generalizes from discriminative models to generative models, and enables models to signal to themselves and observers self-assessment at point-of-inference.

Relevance Structures

Relevance structures are the basic mathematical objects used throughout the paper. We give the definition first, then motivate each component with concrete examples ranging from Boolean triggers and proper binary classification to structured (tree and lattice) universes. We then define exploratory, parsimonious, and balanced relevance scores, explain how they induce empirical competence objectives, and finally connect the framework to version-space algebra [Mitchell, 1979]. For readability, we keep only the diagrams and high-level operational interpretation in the main text, while deferring full operation tables and admissible (X, Y) assignments to the Appendix.

Relevance structures

Universe. Let U denote a universe of elements: hypotheses, contexts, explanations, feature-sets, label-sets, or internal states compared during learning and regulation.

Definition 1 (Relevance structure)**.** *A* relevance structure *is a tuple*

$$\mathcal{R} = \langle U, 0, \preceq, \frown, \setminus, |\cdot| \rangle$$

satisfying:

1) $\langle U, \preceq \rangle$ *is a partial order;*
2) 0 *is least:* $\forall x \in U,\ 0 \preceq x$*;*
3) $\frown$: $U \times U \to U$ *is meet (greatest lower bound), i.e.* $x \frown y = \inf\{z \in U : z \preceq x \ \wedge \ z \preceq y\}$;
4) $\setminus : U \times U \to U$ *is a proto-relative pseudo-complement (PRPC) such that for all* $x, y, z \in U$*:*
   - $(x \setminus y) \frown (y \setminus x) = 0$*;*
   - *if* $y \frown z = 0$*, then* $x \frown z \preceq x \setminus y$*;*
5) $|\cdot| : U \to \mathbb{K}_{\geq 0}$ *is a computable valuation into the nonnegative cone of an ordered field* $\mathbb{K}$ *such that* $x \preceq y \Rightarrow |x| \leq |y|$ *and* $|x| = 0 \Leftrightarrow x = 0$.

Items (1–3) say that U supports a notion of *shared content*: meet behaves like intersection/conjunction/common refinement, which is the standard way constraints combine in logic, set systems, and hypothesis spaces. Item (4) formalizes a computable notion of *remainder*: $x \setminus y$ is "x without what is already accounted for by y" (set-difference in inclusion-ordered settings, but defined more generally so it also works for trees/meet-semilattices). Item (5) is a *scalarization* enabling for comparison and optimization.

This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2026.3683332

Four concrete universes: Boolean, binary classifier, tree, lattice

We separate two "binary" regimes that are often conflated, with an appeal towards constructive mathematics where failing to prove a proposition does not mean one has proved the negation of a proposition.

Boolean trigger ($U = \{0, 1\}$).
This models a propositional guard or alarm circuit: once tripped, the output is 1, otherwise 0. Let $0 \prec 1$, take meet $\frown$= min, define PRPC by $1\backslash 1 = 0,\ 1\backslash 0 = 1,\ 0\backslash 1 = 0,\ 0\backslash 0 = 0$, and valuation $|0| = 0$, $|1| = 1$.

Proper binary classifier ($U = \{0, A, B\}$).
A classifier intends to assign each input to one of two classes A or B, where depending on the task, B may be considered "not A". In practice, such a propositional reading may be understood to say that one has *proof* for A or for B, with 0 indicating no proof for either. The primitive model for this design is a "V"-shaped meet-semilattice: $0 \preceq A$, $0 \preceq B$, $A \frown B = 0$, with A, B incomparable. A natural PRPC is $A\backslash B = A,\ B\backslash A = B,\ A\backslash A = 0,\ B\backslash B = 0$, with valuation $|A| = |B| = 1$ (or class-dependent costs).

From binary to graded relevance: trees and lattices ($U = \{0, A, B, C\}$).
Binary settings motivate the axioms but do not yet show why graded relevance is useful. We therefore use a tree (meet-semilattice) and a lattice. The tree emphasizes that meet alone may be the correct primitive; the lattice additionally supports a join $\smile$ needed for symmetric/balanced notions.

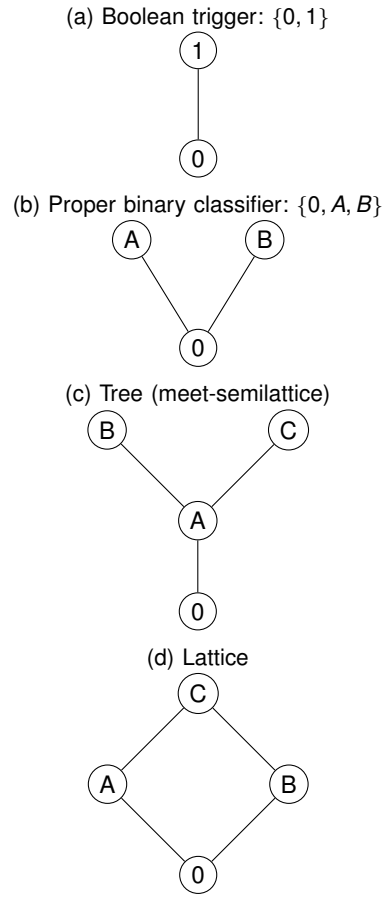

FIGURE 1: Four basic relevance structure patterns: Boolean triggers, proper binary classifiers, tree, and lattice.

Relevance scores

To compare an internal response $s \in U$ against an external target $t \in U$ we use directed relevance scores that separate two error modes: *missing target content* versus *excess response content*:

$$\text{[Exploratory Relv.]} \qquad \delta(s,t) := \frac{|s \frown t| - |t\backslash s|}{|t|}, \quad (1)$$

$$\text{[Parsimonious Relv.]} \qquad \psi(s,t) := \frac{|s \frown t| - |s\backslash t|}{|t|}. \quad (2)$$

Exploratory relevance δ penalizes what is missing in t; parsimonious relevance ψ penalizes what exceeds the scope of t. If U also supports a join $\smile$, we define symmetric difference $x\Delta y := (x\backslash y) \smile (y\backslash x)$ and balanced relevance:

$$\text{[Balanced Relv.]} \qquad \beta(s,t) := \frac{|s \frown t| - |s\Delta t|}{|s \smile t|}. \quad (3)$$

Relevance becomes metacognitively actionable when coupled with an intrinsic certainty score. Let $D_N = \{(x_i, y_i)\}_{i=1}^N$ capture a sample of size N, f denote a learning function, let ς a certainty score function assigning a non-negative real-value to each pair (f, x), and $\rho \in \{\delta, \psi, \beta\}$. Following Berenbeim et al. [2025], Berenbeim and Bastian [2025], we take ς to be the *predictive confidence score* (the difference between the softmax adjusted values for the two most likely disjoint categories, see the Appendix for further details), and define *generalized empirical competence*:

$$\mathcal{K}_\varsigma(f; \rho; D_N) = \frac{1}{N} \sum_{i=1}^{N} \varsigma(f; x_i)\, \rho\big(f(x_i), y_i\big). \quad (4)$$

In general, a convolutional notion of empirical competence may be defined over recognizable groupoid structures, as discussed in the Appendix. Future work in this direction is encouraged, as it will cover a wider range of natural language implication modeling than that which is currently covered in this article.

This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2026.3683332

The formal work supporting competence presupposes that competence is defined relative to a finite directed acyclic graph for the universe defining our relevance structure, and so we can similarly construct component certainty for generative tasks. Recalling the original discriminative setting of CRC, **component competence** is defined from (i) a model pseudo-probability vector $p(x_i) \in \Delta^{d-1}$ over d labels derived from model f outputs and (ii) an empirical label distribution $q(x_i) \in \Delta^{d-1}$ via the averaged inner product with penalty term,

$$\mathcal{CC}(f; D_N) \;=\; -\frac{1}{d} + \frac{1}{N}\sum_{i=1}^{N} \big\langle p(x_i),\, q(x_i) \big\rangle, \qquad (5)$$

where d denotes the number of vertices (answer-classes) in the underlying relevance-structure universe.

We propose the following generative VQA adaptation, with deterministic top-K decoding produces candidates $\{y_{i1}, \ldots, y_{iK}\}$ with sequence scores $\{\alpha_{i1}, \ldots, \alpha_{iK}\}$, which we convert into a within-item pseudo-distribution

$$w_{ik} \;=\; \frac{\exp(\alpha_{ik})}{\sum_{\ell=1}^{K} \exp(\alpha_{i\ell})}. \qquad (6)$$

Candidates then collapse into semantic answer classes $\{\mathcal{C}_{ij}\}_{j=1}^{J_i}$ (via normalization and equivalence/entailment checks) and pool probability mass by

$$\pi_{ij} \;=\; \sum_{k:\, y_{ik} \in \mathcal{C}_{ij}} w_{ik}, \qquad \text{so that} \qquad \sum_{j=1}^{J_i} \pi_{ij} = 1. \qquad (7)$$

To generalize the hard label match implicit in (5), we replace the indicator alignment with a relevance-structure similarity $s(\mathcal{C}_{ij}, g(i)) \in [0, 1]$ between class $\mathcal{C}_{ij}$ and the gold answer-class $g(i)$ (e.g., overlap/meet under the chosen valuation). This yields the per-item *complete* component certainty

$$\mathcal{CC}^*(x_i) \;=\; \sum_{j=1}^{J_i} \pi_{ij}\, s(\mathcal{C}_{ij}, g(i)). \qquad (8)$$

We then define the adjusted per-item component certainty by subtracting a uniform-guess baseline:

$$\mathcal{CC}^*(x_i) \;=\; \mathcal{CCC}^*(x_i) \;-\; \frac{1}{d_i}, \qquad (9)$$

where $d_i \geq 2$ is the number of semantic answer-classes for item i (at minimum *accept* vs. *reject* relative to gold). Finally, the (dataset-level) component competence of a model f on a sample D_n is the sample average

$$\mathcal{CC}^*(f; D_N) \;=\; \frac{1}{n}\sum_{i=1}^{N} \mathcal{CC}^*(x_i), \qquad (10)$$

and

$$\mathcal{CCC}^*(f; D_N) \;=\; \frac{1}{n}\sum_{i=1}^{N} \mathcal{CCC}^*(x_i). \qquad (11)$$

As in Berenbeim et al. [2025], we reproduce the competence hierarchy for a model f from the possible orderings of $\mathcal{K}$, $\mathcal{CC}^*$, and $\mathcal{CCC}^*$, defined relative to samples D_N and *knowledge* parameter $\alpha \in [0, 1)$. We chiefly recall that whenever $\mathcal{K} < 0$, f is **incompetent**; whenever $\mathcal{CC}^* \leq 0$, f is **uninformed**; whenever $\alpha < \mathcal{K}$ is **relatively α-competent**, and whenever $\alpha < \mathcal{CC} \leq \mathcal{K}$, f is relatively $\alpha-$expert. See Appendix for further details.

Version space algebra

Version-space learning maintains a set of candidate hypotheses consistent with observed constraints, ordered by generalization–specialization [Mitchell, 1979]. The same order structure supports a relevance-structure presentation: meet corresponds to shared constraints, PRPC quantifies "what remains to be covered/explained," and valuations enable scalar preference. Consequently, δ and ψ can be used to bias updates toward exploration (reducing missing mass $|t \backslash s|$) or parsimony (reducing overreach $|s \backslash t|$), while β balances both when uncertainty/noise is high. We summarize a relevance-guided version-space update rule featured in the Appendix.

Necessity of Relevance Structures

We first refer the unfamiliar reader to the Appendix where we review several definitions from computatability theory and descriptive set theory in order to state our necessity result in standard settings used to model learning, regulation, and probability.

Modeling assumption. We treat AI learning machines f as computable maps between computable extended metric spaces. Thus f may be studied with the broader Polish/Borel/measurable setting used in descriptive set theory. In the concept-learning view, $\mathcal{X}$ is a Borel subset of an extended Polish space $\mathbb{X}$, $\mathcal{Y}$ is Polish, $\mathcal{H}$ is a class of Borel functions $\mathcal{X} \to \mathcal{Y}$, and training seeks a computable loss function ℓ-minimizing hypothesis under a Borel measure $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. A computable learner maps finite samples $(\mathcal{X} \times \mathcal{Y})^{<\omega}$ to hypotheses in $\mathcal{Y}^{\mathcal{X}}$, equivalently (by currying) $A : (\mathcal{X} \times \mathcal{Y})^{<\omega} \times \mathcal{X} \to \mathcal{Y}$ with $\tilde{A}(D_N)(x) \equiv A(D_N, x)$. With parameters Θ, training selects θ so that $f_\theta = \tilde{A}(D_N)$ is the fitted instance. We write f and A interchangeably for the learner.

Lemma 1. *Any Boolean algebra of sets, including point-set topologies closed under set-difference, can be given a natural relevance structure. Thus, any $\sigma-$algebra naturally gives rise to a relevance structure.*

Lemma 1 supplies the canonical "set-like" source of relevance primitives (shared part vs. remainder),

including from the Stone-theoretic viewpoint where Boolean algebras of distinctions correspond to clopen structure. Importantly, valuation can be taken as trivial (indicator) or refined (e.g. measure-based), with refinements discussed in the Appendix. A key fact of relevance structures is that they occur across numerous mathematical structures of interest found in descriptive set theory:

Lemma 2. *Both $\mathcal{C}$ and $\mathcal{N}$ can be given relevance structures. Further, every separable and completely metrizable topological space (X,τ), i.e. Polish space, can be extended to a natural relevance structure.*

Lemma 2 lifts relevance beyond "class labels", ensuring in standard Polish (Borel) environments used to model learning and regulation that relevance structures are an ambient feature of the space in which hypotheses and targets live, and cannot be treated as a superfluous design choice. Moreover, we observe that relevance structures themselves are objects in a relevance structure partially ordered by refinements of: the underlying poset universe; of the proto-relative pseudocomplement operator; or of the partial-order preserving map.

Lemma 3. *If $\mathcal{L}$ is a relatively complemented lattice, then $\mathcal{L}$ can be endowed with a $\setminus$ function and value map such that $\mathcal{L}$ is a relevance structure.*

Lemma 3 covers logic- and constraint-driven settings (including Heyting-style use cases), while allowing both trivial and nontrivial "without" operators; we catalog practical choices and edge-cases in the Appendix.

As a consequence of Lemmas 1-3:

Theorem 1. *Suppose $\mathcal{X}, \mathcal{Y}$ and Θ are all computable extended metric spaces. Let $f : (\mathcal{X} \times \mathcal{Y})^{<\omega} \times \mathcal{X} \times \Theta \to \mathcal{Y}$ be a computable learner, such that for each sample $D_N \in (\mathcal{X} \times \mathcal{Y})^{<\omega}$, and any computable loss function ℓ, there is a corresponding $\theta_D \in \Theta$ parameterizing f so that for all hypotheses h, $\ell_D(h) \geq \ell_D(f_{\theta_D})$. Then the learner f can be also optimized by empirical competence with respect to the relevance structure of the underlying computable metric space $\mathcal{Y}$ and a choice of computable confidence scoring function ς.*

Similarly, any computable metacognitive process g with respect to monitoring, controlling, or optimizing f can be scored with respect to a corresponding relevance structure obtained from the codomain of g.

We proffer two related interpretations of this theorem along with our position that "relevance structures are a necessary prerequisite for metacognition":

Interpretation (necessity, not uniqueness). Theorem 1 is an existence result demonstrating that any practical implementation of learning and regulation admits at least one relevance structure suitable for defining under-coverage vs. overreach scores and competence-based control. This does not mean that one unique relevance structure exists up to isomorphism for every task specification, as it does for the case of a Boolean trigger. The multiplicity of relevance structures is the engineering degree of freedom, with refinements corresponding to the distinctions that are made explicit for control and for human-facing intelligibility.

Interpretation (control-theoretic reading). The theorem does not explicitly state what form the operators of the relevance structure must take. Rather, any computable, order-aware regulator g that monitors, controls, or optimizes f must compute quantities that factor through shared-content and remainder operations on the relevant hypothesis/context space (e.g. terms of the form $|s \frown t|$, $|t \backslash s|$, $|s \backslash t|$). Both relevance and empirical competence provide a concrete objective for meta-selection and regulation even when ℓ is differentiable and relevance judgments are not. Differentiable training supplies confidence surrogates (such as margins, logit gaps, or calibrated probabilities) that instantiate ς, while relevance supplies the task-aligned notion of "what counts as recognizable and usable" relative to a target space, be it a human-in-the-loop, auditor, or downstream agent.

Medical Visual Question Answering Case Study

In this study, we examine two metacognitive capabilities,*explainability* and *adaptation/perception*, under CRC. Our goal is not to maximize task accuracy alone, but to test whether CRC signals enable (i) transparent diagnosis of persistent reasoning failures and (ii) deployable confidence-conditioned policies whose reliability improves with empirical competence. We instantiate these ideas in a multimodal VQA pipeline on SLAKE [Liu et al., 2021], where relevance structures appear ubiquitously as soon as one asks *which* semantic commitments an answer makes and *how* those commitments align with the ground truth.

Pipeline summary

Each example is normalized into $(x_i, q_i, a_i, \mathcal{A}_i)$ with open-ended answer space $\mathcal{A}_i$. We evaluate metacognitive behavior on non-training, non-holdout splits (validation and test), treating them uniformly and excluding the Chinese holdout split by default. We fine-tune *LLaVA-v1.6-Mistral-7B* [Liu and contributors, 2023] and *Qwen2-VL-7B-Instruct* (VLLaMa) [Team, 2024] with LoRA adapters in the QLoRA regime [Dettmers et al., 2023] under `Accelerate` [Face, 2024]. We

trained across 7 different loss functions, with standard cross-entropy loss as our baseline against *shallow CRC*, where we optimize *weighted* cross-entropy, with certainty/relevance/competence computed post hoc and used for per-example weighting/selection and checkpoint selection (see Appendix for details).

Certainty, relevance, and competence calculation
Certainty is estimated from deterministic decoding with T=0 and beam width K=3, converting beam scores into within-example masses and defining $C := p_1 - p_2$ (top-two mass gap) [Berenbeim et al., 2025]. Relevance is computed by mapping generated text into a fuzzy Ł3 valuation over an atom inventory U induced by the question schema (entities, relations, biomedical concepts). Using MNLI (`roberta-large-mnli` [Liu et al., 2019]), we produce polarity/information pairs $v(a) = (\tau(a), i(a))$, aggregate meet/without magnitudes, and substitute them into exploratory (δ), parsimonious (ψ), and balanced (β) relevance formulas to obtain $\rho_{\beta,\delta,\psi}$, and compute empirical according to Equation 4 (see Appendix for details of the NLI scoring methods).

Accuracy And Its Relation to CRC
Accuracy is computed per item by comparing the generated answer to the gold answer using the pipeline's item-level scoring (exact match for closed-set cases like yes/no or numeric-like answers, and a softer similarity-based correctness score otherwise), and then averaging across items. Cross-entropy loss is the negative log-likelihood of the gold answer tokens under the model, and minimizing it increases the probability mass assigned to the gold answer, which tends to raise accuracy and also increases the chance that a response clears the directional-correctness threshold D, which are derived from the item-level correctness/accuracy score. In contrast, the ρ scores measure semantic alignment between the model answer and gold answer organized with respect to an inferred relevance structure inferred from both, scoring overlap and discrepancy in the underlying atom commitments. Cross-entropy primarily targets probability mass on a correct hypothesis, which raises the certainty signal C, and shifts generations toward the correct semantic region, raising ρ and competence indirectly unless the semantic structure is explicitly represented or weighted.

Metacognitive capability I: explainability (reasoning trajectories).
For each model variant, we track per-question trajectories across 30 epochs on validation and test, assigning each response to one of 3×3 segments defined by certainty bins ($C \le .1$, $.1 < C < .5$, $C \ge .5$) crossed with relevance bins ($\rho < 0$, $0 \le \rho < .5$, $\rho \ge .5$) for $\rho \in \{\beta, \delta\}$. We summarize trajectories into interpretable buckets (always-correct, always-incorrect, improved/worsened, persistent/monotone) and isolate a critical failure mode: *high-certainty/low-relevance* (confident irrelevance), which is precisely the kind of failure a relevance-structured design is meant to expose and target.

Metacognitive capability II: adaptation/perception via confidence–correctness coupling.
We operationalize adaptation/perception as the reliability of certainty as a control signal, and examine the generalization of Fundamental Theorem of the Competence Hierarchy [Berenbeim et al., 2025] to the generative task, expecting to see that as empirical competence increases, certainty distributions for directionally correct versus incorrect responses separate, improving confidence-conditioned log-odds. For each epoch and run we estimate

$$\mathrm{odds}(c) = \frac{\Pr(D = 1 \mid C \ge c)}{\Pr(D = 0 \mid C \ge c)}, \qquad c \in [0, 1),$$

and track confident-error rates $\Pr(D = 0 \mid C \ge c)$ at $c \in \{.3, .6, .9\}$. This yields a deployable gating rule: low-certainty generations can be routed for review or re-prompting, while high-certainty outputs are trusted only insofar as training improves the odds curve and reduces confident error.

Functional regression of odds curves (global tests).
To assess how training choices shift the *entire* confidence–correctness relationship (not a single threshold), we treat $y_{re}(c) := \log(\mathrm{odds}_{re}(c))$ as a functional response. Using FPCA,

$$y_{re}(c) = \mu(c) + \sum_{k=1}^{K} \xi_{re,k}\phi_k(c) + \varepsilon_{re}(c),$$

$$\xi_{re,k} = x_{re}^{\top}\beta_k + u_r + \eta_{re,k},$$

where x_{re} encodes base-model indicators, loss categories, competence type, and empirical competence (and u_r is an optional run effect). The implied conditional mean curve

$$\widehat{\mathbb{E}}[y_{re}(c) \mid x_{re}] = \mu(c) + \sum_{k=1}^{K} \left(x_{re}^{\top}\widehat{\beta}_k\right)\phi_k(c)$$

lets us visualize how increasing competence shifts log-odds across thresholds.

Results

Table 1 reports best-epoch performance per model/objective on the test split. Across runs, F_1/precision/recall closely track accuracy (typically within a few points), consistent with the metrics largely reflecting exact-match correctness. BLEU is consistently

Model	Loss	Acc.	ρ_β	ρ_δ	ρ_ψ	κ_β	κ_δ	κ_ψ	$\mathcal{CC}^*$	$\mathcal{CCC}^*$	F_1	Prec.	Rec.	BLEU	METEOR	ANLS	Val.	Plaus.
LLaVA	L_{acc}	.608	.395	.585	-5091575.076	.229	.304	-2227944.326	.108	.456	.608	.612	.608	.741	.608	.616	1.000	1.000
LLaVA	$L_{\mathcal{K}_\beta}$	.508	.324	.556	-631328.876	.251	.348	-274989.127	.041	.378	.494	.491	.508	.674	.503	.474	1.000	.997
LLaVA	$L_{\mathcal{K}_\delta}$	.484	.321	.564	-5744684.152	.247	.371	-3189101.878	.004	.360	.486	.496	.484	.678	.484	.459	1.000	.981
LLaVA	$L_{\mathcal{K}_\psi}$	.482	.324	.580	-6518649.934	.252	.344	-1781259.233	.041	.402	.478	.478	.482	.675	.481	.463	1.000	1.000
LLaVA	L_{ρ_β}	.501	.348	.580	-558491.866	.188	.266	-3522109.135	-.004	.317	.498	.497	.501	.685	.500	.493	1.000	.994
LLaVA	L_{ρ_δ}	.508	.344	.575	-668819.594	.169	.237	-2728728.735	.004	.345	.504	.504	.508	.690	.506	.489	1.000	.990
LLaVA	L_{ρ_ψ}	.435	.286	.568	-5253719.746	.191	.295	-2278984.998	-.016	.306	.436	.442	.435	.655	.435	.446	1.000	1.000
VLLaMa	L_{acc}	.388	.298	.556	-6912032.610	.164	.245	-2575966.840	-.029	.303	.377	.377	.388	.605	.384	.371	1.000	.984
VLLaMa	$L_{\mathcal{K}_\beta}$	.369	.251	.525	-6472097.191	.161	.244	-2848815.106	-.051	.280	.365	.369	.369	.606	.366	.358	1.000	.997
VLLaMa	$L_{\mathcal{K}_\delta}$	.373	.287	.559	-6649124.738	.162	.265	-3908143.824	-.049	.305	.367	.367	.373	.606	.371	.369	1.000	1.000
VLLaMa	$L_{\mathcal{K}_\psi}$	.361	.258	.557	-6293683.951	.140	.207	-1861154.663	-.062	.298	.358	.358	.361	.608	.360	.357	1.000	.984
VLLaMa	L_{ρ_β}	.384	.298	.565	-626205.426	.175	.302	-3446709.487	-.035	.323	.380	.384	.384	.606	.382	.380	1.000	1.000
VLLaMa	L_{ρ_δ}	.370	.304	.579	-6399203.787	.189	.278	-2602047.731	-.036	.308	.366	.370	.370	.605	.368	.365	1.000	.994
VLLaMa	L_{ρ_ψ}	.356	.268	.570	-6253626.677	.159	.241	-2471901.048	-.058	.302	.353	.356	.356	.608	.355	.354	1.000	.990

TABLE 1: Best-validation epoch summary per model/objective with respect to test split only. For each run, we select the epoch that maximizes the run's optimized criterion on validation split. $\rho_{\beta,\delta,\psi}$ are relevance scores, $\kappa_{\beta,\delta,\psi}$ are empirical competences.

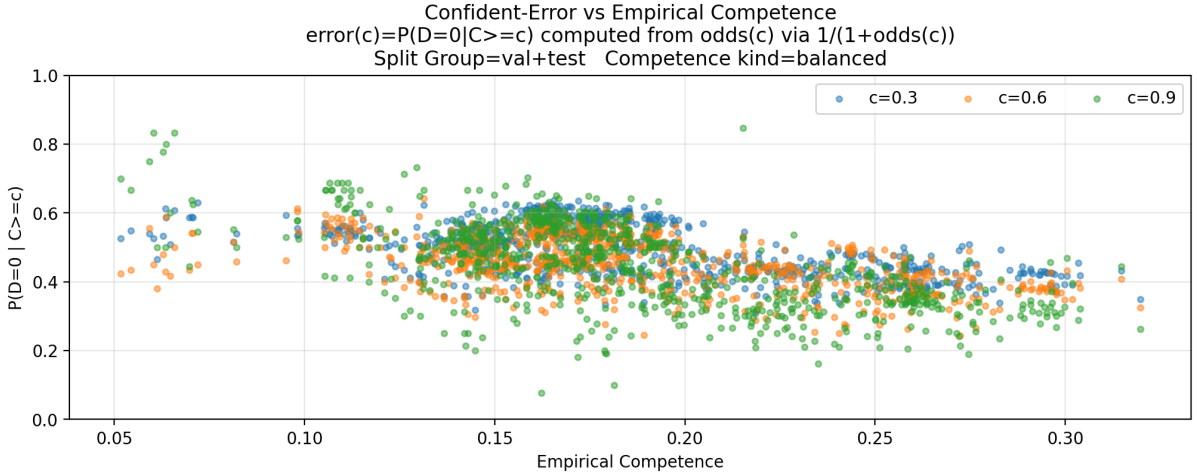

FIGURE 2: Global confident-error vs empirical competence of model relative to validation+test split for thresholds $c \in \{.3, .6, .9\}$.

higher than accuracy for both models/objectives, while METEOR is generally comparable to (and often near) accuracy and ANLS is mixed—sometimes slightly above but often below accuracy for CRC-weighted objectives—suggesting that generations frequently share surface n-grams with the reference, but that semantic/exact alignment remains a limiting factor. Validity is saturated (1.000 for every run) and plausibility remains very high ($\approx$ 0.981–1.000), indicative of well-formed, medically "reasonable" outputs, but also signalling that neither scores provide a sound reliability guarantee for the Medical VQA task.

Turning to the CRC signals, the key failure mode is the parsimonious relevance score ρ_ψ, which remains extremely negative across every objective, indicating models remain incapable of responding tersely to medical questions while introducing extraneous material. The derived parsimonious competence κ_ψ is likewise extremely negative. This reinforces that ψ-based quantities are numerically unstable under the current setup, so ψ-based selection should be treated as diagnostic rather than as a model-selection criterion. Component competence further sharpens the reliability picture. Between all models, the LLaVa model fine-tuned with respect to cross-entropy loss is the 'least bad' and 'most informed' per the competence hierarchy. Empirically, shallow CRC tuning does not yield consistent improvements over the accuracy baselines under the current wiring/weight transform. Overall, the current evidence remains consistent with a regime in which CRC signals are most useful for logging/diagnosis and checkpoint selection, rather than providing a reliably beneficial per-item training gradient under the present experimental configuration.

M	Obj	A	B	C	D	E	F	G	H
LLaVa	$\mathcal{L}_{\text{acc}}$	.173	.208	.055	.440	.303	.257	.013	.010
LLaVa	$\mathcal{L}_{\mathcal{K}_\beta}$	.176	.257	.000	.345	.362	.277	.000	.000
LLaVa	$\mathcal{L}_{\mathcal{K}_\delta}$	.195	.248	.007	.342	.362	.264	.000	.000
LLaVa	$\mathcal{L}_{\mathcal{K}_\psi}$	.267	.277	.013	.430	.407	.296	.003	.003
LLaVa	$\mathcal{L}_{\rho_\beta}$	.293	.293	.029	.466	.414	.329	.029	.023
LLaVa	$\mathcal{L}_{\rho_\delta}$	.270	.303	.036	.459	.423	.355	.007	.000
LLaVa	$\mathcal{L}_{\rho_\psi}$	.150	.179	.059	.362	.384	.287	.000	.000
VLLaMa	$\mathcal{L}_{\text{acc}}$	.195	.459	.049	.287	.557	.489	.036	.016
VLLaMa	$\mathcal{L}_{\mathcal{K}_\beta}$	.215	.489	.078	.332	.580	.518	.049	.029
VLLaMa	$\mathcal{L}_{\mathcal{K}_\delta}$	.215	.453	.075	.342	.580	.495	.049	.029
VLLaMa	$\mathcal{L}_{\mathcal{K}_\psi}$	.228	.466	.016	.293	.577	.502	.046	.023
VLLaMa	$\mathcal{L}_{\rho_\beta}$	.215	.495	.091	.345	.586	.547	.059	.033
VLLaMa	$\mathcal{L}_{\rho_\delta}$	.205	.482	.081	.339	.603	.528	.042	.026
VLLaMa	$\mathcal{L}_{\rho_\psi}$	.228	.495	.052	.329	.586	.541	.059	.029

TABLE 2: Trajectory buckets, persistence, and high-certainty/low-relevance persistent failures. Column key: A: Always Correct; B: Always Incorrect; C: Sustained Improvement; D: Persistently Right; E: Persistently Wrong; F: Never Recovered After Failure; G: Wrong with High Certainty and Low β-Relevance; H: Wrong with High Certainty and Low δ Relevance.

Reasoning trajectories.

We operationalize transparency by tracking per-question trajectories over the 30 epochs on all non-holdout, non-training splits. Each epoch's response is assigned to one of 3×3 segments defined by certainty bins ($C \leq .1$, $.1 < C < .5$, $C \geq .5$) crossed with relevance bins ($\rho < 0$, $0 \leq \rho < .5$, $\rho \geq .5$) for $\rho \in \{\beta, \delta\}$. Table 2 summarizes trajectory buckets and persistent failure regions. The base-case contrast is sharp: LLaVA under $\mathcal{L}_{\text{acc}}$ has a smaller always-incorrect mass and larger persistently-right mass than VLLaMa, whereas VLLaMa exhibits a substantially larger always-incorrect region (always-incorrect .208 vs .459; persistently-right .440 vs .287), while VLLaMa is persistently wrong much more often (.557 vs .303). Persistent high-certainty/low-relevance failures ("confident irrelevance") remain visible but are a smaller share of the total set in the current snapshot. The dominant reliability issue is therefore broad persistence of error, especially with VLLaMa, rather than errors concentrating only in the high-certainty/low-relevance corner. Nonetheless, we still isolate a distinct failure mode that plausibility alone will miss.

Across the accuracy runs, both models are consistently correct on many straightforward, visually anchored questions (scan plane, organ presence/health, organ system, and simple cues), while both remain persistently wrong on knowledge-heavy modality/weighting/type prompts and on disease/abnormality identification. VLLaMa continues to show stronger persistent failures on relational/comparative items and exhibits more "never recovered" behavior in this snapshot (for $\mathcal{L}_{\text{acc}}$, $F = .489$ vs $F = .257$ for LLaVA), indicating more late-epoch degradation or unrepaired mistakes. Table 3 reports segment-group error rates for the balanced-relevance base case. Both models have very high error in low-certainty and low-relevance regions (e.g., low relevance with $C \geq .1$: LLaVA .824; VLLaMa .850) and lower error when balanced relevance is high, but the residual error remains material, particularly for VLLaMa. This motivates abstention/re-query policies keyed to certainty scores, relevance-structured competence, and model-specific calibration, over reliance on plausibility alone.

TABLE 3: Segment-group error rates for base-case runs (balanced relevance over 30 epochs).

model	group	n	error
LLaVa	$C \leq .1$, any relevance	25	.840
LLaVa	Low relevance (< 0), $C \geq .1$	68	.824
LLaVa	High relevance ($\geq .5$), $C \geq .1$	128	.234
VLLaMa	$C \leq .1$, any relevance	38	.737
VLLaMa	Low relevance (< 0), $C \geq .1$	80	.850
VLLaMa	High relevance ($\geq .5$), $C \geq .1$	125	.440

Confidence–correctness coupling.

We evaluate post hoc reliability by estimating the certainty-conditioned odds curve and confident-error rates $\Pr(D = 0 \mid C \geq c)$ at $c \in \{.3, .6, .9\}$. Figure 2 shows that confident error decreases as empirical competence increases: higher-competence checkpoints are less often wrong overall, and also less often wrong *conditioned* on high certainty. The concavity and clustering are consistent with mixture effects (base-model family, objective class, and epoch), suggesting

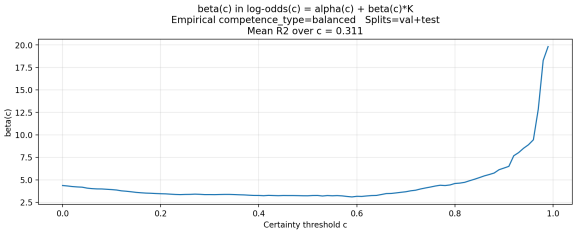

FIGURE 3: Functional linear regression: $\beta(c)$ in $\log \text{odds}(c) = \alpha(c) + \beta(c) \cdot \mathcal{K}$, illustrating increasing competence effect with certainty threshold.

that training choices induce distinct calibration regimes rather than a single smooth continuum.

Functional regression (global test of the odds function). Figure 3 reports the functional linear regression coefficient $\beta(c)$ in $\log \text{odds}(c) = \alpha(c) + \beta(c) \cdot \mathcal{K}$, relating empirical competence to the *entire* log-odds function. The increase of $\beta(c)$ with c indicates that competence matters most in the high-certainty regime, precisely where abstention policies operate. Throughout, the log-odds becomes proportionally larger relative to the empirical competence of the model, providing further empirical validation for the fundamental theorem of the competence hierarchy laid out in Berenbeim et al. [2025]. This points to a two-part reliability design: (i) use relevance-structured competence to gate trust/abstention, and (ii) explicitly target the persistent lacunae (confident irrelevance and stable anatomy/presence failures) with relevance-aware training and calibration rules keyed to competence rather than accuracy alone.

Conclusion

As predicted by the theory of CRC presented in Berenbeim et al. [2025], models with higher empirical competence exhibit larger separation between the certainty distributions of directionally correct and directionally incorrect responses, with relevance furnishing the control mechanism that mediates competence. In our empirical study, the metacognitive signal is two-fold. First, during training and model comparison, relevance supports *discovery and transparency*: by localizing failures into relevance–certainty bins and tracking which question types and relevance-structure kinds do (or do not) improve across epochs, we obtain interpretable evidence about what the model is learning and where it remains confidently misaligned. Second, holding a model's competence profile fixed, certainty supports *adaptation/perception* in production: confidence thresholds can be used to route uncertain outputs for review or to trigger corrective policies. Future work should internalize relevance-structure operations as differentiable components to enable end-to-end metacognitive learning. Even now, our shallow instantiation already separates accuracy from competence and demonstrates that competence- and relevance-oriented objectives can shift the confidence–correctness relationship in systematic, measurable ways.

REFERENCES

John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.

Alexander M. Berenbeim and Nathaniel D. Bastian. Relevance scoring as a feature of metacognitive artificial intelligence. *Proceedings of the 2025 SIAM International Conference on Data Mining 2nd Workshop on Metacognitive Predictive of AI Behavior*, pages 1–6, 2025.

Tom Michael Mitchell. *Version spaces: an approach to concept learning.* Stanford University, 1979.

Alexander M Berenbeim, Adam D Cobb, Anirban Roy, Susmit Jha, and Nathaniel D Bastian. Applications of certainty scoring for machine learning classification and out-of-distribution detection. *ACM Transactions on Probabilistic Machine Learning*, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

Haotian Liu and contributors. Llava-v1.6-mistral-7b (model card). https://huggingface.co/liuhaotian/llava-v1.6-mistral-7b, 2023. Accessed 2026-02-03.

Qwen Team. Qwen2-vl-7b-instruct (model card). https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct, 2024. Accessed 2026-02-03.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.

Hugging Face. Accelerate (documentation). https://huggingface.co/docs/accelerate/index, 2024. Accessed 2026-02-03.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Alexander Berenbeim is a Senior AI Researcher within the Robotics Research Center at the United States Military Academy at West Point, NY, 10996, USA. His research interests include combinatorial game theory, applied model theory, and applications of descriptive set theory. He received his Doctorate in Pure Mathematics from the University of Illinois at Chicago. Contact him at alexander.berenbeim@westpoint.edu.

Ramneet Kaur is an Advanced Computer Scientist at SRI International, Menlo Park, CA, 94025, USA. Her core research areas are AI, Trustworthy Machine Learning, and Open world Learning. She received her Doctorate in Computer & Information Science from the University of Pennsylvania. Contact her at ramneet.kaur@sri.com

Adam Cobb is a Senior Computer Scientist at SRI International, Arlington, VA, 22209, USA. His research interests include uncertainty quantification in machine learning, Markov chain Monte Carlo, and quantum computing. Cobb received his Doctorate in Machine Learning from the University of Oxford. Contact him at adam.cobb@sri.com.

Brian Matejek is an Advanced Computer Scientist at SRI International, Arlington, VA, 22209, USA. His research interests include machine learning, cybersecurity, and connectomics. Matejek received his Doctorate in Computer Science from Harvard University. Contact him at brian.matejek@sri.com.

Anirban Roy is a Principal Scientist at SRI International, Arlington, VA, 22209, USA. His research interests are generative models, assured machine learning, and AI for creativity and design. Roy received his Doctorate in Computer Vision from Oregon State University. Contact him at anirban.roy@sri.com

Susmit Jha is the Technical Director of the NuSCI Research Group at SRI International, Arlington, VA, 22209, USA. His research interests combine formal methods, machine learning, and control theory for trustworthy AI. Jha received his Doctorate in Computer Science from University of California Berkeley. Contact him at susmit.jha@sri.com

Nathaniel D. Bastian is Deputy Director of the Robotics Research Center at the United States Military Academy at West Point, NY, 10996, USA, and he serves as Principal Investigator of the Laboratory for Artificial Intelligence Research & Engineering (LAIRE). His research interests combine optimization, decision theory, machine learning, and statistical computing for secure, robust and resilient AI-enabled autonomous C5ISRT systems. Bastian received his Doctorate in Industrial Engineering and Operations Research from Pennsylvania State University. Contact him at nathaniel.bastian@westpoint.edu.