Operadic consistency: A label-free signal for compositional reasoning failures in LLMs

Nathaniel Bottman[1] **Yinhong Liu**[2] **Kyle Richardson**[3,4*]
[1]Incubilate [2]University of Cambridge [3]Allen Institute for AI [4]University of Tübingen
nate@incubilate.com, williamliuyinhong@gmail.com, kyle.richardson@uni-tuebingen.de

github.com/natebottman/operadic-consistency-paper

Abstract

Detecting LLM reasoning failures at inference time without ground-truth labels has motivated a wide range of confidence baselines, including self-consistency, semantic entropy, and P(True), built on within-question sampling and self-evaluation. Operad theory, the formalism for systems built by iterated substitution, suggests a complementary diagnostic: a model's direct answer to a compositional query should agree with the answer it produces by composing a stated decomposition of the same query. We instantiate this idea as operadic consistency (OC), a per-question signal. Across twelve instruction-tuned LLMs (4B to 671B parameters, open-weights and closed-source) on four multi-hop QA datasets, OC is strongly correlated with accuracy on every dataset (Pearson $r \in [0.86, 0.94]$, all $p \leq 0.0004$), and is the only signal we evaluate with $r \geq 0.85$ uniformly across all four datasets. Chain-of-thought self-consistency (CoT-SC; Wang et al., 2023) matches OC on HotpotQA and DROP ($r = 0.93, 0.87$) but drops to $r \approx 0.45$ on MuSiQue and StrategyQA. At the per-question level, OC contributes information beyond CoT-SC and semantic entropy on every dataset (cluster-robust $p < 10^{-15}$ for the OC coefficient), and the conclusion is robust to additionally controlling for constructed decomposition-aware baselines ($p < 10^{-12}$). The same signal yields selective-prediction improvements (accuracy at fixed coverage) over a tuned CoT-SC baseline at the equal-cost $K = 3$ budget (AUARC lifts of +0.086 to +0.096 and AUROC lifts of +0.092 to +0.164; 95% CIs exclude zero on every cell). On five thinking models, where the decomposition is extracted from the model's own chain of thought, the same comparison gives positive selective-prediction lift on all 16 (dataset, budget, metric) cells tested, with 95% CIs excluding zero on 12 of the 16.

1 Introduction

Modern LLMs (including the latest generation of explicit "thinking" models) routinely produce extended chains of reasoning that look fluent and confident but do not actually compose into a correct answer [Press et al., 2023, Lanham et al., 2023]. Detecting these failures at inference time, without ground-truth labels, is the practical question behind work on hallucination detection, calibration, and selective prediction. Existing methods cluster into a few families: *self-consistency* [Wang et al., 2023b] aggregates K temperature samples by majority vote; as a continuous confidence score we use mean pairwise agreement across the same samples. *Semantic entropy* [Kuhn et al., 2023] computes Shannon entropy over semantic clusters of samples. $P(\textit{True})$ [Kadavath et al., 2022] elicits the model's own probability that its answer is correct. The first two probe sample diversity ("how variable is the model's answer when re-sampled"); the third probes self-evaluation ("does the model think it

*Work done while at the Allen Institute for AI.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

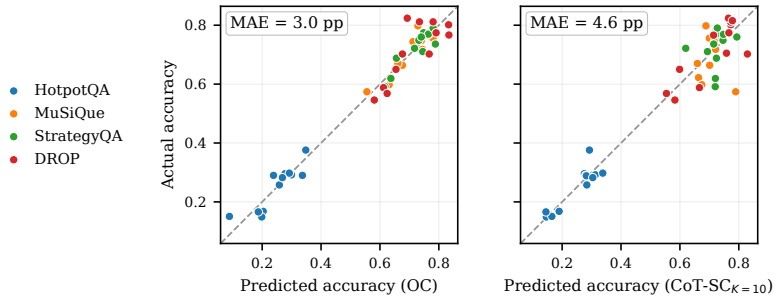


Figure 1: **Operadic consistency yields a label-free cross-model accuracy estimator.** For each non-thinking model and dataset, we hold the model out, fit accuracy $\approx a \cdot s + b$ on the remaining eleven models for a per-(model, dataset) signal-rate s, and predict the held-out model's accuracy. *Left:* $s =$ OC-rate gives pooled MAE 3.0 pp (range 2.1–4.7 pp per dataset) with no calibration set or ground-truth labels for the held-out model. *Right:* $s =$ CoT-SC$_{K=10}$-rate (canonical chain-of-thought self-consistency, Wang et al., 2023b); ten inference calls per question. Pooled MAE 4.6 pp, but with a bimodal per-dataset pattern: CoT-SC matches OC on HotpotQA and DROP (per-dataset MAE 2.4 and 4.9 pp) and loses uniform predictive value on MuSiQue and StrategyQA (5.8 and 5.2 pp).

is right"). A complementary line of work [Jacovi et al., 2024, Akbar et al., 2024] employs external verification, scoring individual reasoning steps or atomic claims against supporting evidence. They typically depend on task-specific tooling or partial supervision.

None of these families asks whether the model's reasoning is *internally compositional* — whether the sub-conclusions, taken seriously, would actually compose into the same final answer the model gives directly; the closest existing checks (§4) require verified gold decompositions.

We make this latter notion precise. Many of the reasoning tasks LLMs are evaluated on (multi-hop question answering, math word problems) have an explicit *compositional* structure: a complex query decomposes into simpler sub-queries whose answers are substituted to produce a final answer. Wei et al. [2022] and a long line of subsequent work [Yao et al., 2023, Khot et al., 2023, Yang et al., 2024a, Besta et al., 2024] use this structure operationally, prompting models to expose intermediate reasoning steps. We use it diagnostically: when a model gives a direct answer to a complex query and *also* operates on a stated decomposition of that query (whether supplied externally or extracted from the model's own chain of thought), the two answers should agree if the model's reasoning is internally consistent. This failure mode is the per-question analog of the population-level *compositionality gap* of Press et al. [2023], who show that LLMs routinely fail to compose correct sub-answers into correct multi-hop answers. We call this check *operadic consistency*, and we show it is a strong, complementary signal of per-question correctness.

We propose that the mathematical framework underlying this signal is the formalism of *operads* [Markl et al., 2002, Loday and Vallette, 2012], a well-studied structure introduced to study systems built by iterated substitution — how many-input operations compose coherently into larger operations — and providing a compact language for tree-shaped composition. We define the *questions operad* $\mathcal{Q}$, in which operations correspond to question templates with blanks and composites correspond to decompositions into sub-queries, and interpret a language model as an algebra over $\mathcal{Q}$. This framework is not just descriptive: it is what *suggests* the notion of operadic consistency — the predicate that direct and decomposed answers agree on a given tree of questions. The framework is developed in App. A, and at greater length in our companion paper [Bottman and Richardson, 2026].

We instantiate the framework on five compositional reasoning benchmarks (HotpotQA [Yang et al., 2018], MuSiQue [Trivedi et al., 2022], StrategyQA [Geva et al., 2021], DROP [Dua et al., 2019], GSM8K [Cobbe et al., 2021]), across a cohort of seventeen models: twelve instruction-tuned LLMs (4B–671B parameters, spanning open-weights and closed-source offerings) and five thinking models.

Contributions. In summary:

1. *Framework.* We define the questions operad $\mathcal{Q}$ and interpret a language model as an algebra over $\mathcal{Q}$. From this we derive *operadic consistency* (the predicate that direct and

decomposed answers agree on every tree of questions), and use its scalar instantiation OC as a per-question signal (§2).

2. *Empirical signal.* Across twelve instruction-tuned LLMs (4B–671B) on four multi-hop QA datasets, OC correlates strongly with cross-model accuracy ($r \in [0.86, 0.94]$) and is the only signal we evaluate with $r \geq 0.85$ on all four datasets; canonical chain-of-thought self-consistency [Wang et al., 2023b] matches OC on HotpotQA and DROP but drops to $r \approx +0.45$ on MuSiQue and StrategyQA. Under a leave-one-model-out fit, the OC-rate of a held-out model predicts its accuracy to within 3.0 pp on average, well below the 4.6 pp pooled obtained from the same procedure with CoT-SC at $K{=}10$ (Fig. 1). At the per-question level, OC contributes information beyond CoT-SC, semantic entropy, and constructed decomposition-aware baselines on every dataset (§3.1).
3. *Deployment lifts.* In a leave-one-model-out protocol (the calibrator is fit on eleven models and evaluated on the held-out twelfth, simulating deployment to a new model), OC combined with CoT-SC$_{K=3}$ improves on selective-prediction metrics over a tuned CoT-SC$_{K=3}$ baseline (the equal-cost comparator at 3 model calls) by $\Delta_{\text{AUARC}} \in [+0.086, +0.096]$ and $\Delta_{\text{AUROC}} \in [+0.092, +0.164]$ on the four datasets, with bootstrap 95% CIs excluding zero on every cell (§3.1).
4. *Thinking models.* On five thinking models, with the decomposition extracted from the model's own chain of thought, the same equal-cost comparison gives positive selective-prediction point-estimate lift on all 16 (dataset, budget, metric) cells tested, with 95% CIs excluding zero on 12 of the 16 (§3.2).

2 Operads

We use *operads* to make the compositional structure of question decomposition precise. An operad organizes operations of varying arities (k inputs, one output) together with composition operations $\circ_i \colon \mathcal{O}(k) \times \mathcal{O}(\ell) \to \mathcal{O}(k + \ell - 1)$ that plug one operation's output into the i-th input slot of another, satisfying *associativity axioms*. Formal definitions are in App. A; for accessible references see Markl et al. [2002], Loday and Vallette [2012].

2.1 The questions operad

The questions operad sits within the space of decomposed-prompting frameworks: chain-of-thought [Wei et al., 2022] and its variants [Khot et al., 2023, Hu et al., 2024, Yao et al., 2023] all surface intermediate sub-questions, and the questions operad organizes the resulting set of templates and their compositions into a single algebraic object. The questions operad organizes *question templates* (natural-language questions with blanks) and their compositions. Crucially, composition is purely an operation on question templates, separate from the algebra step where a model produces answers.

Definition 2.1 (The questions operad)**.** The *questions operad* $\mathcal{Q}$ has $\mathcal{Q}(k)$ the set of question templates with k blanks. For every $k, \ell \geq 0$ and $1 \leq i \leq k$, the composition operation $\circ_i \colon \mathcal{Q}(k) \times \mathcal{Q}(\ell) \to \mathcal{Q}(k + \ell - 1)$ is defined by *insertion of question templates*: the second template (with its ℓ blanks) is substituted in place of blank i of the first, yielding a new template with $k + \ell - 1$ blanks. Composition is associative up to the semantic equivalence $\approx$ of §2.2: the surface form of a composite may vary, but its meaning does not. Under Assumption 2.4 this makes $\mathcal{Q}/\!\approx$ an operad (App. A.4).

Example 2.2. For a concrete instance, let $\texttt{Q1}, \texttt{Q2}, \texttt{Q3} \in \mathcal{Q}$ be the templates "When did World War 2 end?", "Who was President at −?", and "Who was −'s wife?", with $\texttt{Q1} \in \mathcal{Q}(0)$ and $\texttt{Q2}, \texttt{Q3} \in \mathcal{Q}(1)$. Their composite $\texttt{Q3} \circ_1 \texttt{Q2} \circ_1 \texttt{Q1} \in \mathcal{Q}(0)$ is "Who was First Lady when World War 2 ended?", and its depth-2 sub-chain $\texttt{Q2} \circ_1 \texttt{Q1}$ is "Who was President when World War 2 ended?".

Algebras over $\mathcal{Q}$, and language models. An *algebra* over $\mathcal{Q}$ interprets each template as a concrete function on values: fix a value set $\mathcal{V}$, and assign to each $q \in \mathcal{Q}(k)$ a map $\mathcal{V}(q) \colon \mathcal{V}^k \to \mathcal{V}$, compatibly with composition (App. A). A language model m determines a candidate algebra in the natural way: substitute the values into q's blanks, then ask the model. We write $\mathcal{V}_m(q)$ for the operation this assigns to q. The *compatibility condition* $\mathcal{V}_m(q \circ_i q') = \mathcal{V}_m(q) \circ_i \mathcal{V}_m(q')$ would say that $\mathcal{V}_m$ is in fact a $\mathcal{Q}$-algebra (Def. A.3): the model's answer to a question agrees with the result of sequentially answering its parts. Its failure is what we call *operadic inconsistency* (§3).

2.2 Operadic consistency

A *tree of questions* (ToQ) T is a rooted, oriented, planar tree where edges point toward the root: each arity-k vertex has k incoming edges (one per question blank) and one outgoing edge (the answer it produces). Leaves are arity-0 vertices with no incoming edges; the root has a distinguished outgoing edge representing the final answer of the composite. Each arity-k vertex carries a question $q \in \mathcal{Q}(k)$.

The *total composite* of T is the question obtained by composing along every internal edge. Given a model m, two answers to the total composite are available: the *direct answer*, asking m the total composite as a single question; and the *decomposed answer*, evaluating T leaf-to-root: ask m each leaf question, then for each non-leaf vertex substitute its children's answers into its blanks and ask m.

Definition 2.3. m is *operadically consistent on* T if its direct and decomposed answers on T agree. On the depth-2 chain of Ex. 2.2 this is the equation $\mathcal{V}_m(\texttt{Q2} \circ_1 \texttt{Q1}) = \mathcal{V}_m(\texttt{Q2}) \circ_1 \mathcal{V}_m(\texttt{Q1})$: answering the composite agrees with answering the parts in turn.

Making this precise requires saying what "agree" means, for answers and for questions. Two answers are *semantically equivalent*, $a \sim b$, when they have the same denotation; two questions have $q \approx q'$ when the gold-answer maps $\gamma_q, \gamma_{q'}$ they induce on $\sim$-classes agree as partial functions (i.e. have the same domain and take the same values). App. A.4 goes into greater detail. Composition descends to $\approx$ under one assumption.

Assumption 2.4 (Compositionality)**.** $\gamma_{q \circ_i q'} = \gamma_q \circ_i \gamma_{q'}$ for all composable q, q'.

This is the compatibility condition of §2.1 with the gold-answer map γ in place of the model's $\mathcal{V}_m$: we assume the true answers compose, and operadic consistency tests whether the model's do too.

Under Assumption 2.4 the descent is automatic, and sequential and parallel associativity then make Def. 2.3 well defined on a ToQ of any shape, since the total composite is independent, up to $\approx$, of the order in which internal edges are composed (App. A.4).

A proposition then says when the compatibility condition holds exactly, which is what OC measures failure of. Call m *perfect* if it returns a gold answer wherever γ_q is defined.

Proposition 2.5. *If every template in play is well posed, Assumption 2.4 holds, and m is perfect, then m satisfies the compatibility condition up to $\sim$, and so is operadically consistent on every ToQ whose total composite is the question asked.*

An inconsistency is therefore a genuine failure of one of these conditions, or of the surrogate for $\sim$ (proof in App. A.4).

The empirics of §3 evaluate operadic consistency mainly on the simplest non-trivial ToQ, a depth-2 chain (one internal edge; Fig. 2); §3.3 applies the same predicate at depths 3 and 4 and on branching trees. They replace the binary predicate of Def. 2.3 with a continuous score in $[0, 1]$ measuring agreement under a per-dataset semantic equivalence, the *OC score*. The choice of equivalence is itself constrained by the framework: the OC scorer should realize the natural equivalence on the answer space appropriate to the question's expected output type. App. B.14 tests this empirically.

3 Empirical evidence

We present two empirical findings: §3.1 shows that across non-thinking models, operadic consistency is strongly correlated with accuracy where temperature self-consistency is not, and contributes predictive information about per-question correctness beyond standard sample-based, self-evaluation, and decomposition-aware baselines. §3.2 extends the analysis to thinking models, using operadic structure extracted from a model's own chain of thought.

3.1 Operadic consistency as a per-question selection signal

We first study operadic consistency on instruction-tuned (non-thinking) models, where the decomposition tree is supplied by human annotators rather than extracted from the model itself. Figure 2 illustrates the OC computation on the depth-2 chain from Ex. 2.2. We write OC for our signal and compare against three standard uncertainty baselines, defined below: canonical chain-of-thought self-consistency (CoT-SC) [Wang et al., 2023b], semantic entropy (H) [Kuhn et al., 2023], and $P(\text{True})$ [Kadavath et al., 2022].

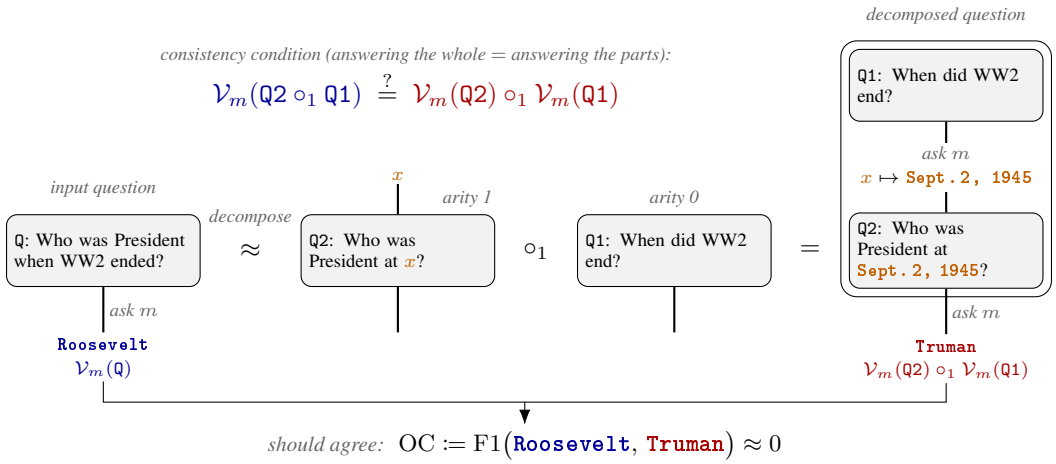


Figure 2: Operadic-consistency (OC) check on a depth-2 chain (Ex. 2.2), where m is the language model being evaluated. The undecomposed question Q (left) is semantically equivalent to the composite Q2 $\circ_1$ Q1 (center). The model is asked two ways: directly on Q, and step-by-step along the tree by answering Q1, substituting the result for x in Q2, and answering again. Model outputs are color-coded: the direct answer to Q in blue, the answer to Q1 (which carries through the substitution) in orange, and the answer to Q2 in red. OC compares the final answers under the score_official scorer, which we abbreviate here as F1; on this example OC ≈ 0, flagging the question.

Setup. We evaluate twelve instruction-tuned models spanning 4B to 671B parameters, open-weights and closed-source, listed with their API identifiers in App. B.1, on four multi-hop QA datasets (HotpotQA, MuSiQue, StrategyQA, DROP; 403–625 questions each), using two-step decompositions from Break QDMR annotations [Wolfson et al., 2020] for HotpotQA and DROP, and native decompositions for MuSiQue and StrategyQA.[2] For each (model, question) pair we compute:

- *Operadic consistency* (OC): semantic-agreement score between the model's greedy direct answer and its greedy decomposed answer (answer step 1, substitute into step 2, answer). Agreement is the per-dataset score_official scorer (App. B.6); the score lies in $[0, 1]$, thresholded at 0.5 where a binary partition is needed. Cost: 3 model calls.
- *Canonical CoT self-consistency* follows the K-sample CoT protocol of Wang et al. [2023b]: sample K chain-of-thought responses at $T{=}0.7$ and extract a short answer from each with a deterministic auxiliary extractor. The continuous score is mean pairwise score_official agreement over the K answers, at $K{=}3$ and $K{=}10$, using the same per-dataset scorer as OC so the two share one equivalence (App. B.4). Cost: K chain-generation calls plus K cheap extractor calls.
- *Semantic entropy* [Kuhn et al., 2023]: Shannon entropy over clusters of the $K{=}10$ samples, where two samples cluster together if their matched-axis pairwise agreement is ≥ 0.5 (App. B.14). Cost: $K{=}10$ calls plus clustering.
- $P(\textit{True})$ [Kadavath et al., 2022]: ask the model whether a proposed answer is correct and read $P(\text{Yes})/(P(\text{Yes}) + P(\text{No}))$ from the next-token logprobs. Cost: 1 model call.

Accuracy for each (model, question) pair is the indicator that the greedy direct answer agrees with the gold answer at per-dataset score_official ≥ 0.5. We report per-question regressions at $K{=}10$ (the literature-default budget at which CoT-SC and $-H$ are at full strength) and additionally include the equal-cost $K{=}3$ budget in selective-prediction comparisons (matching OC's 3 model calls).

Across-model finding. Figure 3 plots each consistency metric against accuracy across the twelve models, per dataset. Operadic consistency is strongly correlated with accuracy on every dataset (Pearson r between 0.86 and 0.94; $p \leq 0.0004$ on all four). Canonical CoT-SC shows a *bimodal* pattern: on HotpotQA and DROP it matches OC ($r{=}{+}0.93$ and $+0.87$, both $p < 0.001$), but on MuSiQue and

[2] GSM8K lacks Break-style two-step decomposition annotations and enters only in §3.2.

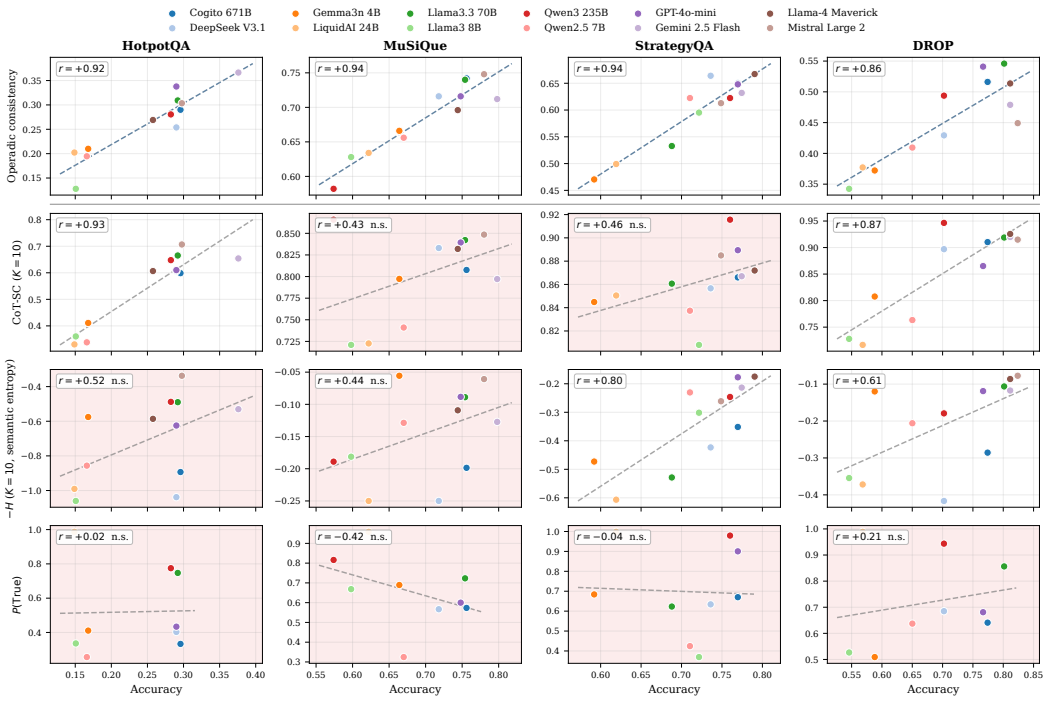


Figure 3: Signal vs. accuracy across twelve models and four datasets, with one row per uncertainty-baseline family. **Row 1:** operadic consistency, the proposal ($r \in [0.86, 0.94]$, all $p \leq 0.0004$). **Row 2:** canonical CoT self-consistency at $K{=}10$ [Wang et al., 2023b], the dominant sample-aggregate baseline. Bimodal: $r{=}{+}0.93, {+}0.43, {+}0.46, {+}0.87$ across the four datasets. **Row 3:** semantic entropy at $K{=}10$ [Kuhn et al., 2023]. **Row 4:** $P(\text{True})$ self-evaluation [Kadavath et al., 2022]. OC is the only row with $r \geq 0.85$ on every dataset; every other row has at least one dataset where $r < 0.85$. Full per-baseline table (including naive direct-answer SC, decomposed-SC, and PRMs) in App. B.13. Each point is one model; trend line and Pearson r with two-sided p shown per subplot; panels whose correlation is not significant at $\alpha = 0.05$ are shaded and marked *n.s.* HotpotQA is evaluated closed-book (no passage), hence the lower per-model accuracies; see App. B.2.

StrategyQA it collapses to $r{=}{+}0.43$ and $+0.46$ (both n.s.). No other sample-based or self-evaluation baseline we test achieves $r \geq 0.85$ on more than two datasets: semantic entropy $-H_{K=10}$ peaks at $+0.80$ on StrategyQA but drops to $+0.44$ on MuSiQue; $P(\text{True})$ shows no significant correlation on any dataset (App. B.13). OC remains the only baseline with $r \geq 0.85$ uniformly across all four datasets, consistent with the broader observation that sample-based methods probe within-question confidence (largely orthogonal to cross-model competence: a confidently-wrong model has low sample variance just as a confidently-correct one does), while OC probes a complementary compositional axis. OC's correlation with accuracy is robust to the choice of correctness scorer: under an LLM-judge correctness scorer (Claude Haiku 4.5 with a paraphrase-tolerant prompt), it gives $r \in [{+}0.69, {+}0.97]$, all $p \leq 0.013$ (App. B.14). Practically, the strong cross-model correlation immediately yields a label-free dataset-level *accuracy estimator*: a leave-one-model-out demonstration on the twelve non-thinking models (App. B.9) achieves mean absolute prediction error of 3.0 pp pooled (range 2.1–4.7 pp across datasets), relevant when ground-truth labels are scarce.

Per-question complementarity. Within a single model, do operadically-consistent questions get answered correctly more often, even controlling for CoT-SC and semantic entropy? Across the same twelve models pooled over four datasets ($n = 24{,}624$ (model, question) pairs), a per-dataset logistic regression of correctness on $(\text{OC}, \text{CoT-SC}_{K=10}, -H_{K=10})$, with standard errors clustered by question, gives the coefficients in Table 1. All three are significant at $p < 10^{-15}$ on every dataset: the signals are complementary and none subsumes another.

Selective prediction. The coefficients translate into gains at roughly one third the inference cost of the strongest sample-based baseline (3 vs. 10 calls). For each (model, dataset) cell we compute AUARC [Geifman and El-Yaniv, 2017] and AUROC of a logistic-regression combiner over $(\text{OC}, \text{CoT-SC}_{K=3})$, fit on the other eleven models and evaluated on the held-out cell, so the calibrator never sees the cell it scores. At the equal-cost $K{=}3$ budget, OC+CoT-SC improves AUARC by $+0.086$ to $+0.096$ and AUROC by $+0.092$ to $+0.164$ over CoT-SC$_{K=3}$ alone (right panel of Table 1); bootstrap 95% CIs over the twelve paired model-level differences exclude zero on every dataset. Lifts against the literature-default CoT-SC$_{K=10}$ budget are uniformly smaller but still positive with CIs excluding zero on every dataset (App. B.12).

dataset	β_{OC} (p)	$\beta_{\text{CoT-SC}}$ (p)	β_{-H} (p)
HotpotQA	$+1.90$ (10^{-38})	$+1.86$ (10^{-27})	$+0.94$ (10^{-30})
MuSiQue	$+2.24$ (10^{-43})	$+1.88$ (10^{-16})	$+1.37$ (10^{-25})
StrategyQA	$+1.69$ (10^{-37})	$+1.93$ (10^{-19})	$+1.29$ (10^{-82})
DROP	$+1.19$ (10^{-16})	$+2.24$ (10^{-23})	$+2.09$ (10^{-53})

dataset	Δ_{AUARC}	Δ_{AUROC}
HotpotQA	$+0.095$	$+0.092$
MuSiQue	$+0.086$	$+0.153$
StrategyQA	$+0.096$	$+0.164$
DROP	$+0.086$	$+0.117$

Table 1: *Left:* three-way logistic regression of correctness on $(\text{OC}, \text{CoT-SC}_{K=10}, -H_{K=10})$ (all continuous), fit per dataset on twelve non-thinking models, with cluster-robust standard errors clustered by question. All three predictors are massively significant ($p < 10^{-15}$) on every dataset, indicating that OC, CoT-SC, and semantic entropy each carry complementary marginal information; no single signal subsumes the others. App. B.10 reports a four-way variant adding $P(\text{True})$. *Right:* selective-prediction lift from adding OC to a logistic-regression combiner over CoT-SC$_{K=3}$, leave-one-model-out across the same twelve models; each cell is the paired mean of [AUARC/AUROC of OC+CoT-SC] minus [AUARC/AUROC of CoT-SC alone] at the equal-cost ($K{=}3$) budget. Bootstrap 95% CIs (full intervals in App. B.12) exclude zero on every cell. Lifts vs. CoT-SC$_{K=10}$ are uniformly smaller but still all positive with CIs excluding zero (App. B.12).

OC outperforms decomposition-aware baselines. The baselines compared so far (CoT-SC, semantic entropy, and $P(\text{True})$) share a common feature: none uses the question decomposition. This raises two concrete concerns about what OC's positive correlation with accuracy is actually measuring. *Concern (i): OC is self-consistency in disguise.* The decomposed-path answer is just one model output, and if the model is unstable on the decomposed path, OC's agreement with the direct answer might track per-question variance on that path rather than any compositional property; the same question signal would be obtainable from K-sample self-consistency restricted to the decomposed-path answer. *Concern (ii): OC is a worse process reward model.* An off-the-shelf process reward model (PRM) scored on the decomposed steps could plausibly capture everything OC captures and more, in which case OC adds no information beyond cheap-PRM features. We test both with constructed decomposition-aware baselines (each combines standard machinery with our specific decomposition protocol; neither is a published off-the-shelf comparison): (i) *decomposed self-consistency* (our continuous pairwise-agreement adaptation of self-consistency [Wang et al., 2023b], applied to $K{=}10$ samples of the decomposed-path answer; see App. B.4); and (ii) *Skywork-PRM*, an open process reward model [He et al., 2024, Lightman et al., 2024] trained on Skywork-o1 step preferences for math and code reasoning, applied off-label to the $(\texttt{Q1}, \texttt{a1}, \texttt{Q2}, \texttt{a2})$ trace.

At the cross-model level, both decomposition-aware baselines show isolated significant correlations on a single dataset each: decomposed-SC on HotpotQA ($r{=}+0.67$, $p{=}0.018$) and Skywork-PRM on DROP ($r{=}+0.82$, $p{=}0.001$). Neither correlates significantly on the other three datasets, leaving OC as the only baseline with $r \geq 0.85$ uniformly. At the per-question level, we re-run the logistic regression with each new baseline added as a fourth predictor alongside OC, CoT-SC$_{K=10}$, and $-H_{K=10}$, fit per dataset on the same twelve non-thinking models with cluster-robust standard errors clustered by question. Across all four datasets the OC coefficient remains massively significant ($\beta_{\text{OC}} \geq +1.11, p < 10^{-12}$; full coefficients in App. B.11). Decomposed self-consistency's coefficient is non-significant or *negative* on every dataset (significantly negative on HotpotQA and StrategyQA), decisively rejecting (i). Skywork-PRM contributes independent information on two of four datasets but never displaces OC, addressing (ii) modulo the off-label use.

3.2 Operadic consistency for thinking models

The chains in §3.1 are annotator-supplied. Thinking models offer an alternative: their chain of thought already exhibits a decomposition into sub-questions and sub-answers, which lets us run the check without external annotations. The reliability of thinking-model confidence is itself an active question [Mei et al., 2026, Yoon et al., 2025, Podolak and Verma, 2025, Kirichenko et al., 2025]; OC supplies a complementary, decomposition-based signal.

CoQ extraction from CoT. For each (model, question) pair we generate one greedy thinking trace and use a lightweight extractor to lift a two-step chain `Q1` $\rightarrow$ `Q2` from it, discarding traces the extractor flags as not factorable (App. B.21 audits these). The resulting depth-2 ToQ is the model's own implicit decomposition, and we then run the check exactly as in §3.1.

Setup. We evaluate five thinking models (DeepSeek-R1, GLM-5.1, Cogito v2.1, Kimi-K2.6, MiniMax-M2.7) on four multi-hop QA datasets (MuSiQue, StrategyQA, GSM8K, DROP), 100 questions per (model, dataset) cell at one seed. After removing extractor-flagged unfactorable traces and small per-cell losses to API errors, we retain roughly 94 questions per cell on average. $P(\text{True})$ is not available for thinking endpoints in our protocol (no top-k logprobs after the ⟨think⟩ tag), so the Kadavath baseline is omitted here.

Because thinking models produce reasoning chains natively, the self-consistency variant compared throughout this section is structurally identical to the canonical CoT-SC of §3.1: K samples at T=0.7, the same Llama-3.3-70B short-answer extractor (App. B.4), and matched-axis pairwise per-dataset `score_official` agreement. For compactness, we abbreviate it SC here.

Equal-cost comparison. Operadic consistency on a depth-2 ToQ costs 3 thinking calls per question (direct, sub-question 1, sub-question 2). $\text{SC}_{K=3}$ also costs 3 thinking calls; $-H_{K=3}$ uses the same three samples and is essentially free additional information on top of $\text{SC}_{K=3}$. We therefore start with an apples-to-apples comparison at the K=3 budget.

Pooling across the five models and four datasets ($n = 1{,}878$ (model, question) pairs after filtering), univariate marginal predictive power orders as $\text{SC}_{K=3}$ (pseudo-$R^2 = 0.174$) > OC (0.128); both are significant at $p < 10^{-32}$ under cluster-robust standard errors clustered by question. Sample-based diversity is the stronger solo signal at this budget. $\text{SC}_{K=3}$ and $-H_{K=3}$ are nearly identical predictors (Pearson $+0.97$ between $\text{SC}_{K=10}$ and $-H_{K=10}$ in this data, with a similar relationship at K=3), because both summarize sample diversity. OC, by contrast, probes *compositional* disagreement and is only weakly correlated with sample diversity (Pearson $+0.39$), so it can contribute information that neither captures.

A bivariate logistic regression of correctness on operadic consistency and sample diversity (using $\text{SC}_{K=3}$ to summarize sample diversity, since $\text{SC}_{K=3}$ and $-H_{K=3}$ are nearly collinear) confirms this:

$$\beta_{\text{OC}} = +1.42\ (p < 10^{-17}), \quad \beta_{\text{SC}} = +2.55\ (p < 10^{-30}), \quad \text{pseudo-}R^2 = 0.226.$$

OC's coefficient remains highly significant after controlling for sample diversity, and pseudo-R^2 rises by 0.052 over sample-diversity alone, a ∼30% relative gain, significant on three of four datasets (Table 2). The conclusion is robust to budget: at K=10 ($n = 1{,}793$ across the 20 thinking-model $\times$ dataset cells; the 85 excluded rows are all Kimi-K2.6 on MuSiQue, where fewer than five of the ten samples parsed, leaving semantic entropy undefined), the three-way regression correct $\sim$ OC + $\text{SC}_{K=10}$ + $(-H_{K=10})$ gives $\beta_{\text{OC}} = +1.06$ ($p < 10^{-7}$), with SC absorbing most of the variance and crowding out semantic entropy. The same complementarity pattern emerges from the binary (OC, SC unanimous at K=10) partition: jointly-consistent questions are answered correctly 43–62 pp higher than jointly-inconsistent ones.

Selective prediction. We repeat the leave-one-model-out AUARC and AUROC analysis of §3.1 on the five thinking models. At the equal-cost (3-call) budget, the OC+$\text{SC}_{K=3}$ combiner improves over $\text{SC}_{K=3}$ alone on all four datasets (Δ_{AUARC} from $+0.017$ on DROP to $+0.042$ on StrategyQA; Δ_{AUROC} from $+0.050$ on StrategyQA to $+0.134$ on GSM8K), with bootstrap 95% CIs excluding zero on every cell (right panel of Table 2). At the 10-call budget, all four datasets retain positive lift; the StrategyQA and DROP K=10 cells have CIs that bracket zero by a small margin (full intervals in App. B.12). All 16 (dataset, budget, metric) cells tested show positive point-estimate lift; CIs exclude zero on 12 of the 16.

dataset	n	β_{OC} (p)	β_{SC} (p)	pR^2
MuSiQue	447	$+1.16$ (10^{-5})	$+0.95$ $(.008)$	0.086
StrategyQA	480	$+0.90$ $(.003)$	$+1.86$ (10^{-7})	0.114
GSM8K	494	$+2.94$ (10^{-5})	$+3.04$ (10^{-5})	0.417
DROP	457	$+0.87$ (ns)	$+2.42$ (10^{-4})	0.053

dataset	Δ_{AUARC}	Δ_{AUROC}
MuSiQue	$+0.033$	$+0.052$
StrategyQA	$+0.042$	$+0.050$
GSM8K	$+0.029$	$+0.134$
DROP	$+0.017$	$+0.071$

Table 2: *Left:* bivariate logistic regression correct $\sim$ OC + $\text{SC}_{K=3}$ at the equal-cost (K=3) budget on the five thinking models, fit per dataset and pooled across models, with cluster-robust standard errors clustered by question. OC contributes statistically significant predictive information beyond sample diversity on three of four datasets, with DROP the lone exception (the joint-positive cell holds 66% of the data, so this is dataset saturation rather than signal failure). *Right:* selective-prediction lift from adding OC to a logistic-regression combiner over $\text{SC}_{K=3}$, leave-one-model-out across the same five models, at the equal-cost (K=3) budget. Bootstrap 95% CIs (see App. B.12) exclude zero on every cell. K=10 lifts (smaller, with StrategyQA / DROP cells bracketing zero) are in App. B.12.

3.3 Deeper and branching trees of questions

Def. 2.3 is stated on an arbitrary ToQ, so we can apply it equally well to finer decompositions, for instance to longer chains. MuSiQue alone supplies 3- and 4-hop annotations; across twelve models at $n = 1800$ per depth the signal persists (protocol and per-topology results in App. B.16). The accuracy gap between OC-consistent and OC-inconsistent questions is $+43.1$ pp at depth 3 and $+36.9$ pp at depth 4 (per-question AUROC 0.768 and 0.667), with β_{OC} significant at both ($p = 2.6 \times 10^{-13}$, $p = 1.3 \times 10^{-9}$). Scored identically at every depth, AUROC is 0.743, 0.741, and 0.682 at depths 2, 3, and 4. 3-hop branching is our strongest cell (0.811) and branching 4-hop is our weakest (0.541), which App. B.16 attributes partly to measurement and rescores to 0.581. These runs predate our move to canonical CoT-SC and use the naive direct-answer SC ablation: against it OC trails at depth 4 (0.667 vs 0.703) while leading at depth 3, and adding OC lifts AUROC to 0.759.

As described in App. B.18, one of the authors produced a 240-case hand-labelling of OC-inconsistent cases, which shows that decomposition noise is not what produces the signal: setting every decomposition-induced case aside leaves the gap intact.

4 Related work

Question decomposition and chain-of-thought. Eliciting intermediate reasoning steps from LLMs has been a dominant strategy for improving performance on multi-step tasks. Wei et al. [2022] introduced chain-of-thought prompting; subsequent work elaborated on the structure of decomposition with zero-shot prompting [Kojima et al., 2022], decomposed prompting [Khot et al., 2023], least-to-most prompting [Zhou et al., 2023], plan-and-solve prompting [Wang et al., 2023a], tree-of-thoughts [Yao et al., 2023], graph-of-thoughts [Besta et al., 2024], buffer-of-thoughts [Yang et al., 2024a], and uncertainty-aware decomposition [Hu et al., 2024]. Question-decomposition annotations on multi-hop QA datasets are formalized in BREAK [Wolfson et al., 2020]. Most directly adjacent is Press et al. [2023], who introduce the *compositionality gap* — the population-level discrepancy between sub-question accuracy and multi-hop accuracy when sub-answers are correct — and propose self-ask prompting as a partial mitigation. Maasch et al. [2025] introduce an evaluation framework and task generator for compositional causal reasoning, with compositional consistency as the central metric. Our framework instead extracts a per-question consistency signal at inference time, so it requires no specialized benchmark and can drive selective prediction on any compositional task. Related checks appear under other names [Gandhi et al., 2022, Chen et al., 2024, Yang et al., 2024b], a convergence we read as evidence that it is natural. Closest to our empirical setting is Madhwal et al. [2026], who use disagreement between a model's direct answer and its execution of a verified gold decomposition as an abstention signal (DBA), with error-detection gains over self-evaluation across nine models and six closed-book benchmarks. Their signal and our depth-2 OC coincide in the annotator-decomposition setting, but we extract the decomposition from the model's own chain of thought, removing the dependence on verified plans they flag as a limitation of DBA, and we read the per-question rate as a label-free estimator of a held-out model's accuracy rather than only as an abstention rule. Our framework treats this entire family as algebras over the questions operad.

Confidence and uncertainty for LLM reasoning. Our baselines come from three inference-time families: sample-and-aggregate [Wang et al., 2023b], cluster-based entropy [Kuhn et al., 2023, Farquhar et al., 2024, Ciosek et al., 2025], and self-evaluation [Kadavath et al., 2022, Lin et al., 2022, Tian et al., 2023]; process-level verification [Lightman et al., 2024, Uesato et al., 2022] is adjacent and complementary. App. C gives the detail.

CoT faithfulness and mechanistic accounts. App. C situates OC against work on chain-of-thought faithfulness [Turpin et al., 2023, Lanham et al., 2023] and mechanistic accounts of multi-hop failure [Biran et al., 2024, Lindsey et al., 2025].

Categorical and operadic structures in language and learning. Algebraic structures have appeared in computational linguistics through semiring parsing [Goodman, 1999, Nederhof, 2003], Hopf-algebraic models of syntactic structure [Marcolli et al., 2023], and compositional distributional semantics [DisCoCat; Coecke et al., 2010]; we are not aware of prior application of operads to LLM question decomposition or to inference-time confidence signals.

5 Further directions

Native structure on chains of thought. The thinking-model results in §3.2 use a lightweight extractor that surfaces a single explicit sub-question from a CoT trace, a crude reduction to a depth-2 chain. A natural direction is to formalize chains of thought *natively* and apply consistency probes to richer per-trace structure, using either depth-≥ 3 extensions of the present framework or structures better suited to the branching and revisitation of long traces. App. B.19 reports a head-to-head against VersaPRM [Zeng et al., 2025], the closest in-domain process reward model, where OC and the PRM prove complementary rather than redundant; DPRM [Wang et al., 2026] does not apply here.

Other algebras over $\mathcal{Q}$. A QA model determines a $\mathcal{Q}$-algebra valued in answer strings, but the same compositional structure supports algebras valued in other quantities, such as a *confidence algebra* composing calibrated probabilities under decomposition.

Limitations

§3.3 extends the depth-2 instantiation to depths 3 and 4 and to branching trees, but on MuSiQue alone; branching 4-hop remains our weakest cell, and the task-breadth pilot of App. B.17 shows OC over-flagging on arithmetic. The decomposition tree is either annotator-supplied (§3.1) or extracted from the model's chain of thought (§3.2); both have failure modes (e.g. traces without explicit sub-questions are discarded). Our DROP scorer is more lenient than canonical Dua et al. [2019] scoring, following the framework prediction that OC compares values on the answer space suited to the question's output type; App. B.14 reports canonical-scorer and LLM-judge checks that corroborate it. The §3.1 cohort spans 4B to 671B-parameter models, but does not span the largest commercial systems (e.g. GPT-5, Claude Opus 4.x, Gemini 2.5 Pro), some of which produce sufficiently verbose answers that token-F1 scoring would require additional calibration. The thinking-model cohort is small, so per-cell statistical power is limited.

Impact Statement

This paper aims to improve LLM reasoning reliability via a label-free inference-time signal for compositional inconsistency. It presents no obvious dual-use concerns and uses only public benchmarks and APIs; we recommend deploying operadic consistency as one component of a multi-signal stack.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) through the Super Intelligence Quantified (SIQ) program, under Cooperative Agreement No. HR00112520028. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of

Defense or the U.S. Government. N.B. thanks the Allen Institute for AI for its hospitality during this collaboration. This was the only third-party funding or support received for any aspect of the work.

Competing interests. Other than the funding acknowledged above and the authors' institutional affiliations, which are stated in the author list, the authors declare no financial relationships that could be perceived as influencing the submitted work, and no author holds a financial relationship with any of the commercial model providers whose systems are evaluated here.

References

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.837. URL https://aclanthology.org/2024.emnlp-main.837/.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://arxiv.org/abs/2406.12775.

Nathaniel Bottman and Kyle Richardson. Operads for compositional reasoning in LLMs, 2026. arXiv:2606.13634.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. Two failures of self-consistency in the multi-step reasoning of LLMs. *Transactions on Machine Learning Research*, 2024.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy. *arXiv preprint arXiv:2504.03579*, 2025.

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

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384, 2010.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2368–2378, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Brendan Fong and David I. Spivak. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge University Press, Cambridge, UK, 2019. ISBN 9781108711821. doi: 10.1017/9781108668804. URL https://www.cambridge.org/9781108711821.

Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5046–5055, 2022.

Bruno Gavranović, Paul Lessard, Andrew Dudzik, Tamara von Glehn, João G. M. Araújo, and Petar Veličković. Position: Categorical deep learning is an algebraic theory of all architectures. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 15209–15241. PMLR, 2024. URL https://proceedings.mlr.press/v235/gavranovic24a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NeurIPS*, 2017.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370.

Joshua Goodman. Semiring parsing. *Computational Linguistics*, 25(4):573–606, 1999.

Jujie He, Tianwen Wei, Rui Yan, Jiacai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. Skywork-o1 open series, 2024. URL https://doi.org/10.5281/zenodo.16998085.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4615–4634, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.254. URL https://aclanthology.org/2024.acl-long.254/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *Proceedings of ICLR*, 2023.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pages 190800–190850, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. Transformer Circuits Thread, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/biology.html`.

Jean-Louis Loday and Bruno Vallette. *Algebraic operads*, volume 346. Springer Science & Business Media, 2012.

Jacqueline R. Maasch, Alihan Hüyük, Xinnuo Xu, Aditya V. Nori, and Javier Gonzalez. Compositional causal reasoning evaluation in language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pages 42325–42365. PMLR, 2025.

Dhruv Madhwal, Lyuxin David Zhang, Dan Roth, Tomer Wolfson, and Vivek Gupta. Decomposed prompting does not fix knowledge gaps, but helps models say "I don't know". In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 36688–36710, 2026. doi: 10.18653/v1/2026.findings-acl.1829.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL `https://aclanthology.org/2023.emnlp-main.557/`.

Matilde Marcolli, Robert C Berwick, and Noam Chomsky. Syntax-semantics interface: an algebraic model. *arXiv preprint arXiv:2311.06189*, 2023.

Martin Markl, Steve Shnider, and Jim Stasheff. *Operads in Algebra, Topology and Physics*, volume 96 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2002. doi: 10.1090/surv/096.

J.P. May. *The geometry of iterated loop spaces*, volume 271 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, Heidelberg, 1972. ISBN 978-3-540-05904-2. doi: 10.1007/BFb0067491.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know? In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3408–3458, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-eacl.178.

Mark-Jan Nederhof. Weighted deductive parsing and Knuth's algorithm. *Computational Linguistics*, 29(1):135–143, 2003.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2080–2094. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.168.

Jakub Podolak and Rajeev Verma. Read your own mind: Reasoning helps surface self-confidence signals in LLMs. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 247–258, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.uncertainlp-main.21.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378.

Michael Sipser. *Introduction to the theory of computation*. PWS Publishing Company, 1996. ISBN 9780534952501.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada, 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.147. URL https://aclanthology.org/2023.acl-long.147/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Xinyi Wang, Yiping Song, Zhiliang Tian, Bo Liu, Tingjin Luo, and Minlie Huang. DPRM: A dual implicit process reward model in multi-hop question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33683–33691, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *Proceedings of ICLR*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. Break It Down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198, 2020.

Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. An implementation of generative PRM, 2024. RLHFlow project; https://github.com/RLHFlow/RLHF-Reward-Modeling.

Ran Xu, Yuchen Zhuang, Zihan Dong, Ruiyu Wang, Yue Yu, Joyce Ho, Linjun Zhang, Haoyu Wang, Wenqi Shi, and Carl Yang. AceSearcher: Bootstrapping reasoning and search for LLMs via reinforced self-play. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37:113519–113544, 2024a.

Qian Yang, Weixiang Yan, and Aishwarya Agrawal. Decompose and compare consistency: Measuring VLMs' answer reliability via task-decomposition consistency comparison. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3613–3627, 2024b. doi: 10.18653/v1/2024.emnlp-main.211.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pages 114971–114998, 2025.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, and Kangwook Lee. VersaPRM: Multi-domain process reward model via synthetic reasoning data. In *International Conference on Machine Learning (ICML)*, pages 74197–74239, 2025.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=WZH7099tgfM`.

A Operad background

This appendix records the formal definitions of operad and algebra over an operad referenced in §2. There are few approachable references on operads; we suggest the curious reader consult Markl–Shnider–Stasheff's *Operads in Algebra, Topology and Physics* Markl et al. [2002], and Loday–Vallette's *Algebraic Operads* Loday and Vallette [2012].

A.1 Operads

Operads model properties of many-in, one-out morphisms. One feature that such morphisms have is that we can compose them. Indeed, consider mappings f and g, which take k and ℓ inputs, respectively. For any $1 \le i \le k$, we can form the composition

$$f\Big(\underbrace{-,\dots,-}_{i-1}, g\Big(\underbrace{-,\dots,-}_{\ell}\Big), \underbrace{-,\dots,-}_{k-i}\Big), \tag{1}$$

i.e. the result of feeding the output of g into the i-th input slot of f. The result is a mapping that takes $k+\ell-1$ inputs, which we denote $f \circ_i g$. Iterated compositions are unambiguous — for instance, when we write the expression

$$f(\dots, g(\dots, h(\dots), \dots), \dots), \tag{2}$$

the meaning is independent of whether we first compose f with g and then feed in h, or first compose g with h and then feed the result into f. These observations lead us to the definition of an operad.

Definition A.1. A *non-unital operad* $\mathcal{O}$ is the following data:

- A collection $\mathcal{O}(0), \mathcal{O}(1), \mathcal{O}(2), \dots$ of sets. (Elements of $\mathcal{O}(k)$ are thought of as arity-k morphisms, i.e. operations with k inputs and one output.)
- For every $k, \ell \ge 0$ and $1 \le i \le k$, a "composition map" $\circ_i \colon \mathcal{O}(k) \times \mathcal{O}(\ell) \to \mathcal{O}(k+\ell-1)$. (When $\mathcal{O}$ is literally a collection of morphisms, and we are given $f \in \mathcal{O}(k)$ and $g \in \mathcal{O}(\ell)$, $f \circ_i g$ is obtained by plugging the output of g into the i-th input slot of f.)

These data must satisfy two associativity axioms. *Sequential associativity*: for $f \in \mathcal{O}(k), g \in \mathcal{O}(\ell), h \in \mathcal{O}(m)$, and $1 \le i \le k$, $1 \le j \le \ell$, the compositions $f \circ_i (g \circ_j h)$ and $(f \circ_i g) \circ_{i+j-1} h$ agree (compare (2)). *Parallel associativity*: for $f \in \mathcal{O}(k), g \in \mathcal{O}(\ell), h \in \mathcal{O}(m)$ and $1 \le i < j \le k$, substituting into disjoint slots of f commutes, $(f \circ_i g) \circ_{j+\ell-1} h = (f \circ_j h) \circ_i g$. (The shifted index $j+\ell-1$ accounts for g's ℓ blanks displacing the j-th slot of f.)

The objects we call "operads" in this paper are, more precisely, non-unital non-symmetric operads in Set. (The standard definition also includes an identity morphism $\mathrm{id}_{\mathcal{O}} \in \mathcal{O}(1)$ satisfying $\alpha \circ_i \mathrm{id}_{\mathcal{O}} = \alpha$ and $\mathrm{id}_{\mathcal{O}} \circ_1 \alpha = \alpha$. We omit it because the questions operad has no natural identity question and we make no use of unitality elsewhere; the rest of the development is unchanged.)

Example: text-processing operad. The *text-processing operad* $\mathcal{F}_\Sigma$ has $\mathcal{F}_\Sigma(k)$ equal to the set of k-ary functions $(\Sigma^*)^k \to \Sigma^*$ on strings over an alphabet Σ, with $\circ_i$ ordinary function composition.

Why operads are helpful. The appeal of operads is that they allow us to separate the abstract forms of composition from any particular interpretation of those forms. Operads were introduced by May [1972] to capture algebraic structure on iterated loop spaces; this gave rise to a central principle that pervades modern algebra, topology, and geometry: *objects with extra structure should be understood as algebras over an appropriate operad.* This is the principle we apply to question decomposition: operads provide the abstract structure of composition, and a QA model is then realized as an algebra over the questions operad $\mathcal{Q}$ defined in §2.1.

Colored operads. A natural extension records the *type* of each input and output. In a *colored* (or *typed*) operad, each operation f has prescribed input colors $c_1, \dots, c_k$ and output color d, written $f \in \mathcal{O}(c_1, \dots, c_k; d)$, and compositions $\circ_i$ are restricted to color-matching pairs. (A category is the special case where all operations are unary; for this reason colored operads are sometimes called *multicategories*.) The questions operad of §2.1 is naturally colored by answer type, and the framework

prediction discussed in §2.2 — that an OC scorer should realize the natural equivalence on the answer space appropriate to the question's expected output type — takes its cleanest form in the colored variant. We work uncolored in the body since none of our empirical analysis depends on per-type structure; the colored extension is a natural direction for future work.

A.2 A second example: derivation trees of context-free grammars

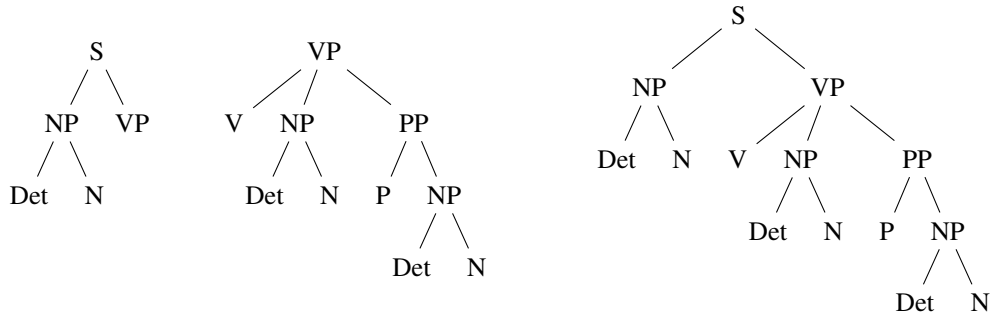
Example A.2. Suppose that $G = (V, \Sigma, R, S)$ is a context-free grammar, as in §2.1 of Sipser [1996]. We define a *derivation tree with non-terminal leaves* for G to be a rooted planar tree T such that:

- Every node is labeled by a non-terminal $A \in V$;
- If an internal node is labeled A, and its children are labeled $B_1, \ldots, B_k$, then $A \to B_1, \ldots, B_k$ must be a production rule in R.

We define D_G to be the colored operad of such derivation trees, with colors V. That is:

- Given $B_1, \ldots, B_k \in V$ and $A \in V$, $D_G(B_1, \ldots, B_k; A)$ consists of the derivation trees for G whose k leaves have labels $B_1, \ldots, B_k$ and whose root has label A.
- Composition in D_G is given by grafting the root of one derivation tree to a leaf of another — a common operation in context-free parsing — subject to the requirement that the labels at the two nodes being identified must match.

For instance, suppose that G is a CFG modeling English, with non-terminals $\{$S, NP, VP, PP, Det, N, V, P$\}$ (standing for "sentence", "noun phrase", etc.). Below, T_1 and T_2 (left) are derivation trees in $D_G(\text{Det}, \text{N}, \text{VP}; \text{S})$ and $D_G(\text{V}, \text{Det}, \text{N}, \text{P}, \text{Det}, \text{N}; \text{VP})$, respectively. Grafting them at their shared VP node yields the composition $T_1 \circ_3 T_2 \in D_G(\text{Det}, \text{N}, \text{V}, \text{Det}, \text{N}, \text{P}, \text{Det}, \text{N}; \text{S})$ (right).



A.3 Algebras over an operad

Operads do not necessarily have to literally be collections of operations. A realization of an operad in terms of actual operations is called an *algebra*.

Definition A.3. Suppose that $\mathcal{O}$ is an operad. An *algebra $\mathcal{A}$ over $\mathcal{O}$* consists of a set $\mathcal{A}$ together with, for every $f \in \mathcal{O}(k)$, a map $\varphi_f \colon \mathcal{A}^k \to \mathcal{A}$. We require that $\varphi_{f \circ_i g}$ is the same operation as inserting the output of φ_g into the i-th input of φ_f. We call this the *compatibility condition*; it is sometimes called "associativity", but we avoid this so that there is no confusion with the similarly named axioms of the operad itself.

For instance, Σ^* is naturally an algebra over $\mathcal{F}_\Sigma$: for each $f \in \mathcal{F}_\Sigma(k)$, the corresponding algebra operation is simply the map $f : (\Sigma^*)^k \to \Sigma^*$.

A.4 Semantic equivalence: remarks, and the proof of Proposition 2.5

We now describe the two relations used in §2.2. Fix a set D of denotations, together with a partial interpretation ι sending an answer string to its denotation in D. Two answers are *semantically equivalent*, $a \sim b$, if $\iota(a) = \iota(b)$; being equality of denotations, $\sim$ is an equivalence relation wherever ι is defined, and a partial equivalence relation on all strings.

Next, we define equivalence of questions. Write $\mathcal{V}$ for the set of $\sim$-classes, the *values*; this is the value set carrying the algebras of §2.1. Say that $q \in \mathcal{Q}(k)$ is *well posed* at $v \in \mathcal{V}^k$ if filling its blanks with representatives of $v_1, \ldots, v_k$ yields a question whose answer is unique up to $\sim$ and independent of which representatives were chosen. The denotation of q is then the partial function $\gamma_q \colon \mathcal{V}^k \rightharpoonup \mathcal{V}$, defined exactly where q is well posed. Two templates are *equivalent*, $q \approx q'$, if $\gamma_q = \gamma_{q'}$ as partial functions, that is, if they have the same domain and agree at every point of it. Well-posedness is exactly what lets γ_q descend to $\sim$-classes.

Two further remarks before the proof. First, $\sim$ is an equivalence relation by construction, being equality of denotations, but the scorers that stand in for it empirically need not be: exact match is an equivalence relation, whereas the token-F1 threshold of §3.1 is reflexive and symmetric but *not* transitive. The empirical OC score therefore measures agreement within a tolerance rather than membership in an equivalence class, and the guarantees below attach to $\sim$ and $\approx$ themselves rather than to these surrogates. Second, the descent of $\circ_i$ to $\approx$-classes needs no well-posedness hypothesis, for the reason given at the end of the argument below.

$\mathcal{Q}/\approx$ is an operad. Assumption 2.4 makes $\approx$ a congruence: if $q \approx q'$ and $r \approx r'$ then $\gamma_{q \circ_i r} = \gamma_q \circ_i \gamma_r = \gamma_{q'} \circ_i \gamma_{r'} = \gamma_{q' \circ_i r'}$, so $q \circ_i r \approx q' \circ_i r'$. The two associativity axioms then descend as follows. Apply Assumption 2.4 twice to each side of the first identity. The two denotations are

$$\gamma_{(q_1 \circ_i q_2) \circ_{i+j-1} q_3} = \gamma_{q_1 \circ_i q_2} \circ_{i+j-1} \gamma_{q_3} = (\gamma_{q_1} \circ_i \gamma_{q_2}) \circ_{i+j-1} \gamma_{q_3},$$

$$\gamma_{q_1 \circ_i (q_2 \circ_j q_3)} = \gamma_{q_1} \circ_i \gamma_{q_2 \circ_j q_3} = \gamma_{q_1} \circ_i (\gamma_{q_2} \circ_j \gamma_{q_3}).$$

These agree, since substitution of partial functions into one another is associative, and equality of denotations is exactly $\approx$. The second identity is the same computation, using instead that substitutions of partial functions into disjoint slots commute.

No well-posedness hypothesis is required. Because γ is partial and $\approx$ is equality of partial functions, a composite that is undefined on some input is undefined on that input on both sides, so undefinedness is carried identically and does not have to be excluded by hypothesis.

Proof of Proposition 2.5. Take a depth-2 chain, so $q \approx q_2 \circ_1 q_1$. Then

$$\mathcal{V}_m(q_2 \circ_1 q_1) \;=\; \gamma_{q_2 \circ_1 q_1} \;=\; \gamma_{q_2} \circ_1 \gamma_{q_1} \;=\; \mathcal{V}_m(q_2) \circ_1 \mathcal{V}_m(q_1),$$

the left-hand side being the direct answer to the composite and the right-hand side the decomposed answer. The outer equalities follow from perfection, and the middle one is Assumption 2.4. Perfection applied to q itself gives $\mathcal{V}_m(q) = \gamma_q = \gamma_{q_2 \circ_1 q_1}$, so the question actually asked has the same answer. Induction along an arbitrary ToQ (by composing one internal edge at a time) gives the general case. □

B Experimental details

This appendix records the experimental protocol behind the empirical findings reported in §3.

B.1 Models and inference infrastructure

All models are accessed via OpenAI-compatible serverless inference APIs (Together AI for the open-weights and Cogito models; OpenRouter for GPT-4o-mini, Gemini 2.5 Flash, and Llama-4 Maverick) at temperature $T \in \{0, 0.7\}$ and a per-call max-token budget of 64. The full set of models, with the API identifiers used, is:

- *Non-thinking models* (§3.1): Llama-3-8B (`meta-llama/Meta-Llama-3-8B-Instruct-Lite`), Qwen2.5-7B (`Qwen/Qwen2.5-7B-Instruct-Turbo`), Gemma-3n-4B (`google/gemma-3n-E4B-it`), LiquidAI-LFM2-24B (`LiquidAI/LFM2-24B-A2B`), Llama-3.3-70B (`meta-llama/Llama-3.3-70B-Instruct-Turbo`), Qwen3-235B (`Qwen/Qwen3-235B-A22B-Instruct-2507-tput`), DeepSeek-V3.1 (`deepseek-ai/DeepSeek-V3.1`), Cogito-671B (`deepcogito/cogito-v2-1-671b`), GPT-4o-mini (`openai/gpt-4o-mini` via OpenRouter), Gemini 2.5 Flash (`google/gemini-2.5-flash` via OpenRouter), Llama-4 Maverick (`meta-llama/llama-4-maverick` via OpenRouter), Mistral Large 2 (`mistralai/mistral-large-2411` via OpenRouter).

- *Thinking models* (§3.2): DeepSeek-R1 (`deepseek-ai/DeepSeek-R1`), GLM-5.1 (`zai-org/GLM-5.1`), Cogito-v2.1 (`deepcogito/cogito-v2.1`), Kimi-K2.6 (`moonshotai/Kimi-K2.6`), MiniMax-M2.7 (`MiniMaxAI/MiniMax-M2.7`). Thinking calls use a per-call budget of 8192 tokens and temperature $T = 0$ for the operadic-consistency direct/decomposed answers and $T = 0.7$ for the K-sample self-consistency runs.
- *Auxiliary extractor models.* Two fixed, cheap non-thinking models serve distinct auxiliary roles. For the operadic-consistency pipeline on thinking models (§3.2), covering both sub-question identification in CoQ extraction and short-answer extraction from verbose thinking outputs, we use Qwen3-Next-80B-A3B-Instruct (`Qwen/Qwen3-Next-80B-A3B-Instruct`). For short-answer extraction from chain-of-thought samples in the canonical CoT-SC baseline (App. B.4) we use Llama-3.3-70B in non-thinking mode (`meta-llama/Llama-3.3-70B-Instruct-Turbo`), in both the non-thinking and thinking cohorts. Each extractor is fixed across every evaluation in its role.
- *Auxiliary scoring models for decomposition-aware baselines* (§3.1; documented in App. B.11): Skywork-o1-Open-PRM-Qwen-2.5-7B [He et al., 2024] (`Skywork/Skywork-o1-Open-PRM-Qwen-2.5-7B`), a process reward model built on Qwen2.5-Math-7B-Instruct, used inference-only on a single 24 GB GPU; and RLHFlow's Llama-3.1-8B-PRM-Mistral-Data [Xiong et al., 2024] (`RLHFlow/Llama3.1-8B-PRM-Mistral-Data`), a Llama-3.1-based process reward model trained on Mistral-generated process-supervision data, also inference-only on a 24 GB GPU. None of these models was fine-tuned for this paper.

B.2 Datasets and per-dataset context

We evaluate on five public benchmarks: HotpotQA [Yang et al., 2018], MuSiQue [Trivedi et al., 2022], StrategyQA [Geva et al., 2021], DROP [Dua et al., 2019], and GSM8K [Cobbe et al., 2021]. The evaluation uses the standard validation/dev splits in each case (524 HotpotQA, 500 MuSiQue, 625 StrategyQA, 403 DROP, 200 GSM8K-thinking-only). In each case we restrict to questions whose annotated decomposition is a length-2 chain (the 2-hop subset of MuSiQue, length-2 native decompositions for StrategyQA, and length-2 BREAK QDMR-high-level decompositions for HotpotQA and DROP), matching the depth of the operadic-consistency check.

Per-dataset context handling varies. For *MuSiQue* and *DROP*, every prompt (direct, decomposed, and self-consistency samples) is preceded by the dataset's gold passage; passages are loaded from `dgslibisey/MuSiQue` and `ucinlp/drop` on HuggingFace. For *HotpotQA*, no passage is included: the original Break QDMR-high-level setup is open-domain at evaluation time, and we follow that convention. This makes our HotpotQA evaluation *closed-book multi-hop* (no retrieval, no passage, parametric knowledge only), which is substantially harder than the passage-grounded MuSiQue / DROP setups; per-model HotpotQA accuracies in Fig. 3 fall in the 0.15–0.38 range as a result, consistent with literature reports for closed-book multi-hop QA. *StrategyQA* is open-domain by construction (no passage). *GSM8K* (used only in §3.2) prepends the problem statement.

Decomposition trees come from three sources: BREAK QDMR-high-level annotations [Wolfson et al., 2020] for HotpotQA and DROP; the dataset's native question decompositions for MuSiQue (`question_decomposition` field) and StrategyQA (`decomposition` field); and a CoT-extracted CoQ for the thinking-model experiments (described below).

B.3 Operadic-consistency protocol

For a depth-2 ToQ comprising sub-questions `Q1` (with answer $[\texttt{A1}]$) and `Q2` (which references $[\texttt{A1}]$), we evaluate operadic consistency in three calls per question, all at $T = 0$:

1. *Direct*: query the model with the unmodified original question.
2. *Step 1*: query the model with `Q1`, obtaining a candidate `a1`.
3. *Step 2*: substitute `a1` into `Q2`'s slot and query the model again, obtaining the decomposed answer.

The continuous score $\mathrm{OC} \in [0, 1]$ is the per-dataset `score_official` agreement between the direct and decomposed answers (§3.1); the binary form thresholds at 0.5. All prompts use the system

message "*You are a concise factual question-answering assistant. Answer with a short phrase or a few words. Do not explain your reasoning.*"

B.4 Self-consistency, semantic entropy, and $P(\textbf{True})$ protocols

Canonical CoT self-consistency. We follow the K-sample CoT protocol of Wang et al. [2023b]: for each (model, question) pair we sample $K = 10$ chain-of-thought responses at $T = 0.7$ using a zero-shot CoT prompt (system: "*Answer the question accurately. Think step by step, then state the final answer.*"; Kojima et al., 2022) with a per-call budget of 512 tokens. Each verbose chain is reduced to a short factoid answer via a deterministic auxiliary LLM extractor (Llama-3.3-70B-Instruct-Turbo at $T = 0$, prompted to "*Extract ONLY the direct answer to the question in 5 words or fewer*"); we use the same extractor and prompt for the thinking-model experiments of §3.2, so the SC procedure is unified across the two cohorts. Wang et al. [2023b] use majority voting over the K extracted answers as a decoding rule; as a continuous confidence score we use instead the mean pairwise per-dataset `score_official` agreement over the $\binom{K}{2} = 45$ unordered pairs (matched-axis with OC and accuracy; see App. B.14). The unanimous binary form requires all pairs to agree at score ≥ 0.5. Statistics for $K = 3$ are computed by truncating to the first three samples of the same draw. Passage handling follows the operadic protocol (passage included for MuSiQue and DROP, none for HotpotQA and StrategyQA).

Naive direct-answer SC (ablation). As an isolation of the CoT contribution, we also evaluated a direct-answer variant: same sampling temperature and K, but the system prompt instead forces a short answer with no reasoning trace ("*...Do not explain your reasoning.*") and the per-call budget is 64 tokens. The naive variant produces consistently lower cross-model r values than the canonical CoT variant on every dataset; full per-baseline cross-model table in App. B.13. We do not report it elsewhere in the main text.

Semantic entropy. We greedily cluster the $K = 10$ samples by matched-axis pairwise agreement ≥ 0.5 (the per-dataset `score_official` scorer; see App. B.14), then compute Shannon entropy over the cluster-size distribution. This is a token-F1 approximation of the bidirectional-NLI clustering used by Kuhn et al. [2023]; we adopt it for consistency with our other F1-based metrics and to avoid an extra NLI model dependency. The negated entropy $-H_{K=10}$ is used as a regression predictor (so that, like the others, larger values correspond to higher confidence).

$P(\textit{True})$. For each (model, question) pair we issue one additional call: the model is presented with the question, the proposed answer (the model's own greedy direct answer from the operadic protocol), and the prompt "*Is the proposed answer correct? Reply with only Yes or No.*" We extract the next-token logprobs (top-5 alternatives via `extra_body={top_logprobs: 5}` on the OpenAI-compatible chat-completions endpoint) and compute $P(\text{Yes})/(P(\text{Yes}) + P(\text{No}))$. For passage-bearing datasets (MuSiQue, DROP), the passage is included in the $P(\text{True})$ prompt for parity with the original answer-generation protocol.

B.5 CoQ extraction for thinking models

For thinking-model experiments (§3.2), the decomposition is not annotator-supplied but rather extracted from the model's own greedy thinking trace. We prompt the auxiliary extractor model (Qwen3-Next-80B-A3B-Instruct) with the original question, the model's full thinking trace, and a request to identify (a) whether the trace is factorable as a depth-2 chain (`Q1` $\rightarrow$ `Q2`), and (b) if so, the explicit text of `Q1`, the model's stated answer to it (`a1`), and the follow-up question template `Q2` with `[A1]` as a placeholder for `a1`. Traces flagged as not factorable are excluded from the analysis (6.1% overall, and from 0% to 32% depending on the model-dataset cell; see App. B.21). For factorable traces we then run the same three-call protocol as App. B.3, with `Q1` and `Q2` taken from the trace rather than from an annotator: we ask the model `Q1` in a fresh call, obtaining a re-derived sub-answer `A1`, then ask `Q2`$[$`[A1]` := `A1`$]$, and compare the resulting decomposed answer against the model's original direct answer under the per-dataset `score_official` scorer to compute OC. The direct call uses a thinking budget of 8192 tokens; the two sub-question calls use 4096 each. Note that `A1` is re-derived rather than read from the trace: the extracted `a1` records what the model asserted while reasoning, and we retain it for analysis but do not substitute it into `Q2`. Each factorable question thus costs three target-model thinking calls, matching App. B.3.

B.6 Statistical methodology

All per-question regressions are logistic regressions of binary correctness on the predictors listed, fit with `statsmodels` in Python. Standard errors are cluster-robust, clustered by `question_id`, to handle the non-independence of observations on the same question across different models. The pseudo-R^2 reported is McFadden's. Reported p-values are two-sided Wald tests on individual coefficients. Across-model Pearson correlations (Fig. 3) and their two-sided p-values are computed with `scipy.stats.pearsonr`.

Token-F1 normalization. We use each benchmark's standard scoring protocol, applied identically to both correctness (prediction vs. gold) and operadic consistency (direct vs. decomposed):

- *HotpotQA, MuSiQue*: SQuAD-style token-F1. Predictions and gold are normalized (lowercase, strip articles `a`/`an`/`the`, strip punctuation, collapse whitespace) and token-F1 is computed in the standard way.
- *DROP, GSM8K*: number-equality fallback plus SQuAD-style token-F1. When both the prediction and the comparison string contain a parseable integer or decimal we compare those numbers exactly (returning 1.0 on match, 0.0 otherwise); otherwise we fall back to token-F1. This approximates the DROP-official scorer for its dominant numerical-answer case (the number-equality path fires on $\sim$79% of DROP accuracy comparisons and $\sim$69% of OC comparisons; on GSM8K, virtually all answers are integer).
- *StrategyQA*: yes/no extraction plus exact match. From each side we extract the first standalone `yes` / `no` / `true` / `false` token (mapping `true`$\rightarrow$`yes` and `false`$\rightarrow$`no`); the pair is consistent / correct if both extractions are non-null and equal. Matches the standard StrategyQA boolean-evaluation protocol; doesn't punish hedged answers like "Yes, with caveats."

The continuous score is thresholded at 0.5 when a binary indicator is needed. Per-dataset rather than uniform scoring is what each benchmark's authors recommend; using the standard scorer on each dataset also avoids spurious cross-dataset coupling between correctness and OC that arises when both metrics share idiosyncratic sensitivities (e.g., both punishing answer-hedging on yes/no questions).

For tables that pool across (model, dataset) cells, we sum within-cell counts; for grand-pool regressions we cluster standard errors by `question_id` (so that the same question observed under multiple models contributes one cluster), but we do not include model-effect random intercepts. This is a deliberate simplification: a fully hierarchical specification with model-effect random intercepts would adjust standard errors further but not the point estimates and would not change qualitative conclusions, given the magnitude of the p-values reported.

B.7 Per-dataset sample sizes

The per-dataset sample sizes for the regressions in Table 1 (twelve non-thinking models pooled per dataset) are: HotpotQA $n = 6{,}287$, MuSiQue $n = 6{,}000$, StrategyQA $n = 7{,}497$, DROP $n = 4{,}828$ (total $n = 24{,}612$). These are the numbers of (model, question) pairs surviving after dropping rows with missing $K{=}10$ self-consistency samples or missing values on any other predictor used in the regression. Sample sizes for the thinking-model tables are reported inline alongside the corresponding tables in §3.2.

B.8 Illustrative decomposition examples

The following examples are sampled verbatim from the Llama-3.3-70B §3.1 runs. Each lists the original question, the supplied two-step decomposition, the model's greedy direct answer, the step-1 answer `a1`, the decomposed answer (the model's answer to `Q2` after substituting `a1` for the blank), the score OC (token-F1 between direct and decomposed answers), and the gold answer.

Example 1: MuSiQue, OC-consistent and correct.

Question	"The island where they filmed Ninja Warrior in Australia is located in what harbor?"
Q1	"Where was Ninja Warrior Australia filmed?"
Q2	"[A1] is located in what harbor?"
Direct answer	"Sydney Harbour."
Step-1 answer	"Cockatoo Island."
Decomposed answer	"Sydney Harbour."
OC	1.00
Gold	"Sydney Harbour." Correct.

The decomposition path correctly identifies the filming island and then locates it in Sydney Harbour, in exact agreement with the direct answer.

Example 2: MuSiQue, OC-inconsistent and incorrect.

Question	"How many players can play in the game based on the show that Steven the Sword Fighter is part of?"
Q1	"What show is Steven the Sword Fighter part of?"
Q2	"How many players can play in the game based on [A1]?"
Direct answer	"Four characters."
Step-1 answer	"Sixteenth episode."
Decomposed answer	"No multiplayer."
OC	0.00
Gold	"seven." Incorrect.

The model fails step 1 (it returns an episode designation when asked for a show), and the decomposed and direct paths disagree. The disagreement correctly flags this as a question on which the model is unlikely to be answering correctly: both paths in fact produce wrong answers.

Example 3: MuSiQue, OC-consistent and incorrect (a false positive).

Question	"What place gets the most rain where Sandy High School is?"
Q1	"Where is Sandy High School?"
Q2	"What place gets the most rain in [A1]?"
Direct answer	"Coastal mountains."
Step-1 answer	"Oregon."
Decomposed answer	"Coastal mountains."
OC	1.00
Gold	"the Coast Range." Incorrect.

The model identifies the state correctly and gives the same answer in both paths, but "Coastal mountains" is a generic descriptor that does not match the more specific gold answer ("the Coast Range") at the token-F1 threshold. This example illustrates the false-positive failure mode of operadic consistency: a model can be self-consistent across paths and still be wrong at the level of detail the task demands. OC provides necessary but not sufficient evidence of correctness, which is why we deploy it as one signal in a multi-signal stack rather than as a standalone gate.

Example 4: MuSiQue, OC-inconsistent and correct (a false negative).

Question	"What mythology is the city where Eutychides was born in a part of?"
Q1	"What city was Eutychides born?"
Q2	"What is [A1] a part of?"
Direct answer	"Greek mythology."
Step-1 answer	"Sicyon."
Decomposed answer	"Corinthia."
OC	0.00
Gold	"Greek mythology." Correct.

The model correctly identifies Sicyon as Eutychides's birthplace, but the open-ended Q2 admits an unintended geographic reading: Sicyon lies in the Corinthia region, which the model returns instead of the mythological category the original question demands. The direct path arrives at the right answer presumably via a higher-level pattern match (any ancient Greek-named entity sits within Greek mythology). This is the false-negative failure mode of operadic consistency: a correct direct answer flagged as inconsistent because the decomposed path is misled at a later step. OC-inconsistent-and-correct is the analogue of Example 3's false-positive mode; together they bracket the two ways the OC signal can disagree with correctness.

Thinking-model examples. For thinking models, the decomposition Q1 $\rightarrow$ Q2 is not annotator-supplied but extracted by the auxiliary model from the thinking trace itself (§B.5); the rest of the protocol is identical. We show two worked Cogito-v2.1 examples below. The full thinking trace is several hundred tokens of reasoning and is omitted for space; only the extracted Q1/a1/Q2 tuple and the two answers (direct, decomposed) are shown.

Example 5: GSM8K (Cogito-v2.1), OC-consistent and correct.

Question	"Paul went to a shop to buy some groceries. He bought some bread for $2, butter for $3, and juice for two times the price of the bread. He had $15 for his shopping. How much money did Paul have left?"
Q1 (extracted)	"How much did Paul spend on groceries in total?"
Q2 (extracted)	"How much money did Paul have left after spending [A1] out of $15?"
Direct answer	"$6."
Step-1 answer	"$9."
Decomposed answer	"$6."
OC	1.00
Gold	"6." Correct.

The model's CoT trace surfaces a clean two-step decomposition (total spending, then subtract from budget); the extractor lifts both sub-questions. Direct and decomposed paths converge on $6, in agreement with gold.

Example 6: MuSiQue (Cogito-v2.1), OC-inconsistent and incorrect.

Question	"Who is the female star in Gone with the Wind married to?"
Q1 (extracted)	"Who is the female star in Gone with the Wind?"
Q2 (extracted)	"Who is [A1] married to?"
Direct answer	"Clark Gable."
Step-1 answer	"Vivien Leigh."
Decomposed answer	"Herbert Holman Laurence Olivier."
OC	0.00
Gold	"Laurence Olivier." Incorrect (direct path).

The model's CoT identifies Vivien Leigh correctly as the female star, then on Q2 lists both of her husbands (Herbert Leigh Holman, her first; Laurence Olivier, her second), concatenated by the short-answer extractor into a single string. The decomposed answer contains the gold answer. The

direct path returns Clark Gable, her famous costar, which is the natural high-prior wrong answer. OC correctly flags this as low-confidence: a selective-prediction rule on OC would defer to the decomposed path or abstain. This pattern — direct path drawn to a high-prior co-occurrence while the decomposed path navigates the actual entity relation — is a common thinking-model OC-inconsistent failure mode in our data.

B.9 Dataset-level accuracy prediction (LOO demonstration)

§3.1 notes that the strong cross-model correlation between OC-rate and accuracy yields a label-free dataset-level accuracy estimator: fit accuracy $\approx a \cdot$ OC-rate $+ b$ on calibration models, predict from OC-rate alone for new models. We demonstrate this concretely with leave-one-model-out (LOO) cross-validation on the twelve non-thinking models: for each (model, dataset) cell, we fit the linear regression on the remaining eleven models and predict the held-out model's accuracy from its measured OC-rate.

dataset	in-sample R^2	LOO MAE	LOO max abs error
HotpotQA	0.86	0.028	0.062
MuSiQue	0.88	0.023	0.069
StrategyQA	0.88	0.021	0.052
DROP	0.74	0.047	0.131
pooled (48 cells)		0.030	0.131

Table 3: Leave-one-model-out demonstration of dataset-level accuracy prediction from OC-rate. For each dataset, we fit a linear regression accuracy $\approx a \cdot$ OC-rate $+ b$ on eleven calibration models and predict the held-out model's accuracy from its OC-rate alone. Across the 48 (model, dataset) held-out cells, mean absolute prediction error is 3.0 percentage points. DROP is the weakest dataset; the other three datasets achieve LOO MAE between 2.1 and 2.8 percentage points.

This is a small but functioning prototype: with eleven calibration models per dataset, a single OC-rate measurement on a new model predicts that model's accuracy on the dataset to within roughly 3.0 percentage points on average, and within 2.1–2.8 percentage points on the three datasets where the across-model correlation is strongest (HotpotQA, MuSiQue, StrategyQA). The protocol is fully label-free at prediction time: only the calibration step uses ground-truth accuracy, and after that, predicting accuracy for a new model requires only the model's own outputs on the dataset's questions (no labels).

B.10 $P(\text{True})$ four-way sensitivity (nine non-thinking models)

The Kadavath $P(\text{True})$ baseline requires top-k logprobs to compute $P(\text{Yes})/(P(\text{Yes}) + P(\text{No}))$ from a single self-evaluation call (§B.4). Of our twelve non-thinking models, top-k logprobs are exposed by the inference API for nine: the eight Together-served models, plus GPT-4o-mini via OpenRouter (which passes through the OpenAI top-k logprob field). For Gemini 2.5 Flash, Llama-4 Maverick, and Mistral Large 2, only the top-1 token logprob is returned in our protocol, so we cannot compute the same Kadavath quantity for those cells. Restricting to the nine cells where $P(\text{True})$ is available, the four-way logistic regression

$$\text{correct} \sim \text{OC} + \text{CoT-SC}_{K=10} + (-H_{K=10}) + P(\text{True})$$

gives the per-dataset coefficients in Table 4 (cluster-robust SE clustered by question). The OC coefficient remains highly significant ($p < 10^{-15}$) on every dataset after adding $P(\text{True})$; the OC story does not depend on whether we control for self-evaluation.

B.11 Decomposition-aware baseline protocols and sensitivity

§3.1 introduces two decomposition-aware baselines we constructed for this paper: decomposed self-consistency and a Skywork process reward model. Neither is published as a stand-alone baseline; each combines standard machinery with our specific decomposition protocol. We document the protocols below.

dataset	n	β_{OC} (p)	$\beta_{P(\text{T})}$ (p)
MuSiQue	4,462	$+2.23$ (10^{-41})	-0.46 (10^{-4})
StrategyQA	5,600	$+1.61$ (10^{-33})	$+0.54$ (10^{-8})
HotpotQA	4,696	$+1.73$ (10^{-29})	$+0.29$ $(.021)$
DROP	3,584	$+1.22$ (10^{-16})	$+0.53$ (10^{-5})

Table 4: Four-way logistic regression with $P(\text{True})$ added (nine non-thinking models with top-k logprobs, our normalization protocol, cluster-robust SE clustered by question). CoT-SC$_{K=10}$ and $-H_{K=10}$ coefficients are omitted for space; they are essentially unchanged from the three-way regression of Table 1. OC remains statistically significant on every dataset ($p < 10^{-15}$).

Decomposed self-consistency. For each (model, question) pair we hold `a1` at greedy (the value already computed in the operadic protocol of §3.1) and sample the model on `Q2`$[$`A1`$] :=$ `a1`$]$ ten times at temperature 0.7. The signal is the mean pairwise per-dataset `score_official` agreement across the ten samples, parallel to the SC pairwise comparison. Cost: ten model calls per (model, question) pair on top of the operadic protocol. Holding `a1` fixed eliminates first-step generative variance from this baseline, restricting the measured diversity to step 2 under a fixed first-step input. A fully-sampled variant (independent K samples of $(\texttt{a1}, \texttt{a2})$ pairs) would be a strictly more demanding test; we use the fixed-`a1` variant for cost reasons. Decomposed-SC has cross-model Pearson $r{=}{+}0.67$ ($p{=}0.018$) on HotpotQA (the strongest cross-model signal among the constructed decomposition-aware baselines; see App. B.13), but is not cross-model significant on the other three datasets. In the per-question logistic regression of Table 5, its coefficient is non-significant or negative on every dataset; this conclusion is robust to the fixed-`a1` choice in the conservative direction (a fully-sampled variant would have more variance and a weaker, more negative or further-from-zero, coefficient).

Skywork PRM. For each (model, question) pair we format the operadic trace as `Step 1: Q1 / Answer: a1 / Step 2: Q2[A1] / Answer: a2` and submit to the Skywork-o1-Open-PRM-Qwen-2.5-7B model [He et al., 2024] via its published HuggingFace inference protocol. The PRM emits a scalar reward at each step boundary; we aggregate to a per-row signal by min, mean, or last-step. The PRM is built on Qwen2.5-Math-7B-Instruct and trained on Skywork-o1 step preferences whose distribution is dominated by mathematical reasoning, so its signal on multi-hop factual QA is weaker than its in-distribution math performance and is most pronounced on MuSiQue / DROP (numerical answers) than on HotpotQA (entity lookup).

RLHFlow PRM (robustness check). We additionally evaluated RLHFlow's Llama-3.1-8B-PRM-Mistral-Data [Xiong et al., 2024] on the same operadic trace as Skywork-PRM. The RLHFlow PRM shows essentially no signal on multi-hop QA in our protocol: its standalone univariate pseudo-R^2 is ≤ 0.001 on every dataset (point-biserial r between min-step PRM score and correctness is -0.057 pooled, vs. $+0.182$ for Skywork-PRM), and its coefficient is non-significant when added as a fourth predictor (HotpotQA -0.06 n.s.; MuSiQue $+0.27$ n.s.; StrategyQA $+0.23$ n.s.; DROP $+0.47$ n.s.; OC unchanged). The contrast with Skywork-PRM is stark and likely reflects the difference in PRM training data (Skywork is trained on Skywork-o1 step preferences spanning math and code reasoning; RLHFlow's PRM is trained on Mistral-generated process-supervision data without an explicit multi-step reasoning curriculum) rather than a deficiency of the off-label evaluation: this is one data point of the broader observation that off-the-shelf PRM signals on out-of-distribution QA are highly PRM-specific.

Per-dataset sample sizes for the joint regressions. The four-way regressions of Table 5 have slightly different per-dataset n for each baseline column due to occasional missing baseline scores: decomposed-SC has $n = 6{,}287/6{,}000/7{,}497/4{,}828$ (HotpotQA / MuSiQue / StrategyQA / DROP), Skywork-PRM has $n = 6{,}260/5{,}985/7{,}463/4{,}794$, and RLHFlow-PRM has $n = 6{,}260/5{,}985/7{,}463/4{,}794$.

Standalone univariate sensitivity. The univariate pseudo-R^2 values for each signal alone:

		+ decomposed-SC		+ Skywork PRM (min)	
dataset	n	β_{OC}	β_{dSC}	β_{OC}	β_{PRM}
HotpotQA	∼6,280	$+2.01^{\diamond}$	-0.49^{*}	$+1.90^{\diamond}$	$+0.69$
MuSiQue	∼6,000	$+2.28^{\diamond}$	-0.26	$+2.20^{\diamond}$	$+1.35$
StrategyQA	∼7,500	$+1.87^{\diamond}$	-0.70^{*}	$+1.68^{\diamond}$	$+2.46^{*}$
DROP	∼4,830	$+1.23^{\diamond}$	-0.24	$+1.11^{\diamond}$	$+2.70^{*}$

Table 5: Per-question logistic regression of correctness on $(\text{OC}, \text{CoT-SC}_{K=10}, -H_{K=10}, X)$ where X is each new decomposition-aware baseline in turn, with cluster-robust standard errors clustered by question (twelve non-thinking models). $\diamond$ marks $p < 10^{-12}$; $*$ marks $p < 0.05$. Coefficients on $\text{CoT-SC}_{K=10}$ and $-H_{K=10}$ are omitted for space (essentially unchanged from the 3-way regression of Table 1). The OC coefficient remains massively significant against every baseline; the decomposed-SC coefficient is non-significant or significantly negative on every dataset (and now significantly negative on *two* datasets, HotpotQA and StrategyQA), ruling out the "OC is just decomposed-path noise" hypothesis.

predictor	HotpotQA	MuSiQue	StrategyQA	DROP
operadic consistency	**0.179**	**0.199**	0.180	0.089
CoT-SC K=10	0.143	0.099	0.088	0.084
naive direct-answer SC K=10 (ablation)	0.141	0.093	0.183	**0.152**
$-H$ at K=10	0.140	0.088	**0.191**	**0.152**
P(True) (9 models)	0.041	0.001	0.015	0.044
decomposed-SC K=10	0.061	0.042	0.010	0.022
Skywork-PRM (min over steps)	0.012	0.033	0.015	0.039
RLHFlow-PRM (min over steps)	0.001	0.001	0.001	0.002

OC is the strongest standalone signal on HotpotQA and MuSiQue; on StrategyQA, $-H_{K=10}$ slightly edges OC (0.191 vs 0.180); on DROP, both $-H_{K=10}$ and the naive direct-answer SC ablation (0.152 each) are stronger standalone than OC (0.089). Notably, canonical CoT-SC is *weaker* standalone than the naive direct-answer SC ablation on StrategyQA and DROP, but the two SC variants reverse roles in the cross-model picture (Table 7): canonical CoT-SC is stronger cross-model on HotpotQA and DROP, where it carries reasoning-chain content the naive variant lacks. In every dataset, OC contributes statistically significant marginal information in the joint regression of Table 5 ($p < 10^{-12}$), even on the datasets where another signal leads standalone.

B.12 Selective prediction: AUARC and AUROC with bootstrap CIs

This appendix gives full bootstrap 95% confidence intervals for the selective-prediction analyses summarized in Tables 1 and 2.

Protocol. For each $(\text{model}, \text{dataset})$ cell we compute two metrics over a confidence signal:

- *AUARC* [Geifman and El-Yaniv, 2017]: area under the accuracy-coverage curve, where coverage runs over $1, 2, \ldots, n$ questions taken in descending score order, and accuracy at coverage k is the mean correctness of those top-k questions. AUARC equals mean accuracy for a random ranker.
- *AUROC*: probability that a uniformly random correct example is ranked above a uniformly random incorrect example, computed via the Mann-Whitney U statistic. Independent of base accuracy.

For combined predictors (e.g., OC+SC), we use *leave-one-model-out* (LOMO) calibration: holding out one model's cell, fit a logistic regression on rows pooled from the other 11 (non-thinking) or 4 (thinking) models with logit-transformed inputs, predict $P(\text{correct})$ on the held-out cell, and report that cell's AUARC and AUROC. Repeat for each held-out model. Δ rows aggregate as paired differences across models with bootstrap 95% CIs from 2000 resamples.

Non-thinking, full intervals. (Twelve paired differences; Δ_{AUROC} = paired mean across models of [AUROC of OC+X combined] minus [AUROC of X alone]; AUARC analogous.)

	vs. CoT-SC$_{K=3}$ (3-call)		vs. CoT-SC$_{K=10}$ (10-call)	
dataset	Δ_{AUARC} [95% CI]	Δ_{AUROC} [95% CI]	Δ_{AUARC} [95% CI]	Δ_{AUROC} [95% CI]
HotpotQA	$+0.095\,[+.074, +.116]$	$+0.092\,[+.067, +.117]$	$+0.068\,[+.047, +.088]$	$+0.061\,[+.039, +.082]$
MuSiQue	$+0.086\,[+.074, +.100]$	$+0.153\,[+.138, +.172]$	$+0.063\,[+.051, +.076]$	$+0.108\,[+.092, +.127]$
StrategyQA	$+0.096\,[+.083, +.111]$	$+0.164\,[+.147, +.181]$	$+0.074\,[+.060, +.090]$	$+0.122\,[+.105, +.139]$
DROP	$+0.086\,[+.069, +.103]$	$+0.117\,[+.095, +.139]$	$+0.074\,[+.062, +.087]$	$+0.090\,[+.066, +.113]$

Table 6: Non-thinking selective-prediction lifts with 95% bootstrap CIs ($n_{\text{boot}} = 2000$).

Thinking, full intervals. The five thinking-model paired diffs are reported with their bootstrap 95% CIs in Table 2 of §3.2; we record the full intervals here for completeness, including the $K=10$ budget that is omitted from the body table. At $K=3$ (equal-cost): MuSiQue $\Delta_{\text{AUARC}}+0.033\,[+.005, +.056]$, $\Delta_{\text{AUROC}}+0.052\,[+.022, +.085]$; StrategyQA $+0.042\,[+.032, +.049]$ / $+0.050\,[+.028, +.072]$; GSM8K $+0.029\,[+.010, +.051]$ / $+0.134\,[+.071, +.180]$; DROP $+0.017\,[+.005, +.030]$ / $+0.071\,[+.023, +.129]$. At $K=10$: MuSiQue $+0.025\,[+.010, +.047]$ / $+0.023\,[+.019, +.028]$; StrategyQA $+0.013\,[-.002, +.028]^{\dagger}$ / $+0.012\,[-.008, +.032]^{\dagger}$; GSM8K $+0.005\,[+.000, +.010]$ / $+0.054\,[+.005, +.108]$; DROP $+0.015\,[-.000, +.032]^{\dagger}$ / $+0.035\,[-.006, +.078]^{\dagger}$. † marks $K=10$ cells whose CI brackets zero on the lower end: StrategyQA on both metrics, and DROP on both metrics. All eight $K=3$ cells and the four MuSiQue / GSM8K-AUROC $K=10$ cells exclude zero cleanly; the GSM8K-AUARC $K=10$ interval $[+.000, +.010]$ touches zero by rounding but is strictly positive.

Reproduction. The full per-cell AUARC and AUROC matrix, including univariate values for every signal in the candidate set, is regenerated by `python -m analysis.selective_prediction` (output: `results/selective_prediction.txt`). Golden tests in `analysis/tests/test_golden.py` pin the headline Δ values reported above.

B.13 Cross-model Pearson r per baseline

§3.1 reports that operadic consistency correlates strongly with accuracy across the twelve non-thinking models on every dataset, while sample-based and self-evaluation baselines do not. For completeness, we report the same per-dataset cross-model Pearson r for every baseline used elsewhere in the paper. Self-consistency uses matched-axis pairwise scoring (the per-dataset accuracy scorer; see §B.14). For each (model, dataset) cell we compute the per-cell rate of the baseline (mean over the cell's questions; for OC binarized at 0.5, the fraction with OC ≥ 0.5); we then take the Pearson correlation between this per-cell rate and the cell's per-question accuracy across the twelve models per dataset.

B.14 DROP sensitivity: scorer choices and an LLM-judge corroboration

The framework underlying operadic consistency makes a non-trivial prediction about scorer design: an OC scorer should realize the natural equivalence on the answer space appropriate to the question's expected output type. For DROP, where most answers are numerical values, the natural equivalence is value-equality: "Two months," "2," "two," and "2 months" all express the same number and should be treated as equivalent. Our DROP scorer (App. B.6) is the number-equality fallback plus SQuAD-style token-F1 with word-to-number conversion, applied identically to correctness and OC; this approximates value-equality on DROP's dominant numerical-answer case (the number-equality path fires on $\sim$79% of DROP accuracy comparisons and $\sim$69% of OC comparisons). It is materially more lenient than the canonical Dua et al. [2019] surface-form scorer, which distinguishes "five" from "5" after light normalization. Under the framework, this is a feature, not a bug: a scorer realizing a too-fine equivalence on the answer space partly measures formatting compliance rather than semantic content.

Matched-axis baselines. The same scorer-design argument applies to every baseline whose construction involves equating two model outputs: if the natural equivalence on the answer space is $\sim$, then "two SC samples agree" (for SC), "two samples cluster" (for semantic entropy $-H$), and "two decomposed-path samples agree" (for decomposed-SC) should all be measured under $\sim$, not under a finer surface-form equivalence. We therefore use the per-dataset `score_official` scorer

signal	HotpotQA	MuSiQue	StrategyQA	DROP
OC (binarized at 0.5, paper's primary)	$\mathbf{+0.92}^{\diamond}$	$\mathbf{+0.94}^{\diamond}$	***+0.94***$^{\diamond}$	$\mathbf{+0.86}^{\diamond}$
OC (continuous)	$\mathbf{+0.92}^{\diamond}$	***+0.96***$^{\diamond}$	***+0.94***$^{\diamond}$	$\mathbf{+0.85}^{\diamond}$
CoT-SC$_{K=3}$ (continuous)	***+0.94***$^{\diamond}$	$+0.45$	$+0.45$	***+0.89***$^{\diamond}$
CoT-SC$_{K=10}$ (continuous)	$\mathbf{+0.93}^{\diamond}$	$+0.43$	$+0.46$	$\mathbf{+0.87}^{\diamond}$
$-H_{K=10}$ (continuous)	$+0.52$	$+0.44$	$+0.80^{*}$	$+0.61^{*}$
P(True) (continuous, 9 models)	$+0.02$	-0.42	-0.04	$+0.21$
naive direct-answer SC$_{K=10}$ (ablation)	$+0.53$	$+0.45$	$+0.70^{*}$	$+0.61^{*}$
naive direct-answer SC$_{K=3}$ (ablation)	$+0.50$	$+0.45$	$+0.67^{*}$	$+0.62^{*}$
decomposed-SC$_{K=10}$	$+0.67^{*}$	$+0.23$	$+0.47$	$+0.21$
Skywork-PRM (min over steps)	-0.02	$+0.38$	$+0.02$	$+0.82^{*}$
RLHFlow-PRM (min over steps)	$+0.48$	$+0.18$	-0.28	-0.31

Table 7: Cross-model Pearson r between per-cell baseline rate and per-cell accuracy, across the twelve non-thinking models per dataset. $\diamond$ marks $p < 10^{-3}$; $*$ marks $p < 0.05$ (others non-significant at $\alpha = 0.05$). P(True) uses nine models; the other rows use twelve. Cells with $r \geq 0.85$ are bolded; per-column maxima are bold-italic. OC is the only baseline with all four cells bolded (uniform strength): CoT-SC takes HotpotQA and DROP point-wise but collapses on MuSiQue and StrategyQA, while OC takes MuSiQue outright (continuous, $+0.96$) and the two OC variants tie for the column maximum on StrategyQA at $+0.94$. Canonical CoT-SC [Wang et al., 2023b] is *bimodal*: it matches OC on HotpotQA and DROP (r=$+0.93, +0.87$) but collapses to $r \approx +0.45$ on MuSiQue and StrategyQA, where its variance fails to track the cross-model accuracy ordering. Other sample-based baselines ($-H$, naive direct-answer SC) reach significance on StrategyQA and DROP but never approach OC's strength on more than one dataset; P(True) fails on every dataset, consistent with its measuring a within-model quantity largely orthogonal to cross-model competence. The naive direct-answer SC ablation is the same K-sample protocol but with a prompt that forces a short answer with no reasoning trace; it lies uniformly below canonical CoT-SC on HotpotQA and DROP, confirming that CoT at sample time is what closes the gap to OC on those datasets. Decomposition-aware baselines (decomposed-SC, Skywork-PRM) show isolated significant correlations on one dataset each: decomposed-SC on HotpotQA (likely a closed-book artifact, where parametric-knowledge variance is the dominant source of temperature-sample variance) and Skywork-PRM on DROP (consistent with the PRM's math-and-code training distribution covering DROP's numerical-answer regime). Neither generalizes to the other three datasets.

(yes/no extraction for StrategyQA, value-equality for DROP, SQuAD-F1 elsewhere) for all such pairwise comparisons throughout the paper, parallel to its use for accuracy and OC. Sensitivity vs. legacy token-F1: HotpotQA and MuSiQue change negligibly (the matched and legacy scorers nearly coincide), but the matched scorer makes a real difference on StrategyQA (SC cross-model r moves from -0.28 to $+0.70$ and $-H$ from -0.30 to $+0.80$, because token-F1 was systematically penalizing more verbose models on yes/no answers) and on DROP (SC $+0.52 \rightarrow +0.62$, because token-F1 distinguished value-equivalent paraphrases like "5" vs. "five"). On no dataset does any matched-axis baseline come within 0.15 of OC's cross-model r.

Number-equality contribution. We compare DROP's across-model Pearson r between OC-rate and accuracy under two scoring choices:

- Number-equality fallback plus SQuAD F1 (paper protocol): $r = +0.86$ ($p = 0.0004$).
- SQuAD F1 alone (no number-equality): $r = +0.67$.

The number-equality fallback restores the OC–accuracy alignment that verbose-correct answers ("Two 1-yard touchdowns" vs. gold "2") and partial-overlap false-positives ("8-yard TD pass" vs. "3-yard TD pass") previously obscured.

Canonical scorer. We rescore both correctness and OC under the canonical Dua et al. DROP scorer (surface-form normalization without word-to-number conversion or value extraction) and re-compute the cross-model Pearson r between OC-rate and accuracy on the same twelve non-thinking models. The result is $r = +0.37$ ($p = 0.24$, n.s.) on DROP, compared with $r = +0.86$ ($p = 0.0004$) under

the paper's value-equality scorer. The collapse is consistent with the framework prediction: the canonical scorer's variance is dominated by a surface-form-compliance dimension that is largely decoupled from reasoning ability for general-purpose LLMs answering in their natural style.

LLM-judge correctness scorer. As an independent sensitivity check, we replace both correctness and OC by an LLM-judge variant: Claude Haiku 4.5 judges, at $T{=}0$, with a paraphrase-tolerant prompt that accepts verbose elaborations expressing the same value or yes/no commitment. The judge serves as a matched-axis equivalence here as well, parallel to the deterministic-scorer choice above: the same judge, prompt, and decision rule are applied to both the direct-vs-gold (accuracy) and direct-vs-decomposed (OC) comparisons, so neither side gets a more lenient or stricter standard. The resulting cross-model r between judge OC-rate and judge accuracy (12 models per dataset) is HotpotQA $+0.97$, MuSiQue $+0.88$, StrategyQA $+0.69$, DROP $+0.72$, all $p \leq 0.013$. The cross-model relationship reported in §3.1 is preserved on every dataset.

Annotator unanimity. A second source of slack on DROP is annotator disagreement: 139 of 403 DROP dev questions (34%) carry alias lists that disagree with the gold answer, so a model's direct answer can match an alias and count as correct even when the decomposed answer takes a different surface form. Restricting to the subset of DROP dev questions where annotator aliases are unanimous with the gold gives a slightly tighter r, but per-model accuracy and OC-rate change by less than 0.03 on this subset and the model ordering is unchanged.

B.15 Compute and cost

The experiments are entirely inference-time API calls; no model training was performed. Approximate counts: the across-model study (§3.1) on 12 models $\times$ 4 datasets $\times$ $\sim$500 questions involves $\sim$24 direct-mode calls per question (3 operadic + 10 naive direct-answer SC ablation + 1 P(True) + 10 decomposed-self-consistency) $\approx$ 580,000 direct-mode calls, plus $\sim$20 chain-of-thought calls per question for canonical CoT-SC (10 CoT samples plus 10 cheap LLM-extractor calls at $\sim$512 + 32 tokens each) $\approx$ 480,000 additional calls, plus inference-only PRM scoring on the same rows. The thinking-model study (§3.2) on 5 models $\times$ 4 datasets $\times$ 100 questions $\times$ 13–23 thinking calls per question is $\approx$ 25,000–50,000 thinking-mode calls. Total token volume is on the order of a few hundred million prompt+completion tokens. LLM generation (operadic-consistency, CoT-SC, P(True), decomposed-SC, thinking traces) is mediated entirely by serverless inference platforms (Together AI and OpenRouter); no local GPU is required for those experiments. The three scorer baselines that load HuggingFace models locally — DeBERTa-v3-large for NLI bidirectional entailment, and the three PRMs (Skywork, RLHFlow, and VersaPRM) — run inference-only on a single 24 GB GPU (we used an RTX 4090). No model training was performed.

Wall-clock latency. Call counts are the primary cost metric above, but they understate one structural asymmetry: OC's calls are sequential, since `a1` must return before `Q2` can be issued, whereas sample-based baselines issue their K calls in parallel. We instrumented end-to-end latency on MuSiQue 2-hop (40 questions per model, served via OpenRouter, questions timed serially to avoid self-contention). For non-thinking models, OC's median per-question latency is 2.84s (GPT-4o-mini) and 2.00s (Llama-3.3-70B). Against canonical CoT-SC, our primary baseline, OC is faster at every budget except GPT-4o-mini at $K{=}3$ (2.84s against 3.35s, overlapping interquartile ranges, better described as parity); at $K{=}10$ the margin is $1.6\times$ to $4.3\times$. The gap is driven by CoT-SC's answer-extraction step, a fixed 2.4–2.7s on both models. Against the naive direct-answer SC ablation the ordering reverses, with OC 1.9–$2.6\times$ slower at equal or lower call count; that gap is intrinsic to sequential composition rather than an artifact of our implementation. For thinking models the comparison inverts again: on GLM-5.1 and Kimi-K2.6 (15 questions each) OC's three sequential thinking calls take a median 33.0s and 27.0s against 14.2s and 5.8s for CoT-SC$_{K=3}$, so OC is 1.8–$4.7\times$ slower, because once per-call latency is large, parallelism across K thinking calls outweighs the fixed extraction overhead. The cost there is concentrated in the substitution call `Q2`, at $1.18\times$ (GLM-5.1) and $2.55\times$ (Kimi-K2.6) the latency of the direct call, while `Q1` is cheaper than direct.

B.16 Deeper and branching trees: protocol and per-topology results

MuSiQue is the one dataset in our suite whose native annotations supply decompositions deeper than two hops, so the depth ladder of §3.3 runs on it alone. We evaluate twelve models at 150

questions per depth per model ($n = 1800$ at each of depth 3 and depth 4). Three of the paper's twelve endpoints had been retired by the time of these runs; we substituted scale-matched models (Llama-3.1-8B for Llama-3-8B, Mistral Large 2407 for Mistral Large 2, and Mistral Small 24B for LiquidAI-LFM2-24B) so that the cohort remains twelve. The protocol is otherwise that of App. B.3, extended to $k + 1$ sequential calls for a depth-k chain.

MuSiQue's deeper questions come in distinct topologies, and pooling them hides the structure. Per-question AUROC by topology is 0.743 for 3hop1 (linear), 0.811 for 3hop2 (branching, merging at the root), 0.736 for 4hop1 (linear), 0.488 for 4hop2 (merging mid-chain), and 0.580 for 4hop3 (merging at the root). The headline "branching 4-hop 0.541" pools 4hop2 and 4hop3, which are two different shapes.

We take the 4hop2 cell to be underpowered and mis-scored rather than a compositional failure, on three measured grounds. First, its effective sample is far smaller than it appears: MuSiQue reuses sub-chains across questions, and 4hop2's 28 questions rest on only 9 distinct opening sub-questions, giving an effective n nearer 9–17 than 28. Second, 32–33% of its gold answers exceed four words, against 3–6% in the clean cells, so token-F1 mechanically marks correct responses wrong. Third, MuSiQue's own step-3 gold is frequently incorrect in this cell, with sub-answer accuracy against it at 0.024. We note for the record that an earlier diagnosis of ours attributed the cell to malformed questions produced by substituting long intermediate answers; on inspection the substitution is correct, and bad intermediate gold is the operative cause.

Rescoring every cell uniformly with the LLM judge of App. B.14 (Claude Haiku 4.5 at $T = 0$) raises 4hop2 from 0.488 to 0.602 and pooled branching-4 from 0.541 $[0.454, 0.630]$ to 0.581 $[0.516, 0.641]$, so the interval no longer crosses chance. Label churn under the judge is 23.2% in that cell against 8.5–10.3% in the clean cells, which independently corroborates the verbose-gold diagnosis.

B.17 Task breadth beyond multi-hop question answering

We ran six Big-Bench Hard tasks [Suzgun et al., 2023] spanning sequential state-tracking, chained truth-value logic, and nested expression evaluation, on the twelve non-thinking models at 100 items each ($n = 7{,}200$). Here the decomposition is the task's own structure: we split the sequence or the expression at its midpoint and take the intermediate state or value as `a1`, so no external annotation is needed. OC-consistent answers are more accurate on all six tasks, with gaps from $+0.123$ to $+0.353$ and per-question AUROC from 0.546 to 0.740 (median 0.630). We report per-task rather than pooled figures because base accuracy ranges from 0.05 to 0.82 across the six, so a pooled gap would mostly track task difficulty.

Two math pilots on the same non-thinking cohort agree in sign but expose a failure mode: the accuracy gap is $+0.199$ on SVAMP [Patel et al., 2021] and $+0.332$ on GSM8K, but both rest on small OC-consistent cells, because OC over-flags on arithmetic — SVAMP is marked inconsistent on 87% of items. We would not deploy OC on math without a value-equality scorer.

B.18 The 240-case annotation study

One author hand-labelled 240 OC-inconsistent cases, 115 drawn from the non-thinking cohort with gold decompositions and 125 from the thinking cohort with trace-extracted ones, into three categories: (a) a model reasoning failure at an identified hop, (b) decomposition-induced, where the decomposition itself is malformed or mis-targeted, and (c) a benign scorer artifact, where the two answers agree in substance but not under the per-dataset scorer. The counterfactual reported in §3.3 recomputes the accuracy gap with every (b) case removed, which is a test of whether decomposition noise manufactures the signal.

Within the thinking half the (b) rate varies sharply by dataset: MuSiQue 18% ($n = 71$), DROP 18% ($n = 11$), StrategyQA 42% ($n = 36$). The driver is not whether the task is multi-hop, since DROP and StrategyQA are both multi-step, but the *answer type* of the intermediate. Among (b) cases the median first sub-answer runs 2 words on MuSiQue and DROP and 8 words on StrategyQA, where 53% exceed six words. StrategyQA sub-answers are whole propositions, and substituting one into the slot of the next sub-question yields a malformed question; entity- and value-valued intermediates substitute cleanly, which is why MuSiQue and DROP sit near the 13% rate observed

for human-written decompositions. This is a type mismatch, as anticipated in the colored-operad remark of App. A.

B.19 Head-to-head against an in-domain process reward model

The Skywork and RLHFlow PRMs of §3.1 are math-trained and applied off-label. VersaPRM [Zeng et al., 2025] is trained for non-math multi-step reasoning, so we run it in-label. A PRM scores steps of a reasoning chain, which the two-step decomposition of the non-thinking setting does not really provide; the thinking cohort does, so we run the comparison there, on the three thinking models with full cached traces ($n = 1360$).

OC and the PRM are complementary, not redundant. Per-question AUROC is 0.758 for OC and 0.685 for VersaPRM, and a combiner over both reaches 0.807 under 5-fold cross-validation, above either alone. In a joint regression β_{OC} remains significant controlling for the PRM score ($p \approx 10^{-41}$). OC is the stronger of the two on three of four datasets (MuSiQue 0.734 vs 0.555, StrategyQA 0.700 vs 0.606, DROP 0.640 vs 0.611), and VersaPRM leads on GSM8K, where both operate near ceiling.

DPRM [Wang et al., 2026], the other in-domain PRM we cite, does not apply to our setting. It is a knowledge-graph QA method whose KG-PRM scores paths of entities and relations and is evaluated on WebQSP and CWQ over Freebase; our benchmarks carry no knowledge graph, so half the model would have no input.

B.20 Cross-model correlations: intervals, scale controls, and scorer choice

Bootstrap intervals. Resampling models, the cross-model Pearson r between accuracy and OC-rate under the paper's scorer is $+0.92$ $[+0.83, +0.99]$ on HotpotQA, $+0.94$ $[+0.82, +0.99]$ on MuSiQue, $+0.94$ $[+0.64, +0.99]$ on StrategyQA, and $+0.86$ $[+0.57, +0.97]$ on DROP. Twelve points is few, and the StrategyQA and DROP intervals are correspondingly wide.

Is OC a proxy for model scale? No. On the ten models with published parameter counts, $\log(\text{parameters})$ explains a mean per-dataset R^2 of 0.389 of cross-model accuracy variance against 0.781 for OC-rate. The partial correlation between accuracy and OC-rate controlling for $\log(\text{parameters})$ remains $+0.81$ (HotpotQA), $+0.97$ (MuSiQue), $+0.88$ (StrategyQA), and $+0.73$ (DROP). Splitting those ten models into capability tiers ($<$10B, 10–130B, $>$130B) leaves $r(\text{accuracy}, \text{OC-rate})$ positive in all twelve tier-by-dataset cells ($+0.12$ to $+1.00$), each resting on three or four models. A leave-one-model-out accuracy predictor built on $\log(\text{parameters})$ is also worse than one built on OC-rate wherever both exist: MAE 5.8 against 1.5 pp on MuSiQue and 9.7 against 3.0 pp on StrategyQA, with no usable parameter predictor on the other two datasets.

Per-question significance under the canonical DROP scorer. §3.1 notes that our DROP scorer is more lenient than canonical surface-form scoring. Re-running the Table 1 three-way regression on DROP with correctness, OC, and the baselines all scored canonically gives $\beta_{\text{OC}} = +1.06$, $p < 10^{-18}$, against $+1.26$ on the same rows under the paper's scorer; under strict exact match, $\beta_{\text{OC}} = +1.53$, $p < 10^{-36}$. The scorer sensitivity reported in App. B.14 is therefore confined to the cross-model aggregate, where variance is dominated by formatting compliance; per-question, OC stays significant under all three scorers. This regression predates the move to canonical CoT-SC and uses the naive direct-answer SC ablation as its sample-based control.

B.21 What the CoQ extractor discards

The factorability filter applies only to the thinking cohort. On the cohort of §3.2 (five models, four datasets, 100 questions per cell) it removes 122 of 2,000 traces, 6.1%, ranging from 0% to 32% across the twenty model-dataset cells, with Cogito-v2.1 on DROP the outlier. The audit reported here was conducted during review on the then-current three-model cohort, where the filter removed 60 of 1,440 traces (4.2%); we hand-audited all 60. Under the standard the analysis requires, that a trace resolve an intermediate answer OC can check, 39 (65%) were rightly dropped: refusals or non-answers (6), no committed bridge answer (10), single-hop answers (16), and parallel aggregation (7), none of which admits an OC measurement. The remaining 21 were genuine two-hop chains lost to surface parsing.

The discards are not significantly harder than what is retained: direct-answer accuracy is 0.650 on discards against 0.754 on retained traces ($z = +1.83$, $p = 0.068$), and that gap is carried entirely by the six refusals and ten unresolved rambles (accuracy 0.06). The 21 factorable discards, the only ones that could be hard cases we lost, match the retained set at 0.762. Imputing all 60 adversarially floors the OC accuracy gap on that cohort at $+0.368$ against $+0.395$ measured, so 93% of the effect survives a worst case. The filter is also not one extractor's idiosyncrasy: two independent extractors agree on factorability for 94.8% of 615 traces, disagreeing symmetrically (17 against 15).

B.22 Where the decomposition comes from, and what happens when it degrades

The source battery and corruption ladder below were carried out during the review period on MuSiQue subsets of 99 and 100 questions respectively, executed by twelve models: the nine models of §3.1 still served at the time, plus three scale-matched substitutes for retired endpoints, on the same principle as App. B.16. The figures below are therefore measured on that subset rather than recomputed on the full cohort of §3.1.

Source battery. Error-detection AUROC per executor on the non-thinking cohort is 0.639–0.824 with annotator decompositions (MuSiQue native, as in the paper), 0.684–0.883 with model self-generated ones, and 0.644–0.811 with a fixed cheap decomposer (GPT-4o-mini). The three are comparable, and the fixed-decomposer arm separates the two variables: with the decomposition held identical across executors the signal still varies by executor, so OC measures the executing model's consistency rather than the decomposition's provenance. Decompositions parse 95–100% of the time, OC is stable across alternatives (65–95% unanimous over three), and the instability itself carries signal, with accuracy 0.884, 0.650, and 0.342 for unanimous-consistent, mixed, and unanimous-inconsistent items.

Corruption ladder. Holding each model's direct answer fixed and re-scoring OC against a clean native decomposition, a corrupted second-hop frame, and a fully random one, AUROC falls from 0.822 to 0.457 and 0.459, that is to chance, while the OC-rate collapses from 0.76 to under 0.03. OC therefore cannot manufacture a signal from an invalid decomposition. The gap tracks quality: $+0.483$ with annotator decompositions on the non-thinking cohort of §3.1 and $+0.365$ with trace-extracted ones on the thinking cohort of §3.2, against nil when corrupted.

B.23 End-to-end call counts and the joint OC-by-SC table

method	generation calls	serial rounds
OC, depth-2, annotated decomposition	3	3
OC, depth-2, self-generated decomposition	4	4
OC, depth-2, fixed cheap decomposer	3 + amortized	3
OC, thinking models (CoQ)	3 + amortized extractor	3
$\text{SC}_{K=3}$ / $\text{SC}_{K=10}$	3 / 10	1 (parallel)
semantic entropy	10	1

Table 8: End-to-end call counts including decomposition generation. The three-call row assumes a supplied decomposition and is free only because a human wrote it; the self-generated row is the more realistic one. OC's calls are sequential while sample-based baselines parallelize, so the latency advantage is smaller than the call-count advantage (App. B.15).

Partitioning the $n = 24{,}624$ non-thinking (model, question) pairs by the binary forms of OC and the naive direct-answer SC ablation gives accuracies of 0.877 where both are consistent, 0.662 and 0.563 in the two off-diagonal cells, and 0.178 where both are inconsistent: a 70 pp spread across the corners. The asymmetry between the off-diagonal cells is what the per-question regressions pick up, and the cell where SC is confident but OC is not is the one a sample-based signal alone would miss.

C Additional related work

This appendix expands the discussion of §4 along three axes: (i) decomposition methods and the question of whether chain-of-thought reflects the model's actual reasoning, (ii) uncertainty

quantification and process-level verification for LLMs, and (iii) operadic and categorical structures in machine learning and computational linguistics.

Faithfulness of chain-of-thought reasoning. The decomposition methods cited in §4 share the working assumption that surfacing sub-steps *improves* a model's reasoning. A complementary line asks whether the surfaced decomposition is *faithful* to the underlying computation. Retrieval systems that learn to decompose by self-play, such as AceSearcher [Xu et al., 2025], treat decomposition quality as the end; we hold it fixed and read the model's disagreement with it. Turpin et al. [2023] construct prompt contexts that bias a model toward a target final answer and observe that the model's CoT trace then rationalizes the (now-biased) answer rather than reflecting the unbiased computation, demonstrating that CoT can be a post-hoc explanation rather than a causal trace. Lanham et al. [2023] measure final-answer sensitivity to several CoT perturbations (truncation, mistake injection, paraphrase) across model scales and tasks, and report that the degree to which final answers actually depend on the trace varies substantially. Operadic consistency probes a third notion in this neighborhood: not whether the trace causally determines the direct answer, and not whether the trace survives perturbation, but whether the direct answer is consistent with the recomposition of the model's own declared sub-answers. This is a structural, not counterfactual, coherence check; an inconsistency flags a question on which the model's direct and decomposed reasoning paths diverge, regardless of which one is correct.

Uncertainty quantification and process-level verification. The dominant inference-time confidence baselines in our setting fall into three groups. (i) *Sample-and-aggregate*: Wang et al. [2023b] introduced self-consistency as majority-voting over K temperature samples; we extend this with a continuous pairwise-agreement variant used as a confidence score. (ii) *Cluster-based entropy*: Kuhn et al. [2023] proposed semantic entropy over NLI-clustered samples, with a discrete-clustering refinement studied in Farquhar et al. [2024] and a Bayesian sample-budget reduction in Ciosek et al. [2025]; we use a token-F1-clustered approximation. (iii) *Self-evaluation*: Kadavath et al. [2022] introduced $P(\text{True})$, the model's own probability of judging its answer correct; Lin et al. [2022] and Tian et al. [2023] study verbalized confidence variants. Process-level verification methods that score *steps* of reasoning rather than aggregate samples [Lightman et al., 2024, Uesato et al., 2022] are conceptually adjacent and complementary to our work. Beyond the three baseline families in §4 (sample-and-aggregate, cluster-based entropy, self-evaluation), the broader uncertainty-quantification literature for LLMs includes paraphrase-consistency tests, surface-form perturbation tests, and reference-free hallucination detectors such as SelfCheckGPT [Manakul et al., 2023], which scores factual consistency across stochastic samples without ground truth. On the process-verification side, the discriminative PRMs we evaluate (Skywork-PRM and RLHFlow-PRM, both math-trained) are step-level descendants of the dense process supervision introduced by Uesato et al. [2022] and Lightman et al. [2024] (PRM800K), with a more recent annotation-free variant in Math-Shepherd [Wang et al., 2024]. In-domain non-math PRMs have since appeared — VersaPRM [Zeng et al., 2025] and DPRM [Wang et al., 2026] — which we flag as natural complements in §5. The math-to-general-reasoning generalization gap visible in our results (Skywork and RLHFlow PRMs underperform OC on multi-hop QA despite extensive math-domain training) is consistent with the limited cross-domain transfer reported in the broader PRM literature.

Operadic and categorical structures in machine learning. The mathematics underlying this work has a long history outside machine learning. Operads originate in algebraic topology, where Boardman–Vogt and May used them to encode multilinear structure on iterated loop spaces; standard modern references are Markl–Shnider–Stasheff [Markl et al., 2002] and Loday–Vallette [Loday and Vallette, 2012]. Beyond pure topology, the *applied category theory* program has used operads and related categorical structures to model the composition of subsystems in robotics, dynamical systems, control, and database design; an accessible book-length introduction is Fong–Spivak [Fong and Spivak, 2019]. The categorical linguistic precedents we already cite in §4 (DisCoCat, Hopf-algebraic syntax, semiring parsing) sit within this broader program. Within deep learning specifically, the recent position paper of Gavranović et al. [2024] argues that essentially all current neural-network architectures can be formalized via structured-functor diagrams, with the categorical structure clarifying which architectural choices preserve which equivariances. That program operates at the level of *network topology* — what kind of object the architecture is. Ours operates at the level of the *semantics of the model's outputs* — what compositional structure can be detected in the answers the model produces under decomposition. The two are complementary rather than competing, and we

are not aware of prior work applying either operadic or related categorical structure to inference-time signal extraction from LLM reasoning, which is the gap the questions operad $\mathcal{Q}$ addresses.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state the contributions we deliver: a mathematical framework (operads + questions operad, §2), a per-question signal (operadic consistency), and an empirical study (§3) showing it contributes statistically significant predictive information about correctness beyond standard sample-based and decomposition-aware baselines on every dataset tested. The scope (twelve non-thinking and five thinking models) is stated in the abstract; the five reasoning benchmarks are named in §3.1 and §3.2.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section before the Impact Statement discusses concrete limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: §2.2 states Assumption 2.4 (Compositionality) and Proposition 2.5, which gives conditions under which a model is operadically consistent exactly. The complete proof is given in App. A.4, which also shows that $\approx$ is a congruence, so that $\mathcal{Q}/\approx$ is a non-unital operad. App. A supplies the standard definitions the statement relies on. The empirical contributions are evaluated statistically and are not derived from the proposition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §3.1 specifies all parameters: model identifiers, datasets, $K{=}10$ sample sizes, sampling temperature ($T{=}0.7$), greedy temperature ($T{=}0$), the per-dataset `score_official` agreement threshold (0.5), regression specification; App. B.4 gives the inference-time prompt for $P(\text{True})$. §3.2 similarly specifies the CoQ extraction protocol, $K{=}3$ and $K{=}10$ budgets, and per-cell sample sizes (100 questions at one seed). App. B.16 gives the depth-3/4 and branching protocol (twelve models, 150 questions per depth per model, $n = 1800$ per depth), App. B.17 the Big-Bench Hard and math pilots, and App. B.18 the annotation codebook and category definitions. All datasets are public.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and per-question result files are publicly available at `https://github.com/natebottman/operadic-consistency-paper`. All datasets used (MuSiQue, StrategyQA, HotpotQA, DROP, GSM8K) are publicly available on HuggingFace; the repository includes scripts that load them directly.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We perform no training. All inference parameters (temperature, K, max tokens, prompts) are specified in §3.1 and §3.2. Prompts and per-dataset context handling (passages for MuSiQue/DROP, no passage for StrategyQA/HotpotQA, problem statement for GSM8K) are documented in App. B.2.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All main claims report two-sided p-values from logistic regression (Tables in §3) or two-sided Pearson r with p-values for across-model correlations (Fig. 3). Statistical tests are computed with `statsmodels` for regressions and `scipy` for correlations.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are inference-only API calls; we trained no models. App. B.15 gives the breakdown: approximately $580{,}000$ direct-mode calls for the across-model study, a further $\sim$480,000 chain-of-thought calls for canonical CoT-SC, $50{,}000$–$100{,}000$ thinking-mode calls for §3.2, and the depth-3/4 and branching runs of App. B.16, at a total on the order of a few hundred million prompt+completion tokens via Together AI and OpenRouter serverless inference. App. B.15 also reports measured wall-clock latency. The locally hosted scorers (DeBERTa-v3-large for NLI and the three PRMs: Skywork, RLHFlow, and VersaPRM) run inference-only on a single 24 GB GPU; no other GPUs were directly used by the authors.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We use only public benchmarks, conduct no human-subjects research, release no models or datasets, and use no PII or scraped data.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Impact Statement discusses the positive use case (improved reliability for deployed LLM reasoning systems via selective prediction) and the recommendation to deploy operadic consistency as one component of a multi-signal stack rather than a standalone training target.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no new models or datasets.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets used (MuSiQue, StrategyQA, HotpotQA, DROP, GSM8K, BREAK QDMR) are publicly released for research use and are cited in §3 and the bibliography. Models are accessed via Together AI and OpenRouter public inference APIs under their respective providers' terms of use; the auxiliary scoring models Skywork-o1-Open-PRM-Qwen-2.5-7B and RLHFlow's Llama3.1-8B-PRM-Mistral-Data are released by their authors for research use and cited in App. B.11, as is VersaPRM [Zeng et al., 2025] in App. B.19.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper introduces no new datasets or models. Analysis code is publicly available in the repository cited above.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing or human-subjects research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper involves no human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Auxiliary LLMs appear in three inference-only roles, all documented in App. B.1. In the thinking-model pipeline (§3.2, App. B.5) Qwen3-Next-80B-A3B-Instruct, in non-thinking mode, identifies sub-questions and their stated answers in the model's chain-of-thought trace. In the canonical CoT-SC baseline (App. B.4) Llama-3.3-70B extracts a short answer from each sampled chain. As a robustness check (App. B.14, App. B.16) Claude Haiku 4.5 is used at $T=0$ as a correctness judge, applied uniformly across cells. All are standard pretrained models used inference-only and held fixed within their role. No model is fine-tuned for this paper.