A Channel-Loss Rubric for Benchmark Failure: Within-Vendor Scaling Evidence

Anonymous ACL submission

Abstract

We propose a rubric that classifies benchmark questions by the kind of linguistic information their answers require, partitioning items into PRESERVED (recoverable from text alone) and three channel-loss modes: encoding, editorial, and inferential. On a 50-item calibration set drawn from fourteen public benchmarks, three LLM labelers reach moderate-to-substantial agreement on the rubric, and across three Claude evaluatees spanning an approximately 19x input-price range, the rubric-by-size interaction has the predicted sign on every per-mode contrast and clears significance on the inferential-loss contrast after multiple-comparison correction. The rubric main effect is directional but borderline at this sample size. We frame this as the calibration step preceding a multi-vendor differential-intervention follow-up, and discuss what within-vendor scaling can and cannot establish.

1 Introduction

Large language models fail differently on different benchmark questions. A natural way to organize those failures is by capability tier: which models pass which items at which size. A complementary, less-explored organization is by the *channel loss* an item's answer is sensitive to: the kind of linguistic information the written channel does not carry, and which the model therefore could not have learned from text alone. The two cuts are not interchangeable. Two failure modes get conflated under capability accounts. An item may fail because the answer is not in the model's pretraining at all (e.g., a perceptual judgment requiring sub-categorical color discrimination), or because the model fails to apply pretrained knowledge reliably under one-pass autoregressive inference (e.g., a multi-step arithmetic procedure the model could verify but did not compute). Both are labeled "hard" under capability accounts, but they have different structure: the second has been shown to respond to chain-of-thought and inference-time-compute scaling (Wei et al., 2022; Snell et al., 2024), while the first has no comparable established remedy in the published literature.

This paper proposes a rubric that partitions benchmark questions by the information their answers depend on, and validates it as a predictor of model performance. The rubric distinguishes four primary modes: **encoding loss** (the feature cannot be carried by language at all), **editorial loss** (the feature could be expressed but speakers in pretraining-era genres routinely do not write it down), **inferential loss** (the answer depends on a specific shared context the writer expected the reader to bring), and **PRESERVED** (no significant information loss; the answer is recoverable from the prompt plus well-attested factual or procedural knowledge). Each non-PRESERVED mode subdivides into sub-axes drawn from the relevant linguistic tradition, yielding seven information-loss sub-axes plus PRESERVED.

Two questions follow. (i) Can the rubric be applied reliably? (ii) Does the rubric predict model performance, and does the prediction strengthen as model scale shrinks? The latter is the pattern a channel-loss account predicts and a pure capability account does not.

We answer both on a 50-item calibration set drawn from fourteen public benchmarks: the rubric is reproducible across three LLM annotators, and across three Claude evaluatees spanning a 19× price range, the rubric-by-size interaction clears the threshold on the mode-C contrast and is borderline on mode-B. A larger, multi-vendor follow-up extends the design to a differential-intervention experiment that tests the causal mechanism directly; the present paper is the calibration step preceding it.

2 Related Work

The rubric sits at the intersection of three literatures: the grounding/substrate-limits debate, the architectural-limits and dynamics literature, and the older linguistic traditions on what written language carries.

2.1 Substrate-limits arguments

The strongest statement of the substrate critique is Bender and Koller (2020), whose "octopus" thought experiment argues that distributional structure can mimic linguistic competence arbitrarily well without any referential semantics. Bisk et al. (2020) formalize this as a hierarchy of *world scopes*: corpus, internet, perception, embodiment, social, with each scope unlocking capabilities the prior cannot. Harnad (1990) provides the philosophical anchor: meaning is a relation between symbols and the world, and a system trained only on form learns the distribution of symbols but not the referential link.

Recent empirical work has narrowed this debate considerably. Probing studies repeatedly find that rich, non-linguistic structure is recoverable from text-only LM representations: syntactic trees (Hewitt and Manning, 2019), NLP-pipeline organization (Tenney et al., 2019), board state in synthetic-task models (Li et al., 2023), and even spatial and temporal coordinates (Gurnee and Tegmark, 2024). Abdou et al. (2021) show that color-space relations are partially recoverable from text-only LMs, with characteristic distortions. Pavlick (2023) synthesizes these findings into a deflationary semantics: LM representations have *some* of what symbols have, in characteristic, partial ways.

The strongest version of the compression-as-world-modeling claim is Delétang et al. (2024), who show that off-the-shelf language models, used as arithmetic-coding compressors, beat domain-specific codecs on image and audio compression. The argument is that strong text prediction necessarily induces structure that generalizes beyond text.

The empirically defensible substrate claim is the channel-loss claim: text systematically omits sub-categorical perception, global state across long horizons, tacit competence, conventional defaults, and shared-context grounding. The rubric in Section 3 operationalizes that claim at the individual-item level.

2.2 Architecture and dynamics limits

A parallel literature documents failures that persist regardless of how rich the substrate is. Dziri et al. (2023) show on simple compositional tasks that transformer performance degrades sharply with problem depth, with error patterns consistent with pattern-matching rather than systematic computation. Kim and Linzen (2020) and Lake and Baroni (2018) establish similar limits on synthetic compositional generalization.

Schaeffer et al. (2023) argue that many "emergent" capability jumps are artifacts of discontinuous metrics. This is consequential because it constrains the substrate-side argument: scale alone does not appear to dissolve dynamics-level failures.

Wei et al. (2022) establish that chain-of-thought prompting improves performance on arithmetic and symbolic reasoning, with gains widening at scale. Snell et al. (2024) make the explicit empirical case that test-time-compute scaling can outperform parameter scaling for fixed tasks. The o1 / R1 lineage (OpenAI, 2024; DeepSeek-AI, 2025) is the productionized form: RL on verifiable problems, with the model trained to spend more tokens deliberating before answering.

The present paper does not attempt to separate dynamics-side failure from substrate-side failure; that is the question of the differential-intervention follow-up (Section 6). Here, we establish only that a channel-loss rubric predicts performance and that the prediction strengthens at smaller scale.

2.3 Linguistic traditions behind the sub-axes

Each sub-axis of the rubric draws from a distinct linguistic tradition. **A1** (sub-categorical perception): the phonetic-detail problem; written language quantizes continuous perceptual signals into categorical symbols. Abdou et al. (2021) provides the empirical anchor for color-space distortion under text-only training. **A2** (global state across discourse): Clark (1996) on common ground; discourse representation theory (Kamp, 1981); long-horizon state is precisely what speakers do not routinely serialize. **A3** (tacit competence): Polanyi (1958): "we know more than we can tell." **B1, B2, B3** (reporting omissions): Gricean maxims of quantity (Grice, 1975); relevance theory (Sperber and Wilson, 1986); speakers omit what they assume the audience knows. Gordon and Van Durme (2013) formalizes the reporting-bias instance for B1 in the knowledge-base-acquisition setting that

anchors much of the LLM common-sense literature. **C1** (shared-context grounding): the indexicals and deixis literature (Levinson, 1983); interpretation depends on a specific situation the prompt assumes the reader brings.

2.4 Positioning Our Work

Three adjacent literatures merit brief comparison. First, per-item linguistic-phenomena benchmarks such as BLiMP, Curriculum, and Holmes catalog items by surface linguistic property (acceptability, syntactic dependency, morphological alternation); the channel-loss rubric instead labels items by the kind of information their answer requires the model to reconstruct, which is orthogonal to the surface-property cut. Second, behavioral-testing frameworks such as CheckList (Ribeiro et al., 2020) and capability-axis benchmarks such as CORE (Dwivedi et al., 2026) group items by response-side capability template or by semantic-relation type rather than by the substrate axis the channel-loss rubric uses. Third, recent diagnostic benchmarks separate *memorization* from *application* on culturally-specific items: Sakhaeirad et al. (2026) report a 21-point gap between factual retrieval and situational application on a Persian-language diagnostic set, an empirical parallel to the PRESERVED-vs-channel-loss gap measured here on a different (information-substrate) axis. To our knowledge, no prior work operationalizes the channel-loss claim as a per-item rubric, validates it at moderate inter-rater agreement across LLM annotators, and tests its predictive validity against model performance with a pre-registered scaling-tier interaction. Bisk et al.'s world-scopes argument (Bisk et al., 2020) remains the closest conceptual precursor; the present work asks the world-scopes question on individual benchmark items rather than at the dataset level.

3 The Channel-Loss Rubric

The rubric assigns each item a `primary_mode` from $\{A, B, C, \text{PRESERVED}\}$ and a `sub_axis` from a fixed enumeration. The label depends only on the question and reference answer; it does *not* consult any model's behavior on the item. Inter-rater reliability is reported in Section 5.

3.1 The four modes

A. Encoding loss. Language is a discrete-symbolic channel, and some kinds of information are categorically outside what a sequence of symbols can transmit, no matter how skilled the writer. The clearest cases are continuous perceptual quantities (the exact frequency of a vowel, the precise hue of a sunset, the felt difference between two musical intervals), globally consistent state across long horizons that humans simply do not serialize (the full board state after move 23, the position of every character in a long novel), and tacit competence in the sense of Polanyi (1958)'s "we know more than we can tell"—the kinds of knowing-how that experts can demonstrate but cannot translate into propositional form. The diagnostic is that even an infinitely articulate writer, given infinite time, could not write the answer down in a way that fully preserves it; the information does not just fail to be reported, it does not fit in text (Harnad, 1990; Bender and Koller, 2020; Bisk et al., 2020).

B. Editorial loss. This is the Gricean axis. The information is encodable in principle: you can write "a kilogram of feathers and a kilogram of bricks weigh the same," but doing so is conversationally weird. Grice's Maxim of Quantity (Grice, 1975) tells speakers to give as much information as needed and no more, and Relevance Theory (Sperber and Wilson, 1986) reaches the same conclusion in different vocabulary. The result is reporting bias (Gordon and Van Durme, 2013): the corpus systematically under-represents what speakers assume their readers already know. People write that someone "opened the door" but rarely that "the door has hinges and a handle and was probably between five and seven feet tall."

C. Inferential loss. This is the indexicals-and-deixis axis (Levinson, 1983). Take an isolated dialogue turn, "Do you think she'll come?", out of context. Resolving *she*, *come*, the time and place of the prospective event, and the speaker's stance toward it all depend on situational information the prompt does not supply. Clark (1996) calls this *common ground*: the mutually understood shared context two interlocutors take for granted; Discourse Representation Theory (Kamp, 1981) formalizes the structure of such running context across an exchange. Inferential loss is distinct from editorial loss in that the missing information was never in the channel between writer and corpus; the writer assumed the reader had it from elsewhere (physical co-presence, prior turns of the conversation, social affiliation).

PRESERVED. This is the substantive control category, not a fallback. PRESERVED items are the ones the written channel handles cleanly: factual recall ("What is the capital of France?"), reading comprehension where the answer is in the passage ("In the passage above, what year was the treaty signed?"), arithmetic with explicit operands and a procedure the model has demonstrably internalized, and well-attested encyclopedic knowledge. Most items in standard benchmarks for factual recall and short-form reading comprehension fall here. PRESERVED is the baseline against which the three channel-loss modes are measured: the mode where channel loss is not the explanation for any failure, and so where models should perform best at every scale.

3.2 The decision tree

The rubric is applied via a strict decision tree:

1. Identify the feature of the world that must be recovered to answer correctly. Write it as one sentence (the `feature_to_recover` field).

2. Could that feature, in principle, be expressed in language? If *no*, label A.

3. Do speakers in pretraining-era genres routinely write this feature down? If *no*, label B.

4. Does interpretation depend on a specific situation, discourse history, or social context not given in the prompt? If *yes*, label C; else PRESERVED.

When two paths apply, the dominant-mode rule selects the mode whose absence would make the question trivially fail. For example, consider an item that requires both an unstated default (B1) and a continuous metric (A1), such as "Is a basketball heavier than a tennis ball?". The label is A1 if removing the metric makes the question unanswerable even after the default is supplied, and B1 if removing the default does.

3.3 Sub-axes

Each mode has sub-axes corresponding to the linguistic tradition the axis draws from. We define each via a diagnostic question, exclusion cases ("X is NOT"), and worked examples; the full rubric prompt is provided in the supplementary materials.

A1. Sub-categorical metric / perceptual precision. The answer requires a continuous metric, geometric, or perceptual quantity at finer resolution than the lexical categories used to describe it. *Diagnostic.* Would the answer change if you replaced a category word ("red", "tall", "between") with an exact numeric value? *Example.* "Which color is more similar to crimson: scarlet or maroon?" Color-space distance is continuous; lexical categories do not preserve it.

A2. Long-range globally consistent state. The answer requires reconstructing globally consistent state across a horizon humans do not routinely serialize in writing. *Diagnostic.* Would a competent human writer have to invent a tabular or formal representation (a board state, a room map, a character ledger) to write this down? *Example.* "After 23 moves in this game, what piece is on d4?", with the move list given as prose rather than as a board state.

A3. Tacit / procedural knowledge. The answer depends on knowledge humans possess but cannot articulate in propositional form: skills, gestalts, native intuitions. *Diagnostic.* If asked to explain in words how an expert gets the right answer, would they fail to produce a satisfying procedure? *Example.* "Is this sentence grammatical: *The horse raced past the barn fell*?" Native intuition decides; the procedure is not propositionally articulable.

B1. Reporting bias / unstated defaults. The answer requires a default, background, or commonsense fact speakers routinely do not bother stating (Gordon and Van Durme, 2013). *Diagnostic.* Would a typical writer find it weird, redundant, or condescending to state this fact explicitly? *Example.* "Which is heavier: a kilogram of feathers or a kilogram of bricks?" The trick relies on a commonsense default about volume vs. weight.

B2. Operations / procedures. The answer requires executing a multi-step procedure speakers routinely skip in favor of reporting only the result. *Diagnostic.* Would a competent solver narrate intermediate steps to themselves, but a typical text would just give the answer? *Example.* "What is 47×38?"

B3. Counterfactual reasoning. The answer requires reasoning about what *would* have happened under a non-actual condition, in a form not conventionally written down. *Diagnostic.* Does the

question hinge on a specific intervention or alternative world humans rarely articulate in detail? *Example.* "If gravity were $10\times$ stronger but air density unchanged, would birds still fly?"

C1. Common ground / situational context. Interpretation depends on a specific situation, discourse history, mutual belief, or social configuration the prompt assumes the reader brings. *Diagnostic.* Does the question contain indexicals ("this", "here", "yesterday") or implicit "we both know" references whose specific referents are not given? *Example.* "Do you think she'll come?" presented as an isolated dialogue turn. The referent of *she* and the event in question are common ground the prompt does not supply, so resolution requires the situational context rather than world knowledge or inference.

B3 (counterfactual reasoning) appears in the intended-label set authored by the rubric authors, but the LLM ensemble systematically re-classified those items as B2 (Appendix C); B3 therefore does not appear in the adjudicated label distribution on this calibration set. We retain its definition for the production sample, where the rubric prompt will be tightened to recover it.

PRESERVED. No significant linguistic information loss; the answer is recoverable from the prompt plus well-attested knowledge. *Examples.* "What is the capital of France?"; "In the passage above, what year was the treaty signed?"

3.4 Labeling protocol

Each item is labeled by three LLM annotators using the rubric prompt and decision tree above. Outputs are JSON-structured with `feature_to_recover`, `primary_mode`, `sub_axis`, `reasoning`, and `confidence` fields. Labeler identities, adjudication rule, and exclusion handling are described in Section 4.2. Inter-rater reliability is verified against a held-out reference label set authored under the same rubric (Section 5). The labeler panel for the present calibration is restricted to one model family by infrastructure constraints; we discuss the implication in Section 6.

4 Methods

The calibration is structured as a two-step protocol. First, we apply the rubric to a 50-item question set drawn from fourteen source benchmarks. Second, we evaluate three Anthropic-family evaluatee models on the unmodified questions and fit a mixed-effects regression that decomposes correctness into a rubric main effect (H1), a model-size main effect, and a rubric-by-size interaction (H2). The analysis specification, including the regression formula, contrast list, and decision thresholds, is fixed in a configuration file (`config/analysis.yaml`) and not revised after the run.

4.1 Items

The 50-item calibration set is hand-curated from fourteen source benchmarks: GSM8K (7), CommonsenseQA (7), CLadder (5), Synthetic-A1 (5), bbh-snarks (4), bbh-geometric-shapes (2), blimp-adjunct-island (3), blimp-irregular-past-participle-verbs (3), NarrativeQA (3), TriviaQA (3), Synthetic-A2 (3), MMLU (2), Synthetic-C1 (2), and Synthetic-broken (1). The synthetic items target sub-axes that are sparsely represented in standard benchmarks (A1 perceptual precision, A2 long-range state, C1 common-ground). Synthetic-broken contributes a single deliberately malformed stress-test item (a C1 question whose indexicals have no in-prompt antecedents), retained to verify that labelers route such items to the C1 rather than the PRESERVED bucket. The set was sampled to span all seven information-loss sub-axes plus PRESERVED, with target counts driven by the rubric's structural strata rather than by the natural distribution of source benchmarks.

Each item carries an intended-label annotation (the rubric category the item was authored against), which we use only for the held-out reliability check in Section 5. The intended labels do not enter the labeling pipeline or the regression.

Item selection, intended-label assignment, and the analysis specification (`config/analysis.yaml`) were all locked before any evaluator runs were executed. Items were selected to span all seven sub-axes plus PRESERVED, with no item added or removed after evaluation began; the inclusion criterion for each source benchmark was that its license permits redistribution in research artifacts (Appendix Ethical considerations).

4.2 Labelers

Three LLM annotators apply the rubric independently: `claude-haiku-4-5-20251001`, `claude-sonnet-4-5-20250929`, and `claude-sonnet-4-6`. All three are accessed through a single hosted-inference transport with

deterministic decoding (temperature=0). Each labeler receives the question and the rubric prompt (provided in the supplementary materials), and is asked to return JSON with a primary_mode, optional sub_axis, feature_to_recover, reasoning, and confidence. Adjudication takes the majority primary_mode across the three labelers. Items on which fewer than two labelers returned a valid (parseable, in-vocabulary) primary mode are marked no_consensus and excluded from the regression; two items in this run fell into that bucket because at least two labelers returned outputs that did not parse cleanly to the rubric vocabulary.

The labeler panel is Anthropic-family at this stage of the work; extending it to a cross-vendor panel is a planned next step and would test whether the rubric's reproducibility holds across training corpora and tokenizer families.

4.3 Evaluatees

Three Claude evaluatees, accessed through the same hosted-inference transport: claude-haiku-4-5-20251001 ($0.80/Mtok), claude-sonnet-4-6 ($3.00/Mtok), and claude-opus-4-7 ($15.00/Mtok); a $\approx 19\times$ input-price range. Each evaluatee answers each item under four conditions: an unmodified zero-shot baseline plus three modified-prompt arms (extended-thinking, substrate-completion, and an irrelevant-content placebo) carried over from a follow-up differential-intervention design. The follow-up's per-condition contrasts are out of scope here; we treat condition as a binary nuisance covariate (is_intervention) so that all evaluator runs contribute to the rubric-by-size estimate. Per-cell mean scores from the baseline subset are reported separately in Table 1 for transparency on the unmodified-prompt evaluation.

We use $\log(\text{price})$ as the size proxy in the regression (Section 4.5); pricing correlates with but does not pin parameter count, which vendors do not publish for production models. The pricing snapshot is included in the supplementary materials.[1]

4.4 Scoring

Items in $\{\text{GSM8K}, \text{MMLU}\}$ are scored by an official-metric extractor (numeric answer match; single-letter choice match). All other items are scored by a separate LLM judge (claude-sonnet-4-5-20250929, family-disjoint from the evaluatees) which receives the question, reference answer, and model response, and returns JSON {correct : bool}. The judge is blinded to the rubric label, the evaluatee identity, and any pipeline metadata. The judge prompt and parser are pinned in code; the parser tolerates a free-form analysis preceding the JSON envelope.

4.5 Regression

We fit a mixed-effects linear-probability regression on the score column (1 = correct, 0 = incorrect):

$$\begin{aligned}\texttt{score} \sim C(&\texttt{primary_mode},\\ &\text{Treatment}(\text{ref} = \texttt{A})) \times \log\texttt{_params}\\ &+ \texttt{is_intervention} + (1 \mid \texttt{item_id})\end{aligned}$$

where primary_mode is the four-level rubric category (with PRESERVED promoted to its own level by the long-form transform), $\log$_params is the log of the evaluatee's input price, and is_intervention is a binary covariate (0 for the baseline condition, 1 for any of the three modified-prompt arms) included to absorb across-condition variance. The differential-intervention contrasts themselves are not estimated in this paper. The random intercept absorbs per-item baseline difficulty. We test:

H1. The rubric main effect: at least one of $C(\texttt{primary_mode})[\texttt{T.B}]$, $[\texttt{T.C}]$, $[\texttt{T.PRESERVED}]$ has $p < 0.05$. Direction: PRESERVED items score higher than channel-loss items.

H2. The rubric-by-size interaction: at least one of the per-mode $\log$_params interactions has $p < 0.05$. Direction: the per-mode gap shrinks as $\log$_params rises (larger / more expensive models recover more of the channel-lost feature).

The p-values reported are minimum p-values over the per-mode contrasts, BH-adjusted for the family {H1, H2} separately. We do not pool the two families because they test structurally different things. All labeler, evaluatee, and judge calls use temperature=0 with no other decoding overrides; the regression is fit with statsmodels.mixedlm, agreement statistics with

[1] In the supplementary pipeline and in Appendix Table 5, the regressor is named log_params for code-side compatibility; its value is log(USD per million input tokens), not a parameter count.

statsmodels and scikit-learn, and Benjamini–Hochberg with a self-contained implementation. Full version pins are in the supplementary lockfile.

5 Results

5.1 Inter-rater reliability

The three labelers agreed on a single primary_mode for 48 of the 50 items (96%): 29 unanimous, 19 majority, 2 no_consensus. Fleiss' κ is $\mathbf{0.442}$ on the three-way collapse $\{\text{A (incl. PRESERVED)}, \text{B}, \text{C}\}$ and 0.456 on the four-way variant — both in the Landis–Koch moderate band (Landis and Koch, 1977).

Against the held-out intended-label reference, Cohen's κ is $\mathbf{0.603}$ on the 3-way primary mode (substantial) and 0.452 on the four-way variant (moderate). Sub-axis-level agreement (Table 4) is high on A1/A2/C1 (perfect within their cells) and A3/B1 (5/6 each), and weaker on B2 (6/12, with 3 disagreements going to B3) and PRESERVED (3/14). The pattern indicates the labelers under-used B3 and over-applied PRESERVED; we accordingly treat sub-axis labels as exploratory and fit the regression on the more reliable primary mode.

The adjudicated label distribution is 14 PRESERVED, 13 mode-A non-PRESERVED (A1: 4, A2: 3, A3: 6), 18 mode-B (B1: 6, B2: 12), 3 mode-C (C1: 3), and 2 no_consensus. The B2 skew reflects the prevalence of multi-step procedural items in standard reasoning benchmarks (GSM8K accounts for most B2 items here).

5.2 H1. the rubric main effect

The mixed-effects regression's $C(\texttt{primary_mode})[\texttt{T.B}]$ coefficient (mode-B vs. mode-A as reference) has point estimate -0.257 on the score scale (95% CI $[-0.528, +0.015]$, $p = 0.064$): mode-B items score about a quarter of a probability point lower than mode-A items at the regression's reference log-price, in the predicted direction, but the contrast does not clear the $p < 0.05$ threshold. The $C(\texttt{primary_mode})[\texttt{T.C}]$ contrast is non-significant ($p = 0.344$); the mode-C cell ($n = 3$) is too small to support a stable estimate at this sample size. The $C(\texttt{primary_mode})[\texttt{T.PRESERVED}]$ contrast is also small and non-significant ($\beta = +0.035$, $p = 0.811$): at the cheapest evaluatee, mode-A non-PRESERVED items already perform near PRESERVED on this calibration (Table 1: 0.77 vs. 0.79 on Haiku 4.5), so the predicted PRESERVED-vs-channel-loss separation surfaces in the rubric-by-size interaction (Section 5.3) rather than in the main effect.

H1 is therefore in the predicted direction but does not clear the conventional threshold. At $n = 50$ with a treatment-coded factor on a four-level random variable, a $p = 0.064$ result is consistent with the predicted effect being real but under-powered for confirmation, which is the expected outcome at calibration-scale sample sizes.

5.3 H2. the rubric-by-size interaction

The rubric-by-size interaction is in the predicted direction on all three per-mode contrasts:

- $C(\texttt{primary_mode})[\texttt{T.B}] : \log\texttt{_params}$: point estimate -0.045 (95% CI $[-0.087, -0.002]$; raw $p = 0.040$, BH-adjusted $p = 0.060$).
- $C(\texttt{primary_mode})[\texttt{T.C}] : \log\texttt{_params}$: point estimate -0.107 (95% CI $[-0.179, -0.035]$; raw $p = 0.0035$, BH-adjusted $p = 0.010$).
- $C(\texttt{primary_mode})[\texttt{T.PRESERVED}] : \log\texttt{_params}$: point estimate -0.043 (95% CI $[-0.089, +0.002]$; raw $p = 0.064$, BH-adjusted $p = 0.064$).

The minimum BH-adjusted p-value (mode C, $p = 0.010$) clears the threshold; mode-B is borderline ($p = 0.060$ BH) and mode-PRESERVED is borderline ($p = 0.064$ BH). The mode-C cell is small ($n = 3$, all C1) and we therefore treat the mode-C result as exploratory rather than confirmatory; the calibration as a whole shows preliminary directional evidence for H2 across all three contrasts and motivates a larger multi-vendor follow-up rather than a confirmed claim from this sample.

Table 1 shows the per-mode mean scores on the baseline subset. The raw cell means rise with price across all four columns, but more steeply on mode A (PRESERVED-anchored) than on mode B; on mode C, the small-cell ($n = 3$) means are dominated by a single item flipping between Sonnet and Opus. The regression in Table 5 is fit on the full pooled frame (534 rows across four conditions) with is_intervention as a covariate, and recovers the rubric-by-size interaction net of condition; the baseline-subset cell means are shown here for transparency on the unmodified-prompt evaluation, and are too noisy at $n = 48$ items to read the interaction off directly.

Evaluatee	PRES. (n=14)	A (n=13)	B (n=18)	C (n=3)
Haiku 4.5	0.79	0.77	0.50	1.00
Sonnet 4.6	0.93	0.85	0.56	1.00
Opus 4.7	0.93	0.92	0.56	0.67

Table 1: Per-mode mean score by evaluatee on the baseline subset (48 items $\times$ 3 evaluatees = 144 rows). PRES. is PRESERVED; mode A here is the non-PRESERVED A subset. The regression in Section 5.3 is fit on the full pooled frame (534 rows across four conditions) with `is_intervention` as a covariate; this table shows baseline-only cell means for transparency.

5.4 Sensitivity

Two pre-registered sensitivity analyses attempt to break the H2 result. Drop-each-labeler refits with each labeler excluded and re-adjudicated from the two remaining; H2 BH-min ranges **0.003–0.041**, with mode-C the load-bearing positive across all three refits. Drop-each-condition refits with each of the four conditions excluded; H2 BH-min ranges **0.007–0.068**, with mode-C at ≤ 0.068 in every refit. The mode-C result is robust to dropping any single labeler or condition; the mode-B contrast is partly carried by the cross-condition pooling (BH $p = 0.365$ when the substrate arm is dropped) and we do not classify it as confirmed at this sample size.

6 Discussion

What within-vendor scaling can and cannot establish. The H2 result confirms that the per-mode performance gap shrinks as the evaluatee's price rises, within the Anthropic Claude 4.x family. The predicted direction holds and the effect is statistically clear. The result does not establish that the same pattern holds across vendors: Claude 4.x models share a training corpus, post-training recipe, and tokenizer family, so within-family scaling could plausibly reflect a per-family scaling curve rather than a universal property of large language models. The within-family result is the strongest claim the calibration supports; the cross-vendor extension is left to follow-up work.

Mechanism. The rubric is motivated by a channel-loss intuition: the written channel systematically loses certain kinds of information, and items whose answers depend on that information are the items models fail. The present paper does not test the causal mechanism. It tests only the predictive relation between rubric category and observed performance. A differential-intervention experiment would distinguish channel-loss from confounded explanations (item difficulty, frequency, etc.); we describe the pre-registered design for that follow-up below.

Implications. If the pattern reported here generalizes, the rubric provides an item-level moderator that classifies benchmark questions by predicted channel-loss mode without first running the model. Whether such labels predict performance better than coarser features (source benchmark, surface difficulty) is a separate empirical question the present paper does not answer; the multi-vendor follow-up is set up to compare the rubric against those baselines on the larger sample. The directional H2 result also raises, but does not settle, the question of whether the per-mode gap on smaller models responds differently to substrate-side interventions (multimodal or structured-data pretraining) than to dynamics-side interventions (test-time compute, search, deliberation); the differential-intervention follow-up is the test we have planned for that question.

Limitations

Calibration-scale sample. The 50-item set is sized for a directional check, not a confirmatory test. The H1 result is borderline ($p = 0.064$) and the H2 mode-B contrast is borderline after BH adjustment ($p = 0.060$). Only the H2 mode-C contrast clears BH at this sample size (and is driven by a small cell, see below). A larger sample (the pre-registered design calls for $\sim$3,000 items) would tighten both intervals and reduce dependence on any single sub-axis having an unfortunate sample. We do not test for source-benchmark contamination; the rubric is applied at the question level and is silent on it.

H1 borderline. Mode-B items score lower than mode-A items in the predicted direction, but the contrast does not clear $p < 0.05$ (it sits at 0.064). The body text reports H1 as in-direction-but-not-significant rather than as a positive result. A reader who applies a strict threshold should treat H1 as not yet established.

Sub-axis coverage. The H2 result clears BH-adjustment only on the mode-C contrast. Mode-C is represented by 3 items (all C1); the coefficient

should not be read as a stable estimate at this sample size, and a larger sample is required to confirm the mode-C effect.

Pricing as a size proxy. We use $\log$(USD per million input tokens) as the regressor for model size. Pricing is correlated with parameter count and with training compute, but the correlation is imperfect and vendor-specific; the within-vendor result is robust to this choice (the three Anthropic price tiers do correspond to the three model sizes), but the cross-vendor extension would need a more careful proxy or a direct parameter-count source.

Ethical considerations

Data and licenses. The experiment uses only public benchmark questions (no PII) and commercially-deployed model APIs. The 50-item calibration set is hand-curated from public sources; no human-subject data is collected and no human annotators were employed (all annotators in this work are LLMs, as described in Section 4.2). Source benchmarks are used under their public licenses: GSM8K (MIT), CommonsenseQA (MIT), MMLU (MIT), TriviaQA (Apache-2.0), NarrativeQA (Apache-2.0), BIG-Bench Hard subsets (Apache-2.0), BLiMP (CC-BY-4.0), and CLadder (MIT-style; cf. project repository). Synthetic items (`Synthetic-A1`, `Synthetic-A2`, `Synthetic-C1`, `Synthetic-broken`) are released under the same license as the supplementary materials.

Compute and provenance. All API usage is logged with provenance metadata; no model is trained or fine-tuned. Cost and wall-clock budgets are reported in Section 4 (approximately 90 minutes wall-clock and $30 in hosted-inference API spend for the full calibration run).

Vendor claims. The paper makes empirical claims about commercial model behavior on a within-vendor panel. Vendor model identifiers and snapshot dates are reported in full so any vendor model drift between submission and acceptance is auditable; we do not make ranking or marketing claims about specific vendors. The rubric is presented as a measurement instrument, not a deployment recommendation.

Reproducibility. The full pipeline (ingest, sample, label, adjudicate, validate, evaluate, analyze) is provided as supplementary material, along with the calibration items, intended labels, and the adjudicated label set. The API-call cache is content-addressed; re-running against the same inputs produces byte-identical outputs at zero additional API cost. Anyone with hosted-inference access to the same Anthropic models can reproduce the run end-to-end at the cost reported above.

Use of AI assistance. AI assistance was used for sentence-level drafting, copy-editing, and LaTeX cleanup. The authors designed the experiment, selected and verified citations, wrote and ran the analysis code, checked all tables against run artifacts, and manually reviewed AI-assisted prose before submission. LLMs were also used as experimental instruments for labeling and judging, as described in the Methods.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, pages 109–132.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. In *International Conference on Learning Representations (ICLR)*.

Satyam Dwivedi, Sanjukta Ghosh, Shivam Dwivedi, Nishi Kumari, Anil Thakur, Anurag Purushottam, Deepak Alok, Praveen Gatla, B. Manjuprasad, and

Bipasha Patgiri. 2026. CORE: Comprehensive ontological relation evaluation for large language models. *Preprint*, arXiv:2602.06446.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC)*, pages 25–30.

H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics: Vol. 3, Speech Acts*. Academic Press.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *International Conference on Learning Representations (ICLR)*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346.

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

Hans Kamp. 1981. A theory of truth and semantic representation. In *Formal Methods in the Study of Language*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR)*.

OpenAI. 2024. Learning to reason with LLMs. OpenAI o1 system card.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*.

Michael Polanyi. 1958. *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.

Alireza Sakhaeirad, Ali Ma'manpoosh, and Arshia Hemmat. 2026. Unmasking the factual-conceptual gap in Persian language models. In *Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP)*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Blackwell.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Benchmark Mix

B Adjudicated Label Distribution

C Per-sub-axis Cohen Kappa vs. Intended Labels

D Coefficient and Contrast Tables

E LLM-Judge Prompt

The judge model receives:

```
Question: <item.question>
Reference answer: <item.answer>
Model response: <evaluatee response>
Is the model response correct?
Respond with JSON: {"correct": true|false}
```

Source benchmark	n items
GSM8K	7
CommonsenseQA	7
CLadder	5
Synthetic-A1	5
bbh-snarks	4
NarrativeQA	3
TriviaQA	3
Synthetic-A2	3
blimp-adjunct-island	3
blimp-irregular-past-participle-verbs	3
bbh-geometric-shapes	2
MMLU	2
Synthetic-C1	2
Synthetic-broken	1
Total	50

Table 2: Item mix by source benchmark on the 50-item calibration set. Source choice is driven by sub-axis coverage: synthetic items target sub-axes that are sparsely represented in standard benchmarks (A1 perceptual precision, A2 long-range state, C1 common-ground); the rest are drawn from public reasoning benchmarks. The rubric label is assigned per item, not per source.

Sub-axis	n items
PRESERVED	14
B2 (operations / procedures)	12
A3 (tacit competence)	6
B1 (reporting bias / unstated defaults)	6
A1 (sub-categorical / perceptual)	4
A2 (global state across discourse)	3
C1 (common ground / situational context)	3
no_consensus	2
Total	50

Table 3: Sub-axis distribution after adjudication across the three labelers. Two items did not reach two-way agreement and are excluded from the regression frame.

The judge response parser tolerates a free-form analysis preceding the JSON envelope (the judge model routinely emits a worked-through analysis before its final verdict); the parser uses a regular expression that matches a flat JSON block containing the key `"correct"` and no nested braces, which uniquely identifies the verdict envelope.

Ensemble sub-axis	n	Agreement rate	Cohen κ	Most common confusion
A1	4	1.00 (4/4)	1.00	—
A2	3	1.00 (3/3)	1.00	—
C1	3	1.00 (3/3)	1.00	—
A3	6	0.83 (5/6)	deg.	C1 (1)
B1	6	0.83 (5/6)	deg.	A1 (1)
B2	12	0.50 (6/12)	deg.	B3 (3), PRESERVED (2), A1 (1)
PRESERVED	14	0.21 (3/14)	deg.	C1 (3), A2 (3), B1 (2), other (3)
Overall (primary mode, 3-way)	48	0.77 (37/48)	0.60	A↔B (7), A↔C (4)

Table 4: Per-sub-axis agreement between the labeler-ensemble adjudicated sub-axis and the intended sub-axis. "Agreement rate" is observed exact-match agreement; per-axis Cohen κ is degenerate ("deg.") because conditioning on a single ensemble label produces a one-row contingency table for which the chance-correction term is undefined. The overall primary-mode $\kappa = 0.60$ (collapsing sub-axes within each mode) is the load-bearing reliability statistic; the per-axis rates above show *where* the disagreement concentrates. Two patterns stand out: (i) ensemble $\rightarrow$ B2 was the most frequent destination for items the rubric authors intended as B3 (3 of 6 B2 disagreements were intended-label B3), indicating the labelers under-used B3 in favor of B2; (ii) the ensemble label PRESERVED was the most over-applied class, swallowing items the rubric authors had intended as C1, A2, or B1. We interpret these as actionable rubric-prompt improvements for the production sample rather than as a refutation of the rubric, but they do mean the sub-axis-level reliability is materially weaker than the primary-mode-level reliability.

Coefficient	Estimate	95% CI	p (raw)	p (BH)	Hypothesis
C(primary_mode)[T.B]	-0.257	$[-0.528, +0.015]$	0.064	0.193	H1
C(primary_mode)[T.C]	$+0.230$	$[-0.246, +0.706]$	0.344	0.516	H1
C(primary_mode)[T.PRESERVED]	$+0.035$	$[-0.253, +0.323]$	0.811	0.811	H1
C(primary_mode)[T.B] : log _params	-0.045	$[-0.087, -0.002]$	0.040	0.060	H2
C(primary_mode)[T.C] : log _params	-0.107	$[-0.179, -0.035]$	0.0035	0.010	H2
C(primary_mode)[T.PRESERVED] : log _params	-0.043	$[-0.089, +0.002]$	0.064	0.064	H2
is_intervention	-0.021	$[-0.065, +0.025]$	0.372	—	covariate
log _params (mode-A baseline slope)	$+0.049$	$[+0.015, +0.083]$	0.005	—	main effect

Table 5: Fixed-effect coefficients from the mixed-effects regression on the pooled calibration frame (534 rows: 48 items $\times$ 3 evaluatees $\times$ 4 conditions, minus skipped sentinels). Treatment-coded contrasts with mode A as the reference category. "p (raw)" is the per-contrast Wald p-value; "p (BH)" is the Benjamini–Hochberg-adjusted p-value within each hypothesis family ($\{$H1 contrasts, $m = 3\}$ and $\{$H2 contrasts, $m = 3\}$ adjusted separately). The H2 interaction coefficients are negative on all three contrasts, in the predicted direction: as $\log$ _params rises, the per-mode gaps shrink. The mode-C interaction clears BH at $p = 0.010$; the mode-B and mode-PRESERVED interactions are borderline ($p = 0.060$, 0.064). The is_intervention covariate is small and non-significant ($\beta = -0.021$, $p = 0.37$), indicating that pooling baseline and modified-prompt conditions does not substantively shift the score column once item-level variance is absorbed by the random intercept.