Cross-Format Representational Similarity Measures Output Form, Not Knowledge: A Negative Result on Diagnosing Format Exploitation

Tanner O'Rourke
The University of Texas at Austin
tannero@live.com

August 2026

Abstract

If a model answers a multiple-choice question by knowing the answer rather than by leaning on the options, its internal state before answering should not depend much on whether the options are present. We operationalize this intuition as FIRS, the cosine similarity between residual-stream activations of multiple-choice (MCQ) and open-ended renderings of the same MMLU question, captured in Gemma-2-2B at the final prompt token and averaged over mid-depth layers. The metric fails construct validity. Its headline correlation with open-ended success swings from AUC 0.698 [0.634, 0.760] to 0.530 [0.471, 0.588] on the same activations, depending only on whether "correct" requires the verbatim gold string or accepts a paraphrase of it; genuinely correct paraphrases carry *lower* FIRS than wrong answers. Four diagnostics locate the failure: the score predicts MCQ accuracy up to three times more strongly than the open-ended outcome it was built for; replacing the correct option's text with "None of the above" moves it by 0.002 while switching template moves it by 0.28–0.39; prompt geometry alone explains 27–28% of its variance; and on the MCQ-correct subset, the only place format exploitation is defined, it sits at chance under every scorer. The mechanism is structural: at the shared final token the two formats demand different output types, a letter versus a content word, so similarity there measures output form. We extract the transferable lessons and outline successor designs that decode answer content instead of comparing representational form.

1 Introduction

Multiple-choice benchmarks reward more than knowledge. Models answer MCQ items at well above chance without seeing the question (Balepur et al., 2024), change their answers when the options are reordered (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024), and lose double-digit accuracy when the same questions are asked open-ended (Myrzakhan et al., 2024). Leaderboard rankings shift under perturbations that leave item content untouched (Alzahrani et al., 2024; Sclar et al., 2024). The standard reading is that the option set does part of the work: it permits elimination, positional priors, and likelihood comparison, and a benchmark score cannot say which items were answered by knowledge and which by format.

A mechanistic test seems within reach. Probing work shows that hidden states carry knowledge the output does not (Azaria and Mitchell, 2023; Gekhman et al., 2025; Park et al., 2025; Gottesman and Geva, 2024; Orgad et al., 2025), so one might hope to read off, per item, whether the answer was present before the options arrived. The intuitive design compares the model's internal representation of a question with and without its options: a stable representation means the model held its

answer from the question alone; a large shift means the options did the work. Versions of this intuition circulate wherever hidden-state similarity is interpreted as consistency or knowledge; to our knowledge no one has built and validated the direct form.

We built it. FIRS, the Format-Invariant Representation Score, is the cosine between residual-stream activations of MCQ and open-ended renderings of the same question, taken at the aligned final prompt token and averaged over mid-depth layers. We evaluated it on 500 stratified MMLU questions under two elicitation regimes, with pre-committed scoring rules, a blinded judge, and a battery of construct-validity checks.

It does not measure format exploitation. The primary correlation is a function of the scoring rule rather than of the representations: strong under exact string match, chance once correct paraphrases are counted. The construct-validity checks fail independently of any correctness label, and they fail identically in both elicitation regimes. What FIRS measures is the shape of the prompt and the form of the required output. The failure has a clean structural cause and, we argue, could not have been repaired by tuning.

This paper is a negative result, and we present it as one on purpose. Hidden-state similarity metrics are cheap to build, intuitive to interpret, and publishable when they appear to work; the failure mode we document here, an apparently significant effect manufactured jointly by an output-space confound and a lexical scoring rule, is exactly the kind that survives peer review. Our contributions:

1. **The metric and pipeline**: a fully specified, reproducible cross-format similarity metric over four prompt renderings per question (§3, §4).
2. **A construct-validity falsification**: four diagnostics that condemn the metric without reference to its headline correlation (§5).
3. **The mechanism**: the score decomposes into output-form and difficulty components; the sharpest evidence is that correct paraphrases score below wrong answers (§6).
4. **The scorer-dependence trap**: exact-match versus semantic scoring flips a significant AUC to chance on identical representations (§5, §8).
5. **Lessons and successors**: transferable checks for representation-based evaluation metrics, and better-posed designs that decode content rather than compare form (§8, §9).

2 Related work

Format sensitivity of benchmarks. LLM benchmark scores are sensitive to surface features that leave content fixed: prompt formatting (Sclar et al., 2024), prompt paraphrase (Mizrahi et al., 2024), option order (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024), and benchmark-specific template choices (Alzahrani et al., 2024; Liang et al., 2023). For MCQA specifically, models exploit option-only artifacts (Balepur et al., 2024), and converting benchmarks to open-style form drops accuracy substantially (Myrzakhan et al., 2024). MMLU (Hendrycks et al., 2021) inherits these issues alongside item-quality problems (Gema et al., 2025); successors harden the format without removing it (Wang et al., 2024). This literature is behavioral: it compares accuracies across renderings. We asked whether the internal representation carries the same information per item.

Consistency of factual knowledge. Elazar et al. (2021) measured factual consistency across paraphrased queries and found pretrained models fluctuate on meaning-preserving rewordings, again at the output level. Our design replaces the paraphrase intervention with a format intervention and moves the measurement inside the model.

Internal knowledge versus expressed knowledge. Linear probes recover information from hidden states that generation does not express (Alain and Bengio, 2017; Azaria and Mitchell, 2023; Orgad et al., 2025); models rank correct answers highly internally while failing to produce them (Gekhman et al., 2025), mispredict MCQ answers they encode (Park et al., 2025), and their internal graded knowledge can be estimated without generation (Gottesman and Geva, 2024; Kadavath et al., 2022). The logit and tuned lenses read intermediate layers directly (nostalgebraist, 2020; Belrose et al., 2023). This line motivated FIRS: if knowledge is readable from the residual stream, its invariance across formats seemed a natural signature. The negative result here does not touch the probing line; it shows that *similarity* is the wrong functional of those same activations.

Faithfulness. Chain-of-thought explanations do not reliably reflect the computation that produced the answer (Turpin et al., 2023; Lanham et al., 2023). Format exploitation is the analogous gap for MCQA, with the option block rather than the rationale as the suspect input.

Representational similarity and its pitfalls. Comparing network representations is a mature subfield (Raghu et al., 2017; Kornblith et al., 2019; Klabunde et al., 2023) with known reliability failures (Davari et al., 2023), and interpretability methods more broadly admit illusions (Bolukbasi et al., 2021; Friedman et al., 2024). Cosine over raw hidden states carries a further, geometric liability: contextual representation space is anisotropic (Ethayarajh, 2019), and a few high-magnitude dimensions can dominate the similarity outright (Timkey and van Schijndel, 2021), so such a functional reads whatever contrast carries the most variance, here surface form, by default. We add a case study at the metric-design level: a similarity score whose variance is dominated by nuisance structure that a natural reading attributes to semantics.

Construct validity. Measurement-modeling work asks whether an operationalization measures its construct (Jacobs and Wallach, 2021; Subramonian et al., 2023; Raji et al., 2021). We apply that lens to a mechanistic metric and show the operationalization fails at every joint, which is the paper's methodological core.

3 The FIRS metric, as proposed

Each MMLU question is rendered in four formats sharing a trailing `Answer:` suffix, so the final prompt token is lexically identical across formats:

Format	Content
`mcq`	`Q: {question}` + four lettered options + `Answer:`
`oe`	`Q: {question}` + `Answer:` (options removed)
`mcq_shuffle`	`mcq` with the correct option rotated off its original position
`mcq_nota`	`mcq` with the correct option's *text* replaced by "None of the above"

`mcq_nota` keeps the correct option's position and the question's own three distractors, which are known-wrong by construction; it destroys exactly one thing, the correct answer's surface content.

At the final token position we capture `resid_post` at every layer l and define, per question q,

$$\textsc{Firs}[q] \;=\; \frac{1}{|W|}\sum_{l\in W}\cos(\mathbf{h}^{\text{mcq}}_{q,l},\,\mathbf{h}^{\text{oe}}_{q,l}), \tag{1}$$

with W the 50–80% depth window (layers 12–19 of 26), a hyperparameter sanity-checked against the layer-wise profile (Figure 1). Two companion scores use the same functional: $\textsc{Firs}_{\text{shuffle}}[q] = \cos(\text{mcq}, \text{mcq\_shuffle})$ for positional invariance and $\textsc{Firs}_{\text{nota}}[q] = \cos(\text{oe}, \text{mcq\_nota})$ for sensitivity to option content.

The claim under test: high Firs (small representational shift) marks genuine knowledge and should predict answering the open-ended rendering correctly; low Firs marks format exploitation. The design leans on one alignment assumption, that the shared suffix makes the final-token comparison clean. We verified that the final *token* matched across formats and treated the comparison as licensed. That verification checked the wrong thing, as §6 shows.

4 Experimental setup

Model and extraction. Gemma-2-2B (Gemma Team et al., 2024), a 26-layer base model ($d_{\text{model}} = 2304$), loaded in bfloat16 under TransformerLens (Nanda and Bloom, 2022). Activations are captured in a single forward pass per prompt, before any generation, so they are independent of decoding choices. Behavioral outputs are greedy: a single token for the MCQ family, up to 32 tokens for open-ended, scored on the first content line.

Data. 500 questions from the MMLU test split (Hendrycks et al., 2021), proportionally stratified by subject (seed 42). Eligibility filters drop items that break under the interventions: wrong option counts, duplicate option texts, and self-referential options such as "all of the above."

Elicitation conditions. Two regimes, all artifacts kept side by side. *0-shot* uses the bare templates. *5-shot* prefixes each prompt with five same-subject `dev`-split exemplars rendered in the target's own format, demonstrations rather than instructions, since instruction text would itself differ by format and become a template confound. The MCQ family shares one identical prefix per subject, so within-family prompts still differ only in the target's option block.

Open-ended scoring. The protocol was fixed before results were seen. Primary scorer: deterministic normalized token-boundary match of the gold answer within the continuation, with empty continuations scored incorrect. A stratified audit (about 25 match-positives for the false-positive rate, about 50 non-empty match-negatives for the paraphrase false-negative rate) decides sufficiency; a false-negative rate above 10% escalates to a hybrid: an LLM judge, blind to Firs and distinct from the subject model, re-scores only the non-empty match-negatives, so it can flip false negatives and nothing else. Both runs tripped the threshold (0-shot: FP 8%, FN 12%, 52 of 430 flipped; 5-shot: FP 4%, FN 14%, 57 of 427 flipped), so all results are reported under both scorers. That duplication was meant as a robustness formality. It became the finding.

Cuts and statistics. Results are reported over three cuts, because "wrong on open-ended" conflates three situations: *all* items, *non-empty* (the model attempted an answer), and *answerable* (self-contained stems; cloze items, Statement-1/2 pairs, and option-referential stems are unanswerable open-ended regardless of knowledge). Statistics: point-biserial r, logistic fit, and AUC with 10,000-sample percentile bootstrap CIs.

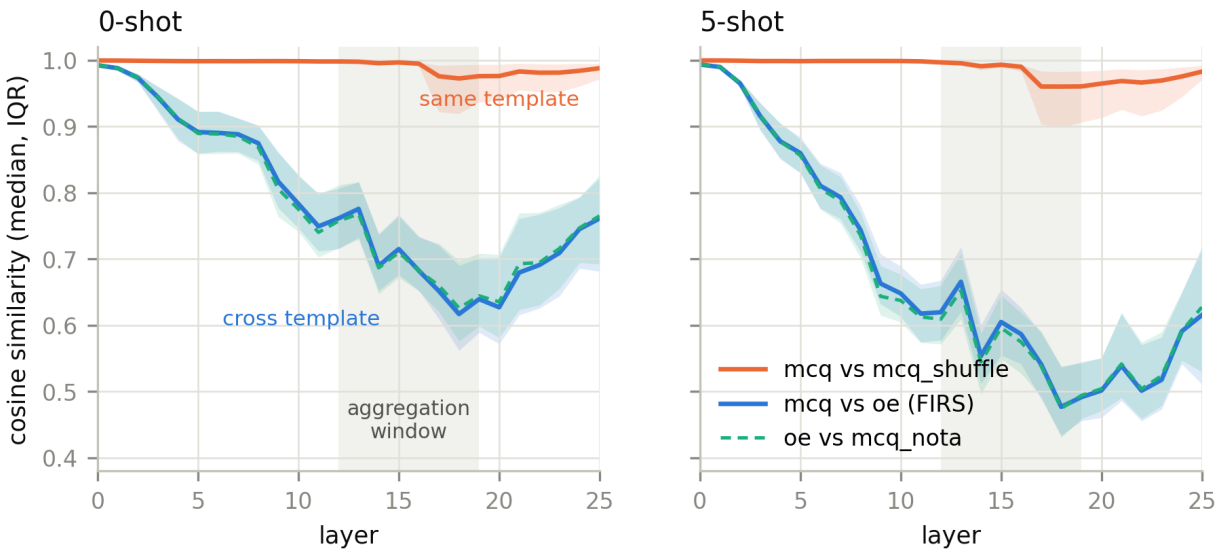

Figure 1: Layer-wise cosine profiles (median, IQR) for the three format comparisons, both conditions. Same-template pairs (orange) stay near 1.0 at every depth regardless of what the options say; cross-template pairs (blue) separate from layer 2 onward. The dashed `oe` vs `mcq_nota` profile is indistinguishable from FIRS despite the correct answer's text being destroyed, a preview of the option-content blindness quantified in §5. The shaded band is the aggregation window.

Behavioral panel. 5-shot elicitation works as intended: MCQ accuracy 0.560 with 100% letter parsing, all 500 open-ended continuations non-empty (0-shot: 0.542, 479 of 500). Shuffling options costs the 0-shot model 4 points of accuracy (0.542 to 0.502) and the 5-shot model nothing. When the correct answer is replaced by "None of the above," the model selects the NOTA letter only 7.8% of the time 0-shot and 18% 5-shot, behavioral evidence that it rarely registers the absence of the answer it supposedly knows.

5 Results: the metric does not measure format exploitation

5.1 The primary correlation is a property of the scoring rule

Table 1 and Figure 2 give the primary validation: FIRS as a predictor of open-ended correctness. At 0-shot the claim is unsupported under both scorers; every CI covers chance, and the two scorers' point estimates sit on opposite sides of it. At 5-shot the claim appears strongly supported under exact match (AUC 0.698 [0.634, 0.760]) and collapses once the judge adds the 57 correct paraphrases (0.530 [0.471, 0.588]); on the answerable subset it lands on the wrong side of chance (0.465). Nothing about the model or the activations differs between those rows. At this effect size, the choice of scoring rule moves the estimate more than any signal in the representations does, and a claim that survives only under the lexical scorer is not a claim about knowledge.

The remaining four results are construct-validity diagnostics, and all four replicate across both elicitation conditions (Table 2).

Table 1: Primary validation across elicitation conditions, scorers, and cuts. The claim predicts AUC above 0.5. The single supported cell (5-shot, exact match) is the one whose labels reward lexical overlap with the gold string; §6 shows that is not a coincidence.

Condition	Scorer	Cut	n	OE-correct	r_{pb}	AUC [95% CI]
0-shot	match	all	500	49	−0.022	0.521 [0.447, 0.594]
0-shot	match	answerable	301	37	+0.070	0.579 [0.492, 0.664]
0-shot	judge	all	500	101	−0.088	0.446 [0.384, 0.508]
0-shot	judge	answerable	301	65	−0.099	0.436 [0.361, 0.513]
5-shot	match	all	500	73	+0.293	**0.698 [0.634, 0.760]**
5-shot	match	answerable	301	39	+0.176	0.652 [0.571, 0.731]
5-shot	judge	all	500	130	+0.098	0.530 [0.471, 0.588]
5-shot	judge	answerable	301	79	−0.043	0.465 [0.394, 0.536]

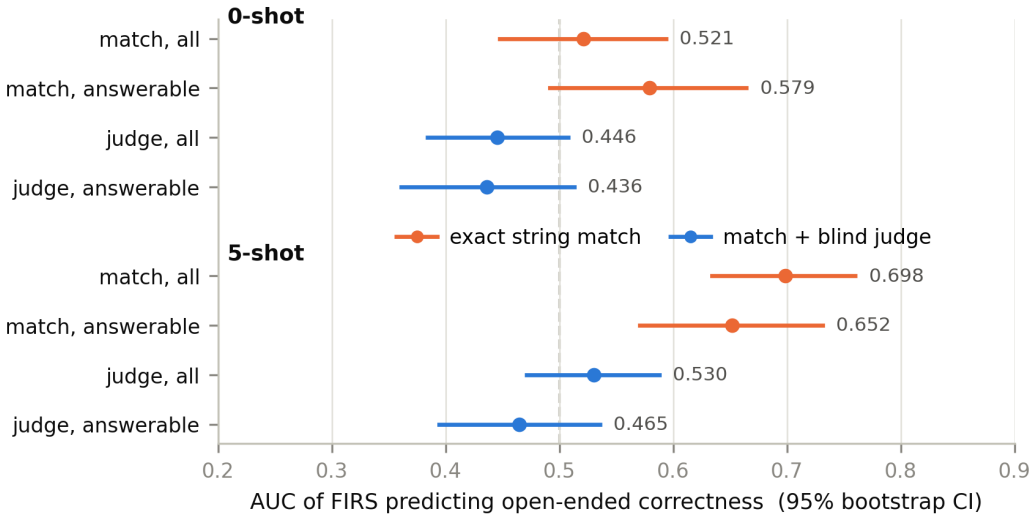

Figure 2: The primary AUC with bootstrap CIs. The apparent 5-shot effect exists only under exact string match; counting correct paraphrases as correct returns it to chance on the same activations.

5.2 Diagnostic 1: not format-specific

A format-exploitation score must track open-ended fragility specifically. Instead FIRS predicts plain MCQ accuracy three times more strongly than the open-ended outcome at 0-shot ($r = -0.252$ versus -0.088), and twice as strongly with the opposite sign at 5-shot (-0.195 versus $+0.098$). It predicts shuffled-MCQ accuracy just as well. A score that anticipates every behavioral outcome about equally is tracking general item difficulty, not the construct.

5.3 Diagnostic 2: template-dominated and blind to option content

Switching template moves the mean cosine by 0.277 (0-shot) and 0.389 (5-shot). Destroying the correct option's text moves it by 0.002 and −0.007 respectively: $\text{FIRS}_{\text{nota}}$ correlates with FIRS at 0.988 and 0.980, and its layer profile lies on top of the FIRS profile in Figure 1. Any two MCQ-shaped prompts sit near 0.97 regardless of what their options say; any MCQ/OE pair sits far below regardless of shared content. The score reads the shape of the prompt, and the one intervention that changes what the options *mean* barely registers. The 5-shot template gap is *larger*

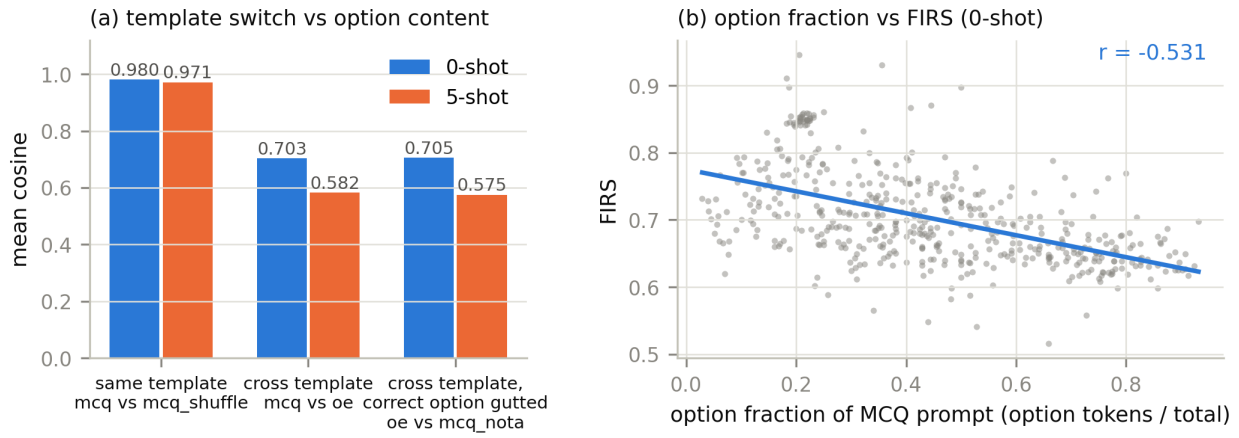

Figure 3: (a) Mean cosine by comparison pair. The template switch dwarfs the option-content intervention in both conditions. (b) FIRS against the option fraction of the MCQ prompt; prompt geometry alone explains 28% of the score's variance.

because the few-shot prefixes also differ by format; cleaner elicitation strengthened the confound.

5.4 Diagnostic 3: contaminated by prompt geometry

The option fraction of the MCQ prompt (option tokens over total tokens) correlates with FIRS at $r = -0.531$ (0-shot) and -0.518 (5-shot), explaining 27–28% of its variance in both conditions (Figure 3b). This is a construct violation independent of any outcome: a knowledge signature should not be predictable from how long the options are. Partialling geometry out of the 5-shot judge-scorer correlation barely moves it (+0.098 to +0.108), so the weak positive there is not itself a geometry artifact; it is simply weak, scorer-fragile, and absent on the answerable cut.

5.5 Diagnostic 4: at chance where the construct is defined

Format exploitation is only defined for items the model answers correctly *with* options: an exploiter is MCQ-correct but open-ended-wrong, a knower is correct both ways. This subset is the metric's entire reason to exist, and on it FIRS separates knowers from exploiters at AUC 0.469–0.491 across every cut, scorer, and condition (Table 3). The marginal 5-shot match effect of §5.1 and this conditional null reconcile cleanly: FIRS distinguishes open-ended-correct from open-ended-wrong only *across* the MCQ-right/MCQ-wrong divide, which difficulty and output form already explain, and not *within* the MCQ-correct set where the construct actually lives.

6 The output-space mechanism

The failure is structural, and it was visible in the design. At the scored position the MCQ prompt must emit a letter and the open-ended prompt must emit a content word. The cosine between their representations therefore measures, in large part, letter-mode versus word-mode. The two prompts are token-identical up to the question and diverge only after it, which leaves every capture position with one of two defects: before the divergence, activations are identical by causal attention and the cosine is exactly 1.0, carrying no information; after it, the required output types differ and the comparison is contaminated. No hook placement escapes this dichotomy. The alignment check of §3 verified the wrong invariant: the final token matched across formats; the final token's *task* did not.

Table 2: Construct-validity diagnostics, both conditions. All quantities except the OE-correct row of Diagnostic 1 are computed without open-ended labels.

	0-shot	5-shot
Diagnostic 1: format specificity		
r(FIRS, MCQ correct)	−0.252	−0.195
r(FIRS, shuffled-MCQ correct)	−0.248	−0.216
r(FIRS, OE correct, judge)	−0.088	+0.098
Diagnostic 2: template dominance		
mean cos, same template (`mcq`, `mcq_shuffle`)	0.980	0.971
mean cos, cross template (`mcq`, `oe`)	0.703	0.582
mean cos, cross template, correct option gutted	0.705	0.575
template gap / option-content gap	0.277 / 0.002	0.389 / −0.007
r(FIRS, $\text{FIRS}_{\text{nota}}$)	0.988	0.980
Diagnostic 3: prompt geometry		
r(FIRS, option fraction)	−0.531	−0.518
variance explained	28%	27%

Table 3: Diagnostic 4: the claim conditioned on MCQ-correct items, where knower (also correct open-ended) must be separated from exploiter (wrong open-ended). Judge labels; the match-scorer cells are equally null.

Condition	Cut	n	knower / exploiter	r_{pb} (p)	AUC
0-shot	MCQ-correct	271	79 / 192	−0.022 (0.72)	0.491
0-shot	+ attempted OE	259	79 / 180	−0.027 (0.67)	0.490
0-shot	+ answerable	168	51 / 117	−0.077 (0.32)	0.470
5-shot	MCQ-correct	280	82 / 198	−0.025 (0.68)	0.485
5-shot	+ answerable	175	54 / 121	−0.032 (0.68)	0.469

Removing the output-mode direction does not rescue the metric. If output mode were a separable additive component, removing it might expose a knowledge signal. In the 0-shot condition the MCQ-to-OE difference is indeed dominated by a single shared direction, which explains 48–61% of the difference energy across the window; projecting it out raises the mean cosine from 0.703 to 0.845 but leaves every correlation in Table 2 essentially unchanged. What remains after the projection is still general difficulty, still not format-specific.

The paraphrase gap. The cleanest evidence comes from partitioning the 5-shot open-ended answers into three groups (Figure 4): answers matching the gold string verbatim (mean FIRS 0.641), correct paraphrases recovered by the judge (0.538), and judge-confirmed wrong answers (0.577). A knowledge metric must score the paraphrase group high; these are answers the model got *right*. FIRS scores them lowest of all, below the wrong answers, because their surface form departs from the gold string. This is the final-token confound reappearing one level up: the exact-match scorer and FIRS reward the same thing, output whose form resembles the gold answer, which is why the AUC 0.698 cell in Table 1 exists at all.

A repair attempt fails. The natural fix is to move the options before the question, so both formats end with the identical suffix `Q: {question}\nAnswer:` and share an output space. The model then regurgitates the option block as its continuation instead of answering. The redesign did not merge the two output spaces; it created a third. We take this as evidence the confound is a

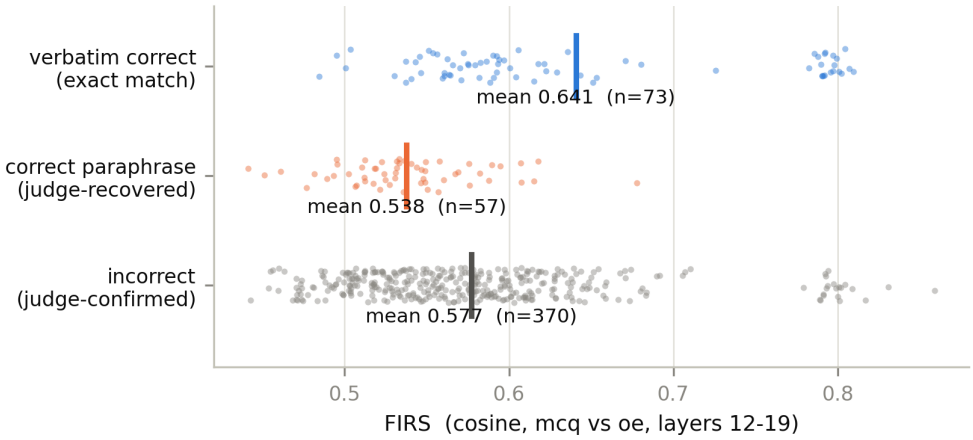

Figure 4: 5-shot FIRS by open-ended outcome group. Correct paraphrases, the one group whose members demonstrably knew the answer without options, carry the lowest scores, below the wrong answers. A metric measuring knowledge cannot produce this ordering; a metric measuring output form must.

property of the metric family, not of our template.

7 Elicitation is a real problem, and a separate one

Gemma-2-2B is a base model, and 0-shot `Q:/Answer:` prompting produces document-continuation pathologies: boilerplate continuations, copied option blocks, and answers opening with a newline. Five-shot demonstrations fixed all of it, which confirms the degeneracy was framing rather than a capability ceiling. It also made the construct failure sharper, not better: the template gap grew (Table 2), and the scorer swing of §5.1 only became visible once the model answered cleanly enough for paraphrases to appear.

The 0-shot condition also supplied a cautionary artifact. Our first extraction truncated open-ended continuations at the first newline; because the model routinely opens with one, this discarded roughly nine in ten answers, and its "non-empty" cut silently meant "answered without a leading newline." That run produced a statistically significant *reversal*, AUC 0.334, which we initially found credible on the reasoning that label noise attenuates effects toward chance and cannot manufacture one. The reasoning is true of noise and was irrelevant here: the bug was not noise but a selection rule correlated with the outcome, and repairing it sent the effect to chance. We kept the episode in the paper because the inference pattern, "the effect strengthened as labels improved, so it is real," is one we expect others to find as persuasive as we did.

8 Methodological lessons

Token identity is not task identity. Aligning the compared position lexically is not the requirement; aligning what the model must *do* at that position is. Any cross-format comparison at a shared suffix inherits the output-space confound whenever the formats demand different continuation types, and the check that would have caught it, asking what token comes next in each format, costs nothing.

A scorer-fragile effect is no effect. Exact-match and semantic scoring produced AUC 0.698 and 0.530 from identical representations. Any result in this genre should be reported under both a lexical and a semantic scorer, and an effect that does not survive the semantic one should be treated as a property of the scoring rule.

Label repair is not evidence. Monotone strengthening under improving labels supports an effect only when the label errors are noise. When the error is a selection rule correlated with the outcome, repair can strengthen, preserve, or reverse the effect, and the direction of movement carries no evidential weight.

Similarity conflates content with form. The question "does the model know the answer" is a decoding question, and similarity functionals answer a different one: "do these prompts put the model in the same output mode." The probing literature already measures knowledge by what is *recoverable* from a representation; our result is a demonstration of what goes wrong when invariance is substituted for recoverability.

Run the construct checks first. All of Table 2 except one row is computable without any open-ended label, in minutes, from the same forward passes that produce the metric. Had we ordered the analysis that way, the template-dominance and geometry results would have condemned the design before the primary correlation was ever estimated.

9 Toward a metric that could work

The reframe that survives this paper: stop asking how similar two representations are and ask whether the answer content is recoverable, and whether the options assist the readout. Four designs follow, in rising order of evidential strength.

Content probing on the options-free representation. Decode the answer *text* from the open-ended representation via logit lens or a trained probe. Our own planned probe targeted the correct *letter* and is invalid by construction, since the open-ended prompt never sees the options and the letter is information the representation cannot contain; the target must be content.

Answer-text likelihood. Score each option's text as a continuation of the bare question and rank. This operationalizes the original intent, needs no representational comparison, and is format-robust because identical strings are scored under identical conditions. It also connects to the finding that direct probability measurement beats metalinguistic prompting (Hu and Levy, 2023) and to symbol-binding analyses of MCQA (Robinson and Wingate, 2023).

Graded format assistance. Bucket MCQ-correct items by the rank of the correct option's text in that options-free distribution: rank 1, the model knew it; ranks 2–3, the options lifted it; rank 4 or absent, position or guessing. This turns the binary construct that failed here into a graded, behaviorally defined mechanism, and it runs on artifacts this pipeline already produces.

Activation patching. Transplant the open-ended hidden state into the MCQ forward pass and test whether the correct letter emerges (Meng et al., 2022; Zhang and Nanda, 2024). This is the causal form of the question FIRS asked correlationally, and it is much harder to fool with output-mode structure.

Whatever the internal metric, its validation anchor should be paraphrase-consistency behavior under a semantic scorer (Elazar et al., 2021; Mizrahi et al., 2024), precisely because of the trap in §5.1.

10 Limitations

The study is one model (a 2B base model), one benchmark, 500 items, and one metric family. The structural argument of §6 does not depend on scale or instruction tuning, and the diagnostics replicate across both elicitation regimes, but the quantitative surface (a 0.98 same-template plateau, a 28% geometry share) may not transfer. The aggregation window and the choice of cosine follow the metric as proposed; alternatives within the same family inherit the same final-token dichotomy. The judge is an LLM, blinded and audited (false-positive rates of 4–8% on match-positives) but not human-verified end to end. Finally, this is a negative result about final-token cross-format similarity; it does not show that no representation-based format metric can work, and §9 is our argument that some can.

11 Conclusion

Final-token residual-stream similarity between multiple-choice and open-ended renderings of a question measures output form, prompt shape, and item difficulty, not knowledge. Built as specified, the score tracks MCQ accuracy up to three times more strongly than the open-ended outcome it targets, is blind to the destruction of the correct option's text, inherits 27–28% of its variance from prompt geometry, and is at chance on the MCQ-correct subset where format exploitation is defined. Its one apparently supportive result is manufactured by an exact-match scorer that rewards the same surface property the metric does. The intuition that representational invariance across formats certifies knowledge is attractive, quietly load-bearing in how hidden-state similarity gets interpreted, and wrong for a reason that generalizes: similarity reads form, and knowledge is a question about content. The successor question worth building is not how far the representation moves when the options appear, but whether the answer was already recoverable before they did.

Reproducibility. Code, prompts, scoring protocol, judge verdicts, and all statistics are available at `https://github.com/tannerorourke/Format-Invariant-Benchmarking`. Every number in this paper regenerates from saved artifacts without re-running the model.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL `https://arxiv.org/abs/1610.01644`.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL `https://arxiv.org/abs/2402.01781`.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. URL https://arxiv.org/abs/2304.13734.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10308–10330, 2024. URL https://aclanthology.org/2024.acl-long.555/.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT. *arXiv preprint arXiv:2104.07143*, 2021. URL https://arxiv.org/abs/2104.07143.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.16156.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. URL https://arxiv.org/abs/2102.01017.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://arxiv.org/abs/1909.00512.

Dan Friedman, Andrew Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2312.03656.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in LLMs. *arXiv preprint arXiv:2503.15299*, 2025. URL https://arxiv.org/abs/2503.15299.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, et al. Are we done with MMLU? In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. URL https://aclanthology.org/2025.naacl-long.262/.

Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://aclanthology.org/2024.emnlp-main.232/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5040–5060, 2023. URL https://aclanthology.org/2023.emnlp-main.306/.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. URL https://arxiv.org/abs/1912.05511.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint arXiv:2305.06329*, 2023. URL https://arxiv.org/abs/2305.06329.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1905.00414.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. URL https://arxiv.org/abs/2401.00595.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-LLM-Leaderboard: From multi-choice to open-style questions for LLMs evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024. URL https://arxiv.org/abs/2406.07545.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.02707.

Yoonah Park, Haesung Pyun, and Yohan Jo. Bridging the knowledge-prediction gap in LLMs on multiple-choice questions. *arXiv preprint arXiv:2509.23782*, 2025. URL https://arxiv.org/abs/2509.23782.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024. URL https://arxiv.org/abs/2308.11483.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1706.05806.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2111.15366.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.12353.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11324.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. URL https://arxiv.org/abs/2305.09022.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://arxiv.org/abs/2109.04404.

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 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.04388.

Yubo Wang, Xueguang Ma, Ge Zhang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.01574.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.16042.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.03882.