ICONIC FORMAT COSTS LANGUAGE MODELS ACCURACY: A PREREGISTERED THREE-FORMAT STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

A representation is iconic when its arrangement mirrors the arrangement of what it represents. We ask whether that difference changes how a language model reasons when the represented information is held fixed. One structure is rendered three ways, as a natural language sentence, as an iconically formatted ASCII diagram, and as a tabular fixed-width table that is discursively structured with superficial diagrammatic characteristics. Each model undergoes the same test battery, consisting of 240 items across four task families for a total of 21,600 preregistered trials over five models. Task performance on iconic renderings is worse than on tabular ones by a wide margin ($b = -0.92$ in log-odds under the registered scorer, $p = 1.6 \times 10^{-36}$; -0.60 when letter-key answers are resolved), while tabular and discursive renderings are indistinguishable ($b = -0.01$, $p = .88$). The deficit holds under all eighteen registered scoring and exclusion variants, is largest in one task family, and narrows to -0.40 log-odds when the models are asked for the answer alone rather than brief reasoning. Because the tabular format is also monospaced, symbol-dense and two-dimensional, whatever makes the iconic task hard is not something it shares with a (discursively structured) table. We also report a methodological result. Two probes were built on one model, Qwen2.5-1.5B, to ask whether it represents spatial content iconically, and neither succeeded in telling apart an iconic representation from a propositional one. This result suggests a larger moral: a linear readout is blind to any difference of organization that amounts to an invertible linear re-encoding, and hypotheses about format can share their content and differ in just this way, in which case linear probes cannot discriminate them. We give a demonstration and advance a simple control for detecting this.

Preregistration filed 7 September 2026; the link is withheld for anonymous review.

1 INTRODUCTION

It is a common claim in the cognitive sciences that mental representations come in different *formats*. A representational format is a way of organizing information that constrains a representation's computational profile, two representations may carry the same content but differ in how they organize that content and consequently differ in how they are used. In the cognitive sciences, much of the discussion concerning formats concerns discursive and iconic representational formats. Discursive representations, paradigmatically those of natural language, have a privileged constituent structure: they decompose into predicates and arguments that recombine in canonical ways. Iconic representations, by contrast, represent by *mirroring*—relations among the representation's parts reflect relations among the parts of what it represents. A typical map is a standard example of a representation that mirrors: distances between marks on the page track, in miniature, distances between the places marked. Iconicity allows an agent to "read off" information from the representation in a relatively cheap fashion, rather than infer it through a series of (possibly expensive) computational steps (Larkin and Simon, 1987).

Probing studies report that language models represent spatial structure linearly (Gurnee and Tegmark, 2024), but whether any such representations are in iconic format has not, to our knowledge, been tested. These studies aim to make progress on these questions. Study 1 evaluates three open-weight and two frontier models on four families of reasoning tasks, `contain`, `grid`, `graph`

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and `order`, each presented in informationally equivalent but differently formatted versions. Study 3 asks the same models to do the same tasks, but with chain-of-thought removed. The studies are numbered as in the registration, so Study 3 is reported before Study 2.

In the case of language models, it is not enough to run a study comparing diagrams with prose, as performance on the diagrammatic task may be confounded. Specifically, a diagram is input to a language model as ASCII art, which differs from prose in being monospaced, symbol-dense and laid out in two dimensions, so a deficit on diagrams could be explained by those surface properties without appeal to iconicity. The design therefore adds a third, *tabular* format that is ASCII-shaped without being iconic: a fixed-width table is monospaced, symbol-dense and laid out in two dimensions, but its arrangement does not mirror what it represents: rather, it is a discursive representation with superficial iconic form.

We find that the five models, pooled, perform worse on iconically formatted tasks than on discursive or tabular ones, by −0.92 log-odds against tables in Study 1 under the registered scorer (−0.60 under key resolution), with −25.1 percentage points in the `graph` family and −6.0 (key-resolved) to −13.0 (registered) pooled over the other three, in the interpreted frame (Appendix D). The frontier models are at or near ceiling outside `graph`, so these tasks are uninformative about them there. That said, the frontier models also do worse on the iconically formatted `graph` tasks. Removing chain-of-thought reasoning narrows the performance gap, pooled across all four task families, from −0.92 to −0.40 log-odds. That the deficit survives a table matched on these surface properties indicates an iconic deficit, though it does not show one conclusively, since the diagrams also differ from the tables in ways other than iconicity (§9).

That said, because Studies 1 and 3 are behavioral, they tell us only whether these models reason successfully on iconic tasks, irrespective of the format the reasoning takes. Study 2 probed the residual stream (the vector of activations that each layer reads from and adds to) of one model, Qwen2.5-1.5B, to address the question of format directly. The study was uninformative about format rather than negative (Appendix D, entry 5), in that the results were determined by the choice of readout, but it still supplied an important methodological result: linear decoding cannot discriminate between a model's having iconic representations and its having a discursive representation of the same spatial facts when the two differ by an invertible linear re-encoding. Consequently, one of the most frequently used interpretability methods—linear probing—cannot by itself settle fine-grained questions about representation-possession and use in language models.

2 Related work

The distinction between iconic and discursive representation has a long history. Peirce (1867, 1903) distinguished icons, which represent by resembling what they stand for, from symbols, which represent by convention, and Goodman (1968) analyzed what makes a system of marks work as a *symbol system*, and how different symbol systems organize their information differently. In cognitive science the same distinction drove the debate over whether mental images are picture-like or are descriptions in a language of thought (Pylyshyn, 1973, 2002), and which formats the mind uses is still debated (Quilty-Dunn et al., 2023; Quilty-Dunn, forthcoming).

Larkin and Simon (1987) draw the fundamental distinction that this study relies on, called *informational equivalence*. Two representations are informationally equivalent when each can be recovered from the other, and computationally equivalent when, in addition, the same conclusions are as easy to draw from each. This study asks language models to solve informationally equivalent but computationally inequivalent tasks. These studies are built on the idea that diagrams and sentences have different computational profiles: drawing some facts into a diagram makes other facts visible for free, so they can be read off rather than worked out (see also Shimojima, 2015). So, other things equal, an iconic advantage or disadvantage would be evidence for the use or non-use of iconic representations on these tasks; §9 lists the ways in which other things are not equal here.

The study also concerns linear probing methods. Probing trains a simple classifier on a model's internal states to see what information they carry. What a probe appears to show also depends on how its performance is measured (Hall Maudslay et al., 2020; Voita and Titov, 2020), and Harding (2023) proposes criteria for when a model can be said to represent something at all. Neuroscience has faced the same questions about decoding brain activity (Kriegeskorte et al., 2008; Naselaris

and Kay, 2015), including whether information that can be decoded is information the brain uses (Ritchie et al., 2019). Lee and Kondor (2026) argue that a probe should give the same answer however a representation's coordinates are chosen. §6.1 raises the other side of that point: a probe that is indifferent to how a representation is coded cannot tell us which format it is in.

3 DESIGN

There are four task families, `contain` (nested containers), `graph` (an undirected network), `grid` (objects on a 2-D lattice) and `order` (objects on a 1-D scale), with three distinct renderings for each. There were sixty items per task family, each rendered in all three formats and posed under three paraphrases of the instruction, giving 2,160 trials per model per study. Every item is generated by a seeded program from a single structure, and all three renderings are functions of that structure, so none can add information to it. The generator was written with Claude Opus 5 from author-provided samples, and its items were hand-checked for accuracy by the study authors. Appendix A gives full prompts.

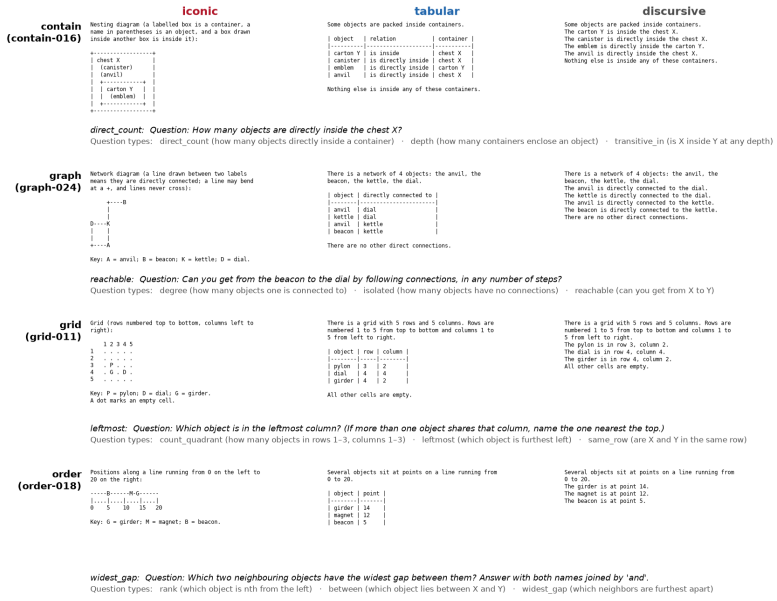


Figure 1: Examples of each task family, and each rendering (iconic, tabular, and discursive). One item per family; the question shown is one of the family's three question types, all of which are listed beneath it.

Information equivalence is built into the generator: one structure per item, each fact stated exactly once, framing and closure shared between the propositional formats, and every diagram object recoverable from the drawing. The registered equivalence tests compared counts and shapes rather than values (Appendix D, entry 16), and Appendix D lists the known asymmetries between renderings (entries 3, 4 and 14). They run in both directions: the legend of the contain diagram and the order diagram's statement of which end is left both favor the diagram.

4 METHODS

Five models were used in the studies: Qwen2.5-1.5B, Qwen2.5-7B and Llama-3.1-8B with open weights, and Claude Opus 4.8 and GPT-5.6 Sol through their APIs. Llama-3.1-8B was loaded with 4-bit quantized weights. The open-weight models used greedy decoding. They also applied a repetition penalty, which discourages a model from repeating words it has already produced, that we did not set: unless told otherwise, the software loads generation settings from each model's published files, and these supply a penalty of 1.1 for Qwen2.5-1.5B, 1.05 for Qwen2.5-7B and none for Llama-3.1-8B. Because the penalty differs by model, it could in principle produce the deficit. It

is not needed to produce it: in Study 1 the iconic deficit against tables is largest for Llama-3.1-8B, which has no penalty (17.3 percentage points under key resolution and 23.4 under the registered scorer, against 4.4 and 7.0 for Qwen2.5-1.5B, which has the largest penalty; Appendix C). Fitted to Llama-3.1-8B alone (unregistered), the contrast is b = −2.17 (SE 0.20) under the registered scorer and −1.61 (SE 0.19) under key resolution. Meanwhile, the hosted models reject sampling parameters and reason internally, so internal reasoning was disabled and repeat-run agreement was measured rather than assuming determinism: on a fixed subset of 144 trials run twice, the scored answer differed on 2 (1.4%) for Claude Opus 4.8 and 4 (2.8%) for GPT-5.6 Sol. Completions were capped at 256 tokens for the open-weight models and 512 for the hosted ones, in both studies. The hosted models were configured as claude-opus-4-8 and gpt-5.6-sol and queried on 7 September 2026 (dates from file timestamps).

Study 1 asks for brief chain-of-thought reasoning and Study 3 asks for the answer alone; the battery, seeds and items are identical and only the wrapper differs. Answers were recovered in four ways: from an `ANSWER:` line, which covers most trials; from the last non-empty line where no such line is present; as a non-answer, scored incorrect; or as a truncation, when generation hit the maximum token cap before giving an answer, which is the only exclusion (Appendix B). Truncation is 11.6 percent on Llama-3.1-8B in Study 1 and near zero elsewhere, and is ordered by format, which triggers the registered sensitivity analysis in which truncated trials are scored incorrect. Both are reported throughout. The model is a mixed-effects logistic regression with crossed random intercepts for item, task family and model; §5.1 also reports the contrast with a by-model random slope, which treats the models as a sample.

5 Results

5.1 Iconic, tabular, and discursive task performance

	iconic	tabular	discursive
Study 1	0.678	0.779	0.781
Study 3	0.644	0.693	0.688

The first major hypothesis of these studies predicted two differences, without predicting their direction. Performance on iconic renderings is worse than on tabular ones by $b = -0.92$ in log-odds (SE 0.073, p = 1.6×10^{-36}; five models, registered scorer, all cells). With a by-model random slope (unregistered), which lets the deficit differ across models, it is −0.77 (SE 0.25, p = .002). Under key resolution the same fit gives −0.15 (p = .77), but only because Claude Opus 4.8's errors on tables all fall in the contain direct-count cell, where the legend asymmetry favors the diagram (Appendix D, entry 4); in the interpreted frame, which drops that cell, it gives −1.17 (SE 0.16, p = 6×10^{-14}), and −1.13 to −1.40 leaving out any one model. Performance on tabular and discursive renderings differs by $b = -0.01$ (p = .88), or 0.17 of a percentage point, so the predicted difference between them was not found.

The choice of scorer changes the size of the deficit but not its sign. Both alternative scoring rules can change only iconic answers, so tabular and discursive performance stays the same at 0.779 to 0.780 and 0.781, while iconic performance moves from 0.678 to 0.719. Six of the original ten percentage points of difference remain after key resolution and name recovery (b = −0.60, SE 0.073, p = 2.7×10^{-16}). Where we interpret a magnitude, we interpret the key-resolved estimate, for the reason given in §7.

5.2 Variation by task family

family	iconic	tabular	discursive
`contain`	0.783	0.751	0.728
`graph`	0.641	0.834	0.843
`grid`	0.696	0.771	0.776

family	iconic	tabular	discursive
`order`	0.587	0.761	0.777

The second major hypothesis of these studies is that the size of the iconic effect varies by task family, and is largest in `grid` and `order`, where an answer can most easily be "read off" an arrangement. The variation (in the table: Study 1, five models, registered scorer) is significant, $\chi^2(3) = 200.7$, p $= 2.9 \times 10^{-43}$, but it is not where it was predicted: the effect is largest in `graph`, at an interaction of −2.38 against contain, with order close behind at −2.13 and `grid`'s −1.16. Under key resolution the raw order deficit falls from 17.4 to 5.6 percentage points, so most of its size under the registered scorer comes from the letter key. So the effect does vary by family, and only partly as predicted. This does not show that the deficit is present in some families and absent in others: under key resolution most of the variation is in `graph`, and whether the remaining families differ from one another depends on the scorer and the exclusion frame, so we do not claim that they do. It should also be said that the apparent iconic advantage on `contain` is an artifact of the legend asymmetry (Appendix D). There is also no capability ordering in the per-model estimates. Outside `graph` and under key resolution, the Study 1 gap runs from +0.0 percentage points for Claude Opus 4.8 to −12.8 for Llama-3.1-8B, and the largest deficit belongs to a mid-capability model.

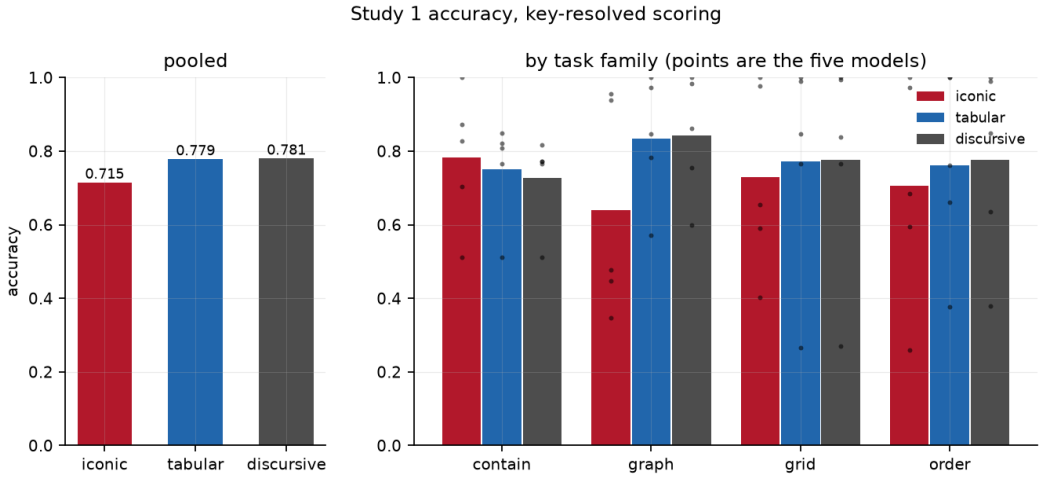


Figure 2: Accuracy by format, pooled and by family, under key-resolved scoring. Points are the five models.

5.3 Prompt length

The third hypothesis was that the format effects are not artifacts of prompt length. Prompt length does not account for the iconic deficit, as predicted. With token count as a covariate the iconic contrast is essentially unchanged, at −0.97 against −0.92. In the registered analysis longer prompts go with higher accuracy (b = +0.31, p $= 4.2 \times 10^{-5}$), but whether that association holds depends on which cells are included, so we do not interpret it.

5.4 Removing reasoning

The fourth major hypothesis of the studies was that removing chain-of-thought reasoning should increase the iconic deficit. The interaction between format and reasoning condition, across Studies 1 and 3, is significant, $\chi^2(2) = 41.0$, p $= 1.2 \times 10^{-9}$, but the predicted direction is not supported—removing reasoning shrank, but did not eliminate, the pooled iconic deficit. Under the registered scorer the deficit is −0.92 in Study 1 against −0.40 in Study 3 (under key resolution, −0.60 against about 0.00). The interaction is −0.52 (SE 0.098, p $= 1.5 \times 10^{-7}$), and the corresponding interaction for the tabular − discursive contrast is null (−0.05, SE 0.10, p = .62).

The same interaction can be read as the value of reasoning in each format. Outside `graph` and under key resolution, asking for reasoning is worth +11.2 percentage points with sentences, +9.3 with a table and +2.3 with a diagram (Figure C2). This keeps its sign and significance under both scorers, in the frame retaining the disputed cells, and under a by-model random slope. With the reasoning channel closed, there is no iconic deficit outside `graph` under key-resolved scoring (+0.10 log-odds, p = .31). It should be said that removing reasoning may shrink the iconic deficit only because it harms the discursive tasks more than the iconic one, so this is not necessarily evidence against there being an iconic deficit.

5.5 THE GRAPH DEFICIT

`graph` carries a Study 1 deficit of −25.1 percentage points against −6.0 outside it (five models, interpreted frame, key-resolved), and it is the one family whose iconic rendering the open-weight models appear not to read. The `degree` question asks how many edges meet a named node, a quantity every rendering supplies. The correlation between the number a model gives and the true degree, under key-resolved scoring, is:

	Claude Opus 4.8	GPT-5.6 Sol	Llama-3.1-8B	Qwen2.5-7B	Qwen2.5-1.5B
iconic	+0.940	+0.951	**+0.110**	**+0.302**	**−0.235**
tabular	+1.000	+0.991	+0.980	+0.924	+0.811
discursive	+1.000	+1.000	+0.945	+0.981	+0.745

Every model recovers degree from either propositional rendering, at $r \geq 0.745$ in all ten cells. From the drawing, two open-weight models do not recover it and the third recovers it weakly. The frontier models recover it well, but less well than from a table, and they undercount high-degree nodes. Whether this is a failure to read the drawing, or a failure in what the model does with what it read, is not identified here.

6 STUDY 2: RESIDUAL-STREAM PROBES

Ridge probes (linear regressions with a penalty on the size of the weights) were fitted from a model's residual stream to structural targets of the prompt, to ask whether spatial content is linearly available and whether its availability differs by format.

It was found that the choice of readout determines the results: mean-pooling over prompt tokens and reading the last token give different results. A target that appears available under one readout is unavailable under the other. Peak R^2 also does not distinguish a representation from a surface feature, so every probe is reported against a ceiling estimated from the prompt's own tokens. Against that ceiling, the only cells in which the diagram clears its ceiling by more than both propositional renderings are the identity-free occupancy maps in `grid` and `order`, the two families whose structure is arranged on a line or a plane. The `grid` cell also passes a harder test, in which the probe must beat a readout of the same content without the arrangement (0.223 against 0.043 R^2 for a bag-of-tokens readout; the tabular and discursive cells fail it, at 0.053 against 0.122 and 0.049 against 0.131). A later follow-up found that content-independent pooling of the prompt's own tokens could supply this cell entirely (bound 1.000), so the cell cannot show that the model computed the map (Appendix E.8). That test was run for `grid` only, and for `order` the layer profile points the other way (Appendix E.7). Study 2 therefore finds occupancy available in `grid` without showing that the model computed it, leaves `order` unresolved, and supports no ordering over formats; Appendix E gives the per-cell results, the positive controls and two limits that apply to every probe result. The study does not locate the accuracy effect. We report it because its limitation applies beyond this design.

6.1 LINEAR PROBING AND LANGUAGE MODEL REPRESENTATION

Linear decodability is one of the most popular methods to inquire into what language models represent. The use of this strategy faces two familiar difficulties: probes can memorize, which control

tasks address (Hewitt and Liang, 2019), and decoded information may be causally inert, which is the concern of most probing critiques (Ravichander et al., 2021; Elazar et al., 2021; Belinkov, 2022). This study reveals a third type of difficulty with interpreting the results of linear probing. The present difficulty concerns what a probe can detect about the format and/or encoding a representation uses, as opposed to the content it carries.

More specifically: a linear readout is blind to differences of encoding that amount to an invertible linear map, sensitive to non-linear ones, and silent on which kind separates two format hypotheses. Suppose every residual-stream vector r, at every layer and position, were replaced by Ar for some invertible matrix A. A linear probe with weights w on the original stream is matched exactly by a probe with weights wA^{-1} on the re-encoded one, so whatever was linearly decodable before is decodable after, and equally well. A linear probe therefore cannot tell apart two codes that differ by an invertible linear map, and hypotheses about iconic and discursive formats can differ in just this way. For example, a discursive hypothesis might say that a model holds each of the 21 points on the `order` line on its own direction (a one-hot code), so that every point is equally far from every other. An iconic hypothesis might say that it holds point k by turning on the first k of those directions (a thermometer code), so that nearer points lie closer together and the distances mirror the line. The second code is the first multiplied by a triangular matrix, an invertible linear map, so an unpenalized linear probe over the whole space does equally well on both, and so do the model's own linear readers; only its normalization layers, which respond to vector length, can register the difference. What separates the two hypotheses is geometry, which a probe does not measure and a comparison of pairwise distances between activation vectors does, as in representational similarity analysis (Kriegeskorte et al., 2008).

The point has precedents: no similarity index invariant to invertible linear maps is meaningful when dimension exceeds the number of data points (Kornblith et al., 2019), mutual information is unchanged by any bijection (Pimentel et al., 2020), and an arbitrarily non-linear alignment map makes causal abstraction uninformative (Sutter et al., 2025).

Two features of the fitted probes break this invariance for non-orthogonal maps: the dimensionality reduction they use retains the directions carrying the most variance, and an anisotropic re-encoding changes which those are; and the ridge penalty on weight size is itself not invariant under such a map. Orthogonal maps (rotations, reflections, permutations) change neither. What sensitivity to format a fitted probe has therefore enters incidentally, through variance ordering and the penalty, so the residual sensitivity is *not* reliable sensitivity to format: a difference in R^2 between two formats may record how variance is distributed in each, rather than saying anything about content organization *per se*.

Consequently, format hypotheses that differ by an invertible linear re-encoding cannot be separated by a linear probe. It should be said that hypotheses that differ non-linearly from one another *are* separable. An actual example of this is Othello-GPT, whose board state linear probes decode with at least 20 percent error under absolute color coding (Li et al., 2023) but with 99 percent accuracy or more from layer four onwards under player-relative coding (Nanda et al., 2023). The two codes carry the board in different encodings, and a linear probe distinguishes them; whether they carry the same content is itself contestable, since "mine" and "theirs" are relative to the player to move. What a probe *cannot* do is tell you, by itself, whether one's hypotheses differ linearly or non-linearly.

The above remarks suggest a control that addresses this. Fit the same probe on an explicit and hand-built encoding of the same content in the representational format you intend to rule out. If that encoding reaches the R^2 you observed, your observation does not discriminate between the two accounts; if it falls well short, that rules out only the encoding tested, since a format admits many encodings. The control can show that a probe does not discriminate, not that it does. For example, we had planned to probe for relations between objects, such as which of two objects lies further left, on the assumption that a diagram displays these relations directly while sentences leave them to be worked out. We abandoned that plan after running the control. A probe, fitted on nothing but the objects' stated coordinates, recovered those relations at R^2 0.793, so finding them in the model would not have shown an iconic representation.

The moral is that whether these models represent spatial content iconically requires methods sensitive to *organization* rather than content. Consequently, answering such questions requires causal intervention on candidate structures, or comparisons designed so that two format hypotheses predict

different orderings rather than different magnitudes. On this argument, probing results about spatial content, such as linear representations of space and time (Gurnee and Tegmark, 2024), bear on content and not on format.

7 Exploratory work and errata

Five exploratory follow-ups were run: a transcoding analysis, probes against the prompt's own tokens, a separately preregistered intervention study on the `graph` deficit, a separately preregistered study of input interventions on the diagram (FU5), and an unregistered three-dimensional rotation battery. The prompt-token probes are reported in Appendix E.8 and the rotation battery in Appendix F. The prompt-token probes qualify the one Study 2 result reported above, the `grid` occupancy readout, which is within what the prompt's own tokens could supply without computation by the model (§6). None of the others, FU5 included, changes a result reported above. The intervention study establishes that a node's own degree is linearly decodable from the iconic prompt when read at the key entry rather than the last token ($\Delta R^2 = +0.420$; above that study's token-bag and 16-token-window ceilings, though not tested against the pooling bound of Appendix E.8), so the failure described in §5.5 is not simply the absence of the information. Whether the models use this information was not estimated, because the erasure the test depends on was incomplete.

FU5, a separately preregistered follow-up, asks which property of the diagram carries its cost, on 494 new items in grid, order and graph, for Qwen2.5-7B and Llama-3.1-8B with reasoning requested. Each arm changes one property of the drawing (names in place of the letter key, the table's rows added beneath it, the labelled layout shuffled, the glyphs spaced one per token), each with a table-side control. The drawn layout was less accurate than the table in every family (A0 = −1.95 log-odds); because battery, wording and decoder differ, this is not a replication of the Study 1 deficit. Its preregistered outcome sentences, given in full with their sensitivities in Appendix G, are these. The letter key did not account for the drawing's cost (C1a = +0.10, within $\pm$0.25 log-odds), but only marginally, and not when truncated trials are scored incorrect, when graph is excluded or when the fit is restricted to compliant trials. Adding the table's rows raised accuracy (C2 = +1.62), which shows the cost is recoverable by stating the facts, not where it arises. With the facts stated in a table, adding the drawing lowered accuracy more than adding the same facts as sentences (C2b′ = −0.55). The shuffled layout was less accurate (C3 = −1.25), but whether it lost more than unsorting costs a table was neither shown nor bounded (C3-DiD = −0.29), so no effect is attributed to spatial alignment. Whether spacing the glyphs changed accuracy was neither shown nor bounded (C4 = −0.08; post hoc wording). FU5 is exploratory relative to the main registration and changes no Study 1 or Study 3 result; Appendix G gives its exclusions, checks, exploratory results and limits.

An audit of the frozen battery found defects that could not be repaired without invalidating the registration, and they are reported in full in Appendix D. One bears directly on the estimates above. The registered scorer is biased by condition, because the single-letter key exists only in the iconic rendering, so all 304 trials that key resolution rescues are in one condition. In FU5, where the key is a manipulated factor, the registered scorer gives the key contrast +0.87 log-odds against +0.10 under key resolution. The registered scorer remains primary as the preregistered analysis, and where we interpret a magnitude we interpret the key-resolved estimate. In general, any format comparison in which one rendering carries a legend, key or abbreviation table is exposed to this bias.

8 Discussion

The result that remains is a format effect that tabular notation does not reproduce. It is not a claim about the capacity for language models to iconically reason. This is because accuracy under a format manipulation cannot distinguish an *inability to represent something* from a *failure to exploit it* under this prompting regime. The deficit itself rules out reading answers off iconic arrangements as the dominant mechanism, since reading off predicts an advantage. The failure of the reasoning prediction rules out successful transcription to a discursive format as a compensating strategy. Failed transcription remains a candidate source of the deficit, and would fit the small value of reasoning with diagrams (§5.4); Appendix B shows one such failure.

Put another way: there are two questions that "representational format" runs together, specifically, whether language models *have* iconic representations and whether they *use* them. These studies

cannot settle the latter question: linear decodability is not use, and behavior, the ordinary evidence about use, is confounded here by the differences listed in §9. Whether they *have* an iconic representation is a harder question to answer, and we argue that linear probing, the method used for this elsewhere in the interpretability literature, cannot answer it by itself: linear decodability is invariant to the re-encodings by which format hypotheses differ, so a probe's R^2 is at best an unreliable signal of format, and evidence for having an iconic representation has to come from its geometry and from intervening on it. Given how much work linear decodability is asked to do in this literature, we take this to be the most transferable result in the study, and the control described in §6.1 is something we would like to see other studies utilize.

9 LIMITATIONS

The iconic rendering differs from both propositional renderings in more ways than iconicity, and the design does not control for them all. For example, the iconic renderings supply a single-letter key in three families of four, require some queried primitives to be derived where the propositional renderings state them outright, and put information below the level of the token, where character positions matter and the tokenizer does not preserve them. Any of these could account for part or all of the measured deficit, as could frequency: fixed-width and markdown tables are common in web and instruction text and ASCII diagrams are not, so the tabular control does not match the diagram in how often such text occurs in training. A separately preregistered follow-up (§7) found, marginally, that the key did not account for the drawing's cost, and attributed the cost neither to glyph spacing nor to alignment. Relatedly, a readable answer is recovered from diagrams less often than from tables or sentences ($b = -1.37$ in log-odds; Appendix C gives the rates for each model and format). The design cannot separate this from the accuracy deficit, so we cannot rule out that diagrams disrupt how models format their answers rather than how they reason. Decoding also differs across the open-weight models, since the models feature unequal repetition penalties, so every cross-model statement is confounded. Within a model the three formats have the same nominal decoder but not quite the same intervention, because a repetition penalty is prompt-sensitive and the prompts differ by format. The Qwen2.5-7B default alone changes 16.2% of answers against no penalty on a follow-up battery (Appendix C). Llama-3.1-8B has no repetition penalty.

The behavioral claim rests on five models, a small sample even though the deficit holds with a by-model random slope in the interpreted frame (§5.1); it concerns models in the 1.5B–8B range, and both frontier models are at or near ceiling outside `graph`, so their nulls there are uninformative. The main causal question remains open: establishing that a model *uses* a decoded quantity requires ablating it and observing a change in behavior. Also, the stimuli are synthetic and ASCII diagrams are just one case of iconicity. Five models were used with one family of prompts in English only and token count is an imperfect proxy for response time.

10 CONCLUSION

For the models tested here, the format in which the same information is presented affects what they get right, across a variety of tasks. The 1.5B–8B models are worse on diagrams than on either sentences or tables when asked to reason, and on `graph`/`degree` their answers from a drawing barely track the true degree. Whatever makes a diagram hard is not a property it shares with a monospaced, symbol-dense, two-dimensional table.

Outside `graph`, the most robust estimate is not the iconic accuracy deficit but this deficit's interaction with reasoning: asking for reasoning is worth +2.3 percentage points with a diagram, against +9.3 with a table and +11.2 with sentences. With the reasoning channel closed there is no iconic deficit outside `graph` under key-resolved scoring, which restricts the claim to prompting in which reasoning is requested, on tasks of this kind.

Whether these models represent iconic structure, and whether they use it, are both out of reach of this design, but this study suggests further ways to test this hypothesis, namely, by a combination of causal intervention via ablation and subsequent studies of language model behavior.

ETHICS STATEMENT

This work involves no human participants and no personal data. All stimuli are synthetic: each item is generated by a seeded program from an abstract structure (§3), and no text was collected from people. The open-weight models were run locally under their published licences, and the hosted models were queried through their public APIs. The study was preregistered, and every departure from the registration is reported (Appendix D). The findings concern how models reason over different renderings of the same information, and we see no direct route from them to harm.

AI USE STATEMENT

We used generative AI tools (Anthropic's Claude) in three ways on this project. The stimulus generator, the evaluation and scoring pipeline, the statistical analysis scripts and the figure code were written with Claude from the author's specifications; the code was checked with the repository's test suite, and new scripts were reviewed by a separate AI instance before use. Claude also ran and monitored the model evaluations. The text was drafted by the author, then subjected to style and clarity edits by Claude. The authors then heavily rewrote the edits made by Claude, and the numbers in the paper were traced to committed analysis output. The research questions, hypotheses, preregistered analysis plan and interpretation of the results are all the authors'. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code, every stimulus battery, every model completion and the saved analysis outputs, anonymized for review, with a README that maps each result in the paper to the script that produces it. The stimulus generator is seeded and rebuilds the registered battery byte for byte (§3). Because all completions are included, every behavioral estimate can be recomputed without querying a model; rescoring the included completions reproduces the scored files byte for byte, and refitting the registered model reproduces the registered estimate of §5.1 exactly. Decoding settings, including the checkpoint repetition penalties a rerun must match, are given in §4 and Appendix C, and the hosted models' repeat-run agreement in §4. Probe settings are in Appendix E; refitting the probes requires the open-weight model and a GPU, and their outputs are included. The unregistered model fits cited in §4 and §5.1 are included with their scripts. The preregistration and all twenty-one departures from it are included (Appendix D).

REFERENCES

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Computational Linguistics* 48(1), 207–219.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic probing: behavioral explanation with amnesic counterfactuals. *TACL* 9, 160–175.

Goodman, N. (1968). *Languages of Art*. Bobbs-Merrill.

Gurnee, W. and Tegmark, M. (2024). Language models represent space and time. ICLR.

Hall Maudslay, R., Valvoda, J., Pimentel, T., Williams, A., and Cotterell, R. (2020). A tale of a probe and a parser. *ACL*.

Harding, J. (2023). Operationalising representation in natural language processing. *BJPS*.

Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. *EMNLP-IJCNLP*, 2733–2743.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. *ICML*, PMLR 97, 3519–3529.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2:4.

Larkin, J. H. and Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* 11(1), 65–100.

Lee, S. H. and Kondor, R. (2026). Deep minds and shallow probes. arXiv:2605.11448.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2023). Emergent world representations: exploring a sequence model trained on a synthetic task. *ICLR*.

Nanda, N., Lee, A., and Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. *BlackboxNLP*, 16–30.

Naselaris, T. and Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences* 19(10), 551–554.

Peirce, C. S. (1867). On a new list of categories. *Proceedings of the American Academy of Arts and Sciences* 7, 287–298.

Peirce, C. S. (1903). *A Syllabus of Certain Topics of Logic*. Alfred Mudge & Son.

Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., and Cotterell, R. (2020). Information-theoretic probing for linguistic structure. *ACL*, 4609–4622.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin* 80(1), 1–24.

Pylyshyn, Z. W. (2002). Mental imagery: in search of a theory. *BBS* 25(2), 157–182.

Quilty-Dunn, J. (forthcoming). Visual icons. *Philosophy and Phenomenological Research.*

Quilty-Dunn, J., Porot, N., and Mandelbaum, E. (2023). The best game in town: the reemergence of the language-of-thought hypothesis across the cognitive sciences. *BBS* 46:e261.

Ravichander, A., Belinkov, Y., and Hovy, E. (2021). Probing the probing paradigm: does probing accuracy entail task relevance? *EACL*, 3363–3377.

Ritchie, J. B., Kaplan, D. M., and Klein, C. (2019). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *BJPS* 70(2), 581–607.

Shimojima, A. (2015). *Semantic Properties of Diagrams and Their Cognitive Potentials*. CSLI Publications.

Sutter, D., Minder, J., Hofmann, T., and Pimentel, T. (2025). The non-linear representation dilemma: is causal abstraction enough for mechanistic interpretability? arXiv:2507.08802.

Voita, E. and Titov, I. (2020). Information-theoretic probing with minimum description length. *EMNLP*, 183–196.

The tables in Appendices A to C and E were generated from the data files by `scripts/make_appendices.py` and have since been edited by hand. No number here is carried over from the body of the paper.

A FULL PROMPTS

One item, `grid-011`, in all three renderings, exactly as the model received it. Figure 1 shows one item from each family, and every prompt in both batteries is in `data/stimuli/stimuli.jsonl`. The blocks below are copied verbatim from that file; nothing is elided or reflowed. Paraphrase 0 is shown throughout, and the wrapper wording of the other two is given at the end of the appendix.

The three renderings of an item share their first line, their question line, their response-type line and their closing instruction, so within a block of three only the representation varies.

GRID — ITEM GRID-011

Question type `leftmost`, answer type `name`, gold answer `pylon`, options `dial`, `girder`, `pylon`.

iconic

```
Read the description below and answer the question.

Grid (rows numbered top to bottom, columns left to right):

    1 2 3 4 5
1   . . . . .
2   . . . . .
3   . P . . .
4   . G . D .
5   . . . . .

Key: P = pylon; D = dial; G = girder.
A dot marks an empty cell.

Question: Which object is in the leftmost column? (If more than one object shares that column,
↪  name the one nearest the top.)
Answer with the object name only.
Give your reasoning briefly, then end your reply with a final line of the form 'ANSWER: <your
↪  answer>'.
```

tabular

```
Read the description below and answer the question.

There is a grid with 5 rows and 5 columns. Rows are numbered 1 to 5 from top to bottom and
↪  columns 1 to 5 from left to right.

| object | row | column |
|--------|-----|--------|
| pylon  | 3   | 2      |
| dial   | 4   | 4      |
| girder | 4   | 2      |

All other cells are empty.

Question: Which object is in the leftmost column? (If more than one object shares that column,
↪  name the one nearest the top.)
Answer with the object name only.
Give your reasoning briefly, then end your reply with a final line of the form 'ANSWER: <your
↪  answer>'.
```

discursive

```
Read the description below and answer the question.

There is a grid with 5 rows and 5 columns. Rows are numbered 1 to 5 from top to bottom and
↪  columns 1 to 5 from left to right.
The pylon is in row 3, column 2.
The dial is in row 4, column 4.
The girder is in row 4, column 2.
```

```
All other cells are empty.

Question: Which object is in the leftmost column? (If more than one object shares that column,
↪  name the one nearest the top.)
Answer with the object name only.
Give your reasoning briefly, then end your reply with a final line of the form 'ANSWER: <your
↪  answer>'.
```

THE THREE PARAPHRASES, AND THE STUDY 3 WRAPPER

A paraphrase changes the opening line and the closing instruction and nothing between them. The opening lines are:

0. `Read the description below and answer the question.`
1. `Below is some information followed by a question about it.`
2. `Consider the following, then answer the question that follows.`

Study 1 and Study 3 differ only in their closing instructions; the items, seeds, structures and renderings are identical. Study 1 asks for brief reasoning:

0. `Give your reasoning briefly, then end your reply with a final line of the form 'ANSWER: <your answer>'.`
1. `Think it through in a sentence or two, then finish with a line reading 'ANSWER: <your answer>'.`
2. `You may work through it briefly. Your last line must be 'ANSWER: <your answer>'.`

Study 3 asks for the answer alone:

0. `Do not explain. Reply with one line only, of the form 'ANSWER: <your answer>'.`
1. `Give no reasoning. Your entire reply must be a single line reading 'ANSWER: <your answer>'.`
2. `Answer immediately, with no working out. Reply with the single line 'ANSWER: <your answer>' and nothing else.`

B THE FOUR SCORING ROUTES

`src/score.py` recovers an answer by one of four routes and records which one in the `answer_source` column. The routes are kept separate because the registration treats them differently, and because a difference between formats in how answers are recovered is a relevant result.

A completion is first searched for an `ANSWER:` line, with markdown emphasis stripped first so that `**ANSWER:** yes` matches. The last such line wins, since models sometimes restate the demanded format before using it. If a tag is found, its content governs and no fallback is applied. If none is found and the completion stopped because it reached the token cap, the trial is excluded. Otherwise the last non-empty line is read instead.

route	tag present	recovered from	scored as
`tag`	yes	the `ANSWER:` line, normalized to the demanded response type	correct or incorrect
`last_line`	no	the last non-empty line of the completion	correct or incorrect
`none`	either	nothing that normalizes	incorrect

route	tag present	recovered from	scored as
truncated	no	nothing; generation stopped at max_new_tokens	excluded

Truncation is the only exclusion. A non-answer is scored incorrect and kept in the analysis, so compliance failures count against the estimate. The last_line fallback exists because models often answer correctly and omit the tag; dropping those trials would discard real data, and scoring them incorrect would mark correct answers wrong.

The none route covers two cases that the tag rate separates but the parse rate does not. A model may put a non-answer under the tag, which happens on 101 Study 1 trials, or it may give no tag and end on a line from which nothing normalizes, which happens on 38. Both are scored incorrect, and the worked example below is of the first.

Route counts over the registered battery:

	n	tag	last_line	none	truncated
Study 1	10,800	10,200	184	139	277
Study 3	10,800	10,287	416	97	0

The same counts by format, where the compliance asymmetry of Appendix C first appears:

study	format	tag	last_line	none	truncated
Study 1	iconic	3,319	61	99	121
Study 1	tabular	3,460	43	10	87
Study 1	discursive	3,421	80	30	69
Study 3	iconic	3,410	106	84	0
Study 3	tabular	3,401	197	2	0
Study 3	discursive	3,476	113	11	0

The tag route is the ordinary case, covering 94.4 percent of Study 1 answers. One worked example of each of the other three routes follows; all three are real Study 1 trials, and the last is quoted in full from data/raw/.

LAST_LINE — QWEN2.5-1.5B ON GRAPH-024-TABULAR-P2

There is no tag anywhere in the completion, and generation stopped before reaching the token cap, so the last non-empty line is read. Here it is the single word YES, which normalizes to yes and is correct. Excluding this trial would discard a right answer and scoring it incorrect would record a right answer as wrong.

finish_reason stop, 95 completion tokens. Tag content none; parsed answer yes; gold yes; scored correct.

NONE — LLAMA-3.1-8B ON ORDER-022-TABULAR-P2

A tag is present and the model complied with the demanded form of reply, but its content is not an answer. The response type is a pair, and this reply names four objects because the model computed a tie between two pairs. The item does not in fact tie (the points are 5, 6, 12 and 19, so the neighbouring gaps are 1, 6 and 7), and the tie comes from the model's own arithmetic. The fallback is not applied, because a malformed tagged answer is treated as a non-answer, and the trial is scored incorrect.

finish_reason stop, 139 completion tokens. Tag content dial and emblem and emblem and anvil; parsed answer none; gold anvil and emblem; scored incorrect.

TRUNCATED — LLAMA-3.1-8B ON ORDER-014-ICONIC-P2

Generation reached `max_new_tokens` before the model gave an answer. The completion breaks off inside `(D`, so the last non-empty line is an incomplete fragment and the fallback would produce a spurious answer. This is the one route that is excluded. `run_eval.py` records `finish_reason` and `completion_tokens` on every trial; without them a completion that stopped because the model finished is indistinguishable from one that hit the cap.

```
To find the objects with the widest gap, we need to consider the positions of the objects.

The positions are:
- 0 (canister, C)
- 5 (dial, D)
- 10 (hinge, H)
- 15 (kettle, K)
- 20 (no object)

The gaps between the objects are:
- 5 - 0 = 5
- 10 - 5 = 5
- 15 - 10 = 5
- 20 - 15 = 5

However, we also need to consider the gaps between the objects and the ends of the line.
- 0 - 5 = 5
- 5 - 10 = 5
- 10 - 15 = 5
- 15 - 20 = 5

Since the gaps are the same, we need to look at the positions of the objects. The canister (C)
↪  is at position 0, and the dial (D) is at position 5. The hinge (H) is at position 10, and
↪  the kettle (K) is at position 15.

The widest gap is between the canister (C) and the dial (D
```

`finish_reason` `length`, 256 completion tokens. Tag content none; parsed answer none; gold `hinge and kettle`; scored excluded.

C PER-MODEL COMPLIANCE

Three compliance quantities are reported by model and format for both studies. The tag rate is the proportion of trials carrying an `ANSWER:` line that the scorer could find. The parse rate is the registered secondary measure, `parsed_ok`, which records whether an answer of the demanded response type was recovered by any route. The truncation rate is the proportion of trials whose generation stopped at the token cap. These are the excluded trials, except where the model gave a tagged answer before reaching the cap, which is scored on that answer: in Study 1, 283 trials reached the cap and 277 were excluded (Appendix B).

The rates are computed from `analysis/frames/compliance-frame.csv`, which `analysis/prepare.py` builds from `data/processed/scored.csv` and `data/processed/scored-nocot.csv`.

STUDY 1, BRIEF REASONING

model	format	n	tag rate	parse rate	truncation
Claude Opus 4.8	iconic	720	0.996	0.978	0.004
Claude Opus 4.8	tabular	720	1.000	1.000	0.000
Claude Opus 4.8	discursive	720	1.000	1.000	0.000
GPT-5.6 Sol	iconic	720	1.000	1.000	0.000
GPT-5.6 Sol	tabular	720	1.000	1.000	0.000
GPT-5.6 Sol	discursive	720	1.000	1.000	0.000
Llama-3.1-8B	iconic	720	0.839	0.832	0.153
Llama-3.1-8B	tabular	720	0.867	0.865	0.121
Llama-3.1-8B	discursive	720	0.917	0.914	0.074
Qwen2.5-1.5B	iconic	720	0.903	0.910	0.015

model	format	n	tag rate	parse rate	truncation
Qwen2.5-1.5B	tabular	720	0.942	1.000	0.000
Qwen2.5-1.5B	discursive	720	0.846	0.949	0.025
Qwen2.5-7B	iconic	720	0.999	0.975	0.001
Qwen2.5-7B	tabular	720	1.000	1.000	0.000
Qwen2.5-7B	discursive	720	1.000	1.000	0.000
all five	iconic	3,600	0.947	0.939	0.035
all five	tabular	3,600	0.962	0.973	0.024
all five	discursive	3,600	0.953	0.973	0.020

STUDY 3, ANSWER ONLY

model	format	n	tag rate	parse rate	truncation
Claude Opus 4.8	iconic	720	1.000	0.978	0.000
Claude Opus 4.8	tabular	720	1.000	1.000	0.000
Claude Opus 4.8	discursive	720	1.000	1.000	0.000
GPT-5.6 Sol	iconic	720	1.000	1.000	0.000
GPT-5.6 Sol	tabular	720	1.000	1.000	0.000
GPT-5.6 Sol	discursive	720	1.000	1.000	0.000
Llama-3.1-8B	iconic	720	1.000	0.972	0.000
Llama-3.1-8B	tabular	720	1.000	1.000	0.000
Llama-3.1-8B	discursive	720	1.000	1.000	0.000
Qwen2.5-1.5B	iconic	720	0.853	0.982	0.000
Qwen2.5-1.5B	tabular	720	0.726	1.000	0.000
Qwen2.5-1.5B	discursive	720	0.843	0.985	0.000
Qwen2.5-7B	iconic	720	1.000	0.951	0.000
Qwen2.5-7B	tabular	720	1.000	0.997	0.000
Qwen2.5-7B	discursive	720	1.000	1.000	0.000
all five	iconic	3,600	0.971	0.977	0.000
all five	tabular	3,600	0.945	0.999	0.000
all five	discursive	3,600	0.969	0.997	0.000

Two features of these tables matter.

The parse rate is lower for iconic prompts in both studies, at 0.939 against 0.973 and 0.973 in Study 1, and 0.977 against 0.999 and 0.997 in Study 3. It is lower for iconic prompts in every model but one, and the exception is the model that is at ceiling in all three formats.

Truncation is confined almost entirely to Llama-3.1-8B in Study 1, at 11.6 percent against 0.14 percent for Claude Opus 4.8 and zero for GPT-5.6 Sol. Within that model it is ordered by format, at 15.3 percent iconic against 12.1 tabular and 7.4 discursive, and this ordering triggers the registered sensitivity analysis in which truncated trials are scored incorrect rather than excluded. No trial in Study 3 was truncated in any cell.

EIGHTEEN SCORING AND EXCLUSION VARIANTS

The iconic − tabular contrast is negative and significant in every combination of scoring rule, truncation handling and cell exclusion, and the tabular − discursive contrast is null in every one. On the scoring dimension alone, in Study 1:

scoring rule	iconic − tabular (log-odds)	tabular − discursive (log-odds)
registered	−0.916 (SE 0.073, p = 1.6 × 10^{-36})	−0.011 (p = .88)
+ name recovery	−0.894 (SE 0.072, p = 5.3 × 10^{-35})	−0.001 (p = .99)
+ key resolution	−0.598 (SE 0.073, p = 2.7 × 10^{-16})	−0.014 (p = .85)
+ both	−0.580 (SE 0.073, p = 2.0 × 10^{-15})	−0.004 (p = .96)

Scoring truncated trials incorrect rather than excluding them moves the iconic − tabular contrast by at most 0.02 log-odds in any row, so the scoring × truncation grid spans −0.58 to −0.92. The full range across the eighteen variants, −0.58 to −1.37, comes from the exclusion dimension: dropping the three cells that cannot discriminate strengthens the Study 1 iconic deficit from −0.92 to −1.12 under the registered scorer, and the frame that retains them is reported beside every interpreted estimate. Retaining them changes the sign of no model-based estimate and changes significance in eight places, all of which are listed in `analysis/results/`.

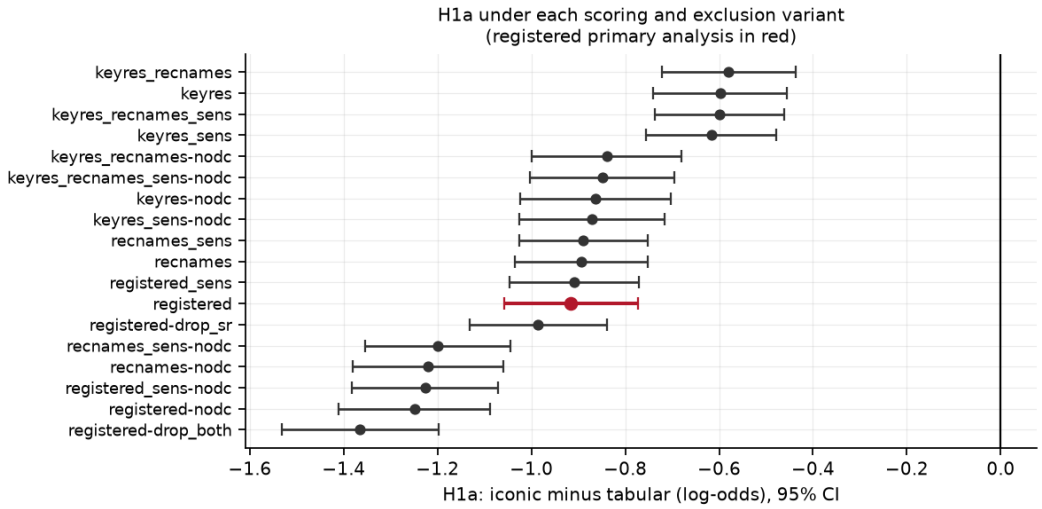


Figure C1: The iconic − tabular contrast under all eighteen scoring and exclusion variants, with 95% confidence intervals. The registered primary analysis is shown in red.

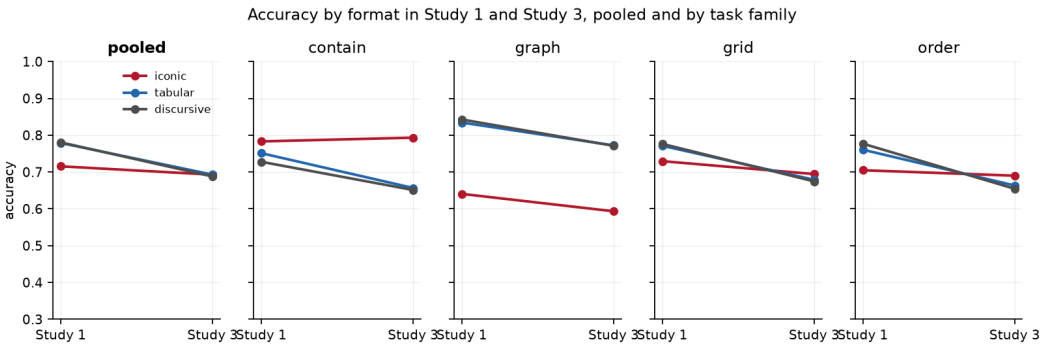


Figure C2: The reasoning-channel interaction: accuracy by format in Study 1 (reasoning) and Study 3 (no reasoning), pooled and by task family.

THE DECODING CHECK

Section 4 gives the checkpoint repetition penalties and the first check on them. Two details matter for reproducing the runs. Selection remained argmax, so the runs are greedy in the sense the registration intended, but a rerun with an explicit `repetition_penalty=1.0` will not reproduce these completions. Generation is one prompt at a time, because batching changes which token is argmax.

The ordering given in Section 4, a larger deficit for Llama at no penalty (17.3 percentage points under the most lenient scorer) than for Qwen2.5-1.5B at 1.1 (4.4), holds under all eight scorers (the four scoring rules, each with truncated trials excluded or scored incorrect). A second check concerns truncation. Llama-3.1-8B's truncation, at 11.6 percent and ordered by format, is not a decoding pathology: 23 of its 250 truncated completions are degenerate repetition, against 89 of 89 in the rotation battery of Appendix F. Its truncated replies are long rather than looping, which is what a 256-token cap on a verbose model produces.

The effect of the checkpoint defaults on answers was measured afterwards, on the follow-up study's pilot battery (680 trials, Qwen2.5-7B, cap 1,536), against a baseline that sends `repetition_penalty=1.0` and so applies no penalty. The Qwen2.5-7B default of 1.05 changes the answer on 110 of the 678 trials that terminate under both settings, or 16.2%, comparing answer content with case, punctuation and pair order normalized. This is a different battery at a longer cap, so it indicates the size of the effect on this study rather than measuring it. It does not make the penalty an explanation of the deficit, for the reason given in Section 4: the deficit is larger for Llama, which runs with no penalty, than for Qwen2.5-1.5B, which runs with the largest.

D DEPARTURES, CORRECTIONS AND ERRATA

`prereg/departures.md` is reproduced in full in the repository, all twenty-one entries, without editing. Section 7 of the paper states the erratum that generalizes beyond this study. The two stimulus defects referred to in §3 are described below.

The table lists all twenty-one entries. The three that bear most directly on the estimates are described in more detail below.

§	kind	entry
1	deviation	Key-resolved scoring: a second scorer maps a single-letter key label to the name it stands for; reported beside the registered scorer.
2	correction	Contrast coding: recoded so each coefficient is the named difference; the two contrasts correlate at r = +0.5.
3	erratum	The registration says all three renderings share their framing and closure sentences verbatim; the iconic renderings replace the framing with a notation legend.
4	erratum and deviation	The `contain` legend states a convention the sentences do not; the iconic − tabular contrast is also reported without that cell (below).
5	erratum	Study 2 is reported as uninformative rather than null, because its positive controls do not all pass (Appendix E.4).

§	kind	entry
6	not a departure	Truncation exceeds five percent in eight Llama-3.1-8B cells, so the registered sensitivity analysis is reported (Appendix C).
7	erratum	`grid`/`same_row` has the gold answer "no" on every item; a question-only arm scores 0.905 on it in Study 1 (below).
8	exploratory addition	A question-only arm with the facts removed: 0.134 for the two hosted models in Study 1, against 0.947 on the full tabular task.
9	correction	The scorer excluded untagged non-answers; fixed so that truncation is the only exclusion.
10	deviation	Answers phrased as sentences were scored on their first word; a name-recovery scorer is reported beside the registered one.
11	erratum	`graph` topology was selected by what the diagram could draw (below).
12	erratum	Two cells are close to degenerate (below).
13	erratum	9 of 17 `order`/`between` items can be answered by elimination; reported split by object count.
14	erratum	Two further asymmetries between renderings are missing from the registered list; for example, only the `order` diagram states which end of the line is the left.
15	erratum	The registered prompt-length means hold for GPT-5.6 Sol only; for Claude Opus 4.8 the iconic prompts are shorter than the tabular ones.
16	correction	The equivalence tests compare counts and shapes, not values: 5 of 6 mutations that make a rendering state false facts pass them.
17	correction	Three outputs reported random effects as dropped for models that had converged with a variance component at zero.

§	kind	entry
18	correction	The compliance model was fitted to the wrong column; the corrected iconic compliance deficit is about three and a half times larger.
19	deviation	The key-resolved estimate is the one interpreted (§7), which roughly halves the headline magnitude.
20	deviation	The three cells that cannot discriminate are dropped from the interpreted analysis, and the frame retaining them is reported beside every interpreted estimate (below).
21	decision	A control adding the letter key to the propositional renderings is exploratory and does not gate the paper.

GRAPH TOPOLOGY

Edges were rejection-sampled according to whether the ASCII renderer could lay them out, so a constraint belonging to one format determined the structures that all three formats describe. The nominal mean of 5.00 edges is 3.50 in the frozen battery, no item exceeds 5, and 26 of 60 items contain an isolated node. The generator is frozen, and this could not be repaired without invalidating the registration, so the iconic − tabular contrast for `graph` is also reported within edge-count stratum. The deficit is present at every density, at −1.91, −1.65 and −1.59 log-odds for three, four and five edges, and does not track density. The defect is real but does not explain the `graph` result.

THE CONTAIN DIAGRAM CONVENTION

The discursive rendering opens by saying that some objects are packed inside containers. The diagram as first written explained box-in-box nesting but did not state that a labeled box is a container and a name in parentheses is an object. The `direct_count` question turns on that distinction, and its gold answer excludes nested containers, so on 14 of 22 items the sentences stated a fact that the picture only implied. The repair made before registration added a legend stating the convention to the diagram, and it overcorrected, since the diagram now states the convention and the sentences do not. A substitution control puts the legend's value at 19.9 percentage points on that question in Study 1 and 27.3 in Study 3, and at approximately zero on the other two. With the legend removed, no family shows a reliable iconic advantage in Study 1.

FU5 DEPARTURES

`studies/fu5/DEPARTURES.md` records twelve dated entries after the tag `fu5-frozen`. None changes a registered estimate.

#	entry
1	The Qwen2.5-7B run was interrupted after 2,661 trials and resumed with no gap or overlap.
2	The run guard stopped Llama-3.1-8B at its 1,200-minute limit after 8,077 trials; it was relaunched from a commit that changed only paper files, and resumed with no gap or overlap.

#	entry
3	Generation took 27.5 GPU-hours against an estimate of 13.5, within the 30-hour budget.
4	The plan text was not updated for the run's scale (494 items, 9,296 trials per model) or decoder (the 20-token no-repeat rule, no penalty).
5	The prompt-length table and one tokenisation record were not committed; the verifier re-ran the tokenisation checks, which hold.
6	The truncated-as-incorrect frames score every cap-length completion incorrect; the two tagged ones were already incorrect, so nothing changes.
7	The battery and raw completions were untracked; resolved by committing them, with checksums matching the verification.
8	Rescue counts by arm were produced after accuracy had been read, not before.
9	Four reported standard errors came from degenerate fits (printed p = 0); the labels were checked post hoc over 21 row orders and every reliable fit gives the reported label.
10	The sign-disagreement flag of C3-DiD was added to the C3 sentence post hoc.
11	The C4 "not detected" sentence was written after the results; the plan fixes none for that pattern.
12	The rescue, dereference and exploratory items the plan promises were produced after the results were read; the I+T quote-versus-describe share was not produced, because its heuristic was not defined before the run.

THREE PROBLEMATIC CELLS

`contain`/`transitive_in` is 82.4 percent "yes" and largely entailed by its framing; `graph`/`isolated` is worth 0.012 over a guessing baseline; and `grid`/`same_row` has the gold answer "no" on all 14 items, with a question-only arm scoring 0.905 in Study 1. All three are dropped from the interpreted analysis. That decision was taken after the data were seen, and the frame retaining them is reported throughout.

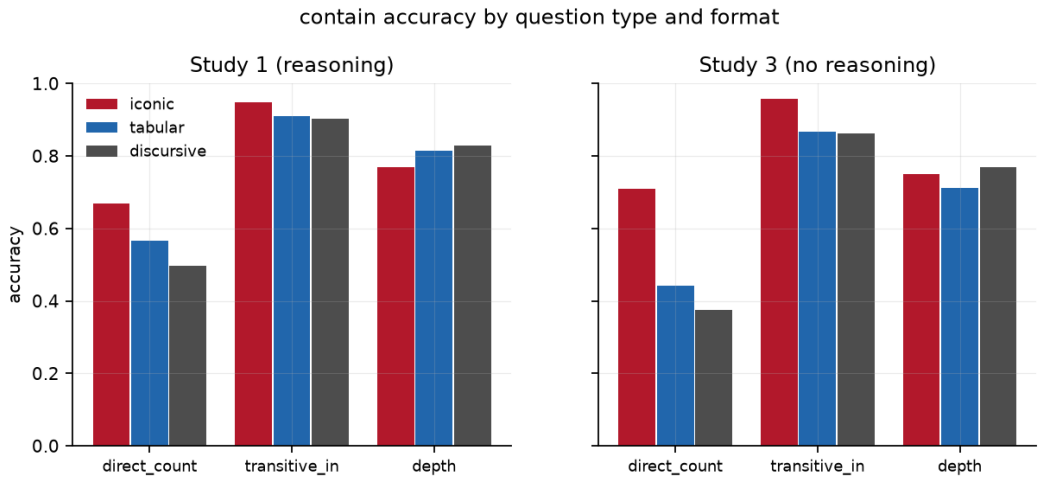


Figure D1: `contain` accuracy by question type and format, in Study 1 (reasoning) and Study 3 (no reasoning).

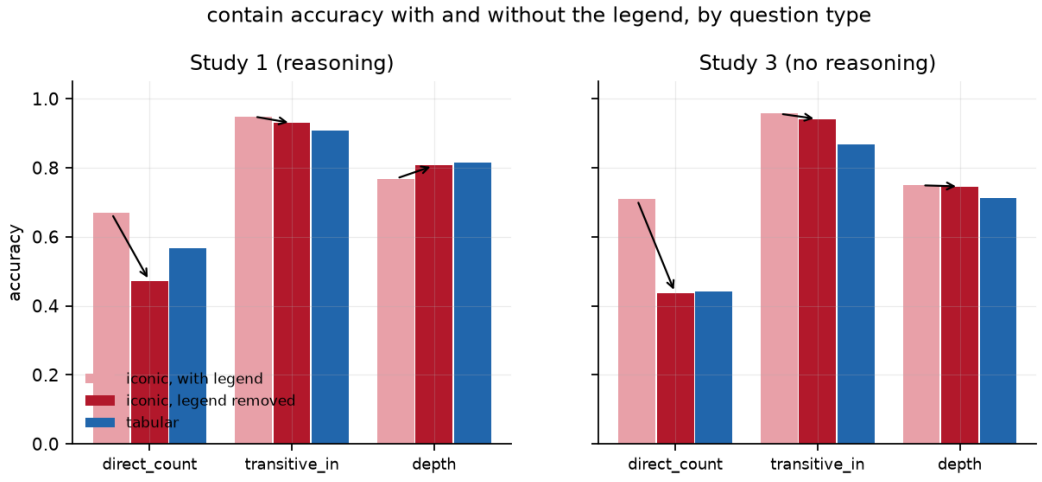


Figure D2: The effect of removing the `contain` legend, by question type, in both studies.

E PROBE DESIGN

Study 2 fits ridge regressions from a model's residual stream to structural targets, and asks whether a structure is more linearly decodable from one rendering than from another. All of it is exploratory: it is not powered, the layer sweep is not corrected for multiplicity, and `src/probe.py` writes an `exploratory` column on every row so that no number can be taken from the file and quoted as confirmatory.

E.1 WHAT IS FITTED

All reported probes are fitted on Qwen2.5-1.5B, entered as `qwen-1.5b` in `configs/models.yaml`, whose residual stream is 1,536 dimensional across 28 blocks. Activations are read at `blocks.L.hook_resid_post` for L in [0, 4, 8, 12, 16, 20, 24, 27], every fourth block plus the last. No probe on Qwen2.5-7B is reported: a 20-item smoke test in the repository is degenerate at that item count, and the full model does not fit in the memory the probing library needs. The prompt is passed through the same chat template `run_eval.py` uses, so the probe reads the activation the evaluated model actually had.

Each fit is a ridge regression with the penalty chosen on an inner item-grouped split of the training fold only, scored by cross-validated R^2 over 5 held-out folds, with predictions pooled across folds and scored once rather than averaged per fold. Folds are assigned per item and every row inherits its item's fold, so that two renderings of the same structure never fall on opposite sides of the split; splitting rows at random would let the probe score well by recognizing the item. Activations are projected to 256 principal components, fitted on the training fold only, before the ridge.

Each cell carries a shuffled baseline: the same fit, with the same fold seed, over 5 permutations of which item each target belongs to. The permutation keeps an item's rows together, so the baseline breaks the link between activation and structure and nothing else. Across the reported cells the shuffled baseline never exceeds -0.004, so the baseline is at zero as expected, and the quantity to interpret is the level rather than the difference from the baseline.

The probe battery is generated by the frozen generator at a different seed from the registered battery, with no structural overlap. Item counts are 500 for `contain`, 508 for `graph`, 500 for `grid`, 500 for `order`; `graph` differs because its node count varies and the probe holds it fixed at five, a fixed-width target over a varying object count having either to pad, which leaks the count, or to truncate.

E.2 THE TARGETS

Each family has one identity-bound target and one identity-free target. The identity-bound target cannot be recovered without knowing which object is which; the identity-free one is the same structure with the names stripped out. Only this contrast is compared across families, since R^2 is not comparable between targets of different dimensionality.

family	identity-bound	identity-free	positive controls
`grid`	`coords`, the (row, column) of each object	`occupancy`, the flattened 5×5 map of filled cells	`nobj`, `firstrow`
`order`	`positions`, each object's point on the line	`occupancy`, which points on the line are filled	`nobj`, `first`
`contain`	`depths`, each item's nesting depth	`profile`, how many items sit at each depth	`nobj`, `first`
`graph`	`adjacency`, the upper triangle in node order	`degseq`, the sorted degree sequence	`nedges`, `first`

Objects are always taken in `structure` order rather than in the order in which they appear in a rendering. This is the one order in which all three renderings introduce the objects; reading order on a canvas differs from reading order in a list, and using it would confound the comparison.

E.3 THE READOUT

The results depend on the choice of readout, and two of the three readouts tried are not format-neutral. Mean-pooling over prompt positions discards position, so it favors facts carried by lexical co-occurrence, that is, the propositional formats. Anchoring at each object's own token reads position directly, and on a canvas an object's coordinates are the position of its letter in the token sequence, so it favors the diagram. The propositional formats score at the floor under this second readout for a reason unrelated to representation: in `The dial is in row 4, column 2` the token for *dial* precedes its own coordinates, and causal masking prevents it from carrying them.

The table gives peak observed R^2 in the grid family under each readout. The rows are not on one scale, since their targets differ in width, but the ordering of the formats within a row is comparable, and it reverses:

readout	target	iconic	tabular	discursive	what it reads
mean-pooled	`coords`	0.486	0.697	0.689	lexical co-occurrence, which the diagram does not have
node-anchored	`own`	0.992	-0.004	-0.004	token position, which on a canvas is the coordinate itself
last prompt token	`coords`	0.237	0.263	0.280	the end of a wrapper shared by all three, preceded by every fact

The node-anchored result is the easiest of the three to diagnose. Iconic `grid` reaches 0.969 at layer 0 and 0.992 at layer 20, so the fit is essentially complete at the embedding, before the network has done any processing. The probe is therefore reading the positional encoding rather than a representation. We set both readouts aside, and with them the crossover they appeared to support in opposite directions.

The readout we use is the last prompt token. It is the end of the closing instruction, which is byte-identical across the three renderings of an item, and every fact in every format precedes it, so causal masking and positional encoding affect all three equally.

E.4 THE POSITIVE CONTROLS

Two controls were run on quantities the prompt states outright, on the principle that if a probe cannot recover a stated fact then nothing else it reports can be interpreted. Object count is stated by every rendering as one sentence, one table row or one key entry. The row of the first object in `structure` order is one dimension of the coordinate target on its own.

The first control was run on `grid`, with 60 items and 180 rows against 1,536 raw dimensions and no projection:

control target	iconic	tabular	discursive
`nobj`, object count	+0.948	+0.973	+0.979
`firstrow`, row of the first object	-0.012	-0.012	+0.014

Object count decodes at layer 0 in all three formats, at 0.948, 0.973 and 0.979, which marks it as a surface count rather than something the model computes. The row of a named object reads at approximately zero everywhere. Taken at face value, the second result would mean that the probe cannot read identity-bound content at all.

That reading is too pessimistic, because the sample is small. The same two controls at the same last-token readout, on 500 items rather than 60, put `firstrow` at 0.449 iconic, 0.557 tabular and 0.688 discursive, each rising from approximately zero at layer 0. At 60 items the control is therefore underpowered. That run is recorded in `data/processed/probe-c500-last.csv`. It matches the reported probes in model, readout and item count but not in projection, since it was fitted on the raw 1,536 dimensions rather than on 256 components; no control has been run at the projection the reported probes use.

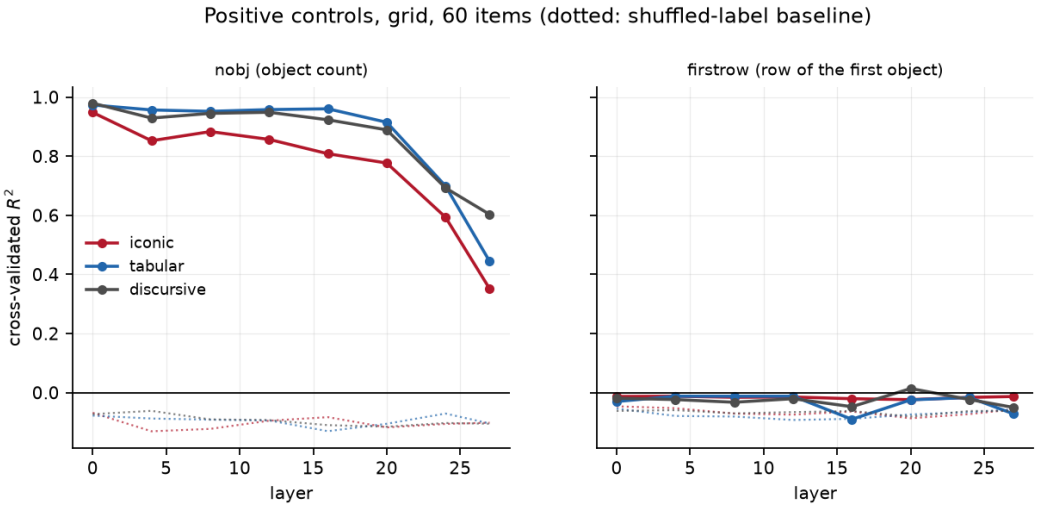


Figure E1: The positive controls on `grid` at 60 items, by layer: object count (left) and the row of the first object (right), with the shuffled-label baseline dotted. At 500 items the row of the first object decodes at 0.449 to 0.688 (see text).

One family fails its control in one condition. In `graph`, the degree of the first node decodes at 0.563 tabular and 0.678 discursive and at 0.040 iconic, while edge count, a global quantity, decodes at 0.342 iconic. A representation that carries how many edges there are but not which node has which degree is what weak identity binding would produce, and it is also what an underpowered probe for that condition would produce; this design cannot separate the two. The rule applied throughout is that a cell whose control does not clear its baseline is not interpreted, so the iconic `graph` numbers are reported but not interpreted. There is a circularity here, since the control is the same kind of node-specific quantity as the targets, so its failure is not an independent check on them.

E.5 The prompt-shape control

A probe R^2 shows that a structural target is linearly decodable from the residual stream. It does not show that the model computed anything: if a format's prompt grows with the target, any faithful encoding of the prompt carries the target, and length alone would score as well. `scripts/probe_length_control.py` fits the same ridge regression, with the same item-grouped folds, the same inner penalty selection, the same projection, the same standardization and the same shuffled baseline, and replaces only the residual-stream vector with surface counts. Everything except the feature matrix is held fixed, so the two R^2 values are on one scale. The script reads

its settings from the reference probe CSV and refuses to write any result whose item count or target dimensionality disagrees with it, because a control that answers a different question from the one it is compared against is misleading, which is worse than having no control.

Two feature sets are fitted. The ceiling quoted for a cell in the tables below is the larger of the two, which is the conservative choice, since a probe then has to beat whichever surface fit does better.

feature set	features
`tokens`	prompt token count alone. At one feature the projection is a no-op, so the number is robust to the projection setting
`surface`	token count plus characters, lines, longest line, digits and non-whitespace characters

Some of these features are constant within a format for some families; standardization removes a constant column, so in `contain` and `graph` the wider control has four effective features rather than six. `probe_length_control.py` reports every constant column at startup.

E.6 Probe, ceiling and clearance

The table gives peak observed R^2 over layers, the shape ceiling for the same cell, and the difference between them. The difference is the quantity to interpret: an R^2 that does not exceed its own ceiling is not evidence of a representation, whatever its level.

family	target	format	probe	ceiling	clearance
`contain`	`depths`	`iconic`	0.128	0.299	-0.171
`contain`	`depths`	`tabular`	0.351	0.110	+0.241
`contain`	`depths`	`discursive`	0.402	0.097	+0.305
`contain`	`profile`	`iconic`	0.880	0.660	+0.220
`contain`	`profile`	`tabular`	0.633	0.337	+0.296
`contain`	`profile`	`discursive`	0.718	0.290	+0.428
`graph`	`adjacency`	`iconic`	-0.005	-0.005	+0.000
`graph`	`adjacency`	`tabular`	0.083	0.008	+0.075
`graph`	`adjacency`	`discursive`	0.086	0.005	+0.081
`graph`	`degseq`	`iconic`	0.094	0.019	+0.075
`graph`	`degseq`	`tabular`	0.477	0.367	+0.110
`graph`	`degseq`	`discursive`	0.495	0.370	+0.125
`grid`	`coords`	`iconic`	0.237	0.005	+0.232
`grid`	`coords`	`tabular`	0.263	-0.009	+0.272
`grid`	`coords`	`discursive`	0.280	-0.009	+0.289
`grid`	`occupancy`	`iconic`	0.223	0.005	+0.218
`grid`	`occupancy`	`tabular`	0.053	-0.005	+0.058
`grid`	`occupancy`	`discursive`	0.049	-0.005	+0.054
`order`	`occupancy`	`iconic`	0.258	-0.005	+0.264
`order`	`occupancy`	`tabular`	0.211	0.033	+0.178
`order`	`occupancy`	`discursive`	0.199	0.033	+0.166
`order`	`positions`	`iconic`	0.246	-0.004	+0.250
`order`	`positions`	`tabular`	0.543	0.146	+0.396
`order`	`positions`	`discursive`	0.533	0.148	+0.384

The two `graph` iconic rows are given for completeness and are not interpreted, for the reason given in E.4. The one cell that falls clearly below its own ceiling is `contain`/`depths` in the iconic condition, at −0.171; `graph`/`adjacency` iconic is at its ceiling, with probe and ceiling both at the floor. The two identity-free cells in which the diagram exceeds its ceiling by more than either propositional rendering are `grid`/`occupancy` and `order`/`occupancy`, the two families whose structure is arranged on a line or a plane. Everywhere else a propositional rendering has the largest clearance.

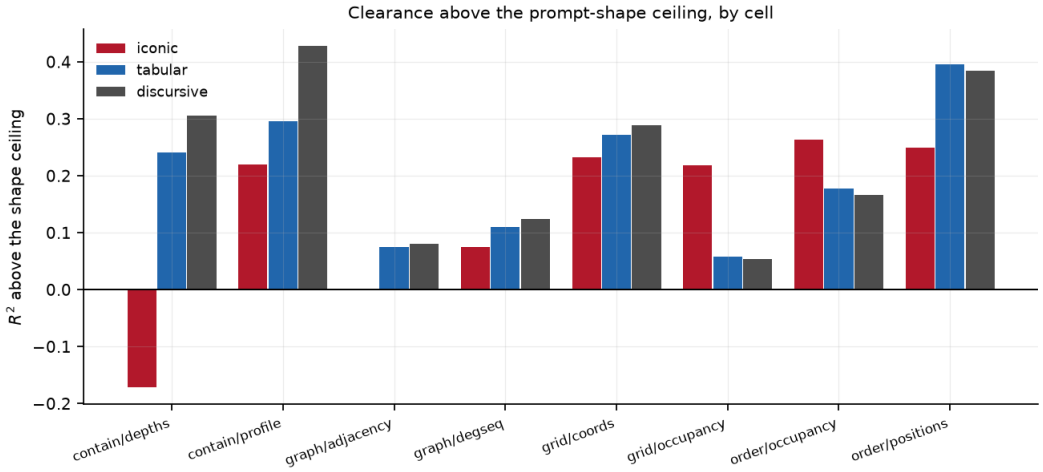


Figure E2: Clearance above the prompt-shape ceiling for every cell: peak probe R^2 minus the ceiling for the same cell.

E.7 Layer profiles

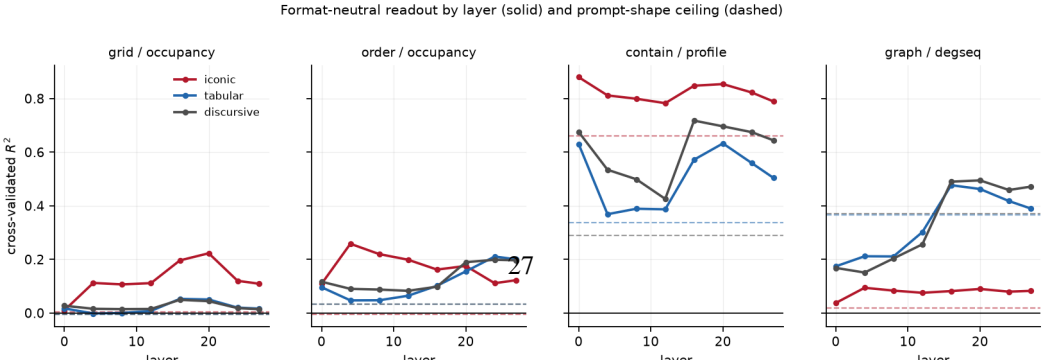
Clearance and layer profile answer different questions and are reported separately. Clearance asks whether prompt counts alone would have scored as well. The layer profile asks whether the fit is already present at the embedding, which would make it surface encoding however large its clearance.

family	identity-free target	layer 0	peak	at layer	ceiling
`contain`	`profile`	+0.880	0.880	0 of 27	0.660
`graph`	`degseq`	+0.037	0.094	4 of 27	0.019
`grid`	`occupancy`	+0.010	0.223	20 of 27	0.005
`order`	`occupancy`	+0.109	0.258	4 of 27	-0.005

`grid`/`occupancy` in the iconic condition runs +0.010, +0.112, +0.107, +0.112, +0.196, +0.223, +0.120, +0.109 across the probed layers: essentially zero at the embedding, rising through the middle layers, then declining. That is the profile expected of a representation the network computes, and the two propositional formats stay flat throughout. It is also what attention moving the prompt's own code to the last token would produce, and E.8 shows that the values are within what such pooling could supply.

`order`/`occupancy` runs +0.109, +0.258, +0.220, +0.199, +0.162, +0.176, +0.111, +0.123, and the two checks give different answers for it. It exceeds its ceiling by more than any other iconic cell, so it is not a length artifact, but it peaks at layer 4 with a nonzero fit already at the embedding, which is not the profile of a computed representation. Whether a fit that early counts as a representation is a question this design does not settle, and we leave the disagreement between the checks unresolved.

`contain`/`profile` in the iconic condition runs +0.880, +0.812, +0.799, +0.783, +0.848, +0.854, +0.823, +0.789, which is the opposite profile: its highest value is at the embedding and it declines from there. It is the largest value anywhere in Study 2, and it reflects prompt shape. Once the ceiling is subtracted, the iconic lead in that cell reverses and discursive has the largest clearance. This is the third analysis in which the identity-free crossover does not hold.



refitted on a separate battery of 500 `grid` items at layer 20. A cell counts as exceeding the bound only if the lower end of its 99% bootstrap interval is more than 0.03 above it.

format	occupancy map	coordinates	occupied cells, sorted
iconic	0.251 (bound 1.000)	0.163 (0.203)	0.430 (0.572)
tabular	0.051 (0.318)	0.253 (1.000)	0.139 (0.452)
discursive	0.045 (0.322)	0.224 (1.000)	0.175 (0.455)

No code in any format exceeds its bound. The occupancy the probe reads from the diagram at layer 20 is therefore within what content-independent pooling of the prompt's tokens could supply, and is not evidence that the model computed the map; this qualifies the `grid` result of E.6 and E.7. The test is one-directional: falling within the bound does not show that the model only pools.

A second analysis trained a probe on one format's stream and tested it on another's. For both codes and every pairing of the diagram with a text rendering, there is no shared linear readout above chance (squared correlation 0.0018 to 0.0068). This does not show that the codes differ, since a shared code in a different basis would also fail to transfer.

Two limits apply to every number in this appendix, and none of the controls above removes either. First, decodability does not imply use: a probe that recovers occupancy shows the information is linearly available at that point in the residual stream and says nothing about whether any later computation reads it. Second, decodability does not imply iconicity, since a linear probe cannot distinguish a representation whose structure mirrors what it represents from a propositional representation carrying the same content.

F THE THREE-DIMENSIONAL ROTATION BATTERY

This battery is unregistered and exploratory throughout. It asks whether one isometric block figure is a rotation of another, in five conditions — `iconic_axes`, `iconic_bare`, `iconic_unlabeled`, `tabular` and `discursive` — divided between two questions, `is_rotation` (yes/no) and `angle` (90/180/270). It ran to completion on five models at 1,200 trials each, and none of its positive results held up. We report it because the way each of its three findings failed is informative, and the third failure has general consequences.

F.1 ANGLE IS RESPONSE BIAS

The gold answers are nearly uniform: 90 in 35.0 percent of items, 180 in 32.5 and 270 in 32.5, so answering "90" every time scores 0.350.

model	modal answer	share	accuracy
Qwen2.5-7B	90	96.5%	0.356
Llama-3.1-8B	90	91.8%	0.348
Claude Opus 4.8	90	67.9%	0.521
GPT-5.6 Sol	90	66.1%	0.427
Qwen2.5-1.5B	90	66.8%	0.343

Three models are within a percentage point of the always-90 baseline, having answered 90 on nine trials in ten. The question elicits a default answer rather than a computation, and `angle` is excluded from analysis. A larger token budget does not change this.

F.2 THE HOSTED MODELS ARE EXCLUDED

Claude's `max_new_tokens` was 768 and its median completion in the two propositional conditions was exactly 768, with 73.3 percent of `discursive` and 68.3 percent of `tabular` trials

truncated against 0.4 percent of `iconic_bare`. Truncated trials carry no answer, so its propositional cells are computed on the roughly 27 percent of trials short enough to finish, which is a selection on completion length. The conditions whose answers can be read show no ability on the task: `iconic_bare` 0.500, `iconic_axes` 0.442, `iconic_unlabeled` 0.308. GPT's median completion was 56 tokens with no truncation and accuracy at or below chance in all five conditions, so it is not attempting the task. A rerun at a cap large enough for Claude was not made.

F.3 IS_ROTATION ACCURACY IS AT ITS OWN BASELINE

The `is_rotation` gold is 345 yes to 255 no, so the majority baseline is 0.575, not 0.5. Results for Llama-3.1-8B at the 2048-token cap:

condition	accuracy, all trials	accuracy, answered trials only
`iconic_bare`	0.625 [0.538, 0.712]	0.630 [0.544, 0.717]
`iconic_unlabeled`	0.600 [0.512, 0.688]	0.600 [0.512, 0.688]
`iconic_axes`	0.583 [0.495, 0.672]	0.609 [0.519, 0.698]
`tabular`	0.417	0.543 [0.442, 0.645]
`discursive`	0.367	0.500 [0.396, 0.604]

Every interval contains 0.575. An earlier draft reported this model as above chance in all three iconic conditions and at chance in both propositional ones, and read that as a reversal of the main study's direction; both claims were computed against 0.5 and are withdrawn.

The apparent iconic advantage is also substantially an artifact of how censored trials are scored. Counting a non-terminating trial as incorrect puts `discursive` at 0.367; scoring only trials that produced an answer puts it at 0.500. The largest iconic-minus-propositional difference falls from 0.258 to 0.130 on that choice alone.

F.4 EXPLAINING NON-TERMINATION

On a 768-token cap, Llama-3.1-8B gave no answer on 17.5 percent of discursive and 13.3 percent of tabular trials, against 1.25 to 3.33 percent in the three iconic conditions. This is a decoding pathology, not a property of the renderings. Decoding is greedy, and at a 2,048-token cap all 1,111 terminating trials reproduce exactly while none of the other 89 terminates; all 89 are degenerate repetition loops, and none of the 1,111 is. A repetition penalty of 1.15 applied to generated tokens only removes all 89, so a format comparison run under a single decoding rule can produce what looks like a format effect. The penalty also changes answers: on 40.9% of the trials that terminate under both decoders, unevenly by format (29.8% discursive to 45.8% iconic), and on 23.3% of Qwen2.5-7B's answers on the follow-up study's pilot battery. No accuracy comparison is therefore made across it.

A less invasive control is available. A no-repeat rule that forbids an exact repeat of any 20-token sequence in the model's own output, excluding the prompt, halves Llama's non-termination on the follow-up battery (from 14 of 136 trials to 7). It changes 0.8% of Llama's answers and 0.3% of Qwen2.5-7B's, and leaves the type-token ratio of terminating output essentially unchanged (0.591). It does not remove the pathology: every remaining truncation is still degenerate, because the model varies the loop enough to avoid an exact match. It does reduce the cost of controlling non-termination by a factor of forty or more, and the follow-up study uses it for that reason. In that study's full run, 426 of Llama-3.1-8B's 9,296 completions still reached the cap, and 390 FU5-wording trials were excluded as truncated, unevenly across arms (17.0% in the shuffled `order` layout against 4.5% overall). A further 167 Llama completions ended without an answer line, 113 of them after more than 300 tokens, consistent with the rule closing a loop with an end-of-sequence token; these are scored, not excluded, and 109 of the 113 long ones are incorrect.

G FU5 OUTCOME SENTENCES AND SENSITIVITIES

The preregistered outcome sentences, as the frozen analysis printed them. Each ends with the plan's tail "(open-weight models 7B–8B, reasoning requested, key-resolved scorer; registered scorer: y)" and a length bracket "[Length sensitivity: body-token-adjusted estimate z, VIF v.]".

contrast	sentence	registered scorer	length-adjusted (VIF)
C1a, null	The letter key did not account for the drawing's cost in `grid`, `order` and `graph` in this battery (C1a within ±0.25 log-odds). Per-family estimates disagree in sign (graph −0.08, grid +0.28, order −0.10).	+0.87	+0.09 (1.09)
C2, holds	Adding the table's rows beneath the drawing raised accuracy in `grid`, `order` and `graph` (C2 = +1.62 log-odds). This shows the cost is recoverable by stating the facts, not where it arises.	+2.33	+2.43 (13.96)
C2b′, holds	With the facts stated in a table, adding the drawing reduced accuracy in `grid`, `order` and `graph` more than adding the same facts as sentences (C2b′ = −0.55 log-odds).	−0.56	−0.27 (1.25)
C3 < 0, C3-DiD neither	The shuffled layout was less accurate (C3 = −1.25); whether it lost more than unsorting costs a table was neither shown nor bounded within ±0.25 log-odds (C3-DiD = −0.29), so no effect is attributed to spatial alignment. [Post hoc, 24 Sep 2026:] Per-family estimates of C3-DiD disagree in sign (grid −1.10, order +0.85).	−1.05	−1.15 (1.06)*

contrast	sentence	registered scorer	length-adjusted (VIF)
C4, not detected	[Post hoc wording; the plan fixes no sentence for this pattern.] Whether spacing the glyphs into single tokens, which also lengthened the drawing by about 18 tokens, changed accuracy in `order` and `graph` was neither shown nor bounded within ± 0.25 log-odds (C4 = −0.08 log-odds). Per-family estimates disagree in sign (graph +0.03, order −0.12).	−0.02	+0.43 (2.78)

The key-resolved scorer is primary in FU5 because the registered scorer marks letter answers wrong in exactly the keyed arms. *The C3 length-adjusted value is from a post hoc reliable refit (−1.149, SE 0.112); the frozen analysis printed −1.13 from a degenerate fit, and its sentence carries that value. The full sensitivity set, the per-cell estimates and the post hoc checks are in `studies/fu5/results/`, and the rulings on what these results license are in `studies/fu5/verification/results-ruling-2026-09-24.md`.

Design, exclusions and checks. FU5 is registered at studies/followups/PLAN-fu5.md (tag fu5-frozen) and used greedy decoding with a 1,536-token cap and the no-repeat rule of Appendix F.4. Per family, A0 is −1.48 (grid), −1.35 (order) and −4.35 (graph). The C1a equivalence is marginal: C1a = +0.10, 90% CI [−0.05, +0.25], inside the margin by 0.004, and the upper limit is +0.27 to +0.32 in the three sensitivities that break it. The alignment account predicted C3 > 0; because a shuffled labelled layout also requires looking up every label, the negative C3 neither supports nor counts against it. With truncated trials scored incorrect, C4 is −0.04 (90% CI [−0.21, +0.14]). The key × format interaction (C1 = −0.12, declared underpowered) and labelling every point under the order lane (C5 = −0.08) were not detected. Truncation, the only exclusion, occurred on Llama-3.1-8B alone (390 of 8,628 trials, 4.5%), mostly in the order drawn arms; scoring those trials incorrect changes C1a to not detected and C4 to null, and leaves C2, C3 and C2b′ unchanged. A further 167 Llama completions ended without an answer line and are scored, mostly incorrect (Appendix F.4). Standard errors were checked post hoc over 21 orderings of the data, and every reliable fit gives the reported label. Exploratory, produced after the results were read: on graph/degree the plain and spaced drawings drew the answer "2" on 74–89% of trials, and answers tracked the true degree at r = 0.11–0.27, against 0.66–0.86 in every arm that states the facts, which is consistent with §5.5; and on the same 223 items FU5's question wording gave a larger gap than Study 1's (−1.58 against −1.07 log-odds; interaction +0.55, SE 0.25), so the two A0 values should not be compared. FU5 does not test alignment in graph, tests derivation only in order, treats length as a covariate rather than a control, apportions no share of A0, and says nothing about representation inside the model. Its departures are in Appendix D, and the decoding measurements in Appendices C and F come from its pilot runs.