Ex 2.1.5: disjoint vocabularies and more named colors

> Names and hex codes. Both sublanguages train, and each builds its own linear color geometry. But the two geometries resist being combined, even under pressure from a narrow residual stream. Keeping them apart apparently costs less than merging them would.

Do two surface languages for the same domain converge on one internal geometry?

The corpus for this experiment holds two sublanguages that are never seen together: color-mixing equations written with names, and the same arithmetic written as hex codes. For example:

```
melon + ultramarine = …
#e26 + #48a = #958
```

Nothing directly ties a name to a hex value; there are no alias lines and no mixed-form equations. So if the model places both vocabularies in the same latent geometry, the cause is pressure internal to the network, e.g. capacity constraints.

Graded concept labels would be computed from color values (see M1/Ex-1.7), so they can be used with either form. If this experiment finds that the two sublanguages use different latent representations, a later one could test whether anchoring the same concept in both forms pulls the geometries together, and it could quantify how many are needed. It would be remarkable if a single anchor aligned the whole cube.

Lineage: the base language (ex-2.1.1, 2.1.2) bridged names and hex with alias and cross lines, and its 27-name sub-grid turned out to be too sparse to grade (ex-2.1.4). The single-vocabulary experiments (ex-2.1.3, 2.1.4) removed hex entirely. This experiment keeps both forms but removes the bridge, and replaces the coarse named sub-grid with 140 real color names spread through the full 8-bit cube.

Data (the language)

First, let's define two RGB grids (cubes):

Space	Levels	Hex digits	Example ("cyan")	Precision	Colors
hex grid	16	3	`#0ff`	4-bit	4,096
full cube	256	6	`#00ffff`	8-bit	16,777,216

Names. The named palette comes from the xkcd color survey: all 949 names, ordered by farthest-point selection so that the first N form the most uniform palette available at that size. The center condition uses $N = 140$; at that size the minimum and median nearest-neighbor distances are ≈ 41 and ≈ 46 (8-bit Euclidean), against 4 and 27 for the CSS keyword list. Names map to points on the full cube.

▶ **On not filtering by name…**

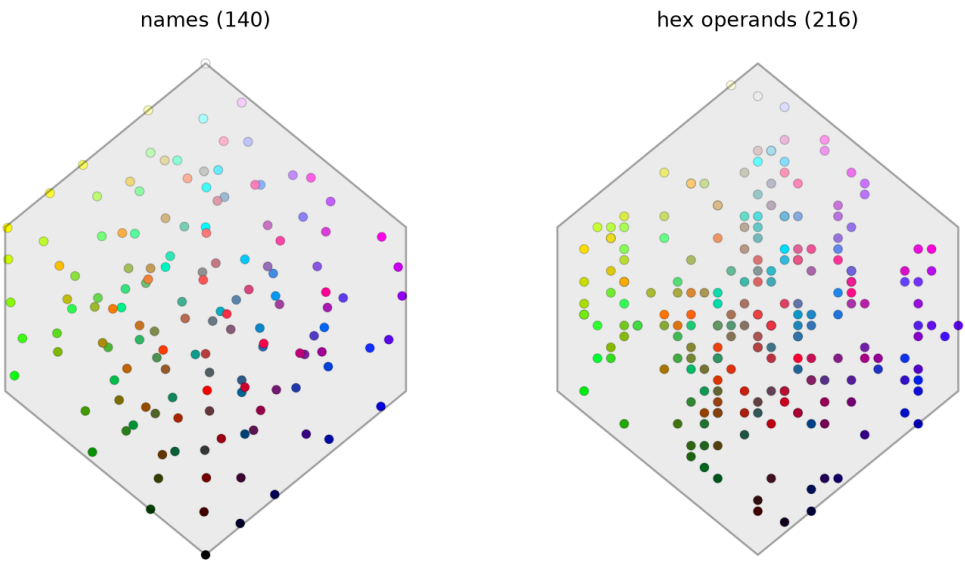
Hex. Hex operands are drawn from a fixed random subset of the hex grid, sampled point-by-point rather than as a regular sub-grid. The subset constrains operands only: the correct completion of a hex equation may be any point on the hex grid.

To prepare training data, we pick two operands from one sublanguage (names or hex) and compute the result. The mix is the channel-wise round-half-up mean, always computed in the full cube; the two forms differ only in how values enter and leave it: - Named colors already sit in the full cube, but the answer must be snapped to the nearest named color. Distance ties break by a coin flip seeded from the mix value. - Hex operands are lifted to it by digit repetition (`#f80` → `#ff8800`) and the answer is snapped back to the hex grid.

Two forms appear in every condition of the sweep, and a third only in the *bridge* arm:

Form	Example	Sweep conditions
named	`melon + ultramarine = <name>`	all
hex	`#e26 + #48a = #958`	all
cross	`melon + #48a = #<hex>`	bridge arm

(Examples are illustrative until the corpus code lands.) Named equations always answer with a name, and hex equations with a hex code, so the answer form is determined by the prompt form; nothing about the result value changes which vocabulary the answer uses.



The two operand sets in the RGB cube (center corpus). Left: the 140 named colors, spaced by farthest-point selection; right: the 216 hex operands, a uniform draw from the 4,096-point grid.

▶ **Is the clumping in the hex panel a problem?…**

corpus	named	hex	cross	held-out named	held-out hex	max line	answer ppl	reachable
n140-h2048	50,040	49,960	0	1,946	419,226	85	86	140
n140-h216	49,990	50,010	0	1,946	4,644	85	86	140
n140-h216-bridge	43,845	44,149	12,006	1,946	4,644	72	86	140
n250-h216	50,076	49,924	0	6,225	4,644	70	153	248

Corpus statistics, one row per corpus (conditions that share names × hex × bridge share a corpus). The numbers from the design study held up: answer perplexity is 86 over 140 names with every name reachable as an answer (the 250-name palette reaches 248), and the longest line is 85 characters, inside the block size of 128. Line counts are the 100,000 lines of the training corpus split by form.

Measurements

Exact-match accuracy is scored per form on seen and held-out operand pairs. On a small palette, raw accuracy can flatter a weak strategy, so each number sits beside two reference guessers. One never reads the prompt: it always answers with the centroid of the training answers. The other is handed the neighborhood of the answer for free: it guesses uniformly among the k palette entries nearest the true mix (the irregular-palette version of the one-step-shell null in earlier reports). The model has learned something only where it beats both.

Geometry is read with ridge probes (leave-one-out, as in `ridge_probe_loo`) fitted for operand and mix values at every layer and every token position, separately per form. The full scan replaces the hand-picked probe sites of earlier reports with a map, so the alignment measures below don't depend on us choosing the right position in advance.

> **Amendment: a second holdout protocol**
>
> The preregistered estimator holds out one *equation* at a time. But the colors and hex digits of the held-out equation still appear in other equations in the fit, so a probe can score well just by memorizing which token has which value. That is a fair way to ask whether a value is present at a site, and one analysis below uses it for that. But it's not a good test of H2, which is a statement about geometry.
>
> So each site now also gets a stricter holdout. To score a value, every equation containing it (as either operand or the answer) is removed from the fit, and the probe must place the unseen value using only the others. The whole-value holdout is needed because the same hex digit serves all three channels, and a color can be operand 1 in one line and operand 2 in another.
>
> Both scores are reported. The preregistered one decides H2 (and it turns out the two agree at the site used for that decision (0.944 against 0.941). We also added two landmarks, the spaces closing each operand, after the strict holdout showed that the value of a named operand sits on them rather than on the characters of the name.

One subtlety: token positions don't naturally line up. Names vary in length, and hex expressions are shorter than named ones. Positions are therefore indexed by grammar landmarks (the last character of each operand, the operator, the pre-answer position, and answer characters counted from the start and end of the answer), and the cross-form measures compare only at landmarks the two forms share.

Alignment between the two forms is scored two ways:

- Transfer ratio ρ: zero-shot cross-form R^2 divided by within-form R^2, at the same layer and landmark. Zero-shot means the probe is fitted on the activations of one form and applied unchanged to the other. Two guards keep the ratio well-behaved. First, ρ is reported only where the within-form $R^2 \geq 0.5$; below that the site isn't measuring geometry, and a small denominator makes the ratio erratic. Second, negative cross-form R^2 clips to zero, so $\rho \in [0, 1]$ with 0 meaning no transfer.
- Principal angles between the row-spaces of the fitted probes of the two forms: a graded measure of whether the two decoders use the same directions of the residual stream.

Hypotheses

- **H1.** The disjoint language trains at the center condition: hex+hex exact match on unseen pairs is comparable to the hex levels of ex-2.1.1, and name+name held-out exact match clears the k-NN analogues of the neighborhood nulls.
- **H2.** Each sublanguage develops latent linear color geometry, with form-specific layout: name-form probes decode operands and mix with mix $R^2 \approx 0.9$ at the pre-answer position in the last layer; hex-form answers assemble just-in-time, channel by channel, with pre-answer full-mix R^2 staying low. An elevated pre-answer hex-mix R^2 would instead be evidence of cross-form coupling (see H3).
- **H3.** The two latent geometries live apart in sweep conditions that have no bridging grammar and width d64: $\rho < 0.2$, and the two probes' row-spaces show large principal angles. ($0.2 < \rho < 0.8$ reads as partial sharing and falsifies the crisp version of both H3 and H5.)
- **H4.** Narrowing the stream aligns the forms: ρ and subspace overlap rise monotonically over d64 → d32 → d16, with d16-L8 the most aligned condition.
- **H5.** Adding the cross form produces alignment: the bridge condition (d64) reaches $\rho > 0.8$. The star design tests this at one width only; bridge × width conditions are candidates for a follow-up round if H4 shows width matters.
- **H6.** At L8, name+name accuracy improves at fixed width and the mix crystallizes before the last layer.

The sweep

A star design around one center condition, three seeds per condition:

Condition	Names	Hex ops	Bridge	Width	Depth
center	140	216	none	64	4
L8	140	216	none	64	8
d32	140	216	none	32	4
d16	140	216	none	16	4
d16-L8	140	216	none	16	8
hex-dense	140	2048	none	64	4
palette-250	250	216	none	64	4
bridge	140	216	cross	64	4

Each arm provides data to score the hypotheses. L8, the width conditions, and d16-L8 score H4 and H6; the bridge condition scores H5. The two density arms attach to H1: *hex-dense* checks that hex accuracy and geometry aren't artifacts of the 216-point operand subset (expected: little change — hex arithmetic in ex-2.1.1 generalized from far sparser coverage), and *palette-250* extends the density axis of ex-2.1.3 to the irregular palette (expected: named held-out accuracy holds or improves, with misses staying neighbor-level).

Attention is held at 8 heads × 8 dims in every condition; only the residual stream and the MLP scale with width. The ngpt-scaling sweep validated widths {32, 64, 128} under this scheme, so d16 sits one step below the tested range. If the d16 conditions train poorly, that is an architecture effect to report, and the width trend in H4 rests on d64 → d32.

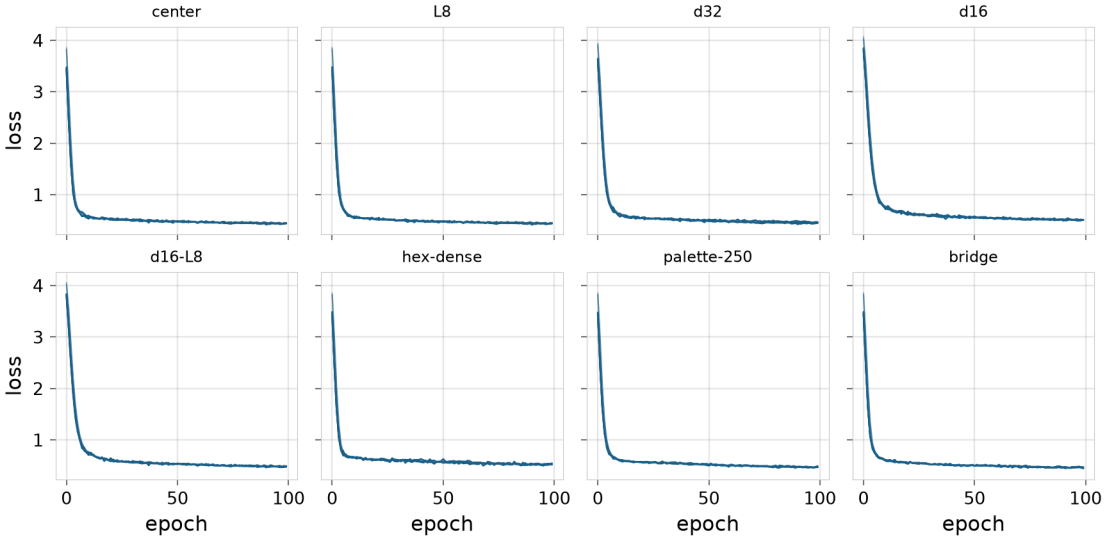
Parameter counts per condition (seeds share a count). The full sweep — 4 corpora, 24 training conditions, 24 evals.

▶ **Runtime…**

condition	width	depth	params
center	64	4	200,713
L8	64	8	397,329
d32	32	4	67,593
d16	16	4	25,609
d16-L8	16	8	50,193
hex-dense	64	4	200,713
palette-250	64	4	200,713
bridge	64	4	200,713

Training

All 24 conditions converge smoothly under the shared 100-epoch schedule — including the d16 conditions, which sit one step below the width range validated by ngpt-scaling. No width-gated instability appeared.



Loss per epoch for every condition: validation solid, training faint, one line per seed. All conditions share the character vocabulary, so per-token losses are comparable across panels.

Exact-match accuracy (H1)

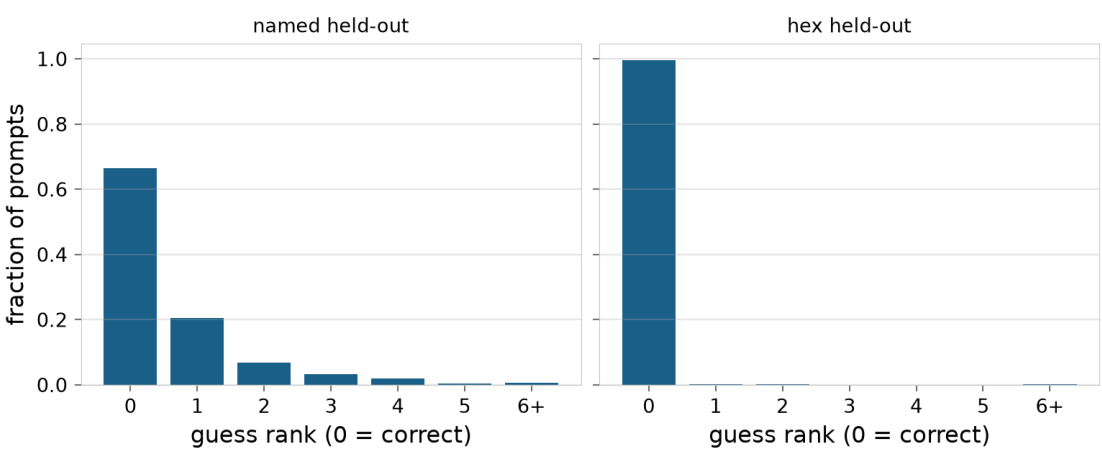
The center condition answers held-out hex equations almost perfectly (0.996 exact) and held-out named equations at 0.667, with no malformed completions in any of its eval sets. The named score sits far above the prompt-blind centroid (0.043). One null turned out to be simple: in this language the true answer is always the candidate nearest the mix, so guessing uniformly among the k nearest candidates scores exactly $1/k$. The hardest version is a coin flip between the two nearest candidates, 0.5, and the model clears that too, by a margin the 1,946 held-out named pairs can resolve.

eval set	exact	guess dist	floor dist	blind null	2-NN null
named seen	0.841	0.115	0.106	0.027	0.500
named holdout	0.667	0.124	0.112	0.043	0.500
hex seen	1.000	0.034	0.034	0.000	0.500
hex holdout	0.996	0.034	0.033	0.000	0.500

Center condition, mean over three seeds. Distances are Euclidean in the unit cube: guess dist from the emitted answer to the exact mix, floor dist from the true answer to the exact mix (the quantization floor). The guesses sit near the floor even where exact match misses.

Named accuracy is well short of 1 even on pairs the model trained on (0.841), which reads at first like a model that never learned the geometry. In contrast, ex-2.1.3 answered a 216-color named vocabulary at ≈ 1.0.

But the ex-2.1.3 colors sat on a regular grid closed under mixing: every answer landed on a vocabulary point, and the runner-up was a full grid step away (0.2 of the unit cube). Here the answer is the nearest of 140 irregularly placed names, and the median gap between the nearest name and the second-nearest is six times smaller. In the exploratory section ("The named ceiling is a resolution limit"), accuracy tracks that gap prompt by prompt, and a single effective-precision number per channel accounts both for named accuracy in this experiment and for the earlier grids scoring near 1. So the geometry is present; what runs out is resolution, as in ex-2.1.3.



Rank of the emitted answer among the candidate vocabulary of the form, ordered by distance to the true mix; rank 0 is the correct answer. Center condition, pooled over seeds.

condition	named held-out	hex held-out	answer ppl
center	0.667	0.996	86
hex-dense	0.582	0.997	86
palette-250	0.564	1.000	153

The density arms, both far above their nulls. Hex accuracy is insensitive to the operand-subset size, as expected. At 250 names the mean named held-out accuracy drops to 0.564, in the predicted direction — the answer perplexity is 153 against 86 — though the three seeds span 0.47 to 0.66, so the effect is soft. The hex-dense mean of 0.582 looked at first like an unpredicted dip under denser hex operands, but it rests on one seed (0.39 against 0.67 and 0.69); the other two match the center condition, so there is little evidence of a real hex-density effect on named accuracy. Three seeds is thin for a difference this size.

More names and fewer correct answers is a strange pairing. Every condition in this table is d64-L4, and the 250-name corpus has the same 100,000 lines, so neither capacity nor data volume changed. Two other things may explain it.

First, answering correctly became harder. The median gap between the nearest and second-nearest name to the exact mix falls from 0.033 to 0.028 of the unit cube, so more prompts ask the model to place the mix inside a tighter boundary. A reference guesser with the precision of the *center* condition, answering the 250-name prompts, predicts 0.613. The observed score is 0.564, so roughly half of the drop from 0.667 can be attributed to the tighter precision demand.

Second, each training pair got less supervision. The same ≈ 50,000 named lines now spread over 21,703 distinct training pairs instead of 7,906, which is 2.3 lines per pair against 6.3. One sign of this is that the gap between seen and held-out accuracy nearly closes at 250 names (0.617 against 0.564), where the center condition runs 0.841 against 0.667. Fitted precision agrees: on *seen* pairs it is 0.033 at 250 names, the same as what the center condition manages on pairs it has never seen (0.033). At 140 names a training pair recurs often enough to be remembered; at 250 it barely does.

Neither reading needs a capacity limit, and the residual gap after standardizing (0.564 against 0.613) is small enough that three seeds cannot separate a genuine precision cost from the seed spread.

We'll look at capacity more closely in the width sweep (H4), below.

Within-form geometry (H2)

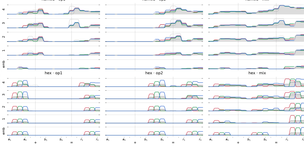
Outcome: the named form matches the prediction, with the mix decodable at $R^2 = 0.94$ at the pre-answer position in the last layer, and hex never assembles a pre-answer mix at all.

The figure below plots probe R^2 at every depth × landmark, for the two operands and the mix, per form; center condition of the sweep only. We used the stricter of the two probe holdout protocols described under Measurements.

To be concrete about what one point on a line is: each combination of {seed, form, target, depth, landmark} is a 'site', and each site gets its own independently fitted ridge probe. The target is the probed quantity (operand 1, operand 2, or the mix). The probe inputs are the residual-stream activations of the N equations at that one token position and depth (an N × C matrix), and its regression targets are the RGB triples of the probed quantity (N × 3). A probe never sees activations from any other depth or position, in fitting or in scoring. The height of a line is the held-out R^2 at that site under the strict holdout protocol. The colored step-lines are the per-channel scores averaged over the three seeds, and the hairlines show the seed spread.

Because the sites are independent, R^2 says nothing about how probe directions relate across sites. A high score means a value-aligned readout direction exists at that site, but the score is the same whichever direction it is, so two sites can score alike with orthogonal readouts. The staggered per-channel bumps in the hex panels, similar as they look, are three separate fits. At the embedding they may well be one readout of the shared digit-embedding axis seen at three positions, but checking that would take a direction comparison (as in the principal-angle measure), which this figure does not make.

A few reading notes. The lines are stepped because each landmark is a discrete character position; a straight line between two would claim measurements that were never made. Risers drawn in lighter ink skip over a variable-length middle of a name (0 to 22 characters, depending on the name), so a named trace that steps up across one of those risers is the name completing; hex lines skip nothing, since all four characters of a hex operand are landmarks. Every minor panel shares the same 0 and 1 gridlines. Scores below zero are drawn at the floor (clamped to zero).



Per-channel probe R^2 under the strict holdout, center condition. Named form on top, hex below; columns are operand 1, operand 2, and the mix. Within a block, depth runs upward from the embedding and the x axis across the grammar landmarks, one step-line per RGB channel. The grey area is the three-channel mean, and the two hairlines are the highest and lowest of the three seeds.

The named form matches the prediction. The mix is decodable at $R^2 = 0.94$ at the pre-answer position in the last layer, and all three seeds land there ($0.93, 0.95, 0.95$) — for mix *values* the probe never saw, which is a stronger claim than the preregistered estimator makes. That estimator gives 0.944 at the same site, so H2 is carried either way and nothing turns on the change of protocol. Where the mix *first* appears is less settled: one seed already reads it at the `=` sign by the second-to-last layer while the other two stay near 0.3 there, visible as the upper hairline lifting away from the outline at `=`.

The value of a named operand is not on its name. At the last character of operand 1 the strict probe scores -0.05, and operand 2 scores 0.00: nothing. The value sits instead on the spaces that follow. It reads 0.23 on the space closing operand 1, rises to 0.47 across the `+`, and reaches 0.64 by the pre-answer position. A variable-length name has no fixed slot, and the model appears to resolve it into the first fixed position available. The value could not have appeared any earlier: the model reads left to right, and the first characters of a color name do not determine its value: `sea green` and `salmon` share a prefix and nothing else, and predicting the RGB of a held-out name from the other names sharing its two-character prefix scores -0.21, below the palette mean.

Part of what the closing space holds is lexical rather than chromatic. 99 of the 140 names are multi-word and the modifier words recur, and holding out whole word families costs that site 0.43 more than a color-matched holdout of the same size does.[1] The site after the `+` is much less lexical, so the 0.47 there is the sturdier number. This is also where the preregistered estimator diverges most: it reads 0.75 at the last character of the name. That reading is name recognition — the identity is recoverable there, but the value is not placed geometrically.

Hex relays its channels and does not retain them. Each channel is legible at its own digit and near zero at the next, a strict relay rather than an accumulation. At the embedding the probe reads 0.42 at the position of the digit itself, so the 16 digit embeddings do carry a real magnitude axis. Depth then improves it, rising to 0.78 at the last digit of operand 2 by the last layer.

Hex reads nothing at the delimiters (-0.23), which is what a form whose three digits sit at fixed offsets should do, because it needs no summary slot.

▶ Why the preregistered estimator reads 1.00 at the embedding…

Hex never assembles the mix, and the named pre-answer site carries everything at once. The hex mix reaches only 0.20 at the pre-answer position and 0.55 mid-answer, and the mid-answer reading sits where the answer digits themselves are in the input (see the surface-text caveat under Exploratory analyses). Nothing in the hex panels looks like a whole pre-answer mix, so the coupling signal that H2 reserved judgment on did not appear. This is likely to matter for anchoring. At the named pre-answer position the mix reads 0.94, operand 1 0.64, and operand 2 0.62, so one site carries the whole equation. Hex has no such site, and a general RGB anchor needs one where the whole concept lives, even if only one color is being anchored.

site	per-equation	strict	difference
named mix_pre-answer (decides H2)	+0.944	+0.941	-0.003
named mix_at the =	+0.603	+0.579	-0.023
named operand 1_its last character	+0.750	-0.047	-0.797
named operand 1_the space after +	+0.683	+0.468	-0.214
hex operand 2_red at its own digit (embedding)	+1.000	+0.416	-0.584
hex operand 2_red at the green digit	+0.821	-0.269	-1.090
hex mix_pre-answer	+0.376	+0.196	-0.180

Sites where the two holdouts disagree, at the last layer except at the embedding. The difference column is how much of a reading was identity recovery: near zero means the site holds a value the probe can place without having seen it, and a large negative gap means the probe was looking the value up. Either estimator decides the H2 site the same way, so adopting the stricter one changes no conclusion.

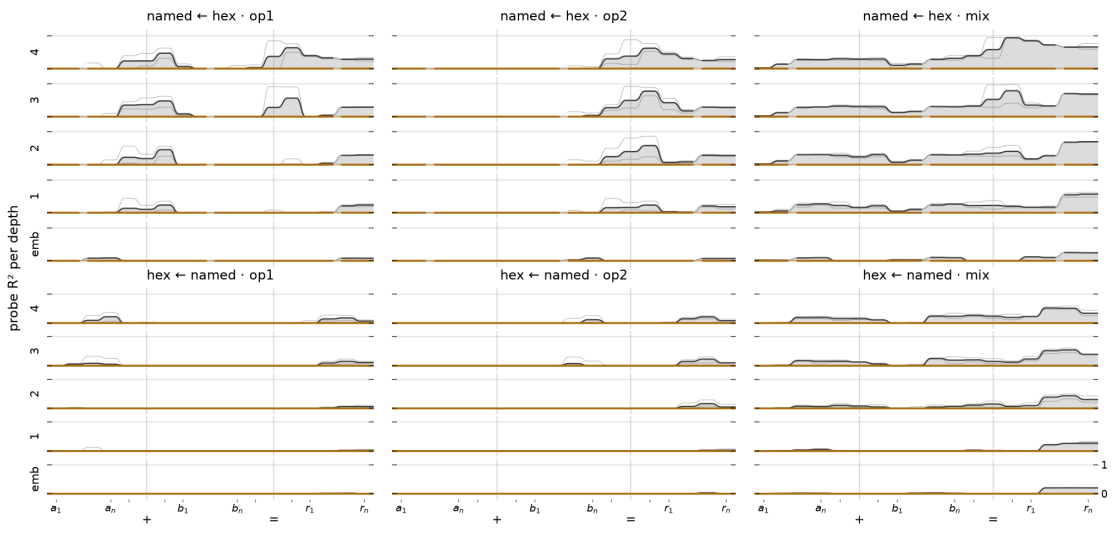
Cross-form geometry separation (H3)

H3 predicted $\rho < 0.2$ and large principal angles. Both hold conclusively: ρ is **zero** at every cell where the guard defines it, in both directions and for all three targets.

Zero is a stronger statement than it looks. ρ clips negative cross-form R^2 to zero, so in principle a floor reading could hide a score just below break-even. Nothing here is close: all 1,620 cross-form cells in the scan are negative, and the best of them is -0.000. Take the pre-answer position in the last layer as an example. The mix probe fitted on the named form reads 0.94 there; the hex probe applied to the same activations reads -27. That is far worse than answering with the training mean, which suggests the decoder is pointing in an unrelated direction.

The figure below shows both readings on the same axes, per form and target. The grey area is what the own-form probe recovers; the amber line is what the other-form probe recovers from the same activations. The amber fraction of the grey is ρ, and it is zero everywhere. Note that grey in the hex row is low because the strict readings for hex itself are low (H2 found it relays its channels rather than holding them), so the transfer claim rests on the named row and on the magnitudes quoted above.

▶ **Which within-form estimate the denominator uses…**



Zero-shot cross-form transfer at the center condition, seed-averaged. Rows are the form the probes are applied to; columns are the probe target; within a panel, depth runs upward from the embedding and the x axis across the grammar landmarks. Grey is the own-form probe under the strict holdout (the same quantity as the per-channel figure under H2), with its three-seed envelope; amber is the other-form probe applied unchanged. Both are floored at zero for drawing.

site (mix)	within named	within hex	cross → named	ρ → named	cross → hex	ρ → hex	angles
depth 3, ae1 (preregistered)	0.70	0.55	-8.95	0.00	-5.91	0.00	66° · 76° · 86°
depth 4, pre-answer (clean ground)	0.94	0.20	-26.82	0.00	-0.28	—	75° · 86° · 88°

The mix probe at two sites, seed-averaged; within-form readings use the strict holdout. The preregistered per-equation ratios give the same reading: 0.00 and 0.00 at the preregistered site, 0.00 and — at the pre-answer site.

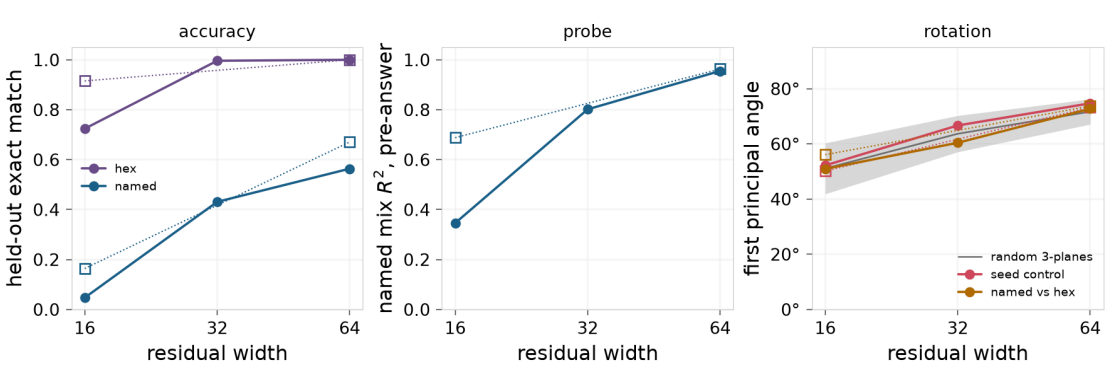
Neither site was picked using ρ. The preregistered site maximizes the within-form R^2 of the weaker form, so the two probes are compared where both have something to carry. The pre-answer site is where the geometry of the named form is strongest and where the answer text cannot contaminate either form. The two sites disagree about where hex is legible: at the pre-answer site, the within-form hex reading falls under the guard on ρ, so the ratio into hex reports nothing there rather than zero. They agree about everything that bears on H3. The pre-answer site is also where the two mix decoders come closest to sharing a direction, and even there the smallest of the three principal angles is 75°.

The guard behind ρ turns out not to matter either. It reports a ratio only where the within-form R^2 clears 0.5. Swapping the per-equation estimator for the strict one changes which cells qualify, from 22–62 of 270 (5 depths × 18 landmarks × 3 seeds) down to 0–37; the hex operand rows drop out entirely, since their strict readings never reach 0.5. The verdict is the same either way, because the numerator is negative in every cell of the map, gated or not. That settles the open question of whether the alignment measures should be gated on the stricter estimator: here it makes no difference.

Alignment under compression (H4)

H4 predicted that narrowing the residual stream would push the two forms together: ρ and subspace overlap rising over d64 → d32 → d16, with d16-L8 the most aligned condition, but that didn't happen. ρ is 0.00 at every one of the 190 cells where the guard defines it across the whole sweep (5,184 cells before gating), in both directions and for all three targets.

What compression does instead is wear the named form down. The figure below tracks three quantities against width: how well each form answers held-out equations, how well the named mix can be read out of the residual stream, and how close the mix decoders of the two forms come to sharing a direction.



Effects of narrowing the residual stream, across the width arms. Filled ● markers joined by lines are the 4-layer conditions; open □ markers are the 8-layer conditions, dashed because there is no d32-L8 condition to run a line through. Right: the first principal angle between the named and hex mix decoders at the pre-answer position (amber), with two references: the angle between two random 3-planes at that width (grey band, mean ± 1 s.d.) and between the named probes of two seeds (red).

condition	width	depth	named held-out	named mix R^2	ρ → named	ρ → hex	ρ best	angle	seed control	random
center	64	4	0.667	0.94	0.00	0.00	0.00	75°	70°	72°
L8	64	8	0.671	0.96	0.00	0.00	0.00	74°	73°	72°
d32	32	4	0.431	0.80	–	–	0.00	60°	67°	64°
d16	16	4	0.047	0.35	–	–	0.00	51°	52°	51°
d16-L8	16	8	0.164	0.69	–	–	0.00	56°	50°	51°

The width arms, seed-averaged. The two ρ columns are at the preregistered site (the depth × landmark where the within-form R^2 of the weaker form is highest, chosen without reference to ρ), and read "–" where the within-form score of that form falls under the 0.5 guard on ρ. "ρ best" is the largest ρ anywhere in the scan for that condition, over both directions and all three targets. The last three columns are the first principal angle between the two mix decoders at the pre-answer position, and the two references the figure plots.

The angle does fall as the stream narrows, from 75° at d64 to 60° at d32 and 51° at d16. That *looks* like the trend H4 predicted, but it's not: two 3-dimensional subspaces drawn at random come just as close once the space they sit in is small enough. The same three widths give 72°, 64° and 51° for a random pair. Every observed angle sits inside one standard deviation of its null, and the seed control (the *named* probes of two seeds, decoding the same quantity in unrelated bases) tracks the same curve. So the cross-form angle at every width is what two unrelated decoders would give.

Under compression the two forms don't merge; one of them gives way first. Hex keeps its held-out accuracy at 1.00 at d32 and loses little even at d16 (0.72), while named falls from 0.67 to 0.43 to 0.05. In terms of the resolution account from H1, the fitted precision σ^2 coarsens from 0.033 of the unit cube at d64 to 0.068 at d32 and 0.253 at d16: roughly an eightfold loss of color resolution over two halvings of width. Hex needs far less resolution, because the runner-up to a hex answer is a whole grid step away, so a coarse read still lands on the right digit.

So the named form clearly lacks capacity at these widths, yet even that pressure causes no sharing. Keeping the two representations apart apparently costs the model less than merging them would.

▶ **Whether d16 is too broken to test…**

Alignment with a bridge (H5)

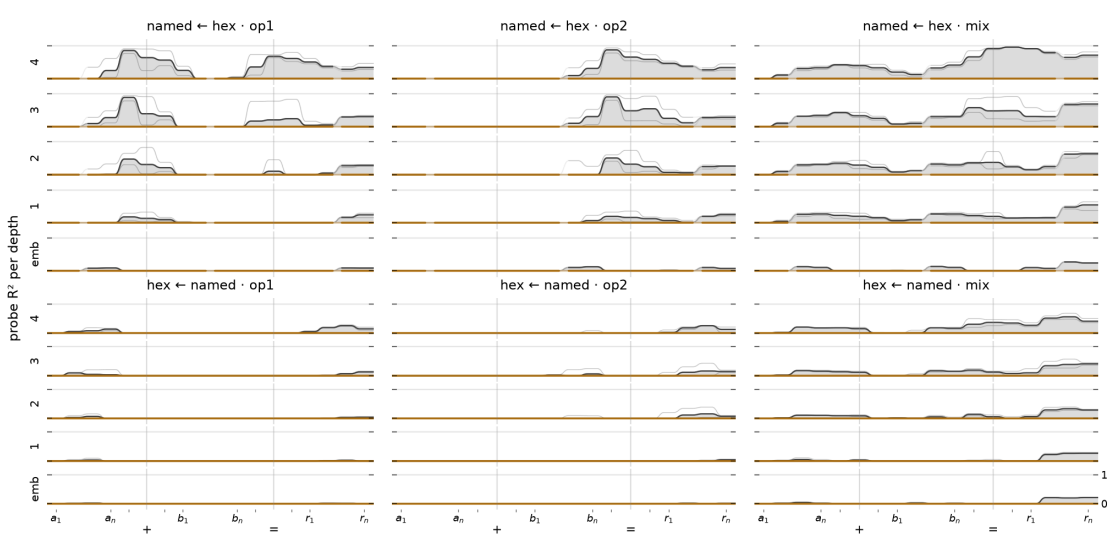
H5 predicted that a little shared supervision would place both forms in the same subspace: cross-form equations lifting ρ above 0.8 at d64. The ρ for the bridge condition is 0.00, the same as for every other condition, so H5 is unsupported on its own terms.

That would be an easy result to dismiss if the bridge had simply failed to learn anything, so let's check whether the supervision worked. Cross lines are 12% of the bridge corpus: 12,006 of them, drawn over 9,880 distinct (name, hex) pairs out of 30,240 possible, so a pair recurs about 1.2 times and two thirds of the cross space is never seen at all. The sweep doesn't grade them, because the form exists to supervise alignment rather than to be scored, so we decoded them afterwards from the published checkpoints.[3]

The model does learn to answer bridge-form equations. Exact match is 0.91 on cross pairs that appeared in training and 0.92 on pairs that did not, with no malformed completions; at 1.2 lines per pair there is not much for the model to memorize, and the two numbers say it didn't. Given the unseen prompt `pale lime + #fc7 =`, all three seeds return `#dd7`, which is the right answer. So the value of a color name does reach the hex answer path, for names and hex codes the model has never seen together.

condition	named held-out	hex held-out	cross unseen	named mix R^2	ρ → named	ρ → hex	ρ best	angle	seed control	random
center	0.667	0.996	–	0.94	0.00	0.00	0.00	75°	70°	72°
bridge	0.635	1.000	0.918	0.96	0.00	0.00	0.00	72°	73°	72°

The bridge condition beside the center condition, both d64-L4, seed-averaged. The columns after the accuracies are the same measures as the H4 table. Adding cross-form equations leaves every alignment measure where it was, and costs the two single-form tasks nothing worth reading.



Zero-shot cross-form transfer at the bridge condition, seed-averaged, drawn exactly as the H3 figure so the two can be laid side by side. Rows are the form the probes are applied to; columns are the probe target; grey is the own-form probe under the strict holdout and amber is the other-form probe applied unchanged.

So the bridge condition converts between the two vocabularies and keeps their geometries apart. Behaviorally the forms are joined: an unseen name and an unseen hex code mix to the right answer 92% of the time, and the answers it emits sit 0.037 from the exact mix on average, against a quantization floor of 0.032. Geometrically nothing moved: ρ stays at zero across the whole scan, and the first principal angle between the two mix decoders is 72° against a random-subspace null of 72° and a seed control of 73°.

Our measurement is fairly narrow: fit a probe on single-form lines, then apply it unchanged to lines of the other form at the same layer and grammar landmark. Suppose the model routes by form, with one pathway for `melon + ultramarine` and another for `melon + #48a`, sharing whatever they need somewhere we didn't look. Such a model would answer cross equations well and still score zero on our metric. What we can say is that a bridge does not make the two single-form geometries interchangeable at matched sites — the property H5 named, and a property a unidirectional anchor might want.

We don't know where the conversion happens. The obvious place to look is the cross lines themselves: fit probes on those activations and ask which geometry — named or hex — the value of a name lands in once the answer must come out as hex. That is a new measurement rather than a different reading of this one, so it goes to the backlog rather than into this report. The same applies to a depth-crossed ρ, which the exploratory section reaches the edge of.

▶ **What a stronger bridge might have done…**

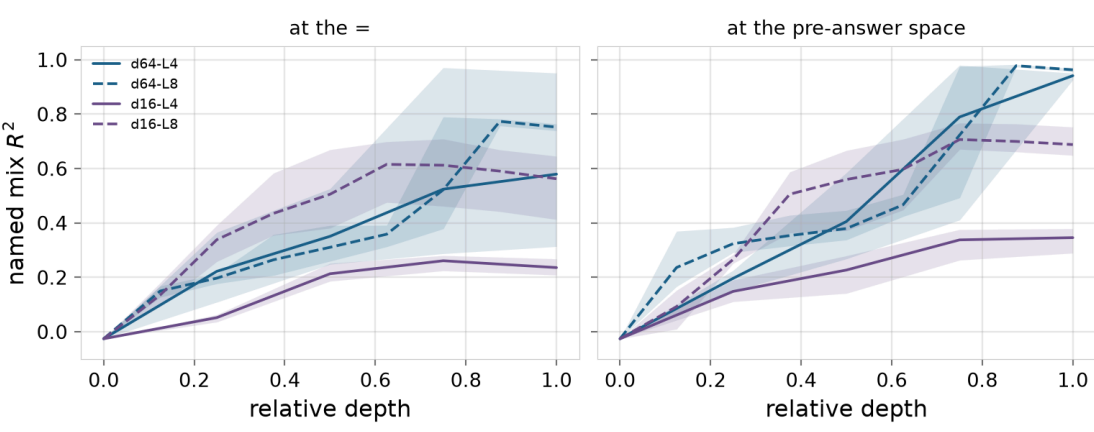
Depth (H6)

H6 had two clauses. Doubling the layers at fixed width should improve name+name accuracy, and the mix should crystallize before the last layer, giving the result concept more than one layer of existence. The preregistration also named a live counter-expectation for the second clause: nothing in the loss rewards computing early, so the mix might stay pressed against the answer instead. The first clause fails at d64 and holds at d16; the second splits, and which way you read it depends on whether you count layers or fractions of the network.

At d64 the extra layers buy nothing. Held-out named accuracy goes from 0.667 at L4 to 0.671 at L8, and accuracy on trained pairs from 0.841 to 0.855. Three seeds cannot resolve differences that small. Fitted precision agrees: σ comes out at 0.032 of the unit cube at L4 and 0.032 at L8, the same to three decimals, so the model reads colors to the same resolution either way. That fits the H1 account, where the named ceiling at d64 is the spacing of the palette rather than anything the model ran out of. Depth cannot buy resolution the task isn't short of.

At d16 it buys a great deal. Held-out named accuracy goes from 0.047 to 0.164, and σ from 0.257 to 0.140. The same eight layers that do nothing at width 64 recover about half the color resolution lost by narrowing to width 16. So depth substitutes for width here, up to the point where the resolution ceiling of the task itself takes over.

The figure below follows the named mix probe up through the stream at two landmarks: the `=` sign, and the pre-answer space one character later where H2 found the mature mix.



Where the named mix becomes readable, at the `=` sign (left) and the pre-answer space (right, the site H2 uses). The x axis is depth as a fraction of the layers in the condition, so the 4-layer and 8-layer curves overlay; 0 is the token embedding and 1 the last layer. Solid lines are the 4-layer conditions and dashed the 8-layer ones, blue for width 64 and purple for width 16; bands span the three seeds.

The mix does crystallize before the last layer at L8, which is what the clause asked for. It reads 0.72 at layer 6, 0.98 at layer 7 and 0.96 at layer 8, so the final layer adds nothing and the concept exists for two layers rather than one. Counting layers over 0.9, L4 averages 1.7 of 4 across seeds and L8 2.3 of 8.

Read as a fraction of the network, though, the mix sits *later* at L8: about 42% of the layers at L4 falls to 29% at L8. The extra layers went in front of the computation rather than after it, and the mature mix stayed where it was, up against the answer. That is the counter-expectation H6 flagged. It is also the reading we'd trust, since it is the one that would generalize to a deeper model.

Depth does move the mix earlier in the sequence. At the `=` sign the named mix reads 0.58 at L4 and 0.77 at L8, so one character before the site H2 measures, the deeper model has already assembled most of the answer. A head start on the token axis is a plausible thing for four extra layers to buy, though we haven't tested that reading.

Exploratory analyses

Analyses conceived after seeing the data land here, marked as post hoc.

Answer positions partly read the answer (post hoc)

Under teacher forcing the answer tokens are part of the input. So a probe at an answer position may be reading the digits off the surface text rather than anything the model computed.

The embedding (depth 0) tells us how much, because it is the token lookup before any attention or MLP has run: whatever is decodable there was already in the input. In the hex form it is a lot. At the embedding, a mix probe scores $R^2 \approx 1.00$ per channel at each answer digit (each hex digit *is* one channel of the answer), and an operand probe scores ≈ 0.48 at the same positions. The answer digit is the round-half-up mean of the two operand digits, so knowing it pins about half the variance of each operand. Both operands echo at the same strength, consistent with mixing rather than one operand being retained.

This suggests that an answer-position reading should be compared against its own depth-0 value, not against zero. For example, the best channel mean of the hex mix, 0.66 (depth 3, middle answer digit), breaks down as 0.22, 0.96, 0.80 per channel, against an embedding baseline of 0.00, 1.00, 0.00: the green channel is read from the input, and the blue is computed a token ahead.

The landmarks that carry the headline claims are clean: at the equals sign and the pre-answer space, every probe scores −0.003 at the embedding in both forms. That is why the named mix result is quoted at the pre-answer space, where the last layer reaches 0.95, 0.94, 0.94. Compared at those clean positions, the two forms separate further than the mid-answer numbers suggest: the best uncontaminated hex mix reading is 0.42 at the equals sign and 0.38 pre-answer, against 0.94 for named.

Similarly, ρ computed at answer positions would be inflated in both directions, since the probes of both forms can read the emitted answer there.

The forms compute at different depths, and ρ compares in place (post hoc)

ρ as preregistered compares the same cell in both forms: same depth, same landmark. That assumes the two forms put the mix in the same place, and H2 showed they do not. Named color mixes are assembled at the pre-answer position in the last layer. The best hex mix reading is mid-answer, a layer earlier, and its pre-answer column never clears 0.20 at any depth (-0.13, -0.02, 0.07, 0.12, 0.20 from the embedding up). So the same-cell comparison may score the mature named mix against a hex representation that hasn't formed at that position, and a depth-crossed ρ could in principle be higher.

We can check half of this from the stored data. The transfer half cannot be answered: applying the hex depth-2 probe to the named depth-4 activations needs the activations, and only the fitted probes and their scores are stored. The subspace half can, from the probe weights alone. The first principal angle between the named mix probe at *any* depth and the hex mix probe at *any* depth, both at the pre-answer position, stays between 66° and 78° across all 16 pairs. Widening the search to every depth × landmark pair where the strict within-form readings of both forms clear 0.3 (a low bar, and one that admits the contaminated answer positions), the closest the two subspaces come is 57°.

So the depth offset is not concealing a shared set of directions: there is no layer at which the hex mix decoder points where the named one does. A depth-crossed transfer R^2 would still be a different test (zero-shot fit rather than subspace angle) and is cheap to add to the eval step; it is banked in todo-science rather than done here, because it is not the measurement H3 was written against.

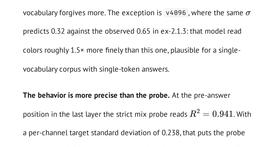
One artifact in the angle map: At the embedding, on the space closing operand 1, the two probes appear to share a direction almost exactly (0.4°). The space character is the same token in both sublanguages, so at depth 0 the probe input is constant and its fitted direction is undetermined. The same degeneracy will show up in H4 and H5, where a small angle would otherwise read as the alignment those hypotheses are looking for.

The named ceiling is a resolution limit (post hoc)

Named exact match tops out around 0.84 on trained pairs and 0.67 on held-out ones, while hex sits at ≈ 1.0, as did the regular 216-color ex-2.1.3 vocabulary. That does not mean the named geometry is weaker; the two vocabularies ask questions of different precision.

What varies is the **snap margin**: how much closer the correct name is to the exact mix than the runner-up. On a regular sub-grid closed under mixing, the answer lands on a vocabulary point and the margin is a constant (one grid step, 0.2 of the unit cube at `v216` in ex-2.1.3). On an irregular palette it varies by two orders of magnitude, and its median here is about a sixth of that. So a raw accuracy pools prompts that need 0.1 of positional accuracy with prompts that need 0.005.

To separate resolution from everything else, we fit a one-parameter reference guesser: read the exact mix with isotropic Gaussian error of per-channel scale σ, then answer with the nearest name (`baselines.precision_limited_acc`). σ is fitted to the *overall* accuracy only, so the shape of accuracy versus margin is free to agree or not.



Exact match against the snap margin (the distance from the exact mix to the runner-up name, minus the distance to the correct one, in unit-cube units). Points are the observed rate in each margin bin, pooled over three seeds, with binomial standard errors; the line is the resolution-limited reference at the σ fitted to that overall accuracy of that panel, so its level is fitted but its slope is not. The shaded profile along the bottom is the margin distribution.

The reference fits. On held-out named prompts the fitted σ is 0.033 of the unit cube (8 of 255 per channel), and from that one number the reference tracks the observed rate across the whole margin range, including the 250-name condition, where the same shape appears shifted toward the tight end. Trained pairs fit less well: at 140 names the fitted σ halves to 0.016, but the model beats the reference at tight margins and falls short of it at wide ones. Precision alone does not describe errors on pairs the model has seen; something more like recall is mixed in.

The earlier rungs are consistent with the same precision. On the regular grids of ex-2.1.3, where a mix lands on a vocabulary point and the margin is the grid step, $\sigma = 0.033$ predicts 0.99 at `v216`, which is about what ex-2.1.3 (≈ 1.0, single-token names) and ex-2.1.4 (0.91, spelled names) actually scored. So the drop from ≈ 1.0 there to 0.67 here needs no change in how well colors are represented; the coarser vocabulary forgives more. The exception is `v4896`, where the same σ predicts 0.32 against the observed 0.65 in ex-2.1.3: that model read colors roughly 1.5× more finely than this one, plausible for a single-vocabulary corpus with single-token answers.

The behavior is more precise than the probe. At the pre-answer position in the last layer the strict mix probe reads $R^2 = 0.941$. With a per-channel target standard deviation of 0.238, that puts the probe residual near 0.058, roughly twice the 0.033 the behavior implies. The two are different estimands (the error of a linear decoder versus an implied end-to-end resolution), so only the ordering is worth reading: the linear probe recovers a coarser version of the value than the model itself uses. For anchor design this means a probe R^2 of 0.94 at a site is a floor on what is there, and anchor supervision is itself a linear readout with the same limitation.

Findings

Both sublanguages train, each builds its own linear color geometry, and the two geometries stay separate under every condition tested.

- **H1.** Supported: top-1 exact match on held-out equations was 0.996 for hex and 0.667 for named, clearing the 0.5 nearest-neighbor null.
- **H2.** Supported: the named mix decodes at $R^2 = 0.94$ pre-answer; hex assembles just-in-time with no pre-answer mix.
- **H3.** Supported: ρ is zero everywhere it's defined, and the cross-form principal angles match a random pair of subspaces.
- **H4.** Not supported: narrowing the stream never aligns the forms — it wears the named form down instead.
- **H5.** Not supported: the bridge supervision works, yet ρ stays zero.
- **H6.** Mixed: depth helps at d16 and does nothing at d64; the crystallization clause depends on whether you count layers or network fraction.

The separation is resilient: even where narrowing the stream costs the named form roughly an eightfold loss of color resolution, the model keeps the two geometries apart, so maintaining separate representations apparently costs it less than merging them would. Perhaps we would see something different at greater depths.

The named ceiling is a property of the palette rather than the model: misses land on the immediate neighbors of the mix, and a single fitted read-noise parameter reproduces the accuracy across conditions.

And the forms seem to compute in different places: named assembles its mix before the answer while hex emits digits just-in-time. So the separation is positional as well as directional.

---

1. A separate refit on the published checkpoints, since it needs folds the stored maps don't carry; the numbers are quoted here rather than recomputed in this notebook. ↩

2. σ is the per-channel noise a guesser would need, reading the true mix and snapping to the nearest name, to match the observed accuracy; see H1. ↩↩

3. `cross_eval.py`, beside this report. It reads the sweep checkpoints and writes its results back to the store, so the numbers here are computed rather than transcribed. It sits outside the experiment DAG because a tick of `experiment.py` would now re-train all 24 conditions: the evidence scheme in `mini` has changed since the sweep ran, and a re-trained sweep would not reproduce the numbers quoted elsewhere in this report. ↩