SMALL TRANSFORMERS TRACK BAYESIAN EVIDENCE FOR LATENT COMMON CAUSES VIA A CONTEXT-INVARIANT MECHANISM

Anonymous authors
Paper under double-blind review

ABSTRACT

We present an in-depth investigation of how a form of Bayesian reasoning about common causes can emerge as a cross-contextual generalization in small, tractable transformers. Incrementing on recent work, our set-up (i) disentangles causal mechanisms in the model from the causal structure of the true data-generating process, (ii) orients more towards natural language prediction by considering inference of latent common causes, and (iii) considers whether and how Bayesian evidence accumulation for latent common causes can be implemented in representations and mechanisms that allow for cross-context generalization to novel test cases.

1 INTRODUCTION

Given the performance of large transformers, it is likely that these models acquire implicit world models which emerges as useful compressions in service of accurate prediction (e.g., Yudkowsky, 2023; Bereska & Gavves, 2024; Shai et al., 2026). Evidence comes from case studies on finite-state problems, such as board games (Toshniwal et al., 2022; Li et al., 2023), taxi driving routes (Vafa et al., 2024), or text-based games (Li et al., 2021). While it remains unclear how pressure for sequence prediction leads to the emergence of world models (Li et al., 2025; Yuan & Søgaard, 2025), it is plausible that world models support the ability to perform probabilistic causal reasoning.

This paper seeks to contribute to the investigation of emergence of world models by an in-depth, mechanistic analysis of generalizable Bayesian (causal) reasoning emerging in a novel, minimalistic experimental setup. Conceptually, we contribute an analytical delineation of different senses of causal information relevant for the analysis of language models and adopt an architecturally aware *ideal-learner analysis*, which supports investigating how Bayesian abductive reasoning (from observed effects to underlying causes) can emerge in a held-out context by exploiting similarity in the representations of the abstract conceptual roles that different tokens play (e.g., Shepard, 1987; Wang et al., 2026). We argue that this kind of cross-context Bayesian generalization amounts to a simple form of *causal representation learning*, i.e., recovery of relevant causal variables, but we also stress the general theoretical limits of causal recovery (Spirtes et al., 2000; Richardson & Spirtes, 2002). Empirically, we demonstrate how small transformers indeed show a form of cross-context "Bayesian generalization" in their behavior by showing that generalizable Bayesian-like behavior is indeed supported by internal "Bayes-like" representations, the systematic manipulation of which with suitable *joint-dependency interventions* leads to downstream effects that are coherent with Bayesian computations, thus satisfying the foundational abstraction requirement of commutativity under intervention (Rubenstein et al., 2017; Beckers & Halpern, 2019; Geiger et al., 2021; Xia & Bareinboim, 2024). Our methods[1] and main results are succinctly summarized in Figure 1.

2 RELATED WORK & NOVEL CONTRIBUTIONS

Causation. For large transformers one important area of research is *represented causality*, i.e., whether models process token sequences which contain information about causality (at least to human observers) in a way that is correct (Jin et al., 2023; Kiciman et al., 2024; Chi et al., 2024) or human-like (Binz & Schulz, 2023). Smaller transformers have been analyzed in-depth with hypothesis-driven

[1] Code for all the experiments in https://anonymous.4open.science/r/bayformer-paper-19FF/

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.

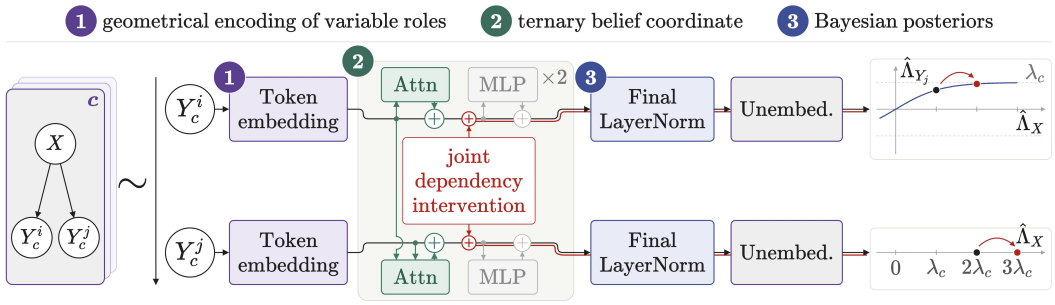


Figure 1: We train small transformers on scrambled token sequences from different contexts, each of which are causal models with a common cause and two effect variables. Held-out contexts require the models to extrapolate Bayesian inference of the causal variable's value from contexts seen in training. We find that "Bayesian generalization" emerges, and we detail its mechanics, finding that (1) token embeddings geometrically encode the relevant roles of the underlying causal variables, that (2) the residual stream encodes a categorical belief representation, which is (3) transformed via the LayerNorm into an approximate Bayesian posterior. Surgical interventions yield systematic downstream effects on the joint distribution over relevant variables, as predicted by Bayes rule.

experiments about causality and its representation and relevance. For example, Rohekar et al. (2024) show that a transformer's self-attention allocation can be used for estimating a structural causal model of the transformer's internal computational mechanism, and Nichani et al. (2024) argue based on emerging attentional mechanisms that transformers can learn the true causal processes that generated the data. Emerging attention patterns track these position-structural dependencies and these, by design, happen to coincide with the true data-generating process. In both cases, the assumed data-generating process is a simple Markov process in which the token at position i depends only on the token at a prior position $i - k$, so attention patterns tracking these positional dependencies coincide with the true process by design. As a results, these experimental designs, while doubtlessly insightful to a fair extent, entangle an important distinction between what we here call *within-model causality*, which involves causal mechanisms operating in the model (abstract properties of the forward pass computation), and *behind-data causality*, which concerns information about causal processes that occur outside of the model (in the data-generating process) but which is nevertheless represented and functionally relevant in the model. Clearly separating *within-model causality* from *behind-data causality* is important, because full natural language generation is a arguably a latent variable problem, in which the next token prediction may depend on abstract, latent variables for which prior tokens at variable positions provide only indirect probabilistic evidence (e.g., Jiang, 2023; Zhang et al., 2023). This is why our experimental setup uses a data-generating process in which there are (partially latent, partially observed) common causes for overt tokens, and where surface-form realization is shuffled to be more position-invariant.

Bayes. While the connection between probabilistic machine learning and (normatively correct) Bayesian inference is an obvious foundational issue (e.g., MacKay, 2003), a recent prominent topic is whether in-context learning in large language models constitutes a form of Bayesian inference (e.g, Xie et al., 2022; Raventós et al., 2023; Falck et al., 2024; Cao et al., 2025). Other work investigated whether Bayesian / causal reasoning abilities emerge that are normatively correct or human-like (e.g., Shwartz & Choi, 2020; Müller et al., 2021; Kauf et al., 2022; Zhu & Griffiths, 2024; Li & Rui, 2024; Gupta et al., 2025). There is also work that explores specific training regimes that induce or enhance Bayesian reasoning abilities on transformers (e.g., Hu et al., 2024; Qiu et al., 2026).

This paper primarily asks how transformers could implement generalizable Bayesian reasoning, contributing to several recent strands of related research. On controlled synthetic tasks, Agarwal et al. (2026) find attention consistent with incremental Bayesian posterior computation along the residual stream. Relatedly, a line of work inspired by *computational mechanics*(Shalizi & Crutchfield, 2001), trace the geometry of latent (Bayesian) beliefs in the residual stream (Shai et al., 2024; Levinson, 2026), and further argue that transformers implement a constrained sequential belief update Piotrowski et al., 2025. So far, however, these studies have focused on cases where transformers'

predictions and internal computations are analyzed on cases that were included in the training sets, and causal-interventionist checks on recovered representations have not always been given.

Incremental contribution. This work goes conservatively beyond mentioned prior works, by the conjunction of: (i) a set-up which disentangles surface position of a token from its potential information about future tokens and its true causal role in the data-generating process; (ii) considering a data-generating process with common causes, which are sometimes overt, sometimes latent; (iii) considering the emergence and mechanistic realization of "Bayesian generalization" for inferring a latent common cause in novel test cases.

3 Experimental design, research questions & approach

The process for generating training and test data consists of repeatedly sampling triplets of token symbols and appending them into a long sequence, separated by a delimiter symbol, e.g., $aCb|feD|Zxy\ldots$. Each triplet of symbols is obtained by sampling a random *context*, and then sampling that sequence exclusively from tokens belonging that context. By designating particular contexts as *held-out contexts* and withholding particular *held-out sequences* from these held-out contexts during training, we are able to investigate cross-context generalization.

Fix $\mathcal{C} = \{C_c\}_{0 \leq c \leq n}$ as a family of *contexts*, where each $C_c = \langle \mathcal{V}_c, G_c, P_c, A_c, M_c \rangle$ consists of a causal Bayes net with three binary variables $\mathcal{V}_c = \left\{X_c, Y_c^1, Y_c^2\right\}$ taking values in $\{+, -\}$, a causal dependency graph G_c in which X_c is a common cause of Y_c^1 and Y_c^2 ($Y_c^1 \leftarrow X_c \rightarrow Y_c^2$), so that the joint distribution P_c is given by the causally disentangled factorization (in simplified notation):

$$P_c\left(x, y^1, y^2\right) = P_c\left(x\right) \times P_c\left(y^1 \mid x\right) \times P_c\left(y^2 \mid x\right) ,$$

with $P_c(x) = \frac{1}{2}$ and $P_c(y^i \mid x) = p_c$ if $y^i = x$ and $1 - p_c$ otherwise. Across contexts, Bayes nets share the variables, their values and the causal graph G_c, but they differ in p_c. Moreover, A_c is an alphabet (a set of six token symbols), so that $A_c \cap A_{c'} = \emptyset$, for all $c \neq c'$. The mapping function $M_c \colon \mathcal{V}_c \times \{+, -\} \rightarrow A_c$ is a bijection, thus associating a unique symbol with each variable and value in C_c. We call X_c the *parent*, and we call Y_c^1 and Y_c^2 the *children*. We say that Y_c^i is a *sibling* to Y_c^j (implicitly assuming $i \neq j$ here and below). The sampling protocol for one triplet sampled from context C_c is: (i) sample a set of values from $P_c\left(X_c, Y_c^1, Y_c^2\right)$; (ii) map these values via M_c to the corresponding token symbols in A_c; (iii) randomly shuffle the three tokens symbols.

For readability, we represent (sets of) sequences of token symbols not in terms of elements of A_c (as the language models will see them), but in terms of the corresponding variable-value pairs, writing $Y_c^{i,-}\, X_c^{+}\, Y_c^{j,+}$ or $Y_c^{i,-}\, Y_c^{j,+}$, where superscripts indicate the instantiated values, e.g., $X_c^+ = M_c(X_c, +)$. We use the symbol $* \in \{+, -\}$ as a variable over values, and use $\bar{*} \in \{+, -\} \setminus \{*\}$ as its negation. When we leave out subscripts or superscripts, we use this as notation for sets of sequences. E.g., $Y_c^i\ Y_c^j$ denotes the set of all sequences of pairs of tokens, each of which is associated with a different child variable in context c, and notation like $\{X_c, Y_c\}$ would be shorthand for context c's alphabet. We say that a sequence is *value-concordant* if all tokens in it are associated with the same value ($+$ or $-$); otherwise we speak of a *value-discordant* sequence. For instance, the set of all value-discordant sequences of child tokens from context c would be written compactly as $Y_c^{i,*}\, Y_c^{j,\bar{*}}$. We consider the context C_0 the *held-out context*; all other contexts are referred to as *supervised contexts*. For all the *held-out sequences* $Y_{c_0}^i\, Y_{c_0}^j$, the training data does not contain the succeeding parent token X_{c_0}, and we refer to all other sequences as *supervised sequences*.

Our **main research question** is: How could LMs learn a context-general form of (abductive) Bayesian reasoning from effect to cause by extrapolating the functional form of such reasoning observed from observed to unobserved cases? — We formulate a series of more specific research questions based on the assumption that sufficiently trained transformers will have found an approximately optimal solution to the training problem. This perspective is kin to several approaches from different fields, e.g., *rational analysis* in cognitive science (e.g., Anderson, 1990), the information-bottleneck approach in information theory (Tishby et al., 2000), or computational mechanics in physics (e.g., Shalizi & Crutchfield, 2001). But, going beyond a simple ideal-learner analysis, we additionally stress the constraints imposed the by the architectural realization of a computation, which is important in the context of generalization, as argued in the following.

An approach based on computational mechanics has recently been applied to the analysis of transformers' internal representations for reasoning under uncertainty as well (Shai et al., 2024; Levinson, 2026). To do so, computational mechanics identifies the *canonical states* of the data-generating process, which are defined as the sufficient statistics for an optimal predictor, which, in turn, are given by the equivalence class of all sets of sequences that are indistinguishable based on the true probabilities of all future tokens (cf., Vafa et al., 2024, for similar methods). In our setup, all stochastic dependencies are broken by the delimiters and contexts have pairwise disjoint symbols, so that the canonical states are a union of the canonical states for each context c. As, by construction, the only future-relevant aspect within a triplet for context c is the value of the parent variable X_c, the canonical states are defined by tracking beliefs about X_c. Consequently, we can characterize the canonical states via $\mathcal{S} = (\Omega, \Lambda)$ and $\Omega \subseteq \{X_c, Y_c\}$, where Ω is the set of variables already observed in the current triplet and Λ is the true log-odds of the value of X_c. We then only need to distinguish, based on the number of observed tokens so far, the following relevant canonical states:

$$|\Omega| = 0: \ (\emptyset, \ 0) \qquad |\Omega| = 1: \ (\{X_c\}, \ \pm\infty), \quad (\{Y_c\}, \ \pm\lambda_c)$$
$$|\Omega| = 2: \ (\{X_c, Y_c\}, \ \pm\infty), \quad (\{Y_c^1, Y_c^2\}, \ \{\pm 2\lambda_c, \ 0\})$$

where $\lambda_c = \log \frac{p_c}{1-p_c}$ is the belief unit for context c. By an ideal-learner analysis, we expect trained transformers to recover these canonical states, i.e., to track incrementally the log-odds of X_c as the only prediction-critical information. Evidence for this Bayesian-like ideal-learner behavior would come from high accuracy on training sequences, patterns of input-variance and -invariance that is consistent with the canonical states, part of which is the recovery of the patterns of stochastic independence between variables. Moreover, we expect the ideal-learner analysis to be supported in the models' representations and their internal computational mechanisms.

What the perspective of computational mechanics and a pure ideal-learner analysis does not give us, is a prediction about generalization from training to held-out sequences. The held-out sequences in context C_0 coincide with three canonical states ($\{Y_c^1, Y_c^2\}, \ \{\pm 2\lambda_c, \ 0\}$), i.e., with the four value-concordant two-child sequences $Y_{c_0}^{i,*} \ Y_{c_0}^{j,*}$ and the two value-discordant sequences $Y_{c_0}^{i,*} \ Y_{c_0}^{j,\bar{*}}$. As all of these states are never seen in training, and ideal-learner perspective in terms of canonical state recovery does not predict whether or how generalization to these unseen cases may happen. Instead, we hypothesize, based on architectural features, that transformers' hidden representations may capture similarities between canonical states in such a way that Bayes-like generalization is possible (Shepard, 1987). Concretely, we conjecture that such generalization can occur if models capture (i) the mechanism of accumulating evidence (about the value regime) in a way that generalizes across (supervised) contexts, together with (ii) cross-context similarity of the evidential role of tokens/variables. We therefore investigate whether such a similarity-based representation of token roles with a cross-contextually shared Bayesian-like mechanism of tracking of prediction-relevant evidence is attested in our trained models.

4 Models, Training & Behavioral Generalization

We train small GPT-style decoder-only transformers over sequences comprised of 3 contexts with $p_c = \{0.7, 0.8, 0.9\}$. 45 models are trained in total (3 held-out contexts × 15 seeds; see Appendix A for full information). Each model has two pre-norm blocks with one attention head. Residual stream and feedforward layers (with GELU) both have a width of 16, and each model a total of $4,051$ parameters. Token embeddings are learned and untied from the unembedding. Sinusoidal positional embeddings are used. Training uses next-token cross-entropy on one-hot targets, with AdamW and a fresh batch at every step. In the held-out context, the parent is removed from the loss and from the attention keys whenever it follows both children. We keep the checkpoint of each model with the lowest validation (never seen during training) KL-divergence to the exact Bayesian conditionals. All trained models are used in all of the following analyses.

Figure 2a shows that trained models achieve high predictive accuracy on the training data. More importantly, although the held-out KL is an order of magnitude higher than the global KL-Divergence, it remains far below that of a random baseline. The trained models are particularly successful at predicting the correct variable at a held-out position. An optimal learner analysis predicts that for high predictive accuracy the patterns of stochastic dependencies entailed in the true causal graph must be recovered. Figure 2b shows that this is indeed so. In both held-out and supervised contexts, models respect the exchangeability of variables and parents screen off children. This suggests that

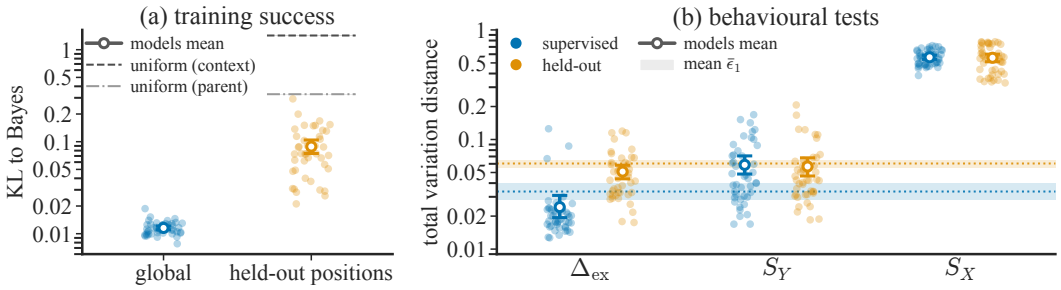


Figure 2: **Training success and stochastic dependencies in models' behavior. (a)** Training success measures; *(global)* averaged over all positions, and *(held-out)* averaged over the masked positions in *held-out sequences*. **(b)** Total variation distances at second token of supervised (blue) and held-out sequences (orange). The three columns represent variable exchange error Δ_{ex} (Eq. 5), child's predictive sensitivity to sibling S_Y (Eq. 6) and child's predictive sensitivity to parent S_X (Eq. 7). Open markers: mean with 95% two-level bootstrapped interval. Shaded bands: $\bar{\epsilon}_1$ (Eq. 8), model's predictive error at the same position. Δ_{ex} and S_Y, are in the range of models' error margin. S_X reaches 0.56, ground-truth is $|2p_c - 1|$ (mean 0.60): an order of magnitude above the floor.

successfully trained models recover general information about the *role* of different tokens in line with the underlying (causal) variables.

5 Representation of the Data-Generating Causal Model

The behavioral results from above motivate subsequent questions about emergent representations. **RQ1:** Do transformers trained on next-token prediction internalize concepts of variables and their values as specified by the data-generating process? **RQ2:** Do transformers carry a latent belief about the common cause? — To address these questions, Section 5.1 shows that token embeddings they encode task-relevant information in a geometry that reflects the functional role of the variables in the true data-generating (causal) model. Subsequently, Section 5.2 shows that the functional attention combines these embeddings such that a belief coordinate carrying the direction of the evidence about the common cause emerges. Finally, Section 5.3 shows that the model counts the number of observed variables utilizing the final LayerNorm. These results are visualized in Figure 1.

5.1 Variable Roles Emerge in Token Embeddings

We ask how each context's token embeddings represent that context's variables and their values, using an orthogonal decomposition of the embedding vectors (Appendix C). Let $\tilde{e}_c(V, *)$ denote the centered embedding of token $M_c(V, *)$ for all variables $V \in \mathcal{V}_c$ and values $* \in \{+, -\}$. Then define:

$$\tilde{e}_c(V, *) = \alpha_c(V) + \beta_c(*) + \gamma_c(V, *)\,, \tag{1}$$

where $\alpha_c(V)$ is the *variable term*, $\beta_c(*)$ is the *value term*, and γ_c is the *interaction term*. Since the children are exchangeable, the variable and interaction planes each split further into a parent-vs-children axis and a child-vs-child axis. To tie these design axes to the embedding geometry, we project the left singular vectors of the centered table (stack of token embeddings minus their context means) onto them and measure the fraction of each vector along each axis.

Each singular vector lies mostly along a single design axis, so the design axes are the principal axes of the embedding (Figure 3d). Among these axes, the variable term dominates, and the value and interaction terms are relatively small. Almost all of the interaction plane is explained by the parent-vs-children axis. The values of each context's variables are therefore represented in a two-dimensional plane, spanned by a common value direction and a parent-vs-children direction (Figure 3b). The two value tokens of a variable lie further apart in contexts with a stronger dependency, i.e., larger p_c (Figure 3c). Taken together, these results answer **RQ1** at the level of token embeddings in two parts: (1) **Token embeddings recover the variables and values of the generative process.** From next token prediction, the model represents each token as a variable (Figure 3a) carrying a value (Figure 3b), akin to the generative process. (2) **The embedding geometry reflects the structure of**

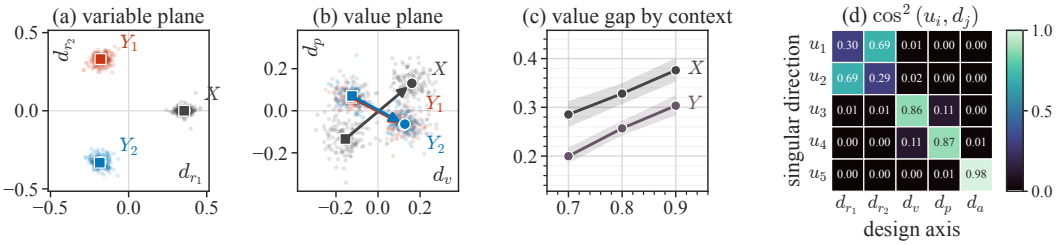


Figure 3: **Context-centered embeddings exhibit a common geometry. (a)** Variable plane encodes the variation among variable identities; d_{r_1} separates parent from children, d_{r_2} children from each other. **(b)** Value plane separates tokens by the regime value; d_v and d_r mostly captures a common value shift among all variables and parent's deviation from children, respectively. **(c)** Child values are exchangeable; their distance in the value plane tracks contexts dependency power (i.e., furthest for $p_c = 0.9$, and shortest for $p_c = 0.7$). **(d)** The design axes align with the natural directions that explain the variance of the raw token embeddings.

the dependency graph. The parent is distinguished from the two children by its value, while the children are exchangeable (Figure 3b).

5.2 A BELIEF COORDINATE EMERGES AFTER ATTENTION

Of the two attention layers in each model, only one is functional, and its pattern is fixed by position (Appendix D). Such a content-invariant attention cannot itself distinguish variables or values. Since the token embeddings already carry this distinction (Section 5.1), what remains is how the functional attention combines them. Attention preserves the design decomposition of the embeddings up to a nearly uniform per-token LayerNorm scale (see Appendix C). Since each model has a single functional attention block, there is a single residual node at which the observed tokens are combined as a weighted sum of their embeddings. This simple attention can still function as part of a circuit that implements a posterior computation. The missing piece is whether the token embeddings supply a form of *evidence* that the attention can combine for prediction.

We study the residual stream of the trained models at the node right after a functional attention block, where the input is a two-token sequence $M_c(V, *)\, M_c(V', *')$ of distinct variables, $V \neq V' \in \mathcal{V}_c$. We call the first token *antecedent* and the second *query* token. We regress the average residual stream at the selected node, $\bar{h}_c$, on a global offset and two token terms over the 24 admissible token pairs:

$$\bar{h}_c(V, *;\, V', *') \;=\; o_c + a_c(V', *') + b_c(V, *) + \varepsilon_c(V, *;\, V', *')\,. \tag{2}$$

For all 45 trained transformers, the cross-term ε_c is negligible, and the two token terms produce almost equal norms, $\|a_c\| \approx \|b_c\|$. This suggests that each observation contributes separate evidence with the same weight as the other, regardless of the emission order. To fix the scale of these terms, we take the residual stream of single-token sequences at the same residual node, $g_c(V, *)$, as reference. We then measure the weight of each token term as $\|a_c\|/\|g_c\|$ and $\|b_c\|/\|g_c\|$. In every context, both weights lie between 0.54 and 0.57 on average. This establishes the *functional* form of the attention block as an approximate average of the two tokens (See Appendix E for more information).

Whether the geometries of the two token terms are aligned is a separate question. We answer it with a principal-angle analysis of their design subspaces (See Appendix E). The principal angles are small for the value axis, both variable plane axes, and the leading interaction direction; only the second interaction direction, which carries little mass, is not aligned. The attention block therefore averages the two tokens within each design subspace as well.

This alignment has interesting implications. First, the three centered variable vectors sum to zero, so the average of any two equals minus one half of the third. The residual of a two-token sequence therefore always points away from the variable still to be emitted (Figure 4a). This alignment is almost exact in both supervised and held-out contexts, which explains why the models identify the variable of the held-out posterior correctly. Second, the children move the residual stream along the same direction in value plane (Figure 4b). Value-concordant sequences shift it toward their shared value and value-discordant sequences cancel out. This direction operates as a candidate *belief*

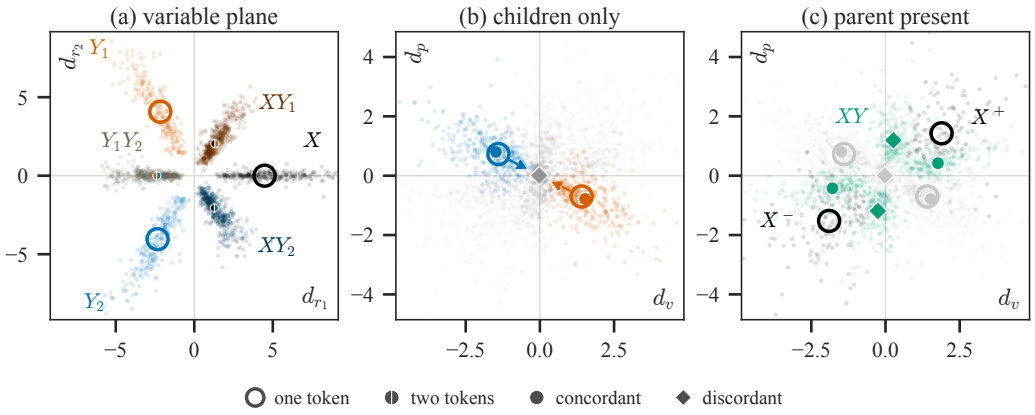


Figure 4: **Geometry of the residual stream.** Residual streams after functional attention, pooled over 45 models and all contexts, projected onto the design planes. Dots are pooled cells; markers are pool medians. **(a)** Each variable and each pair occupies its own ray, a pair lying between its members. **(b)** Value plane; in the absence of parent (gray dots), belief about it lies on the anti-diagonal. A concordant second child leaves the median in place and a discordant one returns it to the origin. Belief encodes the sign of the evidence and is insensitive to its count. **(c)** Value plane (complementary); sequences containing the parent lie off the belief axis.

coordinate, which points at the value regime, i.e., whether current sequence favors one value of the parent over the other. In other words, although the ground-truth belief takes five distinct values, the belief after attention is only ternary: it takes a positive or negative value for the sign of the evidence, and zero when the children disagree.

5.3 Model counts the evidence

A belief coordinate that only carries the direction of the observed evidence cannot by itself account for two-child sequences that produce a posterior of $\pm 2\lambda_c$. Additionally, the held-out sequences are indistinguishable from the supervised ones in the value plane. Therefore, something after the attention has to amplify the belief in case of concordant two-child sequences such that the final read-out matches the ground-truth. We hypothesize that the same mechanism accounts for the prediction deficit of the held-out posterior.

To locate this amplification, we project the output of each component downstream of the attention onto the parent's log-odds direction $w_c = \mathbf{u}(X_c^+) - \mathbf{u}(X_c^-)$, the difference between the unembedding vectors of the two parent tokens of context c. Since the residual stream is the sum of the components' outputs, its projection onto this direction is the sum of each component's output projection. Before the output, the final LayerNorm standardizes the residual h by subtracting its mean $\mu(h)$ and dividing by its scale $\sigma(h)$, and applies a gain γ and a bias b. We can write the parent's log-odds as

$$v_c(h) = \frac{\langle \gamma \odot (h - \mu(h)\mathbf{1}),\, w_c \rangle}{\sigma(h)} + \langle b, w_c \rangle.$$

We then compare the residual of a single-child sequence $x^{(1)}$ with that of a value-concordant two-child sequence $x^{(2)}$. The change in the parent's log-odds between them is

$$\frac{v_c(h^{(2)}) - \langle b, w_c \rangle}{v_c(h^{(1)}) - \langle b, w_c \rangle} = \frac{\langle \gamma \odot (h^{(2)} - \mu(h^{(2)})\mathbf{1}),\, w_c \rangle}{\langle \gamma \odot (h^{(1)} - \mu(h^{(1)})\mathbf{1}),\, w_c \rangle} \cdot \frac{\sigma(h^{(1)})}{\sigma(h^{(2)})}.$$

where the first factor is carried by the model's other components and the second by the final LayerNorm.

Normalization accounts for most of the amplification (Figure 5), which it achieves because the residual shrinks from a one-child sequence to two. Splitting this shrinkage into the part inside the variable plane and the part outside shows the plane carries most of the shrinkage. The amplification

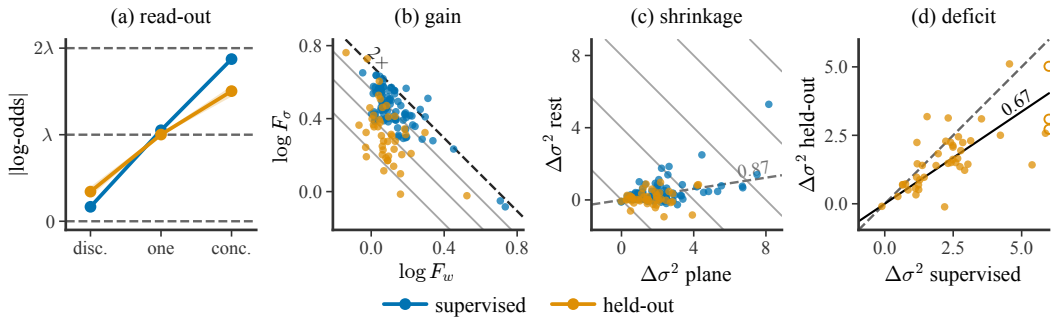


Figure 5: **Accumulation gain. (a)** Bootstrap mean with 95% CI of absolute parent log-odds in units of $|\lambda_c|$. Held-out two-child cases are not exact. **(b)** Amplification of the parent's log-odds from one child to two concordant children, split into the factor from the components, F_w, and from the final LayerNorm, F_σ. Grey diagonals mark total gains of 1.25, 1.5 and 1.75×; the dashed diagonal marks 2×. **(c)** Shrinkage of the final LayerNorm input from one child to two concordant children, $\Delta\sigma^2$, split into the part inside the variable plane and the rest. Grey diagonals mark constant total shrinkage; the dashed line marks the pooled in-plane share, $\sum \Delta\sigma^2_{\text{plane}} / \sum \Delta\sigma^2$. **(d)** $\Delta\sigma^2$ in the held-out context against the mean over the two supervised contexts, one point per model. Dashed: identity; solid: pooled ratio $\sum \Delta\sigma^2_{\text{held-out}} / \sum \Delta\sigma^2_{\text{supervised}}$; open markers are clipped at the axis limit.

is therefore set by the count of observed variables. We call this mechanism the *accumulation gain*. In the held-out context, the one-child log-odds match λ_c but the two-child log-odds fall short of $2\lambda_c$ (Figure 5a). This is explained by the residual shrinkage of the held-out states, which is only 67% of that in the supervised contexts (See Appendix F.2). Smaller shrinkage means a larger activation magnitude, hence a smaller normalization factor and a mismatch in prediction. Next, we ask what governs this mismatch in shrinkage. Expanding the normalization term $\sigma^2(h)$ over the model's components shows how much each pair of components align with each other, and hence how much they lengthen or shorten the residual (See Appendix F.3). The expansion shows that the held-out deficit results from multiple small misalignments between components. This explains why the approximate geometric analyses could not resolve them.

In sum, we resolve **RQ2** with two results. (1) **The residual stream encodes a belief about the common cause.** The shared value direction is represented as a belief coordinate after the attention which encodes *only* the direction of the evidence about the parent, hence a latent belief. (2) **The model counts the evidence toward the common cause.** The unit evidence direction in the belief coordinate is multiplied by the number of observed child tokens, via the accumulation gain.

6 Joint Dependency Intervention

Thus far, we have established that the trained transformers internalize the structure of the data-generating process and encode a latent belief. We argue that these are only prerequisites for ascribing the adjective Bayesian to the transformers' computation. Building on the causal abstraction framework Geiger et al. (2021), Bayesian inference can only be a faithful causal abstraction of that computation if the represented latent *operates* as a Bayesian latent. A Bayesian latent about the common cause under the fork structure of G_c determines the *joint distribution* over the unobserved variables given one observed child variable token. This means that, given an intervention on the represented latent, both the posterior over the parent and the prediction of the sibling should change consistently with Bayes's rule. Therefore we pose **RQ3**: Does the represented latent determine the joint distribution over the unobserved variables as prescribed by Bayesian inference?

Based on results from Section 5, we can describe outputs in log-odds coordinates as:

$$\hat{\Lambda}_{X_c} = G(|\Omega|)\, \rho_{X_c}\, z, \qquad \hat{\Lambda}_{Y_c^j} = \rho_{Y_c^j}\, z, \tag{3}$$

where $z \in \{0, \pm 1\}$ is the belief coordinate in unit beliefs, $G(|\Omega|) = \sigma(1)/\sigma(|\Omega|)$ is the accumulation gain, with $\sigma(|\Omega|)$ the scale of the final normalization when $|\Omega|$ children are observed, and ρ_{X_c} and

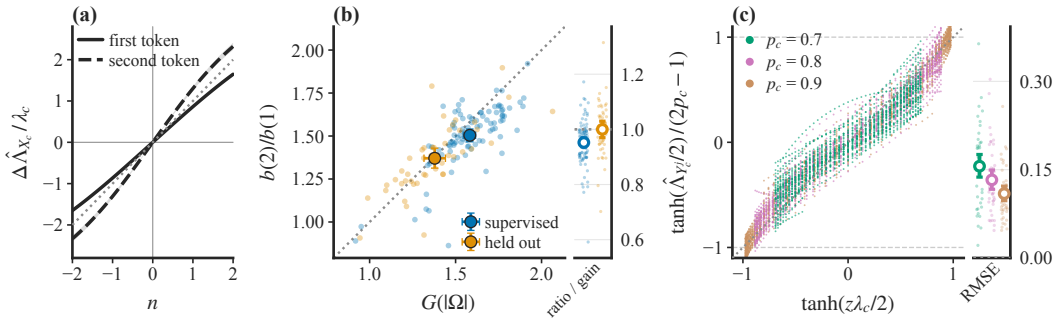


Figure 6: **Joint dependency intervention.** n unit beliefs are added along the belief coordinate after the functional attention; 135 (model, context) cells, means with 95% bootstrap CIs. **(a)** Parent log-odds shift against n at the first (solid) and second (dashed) child positions. **(b)** Slope of the two-child over the one-child line from (a) as *intervention ratio* against the *accumulation gain* $G(|\Omega|)$, supervised and held-out cluster; right: intervention ratio over accumulation gain is close to one, i.e., intervention follows the functional form. **(c)** Sibling against parent in $\tanh$ coordinates, colored by context index p_c; right: residual mean squared error from the Bayes predicted identity line.

$\rho_{Y_c^j}$ are the context-specific readout factors mapping the belief coordinate onto the log-odds of the parent and of the sibling. We use Eq. (3) to fix the alignment between the transformer's states and the variables of the Bayesian model. We then test this by a *joint dependency intervention*. We intervene on the represented latent at the residual node after the functional attention, and add n unit beliefs along z. We compare the resulting shifts in the log-odds of the parent and the sibling with those prescribed by Bayes's rule. We read the parent at one- and two-child sequences, and the sibling at one-child sequences. Bayes's rule couples the log-odds of the sibling to those of the parent, and we linearize this coupling by passing to $\tanh$ coordinates:

$$\tanh\left(\hat{\Lambda}_{Y_c^j}/2\right) = (2p_c - 1)\tanh\left(\hat{\Lambda}_{X_c}/2\right). \tag{4}$$

At one-child sequences, one injected unit belief shifts the parent's log-odds by ρ_{X_c}, slightly below λ_c (Figure 6a). At two-child sequences, the shift is larger by the accumulation gain $G(2)$ (Figure 6b). In $\tanh$ coordinates, the sibling follows the identity line across the steering range (Figure 6c), with a with the minimum error when the dependency is strongest. Steering the represented latent therefore shifts the prediction of the unobserved sibling by the amount Bayes's rule prescribes. The joint dependency intervention thereby answers **RQ3** by showting that **the represented latent operates as a Bayesian latent**, i.e., changing it along the belief coordinate shifts the joint distribution over the unobserved variables consistently with Bayes's rule.

7 Discussion, Conclusions, Limitations & Outlook

We presented a case study of transformers learning generalizable Bayesian abductive reasoning. Conceptually, we separated within-model from behind-data causality, and adopted an architecturally-aware optimal learner perspective to structure analysis. Methodologically, we designed a minimal setup that partially disentangles the two notions of causality. Empirically, we showed, supported by a joint dependency intervention analysis, that "Bayesian generalization" rests on representations shared across contexts, a categorical belief coordinate and a mapping to Bayesian posteriors (see Figure 1). The uncovered mechanism is an implementation of genuine Bayesian inference for the task. However, in our setup, it is not possible to recover the precise causal role of the parent variable due to known theoretical limits of *causal recovery* (Spirtes et al., 2000; Richardson & Spirtes, 2002). Our results do not transfer directly to large models, but suggest that they may also achieve "Bayesian generalization" by exploiting embedding similarity and efficient approximations to Bayesian computation.

Future work could continue the work presented here, e.g., by looking at more substantial (hierarchical) latent generation models, where latent variables are never observed. Models trained to predict tokens that are themselves informative about the data-generating process of other tokens might also make it possible to study "represented causality" in controlled experimental settings.

8 STATEMENTS

AI USE STATEMENT

In this work, we used generative AI tools for coding the experiments, implementing analyses and visualizations, drafting parts of the appendices, and feedback on the experimental design and the interpretation of results. We did not use generative AI tools for developing the conceptual framework, data generation, mathematical proofs, translation, or writing the main text. We have reviewed all AI-assisted work and take full responsibility for the final content, including any text or artifacts produced with the aid of generative AI.

ACKNOWLEDGMENTS

To be added after anonymity is lifted.

REFERENCES

Naman Agarwal, Siddhartha R. Dalal, and Vishal Misra. The bayesian geometry of transformer attention, 2026. URL https://arxiv.org/abs/2512.22471.

John R. Anderson. *The Adaptive Character of Thought*. Lawrence Erlbaum, Hillsdale, NJ, 1990.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):2678–2685, 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33012678. URL http://dx.doi.org/10.1609/aaai.v33i01.33012678.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2218523120.

Yuan Cao, Yihan He, Dennis Wu, Hong-Yu Chen, Jianqing Fan, and Han Liu. Transformers simulate mle for sequence generation in bayesian networks. *CoRR*, abs/2501.02547, 2025. doi: 10.48550/arXiv.2501.02547.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? In *NeurIPS 2024*, 2024. URL https://arxiv.org/abs/2506.21215.

Fabian Falck, Ziyu Wang, and Chris Holmes. Is in-context learning in large language models bayesian? a martingale perspective, 2024. URL https://arxiv.org/abs/2406.00793.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks, 2021.

Ritwik Gupta, Rodolfo Corona, Jiaxin Ge, Eric Wang, Dan Klein, Trevor Darrell, and David M. Chan. Enough coin flips can make LLMs act Bayesian. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7634–7655, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.377. URL https://aclanthology.org/2025.acl-long.377/.

Edward J. Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models, 2024. URL https://arxiv.org/abs/2310.04363.

Hui Jiang. A latent space theory for emergent abilities in large language models, 2023. URL https://arxiv.org/abs/2304.09960.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A Benchmark to Assess Causal Reasoning Capabilities of Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2312.0435010.48550/arXiv.2312.04350. URL https://openreview.net/forum?id=e2wtjx0Yqu.

Carina Kauf, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S. She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. Event knowledge in large language models: the gap between the impossible and the unlikely, 2022.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=mqoxLkX210. Featured Certification.

Matthew Levinson. Finding belief geometries with sparse autoencoders. *arXiv preprint arXiv:2604.02685*, 2026.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In *Proceedings of ACL*, 2021.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *Proceedings of ICLR 11*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Kenneth Li, Fernanda Viégas, and Martin Wattenberg. What does it mean for a neural network to learn a "world model"?, 2025. URL https://arxiv.org/abs/2507.21513.

Shenxiong Li and Huaxia Rui. Dual traits in probabilistic reasoning of large language models, 2024. URL https://arxiv.org/abs/2412.11009.

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

Mateusz Piotrowski, Paul M Riechers, Daniel Filan, and Adam S Shai. Constrained belief updates explain geometric structures in transformer representations. *arXiv preprint arXiv:2502.01954*, 2025.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, 17(1), January 2026. ISSN 2041-1723. doi: 10.1038/s41467-025-67998-6. URL http://dx.doi.org/10.1038/s41467-025-67998-6.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression, 2023. URL https://arxiv.org/abs/2306.15063.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002.

Raanan Y. Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31450–31465, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/642a321fba8a0f03765318e629cb93ea-Paper-Conference.pdf.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf.

Adam Shai, Loren Amdahl-Culleton, Casper L. Christensen, Henry R. Bigelow, Fernando E. Rosas, Alexander B. Boyd, Eric A. Alt, Kyle J. Ray, and Paul M. Riechers. Transformers learn factored representations, 2026. URL https://arxiv.org/abs/2602.02385.

Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander G Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2024.

Cosma Rohilla Shalizi and James P. Crutchfield. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *Journal of Statistical Physics*, 104(3-4):817–879, August 2001. ISSN 0022-4715, 1572-9613. doi: 10.1023/A:1010388907793.

Roger N. Shepard. Towards a universal law of generalization for psychological science. *Science*, 237: 1317–1323, 1987.

Vered Shwartz and Yejin Choi. Do neural language models overcome reporting bias? In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6863–6870. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.605. URL https://aclanthology.org/2020.coling-main.605.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a testbed for language model state tracking, 2022. URL https://arxiv.org/abs/2102.13249.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Proceedings of NeurIPS 38*, 2024. URL https://openreview.net/forum?id=aVK4JFpegy.

Bin Wang, W. Jeffrey Johnston, and Stefano Fusi. A mathematical theory for understanding when abstract representations emerge in neural networks, 2026. URL https://arxiv.org/abs/2510.09816.

Kevin Xia and Elias Bareinboim. Neural causal abstractions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20585–20595, 2024. doi: 10.1609/aaai.v38i18.30044. URL http://dx.doi.org/10.1609/aaai.v38i18.30044.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Yifei Yuan and Anders Søgaard. Revisiting the othello world model hypothesis, 2025. URL https://arxiv.org/abs/2503.04421.

Eliezer Yudkowsky. Gpts are predictors, not imitators. lesswrong.com, 2023. URL https://www.lesswrong.com/posts/nH4c3Q9t9F3nJ7y8W/gpts-are-predictors-not-imitators.

Liyi Zhang, R. Thomas McCoy, Theodore R. Sumers, Jian-Qiao Zhu, and Thomas L. Griffiths. Deep de finetti: Recovering topic distributions from large language models, 2023. URL https://arxiv.org/abs/2312.14226.

Jian-Qiao Zhu and Thomas L. Griffiths. Incoherent probability judgments in large language models. In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, and E. Hazeltine (eds.), *Proceedings of CogSci*, pp. 906–913, 2024.

A TRAINING

Architecture. GPT-style pre-LayerNorm decoder-only transformers with untied embedding and unembedding (Table 1).

Table 1: Architecture configuration common to all 45 models.

Layers / heads per layer	2 / 1
d_{model} / head dimension	16 / 16
MLP hidden size, activation	16, GELU
Dropout	none
Biases	QKV and output projections
Positional encoding	sinusoidal, max length 1000
Vocabulary	19 (1 shared delimiter + 3 contexts $\times$ 6)
Trainable parameters	4,051

Data sampling. Each sequence is sampled by some stratification constraints from the data-generating process to provide a closer match to theoretical distribution. In a 60 triplet long sequence, each context is represented 20 times, and within each context, each value of X_c appears 10 times, and the value counts of Y_c^1 and Y_c^2 match p_c. The emission order is sampled uniformly (approximately if counts permit) per each context. There is no fixed training set. A fresh batch is drawn at every step.

Objective and held-out masking. The loss is cross-entropy against one-hot next-token targets over all content positions; positions whose target is the delimiter are excluded. Each model has one held-out context, C_0. In every triplet of C_0 ordered as $Y_{c_0}^i\, Y_{c_0}^j\, X_{c_0}$, the parent token is removed from both the loss and the attention keys. As a result, the model is never trained to predict the parent from a held-out sequence $Y_0^i\, Y_0^j$. All other positions of C_0 are trained normally.

Training. Table 2 lists the optimization and checkpoint-selection settings. A single seed per model sets both the initialization and the data stream. Checkpoints are selected by $\mathrm{KL}(P_{\text{Bayes}} \,\|\, P_\theta)$ on the validation set, where P_{Bayes} is the exact Bayesian conditional, averaged over contexts.

Table 2: Optimization and checkpoint selection.

Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 0.1
Learning rate	0.05, constant (no warmup or decay)
Gradient clipping	norm 1.0, no accumulation
Batch size	384 sequences (92,160 tokens)
Training budget	20k steps (7.68M sequences)
Validation set	4,800 fixed sequences
Evaluation / checkpoint interval	every 10 steps
Selection criterion	$\min \mathrm{KL}(P_{\text{Bayes}} \,\|\, P_\theta)$
Selected step	2,390–17,480, median 8,450
Selected sequences seen	0.9M–6.7M, median 3.2M

Trained models. We train 45 models, crossing the held-out context parameter $p_0 \in \{0.7, 0.8, 0.9\}$ with 15 seeds $\{42, 100, 110, 120, 130, 140, 150, 200, 300, 400, 500, 600, 700, 800, 900\}$. Seeds repeat across held-out conditions, so each model is identified by its (p_0, seed) pair. Table 3 shows the emergence of a sparse attention layer. The sparse layer is the layer whose attention has the largest mean $\mathrm{KL}(\text{attention row} \,\|\, \text{uniform})$, and the other layer is diffuse; Appendix D shows that the sparse layer is also the functional layer.

Table 3: Sparse attention layer by held-out context.

SparseYour attention	Total	$p_0 = 0.7$	$p_0 = 0.8$	$p_0 = 0.9$
Layer 1	31	8	13	10
Layer 2	14	7	2	5

B Behavioral Metrics

Prediction invariance to emission order. The conditional future should not be a function of the permutation latent over the observed evidence. We measure to what extent this is true by *exchangeability error* Δ_{ex} derived by the total variation distance (TVD) of the model's distributions before and after swapping the two observed tokens in a two-token sequence

$$\Delta_{\text{ex}} = \delta\Big(P_\theta\big(\cdot \mid M_c(v,*)\, M_c(v',*')\big),\ P_\theta\big(\cdot \mid M_c(v',*')\, M_c(v,*)\big)\Big) \tag{5}$$

The parent d-separates the children. The conditional future should only be a function of the parent's value. Two prediction sensitivity measures quantify this claim in models, namely sensitivity of prediction of a child to its sibling S_Y, and to its parent S_X. They are both derived by the TVD of model's two distributions, before and after a value flip. We interpret them up to a predictive error of the model at the same token position as defined below.

$$S_Y = \delta\Big(P_\theta\big(\cdot \mid X_c^* Y_c^{i,*}\big),\ P_\theta\big(\cdot \mid X_c^* Y_c^{i,\bar{*}}\big)\Big) \tag{6}$$

$$S_X = \delta\Big(P_\theta\big(\cdot \mid X_c^* Y_c^{i,*}\big),\ P_\theta\big(\cdot \mid X_c^{\bar{*}} Y_c^{i,*}\big)\Big) \tag{7}$$

$$\bar{\epsilon}_1 = \delta\Big(P_\theta\big(\cdot \mid M_c(v,*)\, M_c(v',*')\big),\ P_c\big(\cdot \mid M_c(v,*)\, M_c(v',*')\big)\Big) \tag{8}$$

C Orthogonal decomposition of the embedding table

For each model and context C_c, the six embeddings $e_c(V,*) = e\big(M_c(V,*)\big) \in \mathbb{R}^{16}$, with $V \in \mathcal{V}_c$ and $* \in \{+,-\}$, form a 3×2 table of vectors. We stack them as the rows of $\tilde{E}_c \in \mathbb{R}^{6\times 16}$ after subtracting their mean, in the order $(X_c,+),(X_c,-),(Y_c^1,+),(Y_c^1,-),(Y_c^2,+),(Y_c^2,-)$. The centered table has rank at most five.

The usual two-way decomposition splits the five-dimensional column space of $\tilde{E}_c$ into three subspaces namely, a two-dimensional variable plane, a one-dimensional value axis, and a two-dimensional interaction plane. Since the children are exchangeable under the generative process, we split each plane once more into a parent-vs-children and a child-vs-child part. This gives five orthonormal axes in $\mathbb{R}^6$, the variable axes d_{r_1}, d_{r_2}, the value axis d_v, and the interaction axes d_p, d_a,

$$\begin{aligned} d_{r_1} &= \tfrac{1}{\sqrt{12}}(2,2,-1,-1,-1,-1), & d_{r_2} &= \tfrac{1}{2}(0,0,1,1,-1,-1),\\ d_p &= \tfrac{1}{\sqrt{12}}(2,-2,-1,1,-1,1), & d_a &= \tfrac{1}{2}(0,0,1,-1,-1,1),\\ d_v &= \tfrac{1}{\sqrt{6}}(1,-1,1,-1,1,-1). \end{aligned} \tag{9}$$

d_{r_1} separates the parent from the children and d_{r_2} the two children from each other. d_v is the value shift common to all three variables. d_p captures how the parent's value shift departs from the common one, and d_a how the two children's shifts differ. Orthogonality gives

$$\|\tilde{E}_c\|_F^2 = \sum_{k\in\{r_1,r_2,v,p,a\}} \big\|d_k^\top \tilde{E}_c\big\|^2. \tag{10}$$

Finally, we ask whether these label-defined axes are also the principal axes of $\tilde{E}_c$. With $u_{c,j}$ the j-th left singular vector of $\tilde{E}_c$, its weight on axis k is the squared cosine

$$\big(d_k^\top u_{c,j}\big)^2, \qquad \textstyle\sum_k \big(d_k^\top u_{c,j}\big)^2 = 1. \tag{11}$$

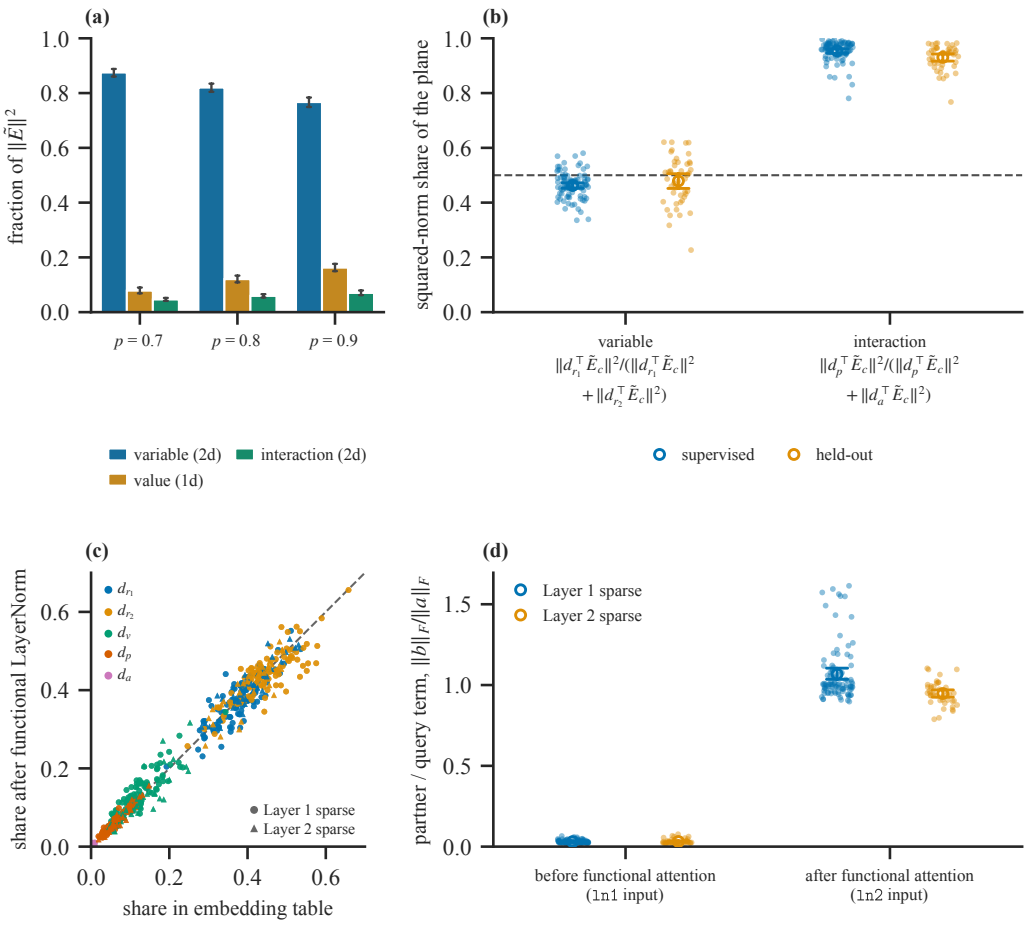


Figure 7: For each model and context C_c, the centered table $\tilde{E}_c \in \mathbb{R}^{6\times16}$, one row per variable and value, is split into the variable plane, the value axis and the interaction plane (Appendix C). **(a)** Share of $\|\tilde{E}_c\|_F^2$ in each subspace, by context p_c. Bars are means over the 45 models, and error bars are 95% bootstrap CIs with the model as resampling unit. This split uses no parent/child structure. Most of the table lies in the variable plane. The value share grows with p_c, and the interaction share stays small. **(b)** Share of the parent-vs-children axis within each plane, $\|d_{r_1}^\top \tilde{E}_c\|^2/(\|d_{r_1}^\top \tilde{E}_c\|^2 + \|d_{r_2}^\top \tilde{E}_c\|^2)$ for the variable plane and $\|d_p^\top \tilde{E}_c\|^2/(\|d_p^\top \tilde{E}_c\|^2 + \|d_a^\top \tilde{E}_c\|^2)$ for the interaction plane. Dots are single (model, context) values, and open markers are means with 95% bootstrap CIs. The dashed line at 0.5 marks a plane with no parent/child asymmetry. The variable plane lies near it, and the interaction plane is almost entirely the parent deviation. Supervised and held-out contexts agree throughout. **(c)** Share of each axis before and after the LayerNorm that precedes the functional attention, for the first, second and third token of a triplet. d_{r_1} separates the parent from the children, d_{r_2} the two children, d_v is the common value shift, d_p the parent's departure from it, and d_a the difference between the children's shifts. Points lie on the identity line, so the LayerNorm leaves the decomposition intact. **(d)** Partner term relative to query term, $\|b\|_F/\|a\|_F$, at the second token of a triplet, where a and b are the query's and the partner's contributions to the cell means of the residual stream. Markers as in (b). Before the functional attention (`ln1` input) the partner term is essentially zero. After it (`ln2` input), the ratio is close to 1 in both model groups, so attention adds the partner's embedding to the query's own at equal weight.

Preservation through the functional attention. The axes are defined on the token index, so they apply wherever the six tokens of a context can be tabulated. Additionally centering removes the positional information. The pre-attention LayerNorm standardizes and rescales each row separately, yet the shares $\|d_k^\top \tilde{E}_c\|^2/\|\tilde{E}_c\|_F^2$ of its output match those of the embedding table (Figure 7, so the decomposition enters the attention intact. The value-output map acts on the residual dimension while the axes act on the token index, hence

$$d_k^\top \left(\tilde{E}_c W_{OV}\right) = \left(d_k^\top \tilde{E}_c\right) W_{OV} \qquad \text{for every } k. \tag{12}$$

Attention thus carries each component through separately, changing only its norm and orientation.

D Attention Localization

We isolate the within-model causal effects of each attention block by activation patching. Let $h_{\text{base}}, h_{\text{donor}}$ be two sampled histories, each followed by a final triplet from context C_c. The two final triplets agree on every token before the last one and differ only in the canonical state $\mathcal{S} = (\Omega, \Lambda)$. We run two separate forward passes, the base and the donor run, and write $\mathrm{att}_L^{\text{base}}, \mathrm{att}_L^{\text{donor}}$ for the layer-L attention output at the last token. The patched run is then inference on the base sequence with substituted activations from the donor, at the last token and layer L. We patch at the first ($|\Omega| = 1$) and the second ($|\Omega| = 2$) token of the final triplet. We additionally draw h_{donor} with an independent history from C_c, and with an independent history from $C_{c'}$, $c' \neq c$.

Let $\ell_{\text{base}}, \ell_{\text{donor}}$ be the logits from the two clean runs and ℓ_{patch} the logits from the patched run, each over the full vocabulary $\{\text{delimiter}\} \cup \bigcup_c A_c$. We define the recovery measure as the fraction of the logit difference between the base and the donor run that the patch recovers,

$$R_L \;=\; \frac{\langle \ell_{\text{patch}} - \ell_{\text{base}},\ \ell_{\text{donor}} - \ell_{\text{base}} \rangle}{\langle \ell_{\text{donor}} - \ell_{\text{base}},\ \ell_{\text{donor}} - \ell_{\text{base}} \rangle}.$$

$R_L = 1$ means att_L carries the entire state distinction, $R_L = 0$ means it carries none.

The models follow the categorization by attention patterns sparsity. In 31 models where the first layer is sparse, that layer recovers a large fraction of the distinction whereas the second layer recovers almost nothing. The reverse holds for the 14 models with the second layer being sparse. (Figure 8a). Recovery is unchanged when the donor history is drawn independently from C_c. Patching both attention layers together yields the same recovery as patching the sparse layer alone, therefore the diffuse attention carries no additional information. The recovery of sparse attention is below one, then the skip connection must carry some of the distinction. Patching the connection alone recovers 0.469 of the distinction, therefore together they account for the entire distinction.

E Functional form of the attention block

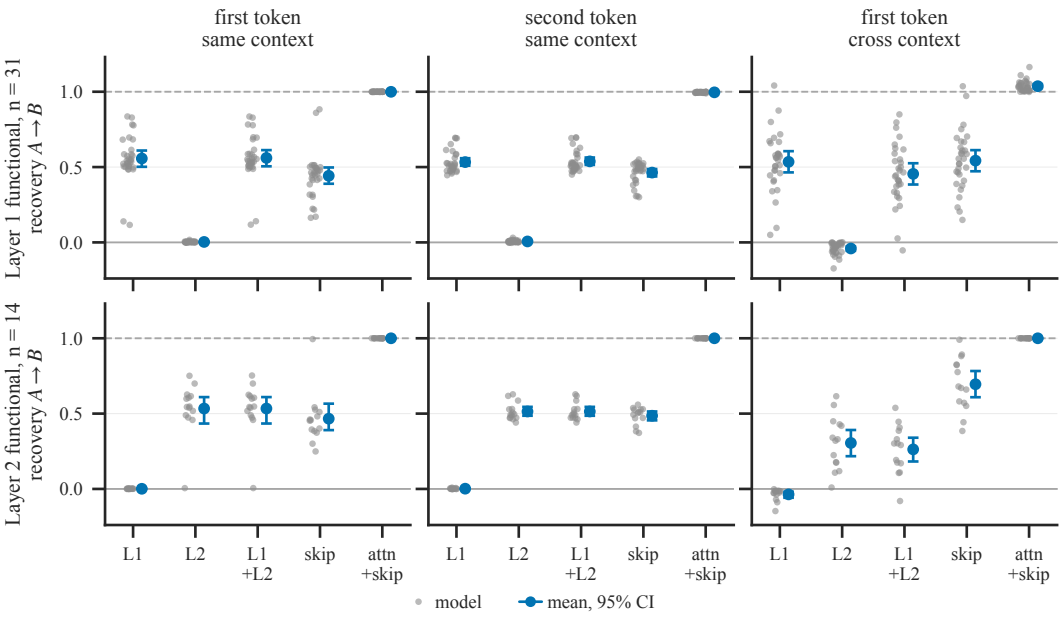
We treat the skip connection and the functional attention as one block and read everything at the residual node that follows it, in unnormalized coordinates. Write h_c for the residual stream at that node.

Regression. For context C_c, consider the second token of a triplet. The two tokens observed so far are $M_c(V, *)\, M_c(V', *')$, with $V \neq V' \in \mathcal{V}_c$ and $*, *' \in \{+, -\}$. Let $\bar{h}_c(V, *;\, V', *')$ be the mean of h_c over all instances of a pair, that is, over all histories that precede it. We arrange these means as a 6×6 table of vectors in $\mathbb{R}^{16}$, with rows indexed by the query token $M_c(V', *')$ and columns by the antecedent token $M_c(V, *)$, both in the row order of $\tilde{E}_c$. Pairs with $V = V'$ do not occur, so the three diagonal 2×2 blocks are empty, and each row holds four cells, 24 in total. Likewise, let $g_c(V, *)$ be the mean of h_c at the first token of a triplet whose token is $M_c(V, *)$.

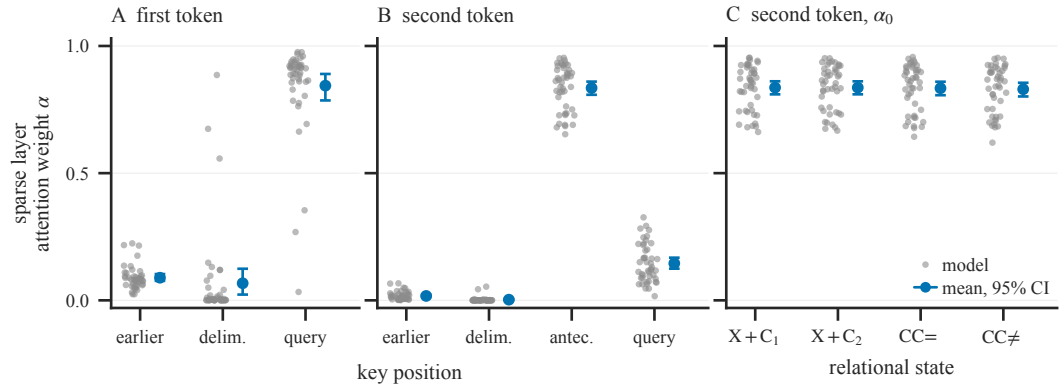
We model the table as a global offset plus a row term and a column term,

$$\bar{h}_c(V, *;\, V', *') \;=\; o_c + a_c(V', *') + b_c(V, *) + \varepsilon_c(V, *;\, V', *'), \tag{13}$$

where a_c is the query term, b_c the antecedent term, and ε_c the cross-term. The terms are made unique by requiring that the means of a_c and b_c, each weighted by how often its token occurs as query or antecedent, are zero; o_c is then the frequency-weighted grand mean of the table.



(a) Caption



(b) Caption

Because the empty cells lie on the diagonal blocks, a row mean averages only over the antecedents of the two other variables. With balanced counts, the antecedent terms in a row sum to $-\big(b_c(V',+) + b_c(V',-)\big)$, so the row mean holds

$$o_c + a_c(V',*') - \tfrac{1}{4}\big(b_c(V',+) + b_c(V',-)\big),$$

a remainder that depends on the query's variable. The same holds for column means. We therefore estimate all terms jointly by minimizing

$$\mathbb{E}\,\big\|h_c - o_c - a_c(V',*') - b_c(V,*)\big\|^2 \tag{14}$$

over all instances of two-token prefixes in the data. This equals the least-squares fit of the 24 occupied cells, weighted by their frequencies.

Additivity. We measure the cross-term by the residual fraction

$$\frac{\mathbb{E}\,\|\varepsilon_c\|^2}{\mathbb{E}\,\|\bar{h}_c - o_c\|^2}, \tag{15}$$

with both expectations over pairs, weighted by their frequencies. On this layout, the fit has 11 free parameters per residual dimension (1 for the offset and 5 for each token term), so even a table without structure leaves a residual fraction bounded away from zero. For balanced counts, its expected value

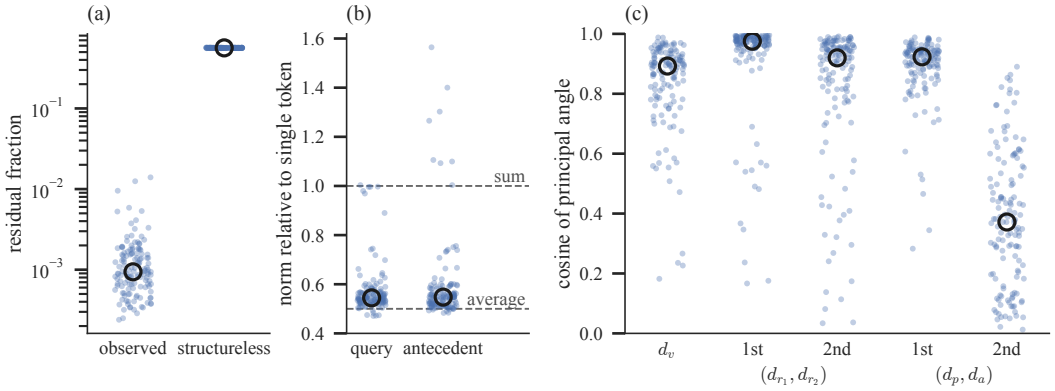


Figure 9: Functional form of the attention block, read at the residual node after the functional attention for the second token of a triplet. Each dot is one model and context; open circles are medians. **(a)** Share of the residual stream that a sum of one query term and one antecedent term cannot explain (log scale). The observed share is compared to that of random tables fitted in the same way. **(b)** Size of the query and antecedent terms relative to the residual of the same token observed alone. Dashed lines mark the values expected if the node summed the two tokens (1) or averaged them (0.5). **(c)** Alignment of the query and antecedent terms within each design subspace: the value axis d_v, the variable plane (d_{r_1}, d_{r_2}) and the interaction plane (d_p, d_a). For each plane, 1st and 2nd are its two principal angles, from most to least aligned; a cosine of 1 means the two terms point in the same direction.

is $13/23$: the residual degrees of freedom over the centered ones. As a reference, we fit independent standard-normal tables with the same layout and frequencies in the same way. The observed residual fraction lies orders of magnitude below this reference (Figure 9a), so nothing at the node depends on the pair beyond what each token contributes alone.

Fixing the scale. The regression leaves open whether the node sums or averages the two tokens. We take the single-token residual g_c as reference. With a_c, b_c and g_c stacked as 6×16 tables in the row order of $\tilde{E}_c$, centered over their rows, and $\|\cdot\|$ the Frobenius norm, summation implies $\|a_c\|/\|g_c\| = \|b_c\|/\|g_c\| = 1$, and averaging implies $\frac{1}{2}$. Both ratios lie near $\frac{1}{2}$ in every context (Figure 9b), so, up to the offset, the node holds approximately the average of the two single-token residuals.

Subspace alignment. Equal norms do not imply that the two terms occupy the same directions. For each design subspace, the value axis d_v, the variable plane $\mathrm{span}(d_{r_1}, d_{r_2})$ and the interaction plane $\mathrm{span}(d_p, d_a)$, we project a_c and b_c onto it. The projections of each term span a subspace of $\mathbb{R}^{16}$: one-dimensional for the value axis and two-dimensional for each plane. Between the two terms' subspaces, we compute the principal angles, the arccosines of the singular values of the product of their orthonormal bases. Principal angles depend on the subspaces and not on the bases chosen for them, so the result is the same for any basis of each design plane (Figure 9c). The angle of the value axis, both principal angles of the variable plane, and the first principal angle of the interaction plane are small, in supervised and held-out contexts alike: their mean cosines differ between the two by at most 0.02 on the value axis and the variable plane. Only the second interaction angle is large, in a direction that carries negligible mass.

F Normalization at the read-out

Read-out. Let $h \in \mathbb{R}^d$ be the residual entering the final LayerNorm, with $\mu(h) = \frac{1}{d}\mathbf{1}^\top h$ and $\sigma(h) = \left(\frac{1}{d}\|h - \mu(h)\mathbf{1}\|^2 + \epsilon\right)^{1/2}$. The LayerNorm maps h to $\gamma \odot (h - \mu(h)\mathbf{1})/\sigma(h) + b$. With

$w_c = \mathbf{u}(X_c^+) - \mathbf{u}(X_c^-)$, the parent's log-odds are $v_c(h) = A(h) + \langle b, w_c \rangle$, with

$$A(h) = \frac{N(h)}{\sigma(h)}, \qquad N(h) = \big\langle \gamma \odot (h - \mu(h)\mathbf{1}),\, w_c \big\rangle. \tag{16}$$

Gain. Let $h^{(1)}$ and $h^{(2)}$ be the residual stream activations before the final LayerNorm of a single-child and a value-concordant two-child sequence, respectively. With $\mathbb{E}^{(1)}$ and $\mathbb{E}^{(2)}$ the means over the single-child and value-concordant two-child sequences of one model and context,

$$F = \frac{\mathbb{E}^{(2)}[A]}{\mathbb{E}^{(1)}[A]} = F_w\, F_\sigma, \qquad F_w = \frac{\mathbb{E}^{(2)}[N]}{\mathbb{E}^{(1)}[N]}, \qquad \log F = \log F_w + \log F_\sigma. \tag{17}$$

Writers. The residual h is the sum of five writer outputs, $h = e + a_1 + f_1 + a_2 + f_2$, with e the token and position embeddings and a_ℓ and f_ℓ the attention and MLP (feed-forward) outputs of layer ℓ. Centering is linear, $h - \mu(h)\mathbf{1} = \sum_m \tilde{m}$ with $\tilde{m} = m - \mu(m)\mathbf{1}$, hence

$$N(h) = \sum_{m \in \{e, a_1, f_1, a_2, f_2\}} \big\langle \gamma \odot \tilde{m},\, w_c \big\rangle. \tag{18}$$

Each writer contributes additively to N, while $\sigma(h)$ depends on all writers jointly and scales every contribution alike. The contribution of writer m to the gain is

$$s_m = \frac{\mathbb{E}^{(2)}\big[\langle \gamma \odot \tilde{m}, w_c \rangle / \sigma\big] - \mathbb{E}^{(1)}\big[\langle \gamma \odot \tilde{m}, w_c \rangle / \sigma\big]}{\mathbb{E}^{(2)}[A]}, \qquad \sum_m s_m = 1 - \frac{1}{F}. \tag{19}$$

F.1 Gain factorization

For each context, 200 sequences of 20 triplets are generated, and the first triplet of each sequence is discarded. The read-out is taken at every position whose next token is the parent and whose earlier tokens in the current triplet are children. Positions with one child give $h^{(1)}$, positions with two value-concordant children give $h^{(2)}$, and positions with two value-discordant children enter only through the magnitude $|v_c|/\lambda_c$, with $\lambda_c = \log\big(p_c/(1-p_c)\big)$. All quantities are computed per model and context, and the models with the sparse layer in layer 1 and in layer 2 are pooled.

F.2 Shrinkage in the variable plane

Drop. The normalization factor F_σ exceeds one because the scale σ of the residual is smaller after two value-concordant children than after one. We measure this shrinkage by

$$\Delta\sigma^2 = \mathbb{E}^{(1)}\big[\sigma^2\big] - \mathbb{E}^{(2)}\big[\sigma^2\big] = \tfrac{1}{d}\Big(\mathbb{E}^{(1)}\big[\|\tilde{h}\|^2\big] - \mathbb{E}^{(2)}\big[\|\tilde{h}\|^2\big]\Big), \tag{20}$$

where $\tilde{h} = h - \mu(h)\mathbf{1}$.

Plane. We fit the regression of Appendix E to the residual at the final LayerNorm input and take the variable plane as the span of the projections of the query term onto d_{r_1} and d_{r_2}. Let $\tilde{h}_\parallel$ be the orthogonal projection of $\tilde{h}$ onto this plane and $\tilde{h}_\perp = \tilde{h} - \tilde{h}_\parallel$. We write

$$\Delta\sigma^2 = \Delta\sigma^2_{\text{plane}} + \Delta\sigma^2_{\text{rest}}, \tag{21}$$

where each term is

$$\Delta\sigma^2_{\text{plane}} = \tfrac{1}{d}\Big(\mathbb{E}^{(1)}\big[\|\tilde{h}_\parallel\|^2\big] - \mathbb{E}^{(2)}\big[\|\tilde{h}_\parallel\|^2\big]\Big) \tag{22}$$

$$\Delta\sigma^2_{\text{rest}} = \tfrac{1}{d}\Big(\mathbb{E}^{(1)}\big[\|\tilde{h}_\perp\|^2\big] - \mathbb{E}^{(2)}\big[\|\tilde{h}_\perp\|^2\big]\Big). \tag{23}$$

The share of the plane is the sum of $\Delta\sigma^2_{\text{plane}}$ over all models and contexts divided by the sum of $\Delta\sigma^2$. The plane carries most of the drop. A cross-model baseline in comparison only captures a small fraction of the drop.

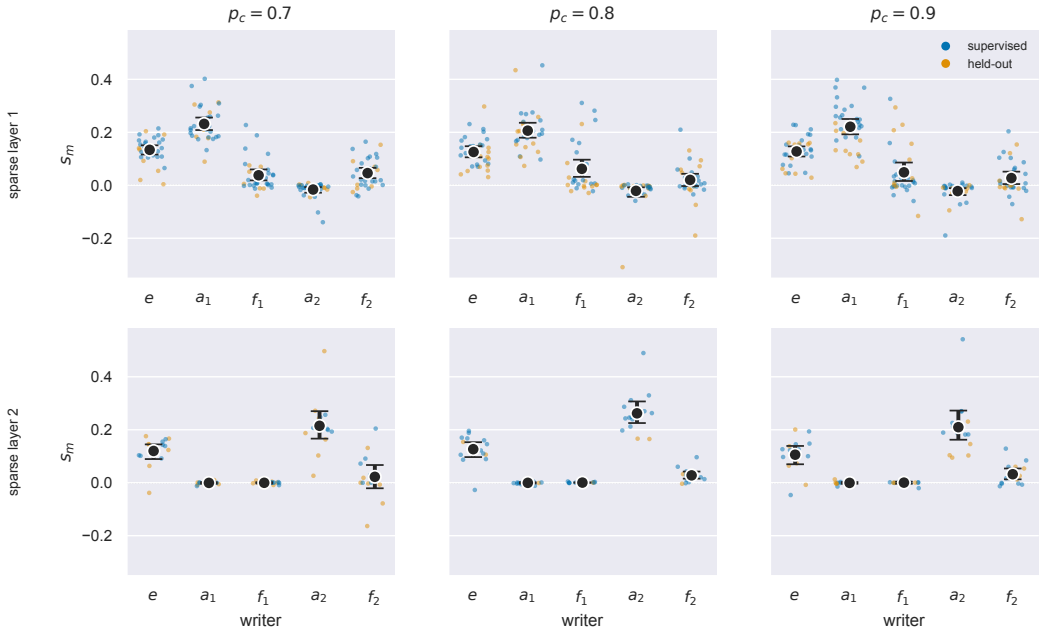


Figure 10: **Writers carrying the change in the readout.** For each writer $m \in \{e, a_1, f_1, a_2, f_2\}$, the share $s_m = \big(\mathbb{E}_2[\langle \gamma \odot \tilde{m}, w_c \rangle / \sigma] - \mathbb{E}_1[\langle \gamma \odot \tilde{m}, w_c \rangle / \sigma]\big) / \mathbb{E}_2[A]$ is the writer's contribution to the change in the readout from one observed child ($\mathbb{E}_1$) to two concordant children ($\mathbb{E}_2$), normalized by the two-child readout $A = v - \langle b, w_c \rangle$. The layer-norm bias term $\langle b, w_c \rangle$ is constant and excluded. For each model and context, the five shares sum to $1 - 1/F$, where $F = \mathbb{E}_2[A]/\mathbb{E}_1[A]$ is the total readout gain. Columns: contexts p_c. Rows: models with sparse attention in layer 1 (31 models) or layer 2 (14 models). Dots: individual models, blue for supervised contexts, orange for the held-out context. Black points: panel mean with 95% percentile-bootstrap CIs (10,000 resamples). The attention writer of the sparse layer carries the largest share, followed by the embedding e. In layer-2 models, the layer-1 writers a_1 and f_1 contribute almost nothing.

F.3 Expansion of the normalization scale

For writers m, m' write $\mathrm{var}(m) = \frac{1}{d}\|\tilde{m}\|^2$ and $\mathrm{cov}(m, m') = \frac{1}{d}\langle \tilde{m}, \tilde{m}' \rangle$, the variance and covariance over the coordinates of the residual. Since $\tilde{h} = \sum_m \tilde{m}$,

$$\sigma^2(h) = \sum_m \mathrm{var}(m) + 2 \sum_{m<m'} \mathrm{cov}(m, m') + \epsilon. \tag{24}$$

Applying $\mathbb{E}^{(1)} - \mathbb{E}^{(2)}$ term by term splits the drop exactly,

$$\Delta\sigma^2 = \sum_m \Delta\mathrm{var}(m) + 2 \sum_{m<m'} \Delta\mathrm{cov}(m, m'), \qquad \Delta(\cdot) = \mathbb{E}^{(1)}[\cdot] - \mathbb{E}^{(2)}[\cdot]. \tag{25}$$

Held-out deficit. For each model, the held-out deficit of a term is its Δ in the held-out context minus the mean of its Δ in the two supervised contexts. The fifteen term deficits sum to the deficit of $\Delta\sigma^2$. The deficits of the variances are close to zero in both families, and the deficit is carried by the covariances ([AP: fig: term deficits]). With the sparse layer in layer 1, it spreads over the covariances of e and a_1 with the MLP writers. The writers of the held-out context thus have the same magnitudes as in the supervised ones but are aligned differently, and no single pair accounts for the deficit.

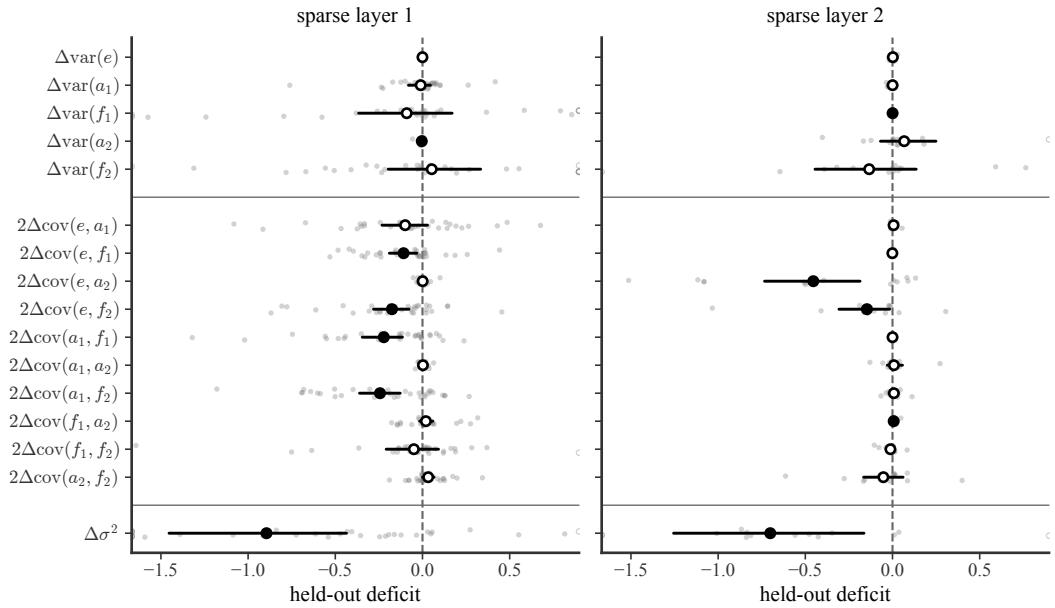


Figure 11: **Held-out shortfall in the drop of σ^2, by term.** For each model and term, Δ is the drop from one observed child to two concordant children ($\mathbb{E}_1 - \mathbb{E}_2$) in the held-out context, minus the mean drop over the model's two supervised contexts. Negative Δ means the held-out context shrinks the term less. Rows are the terms of the exact expansion of σ^2 over the five writers: five variances $\mathrm{var}(m) = \frac{1}{d}\|\tilde{m}\|^2$ and ten covariances $2\,\mathrm{cov}(m, m') = \frac{2}{d}\langle \tilde{m}, \tilde{m}' \rangle$. The bottom row is their sum, $\Delta\sigma^2$. Panels: models with sparse attention in layer 1 ($n = 31$) and layer 2 ($n = 14$). Grey dots are single models. Black points are means with 95% percentile-bootstrap CIs over models (2,000 resamples). A point is filled if its CI excludes zero and open otherwise. The x-axis spans the 2nd–98th percentile of all dots in both panels, padded by 5%. The 24 of 720 dots beyond it are drawn open at the edge. In both families the held-out context shrinks σ^2 less. In layer-1 models the shortfall is spread over the covariances of e and a_1 with the MLP writers f_1 and f_2. In layer-2 models it sits mainly in $2\Delta\mathrm{cov}(e, a_2)$ and $2\Delta\mathrm{cov}(e, f_2)$.