PROTOTYPE TRANSFORMER: TOWARDS LANGUAGE MODEL ARCHITECTURES INTERPRETABLE BY DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

While state-of-the-art language models (LMs) surpass the vast majority of humans in certain domains, their reasoning remains largely opaque, undermining trust in their output. Furthermore, while autoregressive LMs can output explicit reasoning, their true reasoning process is opaque, which introduces risks like deception and hallucination. In this work, we introduce the Prototype Transformer (ProtoT)—an autoregressive LM architecture based on prototypes (parameter vectors), posed as an alternative to the standard self-attention-based transformers. ProtoT works by means of two-way communication between the input sequence and the prototypes, and we show that this leads to the prototypes automatically capturing nameable concepts (e.g. “woman”) during training. They provide the potential to interpret the model’s reasoning and allow for targeted edits of its behavior. Furthermore, by design, the prototypes create communication channels that aggregate contextual information at different time scales, aiding interpretability. In terms of computation scalability, ProtoT scales linearly with sequence length vs the quadratic scalability of SOTA self-attention transformers. Compared to baselines, ProtoT scales well with model and data size, and performs well on text generation and downstream tasks (GLUE). ProtoT exhibits robustness to input perturbations on par or better than some baselines, but differs from them by providing interpretable pathways showing how robustness and sensitivity arises. Reaching close to the performance of state-of-the-art architectures, ProtoT paves the way to creating well-performing autoregressive LMs interpretable by design.

1 INTRODUCTION

Large-scale autoregressive language models have achieved strong performance across various domains, with architectures like GPT-4 and LLaMA (Achiam et al., 2023; Touvron et al., 2023a) demonstrating capabilities on benchmarks spanning mathematical reasoning, code generation, and natural language understanding tasks. However, these models exhibit limited transparency in their reasoning processes, creating challenges for understanding how they arrive at their outputs and potentially limiting their deployment in applications where interpretability is important. For example, it has been observed that there is a large disconnect between models’ explicit reasoning and their internal computational processes (Greenblatt et al., 2024): while language models can generate step-by-step explanations when prompted, research indicates that these explanations may not reflect their actual reasoning pathways (Turpin et al., 2023). This opacity also contributes to hallucination behaviors, where models generate confident but factually incorrect outputs without clear indicators of uncertainty (Zhang et al., 2025).

Current interpretability methods for language models primarily operate as post-hoc analysis tools on architectures not designed with interpretability as a primary consideration. Approaches such as attention visualization (Clark et al., 2019), probing techniques (Tenney et al., 2019), and causal intervention methods (Meng et al., 2022) provide insights into model behavior but face limitations imposed by the underlying self-attention architecture. More recent techniques like sparse autoencoders (Kissane et al., 2024) attempt to disentangle superposed features within existing architectures, though they still operate within the constraints of standard transformer designs.

In this work, we present the Prototype Transformer (ProtoT), an alternative autoregressive language model architecture that incorporates interpretability considerations directly into its design. ProtoT replaces the standard self-attention mechanism with a prototype-based approach, where learnable parameter vectors engage in bidirectional communication with input sequences. This design choice allows prototypes to capture interpretable concepts during training, providing more direct access to the model's reasoning components.

ProtoT offers several characteristics that distinguish it from standard transformer architectures. The prototype-based design enables direct inspection and modification of learned concepts, supporting targeted behavioral adjustments based on identifiable reasoning components. The architecture aggregates contextual information across different temporal scales through prototype communication channels, which facilitates the interpretation of both local and global reasoning patterns. Additionally, ProtoT operates with linear computational complexity relative to sequence length, versus quadratic for the standard self-attention. The explicit prototype representations enable attribution of predictions to the internal pathways that generated them, allowing inspection of how information is routed. Our contributions are briefly as follows:

- We introduce ProtoT, a novel autoregressive language model architecture that replaces self-attention with prototype-based communication. The architecture uses learnable parameter vectors that engage in bidirectional communication with input sequences, achieving linear computational complexity while maintaining competitive performance.
- We demonstrate that prototypes automatically learn disentangled, nameable concepts during training, across abstraction levels, enabling interpretation of model reasoning. We also show the per-prototype time preference, and "predict and consolidate" behavior patterns. We show that targeted behavior edits are possible for a wide range of concepts through selective prototype intervention.
- We provide extensive evaluation showing that ProtoT achieves competitive text generation quality and downstream performance, while offering superior robustness analysis. The architecture demonstrates stability under meaning-preserving perturbations, mediated by the prototypes.

2 RELATED WORK

Interpretability in Language Models. One of the main goals when it comes to interpreting language models is to identify which components—such as heads, layers, or neurons—are responsible for specific behaviors (Zhang & Nanda, 2023). This is non-trivial, as attention magnitude does not necessarily imply causal importance (Jain & Wallace, 2019). Moreover, phenomena like *superposition*, where multiple features are encoded in overlapping directions, make isolating concepts difficult (Elhage et al., 2022). Consequently, this problem is often approached via causal intervention, analyzing activation differences between clean and corrupted prompts (Meng et al., 2022; Geva et al., 2023; Wang et al., 2022). A recent promising direction involves Sparse Autoencoders (SAEs), applied not just to MLPs but also to attention outputs (Kissane et al., 2024), aiming to disentangle superposed features into interpretable units (Rai et al., 2024).

Prototype Methods. Prototype methods seek to render decisions interpretable by relating inputs to learned examples. In computer vision, this often involves comparing inputs to prototypical parts for classification (Chen et al., 2019; Rymarczyk et al., 2022). Recently, ProtoViT (Ma et al., 2024) adapted this to Vision Transformers, using prototypes as deformable parts in the final layer. In NLP, approaches like ProtoAttend (Arik & Pfister, 2020) use attention over entire training examples for decision-making. Other architectures, such as ProtoryNet (Hong et al., 2023) and ProSeNet (Meng et al., 2022), introduce prototype trajectories or sparsity constraints to refine interpretability. Unlike many of these works which place prototypes only at the final classification stage, our architecture integrates prototype routing at every level of the hierarchy. Recent advances also include ProtoLens for sub-sentence span extraction (Wei & Zhu, 2025) and white-box frameworks for sentiment detection (Wen & Rezapour, 2025).

Alternatives to Self-Attention. Recent work has explored replacing standard self-attention with mechanisms using fixed sets of latent vectors. Slot Attention (Locatello et al., 2020) employs a competitive binding mechanism (softmax over slots) to segment inputs, but relies on iterative refinement steps (e.g., GRU) over static inputs. The Perceiver family (Jaegle et al., 2022; Hawthorne et al., 2022) decouples compute from input size by projecting data into a latent space processed by a standard transformer stack. Our ProtoT mixer differs fundamentally in both interaction and state update. Unlike Perceiver, where latents interact globally via self-attention ($O(R^2)$), ProtoT prototypes never

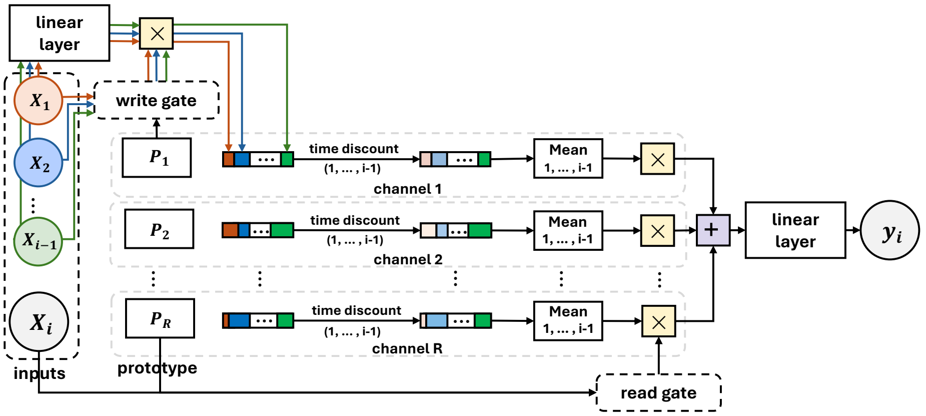


Figure 1: A single autoregressive step of the ProtoT mixer module. Prototypes $P_1, \dots, P_R$ route inputs $x_1, \dots, x_{i-1}$ (past-only – excluding x_i) into R channels via similarity scores at the *write gate*. Time-discount and time-wise mean are applied per channel. The *read gate* reads from each channel via the similarity between its prototype and x_i, followed by aggregation into the output y_i.

interact; they serve as filters for R independent, parallel channels ($O(R)$). And unlike Slot Attention's iterative refinement, ProtoT updates state autoregressively via strict past-only time-discounted aggregation (EMA). This design creates a semantic routing bottleneck rather than a general-purpose processing workspace, encouraging prototypes to capture nameable concepts (Sec. 5.1).

3 PROTOTYPE TRANSFORMER

The prototype transformer (ProtoT) is an autoregressive LM architecture, based on prototypes. It is a transformer stack identical to LLaMA-3 (Grattafiori et al., 2024), apart from the mixer module: ProtoT uses a prototype-based mixer instead of standard self-attention. Like LLaMA-3, ProtoT has L blocks ("layers"), each consisting of a mixer and a SwiGLU (Shazeer, 2020) feed-forward (FFN) module with the same intermediate ratio of ~2.7 as in LLaMA-3, and skip-connections with RMS pre-layernorm (Zhang & Sennrich, 2019) around each of the mixer module and the FFN.

Prototype mixer: This module is a self-attention alternative that uses R prototypes (trainable parameter vectors) to route the communication across the sequence through R corresponding channels (Fig. 1). Communication in and out of the channel is mediated via a write gate (in) and read gate (out) (Eq. 1 and 2). Each prototype is used as a filter via the write gate to aggregate (time-discounted) information from the past, defining a communication channel, and reading back the information via the read gate into the sequence. The prototype mixer follows the following formula applied at any token position i, for linear maps U, V, and W, with full details below:

$$x_i \leftarrow U\Big(\sum_{k=1,\dots,R} \underbrace{\text{Softmax}_k\big((W(x_i)\cdot \mathbf{P}_k)/\tau_r\big)_k}_{\text{Read Gate}} \underbrace{\text{PM}(\mathbf{P}_k, x_1, \dots, x_{i-1})}_{\text{Prefix Mean, channel } k}\Big), \text{ where:} \tag{1}$$

$$\text{PM} = \Big(\sum_{j:j<i} \beta_k^{i-j} \underbrace{\text{Softmax}_k\big((x_j \cdot \mathbf{P}_k)/\tau_w\big)_k}_{\text{Write Gate}} V(x_j)\Big) \Big/ \underbrace{\sum_{j:j<i} \beta_k^{i-j} \text{Softmax}_k\big((x_j \cdot \mathbf{P}_k)/\tau_w\big)_k}_{\text{Mass Normalization}} \tag{2}$$

Communication passes through an R-channel bottleneck, each channel mediated by a prototype via the read and write gate. This bottleneck encourages the network to specialize the channels semantically, which in turn specializes the prototypes semantically. This is likely what leads to the prototypes capturing nameable concepts at the write gate (Sec. 5.1).

Write gate: a cross-attention-like gating mechanism between prototypes P_k and inputs x_j (Eq.2), reweighing the values ($V(x_j)$). Unlike cross-attention, the Write Gate uses softmax over the prototypes, to do channel-aligned rather than sequence-aligned gating. It writes information from the sequence into the R channels, and uses a learned temperature τ_w for added expressivity. At layers 0 and 1 we also add a *local convolution* at the values stream (immediately after $V(x_j)$) – a convolution with kernel size 5, across the 4 past tokens and the current one, with h channels corresponding to the hidden dimension. It adds expressivity at the value stream, by capturing short-term relationships, and we show reduced perplexity and increased utility of layer 0, as measured by increased alpha-gate (Appendix A.7.1, Table 15). We also show kernel-size ablations in Appendix A.7.2, Table 16.

Read gate: a cross-attention-like mechanism that reads information from the R channels according to similarity with the corresponding R prototypes Eq.2. It is identical to the write gate, except for the linear map $W(x_i)$ and separate temperature τ_r, which add expressivity and allow for read/write gate specialization. This specialization may also help with interpretability, by decoupling the reading and the writing mechanisms. At layer 0 we use *shared read/write routing* (removing the W linear map) and sharper τ_r initialization (3.0 vs 1.0) which we show (Appendix A.7.1, Table 15) reduce perplexity and increase utility (higher alpha-gate value, defined below), likely by providing an inductive bias that reduces noise.

Prefix mean (PM): aggregator of past information via R communication channels, each corresponding to a prototype. In PM, a cumulative sum operation (Eq. 2) enforces a *strict autoregressive constraint* for next-token prediction. At position i, the aggregation is only on entries from positions $j < i$, so the representation used for predicting token i is a function of the past context only. In contrast, standard self-attention allows position i to attend directly to itself, providing a vertical shortcut from the input at position i to the output at the same position. By removing this direct path, the prefix mean encourages the write gate to base its updates on earlier positions and to anticipate the needs of the read gate, which we empirically demonstrate in Section 5.1. A *discounted prefix* gives per-channel time preference, allowing aggregation at different time scales. It is defined as exponential moving average (EMA) (time discount) on the Prefix Mean, via $\beta_k = \sigma(\gamma_k) \in (0, 1)$, parametrized by learnable γ_k. It can also be used to interpret the time preference of each prototype, as in Section 5.1. *Mass normalization* then turns the prefix cumulative sum into a prefix mean by dividing it by the sum of coefficients. This theoretically stabilizes the computation, and we have observed reduces perplexity at a small computational cost (Appendix A.7.2, Table 18). We also use *low-rank projection* at $1/2$ of the hidden size (h) at the value stream ($V(x_j)$), which saves up to 50% compute at the mixer module, with similar performance (Table 18). We keep the prototypes and routing (read and write gates) in the full size h as their computational cost is only linear in h.

Alpha Gate: a ReZero-like (Bachlechner et al., 2020) scalar gate applied at the output of each Prototype Mixer module before it merges with the residual stream (skip-connection). Unlike ReZero, which uses it to improve training of extremely-deep models, we use it as a low-computational-cost debugging tool: a low value of α at a given layer (declining rapidly during training) is a strong evidence that the mixer is not contributing to the overall prediction. Due to this role of α, we initialize it at identity (1.0) (vs ReZero's 0.0), which also performs better (see Appendix A.7.2, Table 17).

Compute: The ProtoT computational cost scales linearly in sequence length, as visible by Eq. 1 and 2. Note the recurrence in Eq. 2: the Prefix Mean for x_i depends only on the Prefix Mean for x_{i-1} and on x_{i-1}, both of which can be cached. This means that the model can generate tokens at sequence-wise constant (O(1)) computational and memory cost.

4 EXPERIMENTAL SETUP

Baselines: We compare ProtoT to three representative mixer families while keeping the backbone fixed: depth (6), hidden size (256), FFN ratio (2.7×), RMSNorm, dropout (0.1), and the training recipe. We use the same tokenizer and optimizer across models and do not reuse any pre-trained weights. To isolate mixer effects, we exclude MoE (e.g., Qwen-3) (Yang et al., 2025) and hybrid architectures (e.g., Jamba) (Lieber et al., 2024). We compare against a *LLaMA-style Transformer*: a single-expert, decoder-only self-attention Transformer following LLaMA-3/3.1 (Grattafiori et al., 2024), matched to ProtoT in backbone hyperparameters and training setup. The only architectural difference is the mixer (self-attention vs. prototype); *Mamba* (Gu & Dao, 2023): a modern state-space model (SSM) instantiation with the same dimensionality (6 layers, hidden 256) and FFN configuration as ProtoT; and *DeltaNet (Yang et al., 2024) (delta-rule linear transformer)*: a linear-attention baseline, configured with the same width, depth, and FFN ratio as ProtoT.

Dataset: We use a subset of the FineWeb-Edu dataset (Penedo et al., 2024), a high-quality web crawl dataset specifically curated for language model training. FineWeb-Edu consists of educational and informational web pages, providing diverse, coherent text that is well-suited for training autoregressive language models. The full dataset contains approximately 1.3 trillion tokens (with a less strict "score-2" filtering version containing 5.4 trillion tokens), but we use a manageable 250 million token subset for our experiments. Our sampled dataset contains 360,313 documents, with an average document length of 694 tokens. We use a custom BPE tokenizer (Sennrich et al., 2015) trained on the dataset with a vocabulary size of 16,000 tokens. For training, we use 338,695 documents (234.9M tokens) for the train split, 18,015 documents (12.5M tokens) for validation, and 3,603 documents (2.6M tokens) for testing. We chose this dataset over alternatives like C4 or The Pile because of its focus on high-quality, educational content.

Hyperparameter search: We do hyperparameter search on 18k examples for 10 epochs of the training data, with the default model sizes: hidden size h=256, layers L=6, and context ctx=256, unless otherwise specified. We use automatic search over batch size (32, 64, 128) and learning rate (from interval (3e-5, 3e-2)). For the search, we use Optuna with BoTorchSampler, with 15-trial warmup and 50 total trials, averaging over 3 seeds per trial. We found that batch size of 32 works best for all, but that the best learning rate varies across models. See Appendix A.2 for more details.

Learning rate (LR) and scheduler: We use linear warmup over 2% of training, and cosine annealing towards 10% of the peak learning rate. This is common practice in large LM training (Kalra & Barkeshli, 2024) for two reasons: (1) warmup helps reduce divergence for large LMs (e.g., we observed that LLaMA had convergence issues without warmup when we tried the large-scale setting), and (2) cosine annealing helps with reaching higher peaks and lower lows of the LR (e.g., we found that values more than 1.6e-3 were best, compared to 7e-4 for flat LR), and lead to significantly lower perplexity (2-6 points less, depending on the model, in the default settings). We train all our models with AdamW (Loshchilov & Hutter, 2017), following standard practice in language modeling. Compared to SGD, AdamW is more robust to hyperparameter choice (Zhao et al., 2024).

Dropout: For all models, we use dropout (with probability 0.1) after the token embeddings, at the residual (block output) between blocks, and inside the FFN, because we find that it reduces perplexity for all models (Appendix A.7.2, Table 20). This is likely because it prevents overfitting in the multi-epoch training regime (10 epochs) that we use. For LLaMA, we additionally put dropout inside the self-attention (HuggingFace-supported option), which further decreases perplexity.

Attention heads and prototypes (R): Similar to (Press & Wolf, 2017), we have found that sharing the weights between embeddings and LM head reduces perplexity at the hyperparameter search stage, for all models. This is likely because it provides a good inductive bias aligning the token embeddings between input and final projection. We keep this choice at large-scale experiments as well, for simplicity. We also select attention heads from $\{2, 4, 8\}$, but at both small-scale and large-scale runs we have found that 4 works best for all models with attention heads (LLaMA and DeltaNet), which is what we use. For ProtoT's prototypes (R), we have found diminishing returns in terms of perplexity improvements beyond R=32 (Appendix A.7.2, Table 19), while computation scales linearly with R. Therefore, we use R=32 for all runs.

5 EXPERIMENTS

Large-scale training: In Table 1, we compare ProtoT to the 3 baselines at large-scale training (first vs. last column). We study the effect of simultaneously scaling the hidden size 2x, the layers 2x, the context size by 2x, and the training data $\sim$19x, versus the default training settings. The results show that ProtoT scales well to the large model/data scenario. We show that ProtoT maintains relative performance to LLaMA, or even improves it (15.0 $\rightarrow$ 14.3% worse) with scale. Furthermore, ProtoT outperforms the DeltaNet linear-attention baseline (29.5 vs. 31.5 perplexity, respectively). However, a large gap remains versus LLaMA and the Mamba state-space model (29.5 vs. 25.8

Table 1: Long-context scalability: *Cols. 1–4:* scaling from the default 256 up to 2048; *Cols. 1 & 5:* Default (h=256, L=6, ctx=256, Ex=18k) vs. Large-scale training (h=512, L=12, ctx=512, Ex=339k). Test perplexity (lower is better). Best results in each section are in bold.

Model	Default	512	1024	2048	Large-scale
LLaMA	**78.7**	**70.4**	**65.3**	**63.6**	**25.8**
Mamba	86.0	78.0	70.5	69.5	26.5
DeltaNet	90.4	76.3	70.6	68.9	31.5
ProtoT	90.5	84.8	80.5	81.9	29.5
ProtoT (h=512)	97.2	82.0	**73.7**	**73.0**	–
ProtoT (L=12)	109.6	**79.8**	74.6	76.7	–
ProtoT (R=64)	94.0	83.2	79.1	80.5	–

and vs. 26.5, respectively). While we did our best to optimise ProtoT, this is the first iteration of the model, whereas established LMs like LLaMA have had multiple (Touvron et al., 2023a;b; Grattafiori et al., 2024). We expect with community feedback and further refinement to shrink this gap.

Long-context scalability: The results in Table 1 (columns 1-4) show that ProtoT scales poorly with context length (if other model dimensions are fixed), which suggests that ProtoT is running into a bottleneck. This is likely because the cross-sequence communications pass through the prefix mean (Fig. 1 and Eq. 1), over R channels with h hidden dimensions each, which can be restrictive. We further investigate this issue in the final 3 rows of Table 1, where we compare possible culprits: the hidden size h, the number of prototypes R, and the layers L (which can also play a role). The results show that the hidden dimension is the most restrictive as increasing it is the only one of the 3 that keeps improving with context size beyond 1024. Our model is most affected by this likely because of our choice to project down to $h/2$ at the values ($V(x_j)$ in Eq.1) to save compute, further exacerbating this bottleneck. In practice, this is less of an issue because, in more realistic settings (e.g. *Large-Scale Training*), the larger capacity of the model would allow for larger context lengths.

Text-Generation Performance: To evaluate quality of outputs, we measure open-ended text-generation ability using an LLM-as-a-judge protocol, following the Chatbot-Arena style pairwise comparison setup. For each prompt, two model outputs are evaluated by a frozen judge model under a fixed rubric, providing win/tie statistics that are converted into Elo scores. As shown in Appendix B, ProtoT achieves competitive generation quality, outperforming DeltaNet while landing between LLaMA and Mamba in Elo ranking. See Appendix B for text generation samples from each model.

Table 2: GLUE downstream fine-tuning results (all metrics reported as percentages). For COLA we report Matthews correlation; for SST-2 accuracy; for MRPC F1; for STS-B Pearson correlation; for RTE, WNLI, QNLI, MNLI and MNLI-MM accuracy; for QQP F1. GLUE reports the unweighted average of the nine task-specific scores. Results are averaged over 3 seeds. Best results are in bold.

Model	COLA	SST-2	MRPC	STS-B	RTE	WNLI	QQP	QNLI	MNLI	MNLI-MM	GLUE
LLaMA	**31.5**	**90.8**	**82.7**	**78.3**	**57.8**	**65.1**	**68.0**	**86.0**	**79.8**	**79.6**	**71.6**
Mamba	31.1	88.6	80.3	72.8	54.4	**65.1**	64.8	82.4	74.7	74.7	68.6
DeltaNet	13.8	85.8	80.1	67.0	50.9	**65.1**	62.6	80.1	71.1	71.8	64.5
ProtoT	27.7	90.0	80.1	66.2	53.9	64.6	64.8	81.8	75.3	74.8	67.6

Downstream performance: To comprehensively evaluate the general-purpose language understanding of ProtoT vs baselines, we fine-tune on the GLUE benchmark (Wang et al., 2018) consisting of 9 English NLU tasks spanning sentence- and sentence-pair classification as well as semantic textual similarity (more details in Appendix A.4).

As shown in Table 2, overall, LLaMA achieves the best overall performance, but ProtoT remains highly competitive and often matches or outperforms the dense baselines. In particular, ProtoT consistently attains the second-best scores on MNLI and MNLI-MM, indicating strong cross-domain robustness for large-scale natural language inference. On mainstream single-sentence and sentence-pair classification tasks such as SST-2, QQP and QNLI, ProtoT performs close to LLaMA and on par with Mamba while clearly outperforming DeltaNet, showing that its structured prototype representations do not sacrifice accuracy on high-resource benchmarks. For low-resource tasks such as RTE and CoLA, ProtoT delivers performance comparable to dense models, suggesting that its inductive bias can maintain stable accuracy even when training data is limited. Taken together, these test-set results confirm that ProtoT preserves competitive GLUE performance while offering structural advantages, especially for robust inference under distribution shifts.

Throughput Benchmarks:

We evaluate both training and inference throughput. For training, we use identical conditions across models (same data pipeline, optimizer, BF16 precision, sequence length 256, batch sizes 32 and 128). ProtoT achieves 25.2 and 7.6 it/s (batch 32/128), outperforming Mamba (11.9 and 3.2 it/s) and DeltaNet (3.5 and 1.8 it/s) while lagging behind the optimized LLaMA attention baseline (55.1 and 23.6 it/s). For autoregressive inference with batch size 1 on a single H100 80GB GPU, LLaMA attains the highest throughput at short context lengths, whereas ProtoT scales more favorably with context and surpasses LLaMA at 32k tokens and beyond; DeltaNet maintains the highest throughput at long context lengths. See Appendix A.6 for detailed throughput benchmarks and FLOP analyses.

5.1 INTERPRETABILITY

Prototypes act as representational slots: contextual information is aggregated into R prototype channels via the write gate and then read back through the read gate (Fig. 1). This structure allows features to be stored and reused within each sequence through the prototype-specific prefix means, enabling association of prototypes with identifiable concepts. Each prototype also has an associated decay parameter β_k, applied in the prefix mean (Eq. 2) to discount past activations. Smaller β_k values produce faster decay, while larger values allow information to persist longer. For interpretability, we report the derived half-life $t_{1/2}^{(k)} = -\frac{\ln 2}{\ln(\beta_k)}$, specifying the expected number of steps for the contribution of prototypes to halve and providing a direct way to analyze specialization in short- or long-term dependencies. We analyze read–write interactions to understand how the model integrates and updates contextual information through prototype channels during sequence generation.

Experiments: To investigate interpretability properties, we design four experiments with ProtoT (large-scale model from Table 1). We compute write routing activations across sequences from the FineWeb validation set for each prototype, aggregate them at the sequence level, and rank sequences by total activation strength. This identifies sequences that most strongly activate each prototype and allows us to visually inspect learned concepts and the relation between temporal locality and β_k parameter. We also use collected activations to compute widely adopted metrics (L1 sparsity, Gini Coefficient, Entropy, Mutual Information).

We analyze write and read phases during sequence generation. For a subset of prototypes, we select the most activating sequences and compute write and read routing activations for each token along the same prototype. This enables inspection of the internal dynamics of ProtoT, showing how sequence level information is aggregated and maintained during processing.

To quantitatively assess the presence of disentangled and nameable concepts and the impact of polysemanticity, we introduce an extensive analysis and labeling method inspired by auto-interpretability score metrics (Bricken et al. (2023); Paulo et al. (2024)). For each prototype, we collect its ten most activating sequences and extract, within each sequence, the tokens with the highest activation. This compact summary of prototype usage is submitted to an LLM-based evaluator (GPT-5.1), which is prompted to produce the following:

- **Theme:** the main recurrent theme identified in the most activating sentences.
- **Disentanglement score:** score between 1 and 10 that assesses how well the main theme is disentangled from other themes.
- **Main topic coverage:** count between 1 and 10 that assesses how many of the most activating sentences effectively contain the main theme.
- **Number of themes:** count between 1 and 10 that assesses how many uncorrelated themes are present in the most activating sentences (if the themes are at least 10, assign 10).
- **Explanation:** natural language explanation that describes the labeling and scoring process.

In order to compare the potential for interpretability in ProtoT and Transformer models, we collect activations, compute the same metrics and perform the LLM-aided evaluation experiment also on the trained LLaMA model from Table 1. We present statistics collected with the LLM-aided interpretability experiment in Figure 2 and an example of prototype visualization in Figure 3. The exact prompt used is in Appendix A.10. We show additional results in Appendix A.9 for LLM-aided interpretability experiments with LLaMA (Figure 24) and with multiple model configurations of ProtoT (Figures 19, 20, 21, 22, 23). Interpretability metrics for LLaMA and ProtoT are in Appendix A.8 in Figures 8 and 9.

We probe the functional role of individual prototypes through a targeted intervention experiment. Based on write-gate activations on the FineWeb validation set, we identified three functionally distinct prototypes from Layer 9: *L9 P7*, which encodes a ‘female’ concept; *L9 P18*, which partially encodes a ‘male’ concept; and *L9 P2*, a gender-neutral control. Our intervention consists of disrupting each of these prototypes via parameter re-initialization and measuring the subsequent change in the conditional probability of the target words ‘women’ and ‘girls’. We illustrate these prototypes in Figures 4, 5, and 6. Additional details on the construction of test sentences are in Appendix A.3.1.

Interpreting prototypes at the write gate: Human and LLM-aided evaluation reveal that prototypes capture disentangled concepts across varying levels of semantic abstraction, which naturally emerge as a result of training and encode interpretable patterns. For example, we identify concepts like entity names, functional words, verbs, as well as composite dates, illnesses, or school-related narratives. We also find that these concepts generally reflect the hierarchical organization of the model, with early layers tending to capture more superficial patterns and deeper layers representing composite and abstract semantics. We observe that polysemanticity is present in a few prototypes but remains limited overall and in fact, LLM aided evaluation reveals high disentanglement and a low number of uncorrelated themes for the vast majority of prototypes. Furthermore, we identify a correlation between half-life values and encoded concepts, where lower half-life values tend to capture local elements (such as stop words, or punctuation). Finally, we observe that ProtoT performs significantly better than LLaMA on all LLM-aided evaluation scores, with higher disentanglement and coverage and a lower number of uncorrelated topics. These results show that gate-mediated communication at the write gate forms prototype hubs that can largely be treated as separate, disentangled concept hubs, highlighting their potential for interpretability. A study on the correlation between half-life values and locality can be found in Appendix A.8, Table 21.

Results of the write-read alternation pattern: We observe a consistent temporal pattern in read and write activations, with read activity peaking one step before write activity. For example, in the results shown in Figure 3 (right), for the token ‘protection’, the read gate activates prototype 4 at the preceding token ‘fall’, followed by write activation on ‘protection’. This pattern is consistently seen across the most strongly activating sequences for each prototype and suggests that read and write gates may develop coordinated interactions. This coordination is consistent with a predict-and-consolidate behavior, where the read gate appears to anticipate which prototype may be relevant for the upcoming tokens, and the write gate subsequently updates the memory based on the current token.

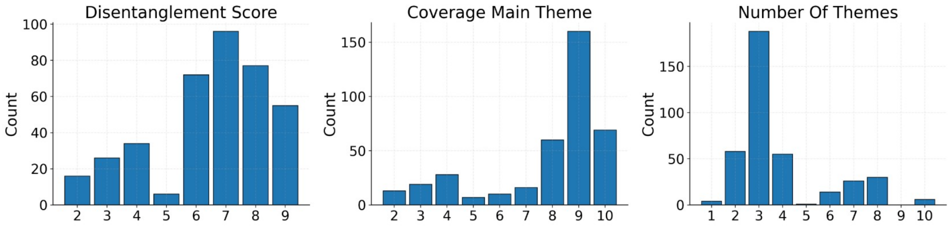


Figure 2: histograms of the numerical scores provided by LLM aided interpretability

Results of the prototype intervention: Our intervention experiments demonstrate that prototypes function as specific and interacting semantic hubs. By employing gate masking (ablating the write/read channels) alongside random re-initialization, we isolated precise mechanistic roles. Disrupting the ‘female’ prototype *L9 P7* significantly decreased the probability of related words (e.g.,

-17.80% for ‘women’ in seed sentence), highlighting its functional importance for this concept. The specificity of this effect was validated by the negligible impact of disrupting the control prototype, *L9 P2*, while disrupting the ‘male’ prototype *L9 P18* consistently *increased* the probability of female-coded words (e.g., $+11.50\%$ for ‘women’ in the seed sentence). These findings indicate that the model learns functionally distinct prototypes and uses them interactively to refine its predictions. This interactive behavior extends beyond gender concepts: we demonstrate similar necessity for geographic entities (e.g., ‘New Zealand’) and abstract states (e.g., ‘Mental Health’). Furthermore, cross-seeds experiments on ‘COVID’ confirm that the emergence of these concept-specific slots is a robust architectural property. Comprehensive results are detailed in the Appendix A.3.

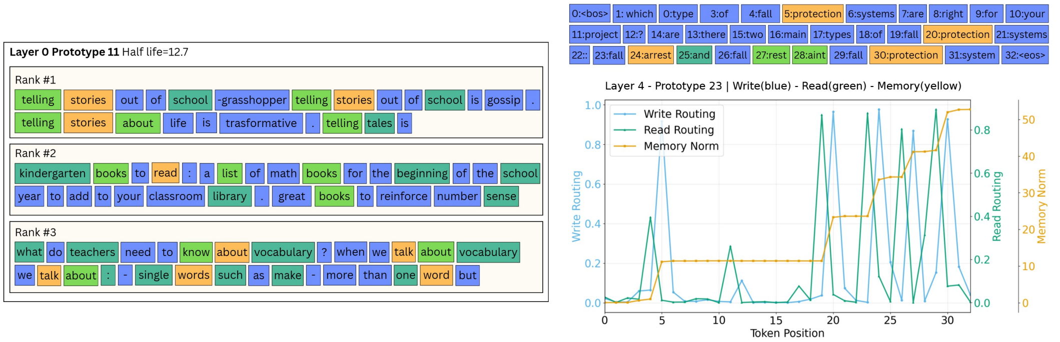


Figure 3: *Left*: Sequences most strongly activating prototype 11 at layer 0, which encodes the concept of narrative in a scholastic context. *Right*: Write-gate, read-gate, and memory curves for a sequence that strongly activates prototype 23 at layer 4. Read-gate peaks precede write-gate activations, spiking on the token immediately before those that trigger write-gate routing.

5.2 ROBUSTNESS

We analyze robustness of ProtoT vs baselines (the large-scale models from Table 1) from three complementary perspectives: (1) robustness to *meaning-preserving noise perturbations*, (2) robustness to *prototype clamping perturbations*, and (3) behavior to *interventions* that alter semantics. This unified view clarifies both stability under benign variations and sensitivity to intended changes.

Noise perturbations: We first consider *black-box, surface-level perturbations* that preserve meaning (e.g., synonyms, typos, contractions). The perturbation benchmark (Appendix A.5) contains 3,500 semantically equivalent sentence pairs across seven categories. Robustness is quantified by the Jensen–Shannon divergence $JS(p(\cdot|x), p(\cdot|x'))$ between next-token distributions for an original input x and its perturbed variant x'. Lower values indicate greater stability. Table 3 shows that Mamba has the overall lowest JS, hence the strongest stability. ProtoT, however, consistently outperforms LLaMA on synonyms, typos, spelling, and morphology. This aligns with ProtoT’s design: prototypes aggregate contextual information into nameable concepts, yielding stability under lexical variation. While ProtoT lags LLaMA on punctuation (where precise attention alignment is beneficial), it reliably surpasses DeltaNet and is overall competitive with strong baselines.

Table 3: Slice-level robustness measured by Jensen–Shannon divergence (lower is better). Abbreviations: *abbr.*=abbreviation, *contr.*=contraction, *morph.*=morphology, *punct.*=punctuation, *spell.*=spelling, *syn.*=synonym, *typo*=typos. Best per column in bold.

Model	**abbr.**	**contr.**	**morph.**	**punct.**	**spell.**	**syn.**	**typo**
DeltaNet	1.0657	0.8310	0.6671	0.5804	0.3547	0.6363	0.6257
LLaMA	0.3325	0.0449	0.2267	**0.1740**	0.0634	0.1450	0.2269
Mamba	**0.1441**	**0.0104**	**0.0476**	0.4428	**0.0054**	**0.0130**	**0.0761**
ProtoAttn	0.4166	0.0823	0.0498	0.3982	0.0260	0.1132	0.2074

Prototype clamping: To test whether robustness is mediated by prototype routing, we compute *Prototype-Mediated Robustness (PMR)*. For a pair (x, x'), let $JS_{\text{base}} = JS(p(\cdot|x), p(\cdot|x'))$. We then clamp the prototype routing weights from x onto x' and recompute $JS_{\text{clamped}} = JS(p(\cdot|x), p^{\text{clamped}}(\cdot|x'))$. We define $PMR = (JS_{\text{base}} - JS_{\text{clamped}})/JS_{\text{base}}$. A positive PMR

indicates that prototypes mediate robustness, while negative values suggest residual pathways dominate. Table 4 shows that while the mean PMR is sometimes slightly negative, for 5 out of 7 slices the positive fraction $PMR_{>0}$ is around 0.5–0.6 and $JS_{\text{clamped}} < JS_{\text{base}}$. This shows that prototypes overall contribute to robustness, providing interpretable routing pathways rather than opaque head-level aggregation.

Table 4: Prototype-Mediated Robustness (PMR). Mean and std of $\mathbf{PMR}$, fraction of positive cases, and average JSDs. Best per column in bold.

Slice	**PMR**$_{\text{mean}}$	**PMR**$_{\text{std}}$	**PMR**$_{>0}$	**JS**$_{\text{base}}$	**JS**$_{\text{clamped}}$	**n**
abbreviation	-0.093	0.367	0.596	0.417	0.415	500
contraction	-0.027	0.104	0.330	0.082	0.083	500
morphology	-0.034	0.176	0.474	0.050	0.051	500
punctuation	-0.000	0.373	0.554	0.398	0.322	500
spelling	-0.033	0.225	**0.610**	**0.026**	**0.025**	500
synonym	**0.013**	**0.075**	0.606	0.113	0.109	500
typo	0.001	0.279	0.533	0.208	0.186	500

Intervention behavior: Finally, we study sensitivity under *interventions* that alter semantics: gender, negation, and number tags. Unlike surface perturbations, these flips should change predictions. We measure JS, top-k overlap (Ov), Spearman correlation (Sp), and top-1 invariance (T1). Higher JS and lower Ov/Sp/T1 indicate greater sensitivity to the intervention. Table 5 shows that while DeltaNet attains the highest raw JS, ProtoT consistently yields lower Ov, Sp, and T1 compared to LLaMA and Mamba. This indicates that ProtoT adapts more reliably under meaning-altering interventions, reflecting appropriate semantic sensitivity through prototype routing. LLaMA and Mamba often remain insensitive to such tags.

Table 5: Intervention robustness on gender (gen), negation (neg), and number (num). Metrics: JS (higher better), Ov/Sp/T1 (lower better). Best values in bold.

Model	**JS** (gen / neg / num)	**Ov** (gen / neg / num)	**Sp** (gen / neg / num)	**T1** (gen / neg / num)
DeltaNet	**0.054** / **0.173** / **0.282**	0.754 / **0.540** / **0.474**	0.610 / **0.176** / **0.033**	**0.616** / **0.388** / **0.330**
LLaMA	0.004 / 0.028 / 0.022	0.946 / 0.875 / 0.843	0.966 / 0.815 / 0.824	0.890 / 0.770 / 0.930
Mamba	0.003 / 0.006 / 0.007	0.936 / 0.935 / 0.907	0.949 / 0.910 / 0.907	0.884 / 0.992 / 0.948
ProtoT	0.037 / 0.081 / 0.083	**0.709** / 0.774 / 0.657	**0.429** / 0.536 / 0.441	0.690 / 0.806 / 0.806

In conclusion, noise perturbation results establish that ProtoT is robust to lexical variation. PMR results show that prototypes actively mediate robustness, exposing interpretable mechanisms. Intervention behavior confirms that ProtoT is more sensitive to meaning-altering changes than Mamba or LLaMA for example. Together, these findings show that ProtoT not only matches or surpasses baselines in robustness but also provides transparent pathways for analyzing where robustness arises.

6 CONCLUSION

We have introduced the Prototype Transformer (ProtoT), an alternative autoregressive language model architecture that replaces standard self-attention mechanisms with prototype-based mixer to enhance model interpretability. Through bidirectional communication between learnable prototype vectors and input sequences, ProtoT demonstrates that architectural design choices can support interpretability with only small compromise in performance. It exhibits strong robustness, better text generation than self-attention, while downstream performance (GLUE) is on par with linear-compute baselines. Furthermore, prototypes automatically learn coherent, nameable concepts during training. The architecture also provides practical advantages through linear computational complexity and enables strong attribution of predictions to specific conceptual components and targeted editability.

Future work will further explore the scope and boundaries of this approach, including broader evaluation across diverse tasks and model scales. In summary, our results show that incorporating interpretability considerations into architectural design may be compatible with maintaining competitive performance. ProtoT contributes to ongoing research toward developing LMs that balance capability with transparency for applications where understanding and correcting model reasoning is essential.

7 REPRODUCIBILITY STATEMENT

We provide full details of the model architecture, training setup, and evaluation protocols in the main paper and appendix. The perturbation benchmark dataset (`perturbation_benchmark.jsonl`), along with its generation and filtering scripts, is included in the supplementary material and will be released publicly upon acceptance. In addition, we introduce a manually constructed intervention benchmark dataset (`intervention_benchmark.jsonl`), which tests semantic interventions on gender, negation, and number. Since the dataset was curated directly rather than generated by scripts, we will release it in full to ensure exact reproducibility of the intervention robustness experiments. We also include the interactive html (`prototype_visualization_word_level.html`). All code to reproduce our experiments will likewise be made available upon acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sercan O Arik and Tomas Pfister. Protoattend: Attention-based prototypical learning. *Journal of Machine Learning Research*, 21(210):1–35, 2020.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W. Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. *arXiv preprint arXiv:2003.04887*, 2020.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning, 2023.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? An analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Curtis Hawthorne, Andrew Jaegle, Catalin Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, et al. General-purpose, long-context autoregressive modeling with Perceiver AR. In *International Conference on Machine Learning*, 2022.

Dat Hong, Tong Wang, and Stephen Baek. Protorynet-interpretable text classification via prototype trajectories. *Journal of Machine Learning Research*, 24(264):1–39, 2023.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Machine Learning*, 2022.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

Dayal Singh Kalra and Maissam Barkeshli. Why warmup the learning rate? underlying mechanisms and improvements. *Advances in Neural Information Processing Systems*, 37, 2024.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Shuaipeng Li, Penghao Zhao, Hailin Zhang, Xingwu Sun, Hao Wu, Dian Jiao, Weiyan Wang, Chengjun Liu, Zheng Fang, Jinbao Xue, Yangyu Tao, Bin Cui, and Di Wang. Surge phenomenon in optimal learning rate and batch size scaling. *Advances in Neural Information Processing Systems*, 37, 2024.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. Interpretable image classification with adaptive prototype-based vision transformers. *Advances in Neural Information Processing Systems*, 37, 2024.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

George A Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically Interpreting Millions of Features in Large Language Models. *arXiv preprint arXiv:2410.13928*, 2024.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37, 2024.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 2017.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Dawid Rymarczyk, Łukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor, and Bartosz Zieliński. Interpretable image classification with differentiable prototypes assignment. In *European Conference on Computer Vision*, 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020.

Bowen Wei and Ziwei Zhu. Protolens: Advancing prototype learning for fine-grained interpretability in text classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Ximing Wen and Rezvaneh Rezapour. A transformer and prototype-based interpretable model for contextual sarcasm detection. *arXiv preprint arXiv:2503.11838*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in Neural Information Processing Systems*, 37, 2024.

Chris Ying, Sameer Kumar, Dehao Chen, Tao Wang, and Youlong Cheng. Image classification at supercomputer scale. In *Proceedings of the Systems for Machine Learning Workshop at NeurIPS 2018*, 2018.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems 32*, 2019.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, pp. 1–46, 2025.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for language models. *arXiv preprint arXiv:2407.07972*, 2024.

A APPENDIX

A.1 AI USAGE

We have used LLMs for proofreading the paper and to polish writing, for retrieval and discovery of related work, and for low-level coding help, e.g. to help us produce the prototype interpretability html. We have checked all AI output, and have verified that the resulting code is correct and works as expected.

A.2 ADDITIONAL DETAILS ON EXPERIMENT SETUP FOR LONG-CONTEXT SCALABILITY AND LARGE-SCALE TRAINING

Batch size: We have found that batch size of 32 works best for training among 32, 64, 128, for all models. Lower batch size values were not considered to preserve parallelisability and reduce number of training steps. We keep this batch size (32) in larger experiments as well, for simplicity, and only select the learning rate from a handful of scaling options. Furthermore, smaller batch sizes generalize better than large batch sizes even with large-scale data (Masters & Luschi, 2018); large batch sizes are mainly used for hardware utilization and training speed-up as they require fewer steps to finish training (Ying et al., 2018).

Learning rate: The best learning rates found via the automatic hyperparameter search for the default model sizes are: LLaMA: 1.6e-3, Mamba: 3.8e-3, DeltaNet: 6.8e-3, and ProtoT: 2.0e-3.

For the *long-context scalability experiment*, we have tried increasing the learning rate accordingly (by square root of context size ratio), as per AdamW scaling laws (Li et al., 2024), because extended context is computationally-similar to a larger batch size. However, we have found that scaling the learning rate helps only for DeltaNet and only in the large-scale model/data setting. In the results, we report only the best value from scaled vs non-scaled LR for all models.

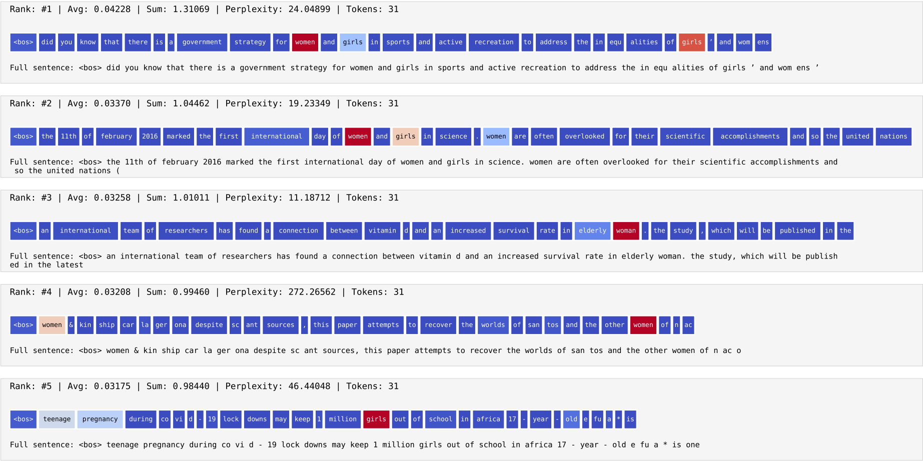
For the *large-scale training experiment*, we ran each model with the best hyp-s from the hyp search, and with scaled version thereof. We observed instability with Mamba, so we reduced the LR until it reached stability (from 3.8e-3 down to 2.3e-3). For all other models, we report results with the best-found learning rates (above).

A.3 PROTOTYPE INTERVENTION EXPERIMENTS

To move beyond correlational observations, we designed an intervention experiment to probe the functional role of individual prototypes within the model's predictive process. This methodology involves systematically manipulating a single prototype by either re-initializing it with random noise or zeroing-out the output of the write/read gate corresponding to the prototype, which is equivalent to zeroing-out/ablating the entire communication channel corresponding to it. We then measure the resulting impact on the model's output probabilities for a targeted linguistic task. By quantifying this change, we can assess the prototype's influence and determine its functional importance for a specific prediction.

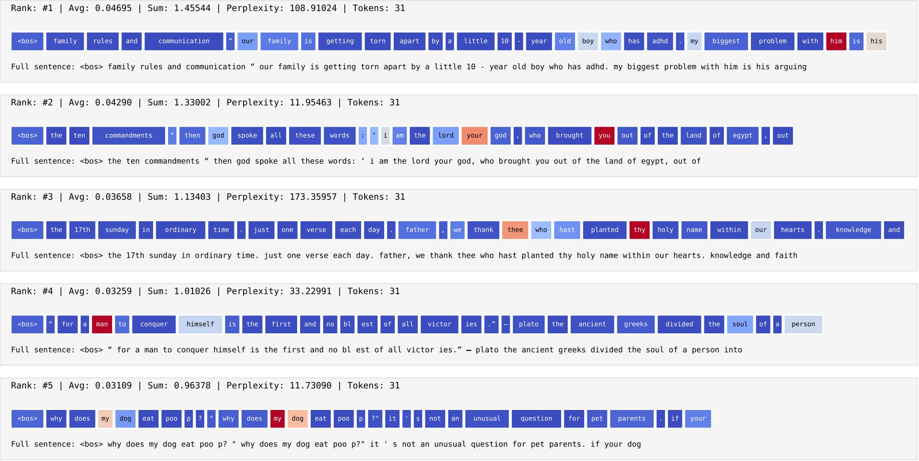
A.3.1 IDENTIFYING AND TARGETING CONCEPT-SPECIFIC PROTOTYPES

To identify prototypes that appear to encode distinct, human-understandable concepts, we analyze the top-activating sentences for each prototype from the visualization introduced in Sec. 5.1. Based on this analysis, we selected three prototypes from Layer 9 for our study. The prototype *L9 P7*(Fig. 4), which consistently activates on sentences containing words such as *'women'* and *'girls'*, we hypothesize that *L9 P7 is a key causal component in the model's representation of the 'female' concept*. Similarly, we identified prototype *L9 P18*(Fig. 5) as a representation for the 'male' concept, as it shows high activation for words like *'man'* and *'boy'*. Finally, prototype *L9 P2*(Fig. 6) was selected as a control, as it did not exhibit a clear, gender-coded semantic preference.



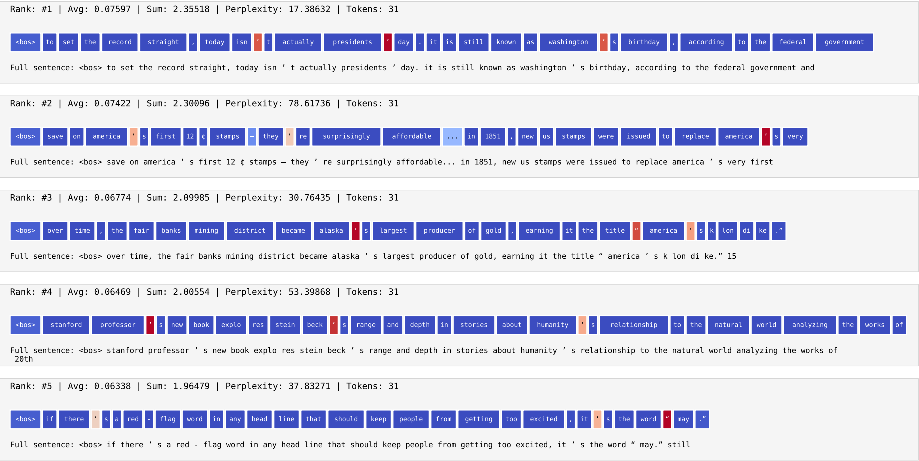
```
Rank: #1 | Avg: 0.04228 | Sum: 1.31069 | Perplexity: 24.04899 | Tokens: 31

<bos> did you know that there is a government strategy for women and girls in sports and active recreation to address the in equ alities of girls ’ and wom ens

Full sentence: <bos> did you know that there is a government strategy for women and girls in sports and active recreation to address the in equ alities of girls ’ and wom ens ’

Rank: #2 | Avg: 0.03370 | Sum: 1.04462 | Perplexity: 19.23349 | Tokens: 31

<bos> the 11th of february 2016 marked the first international day of women and girls in science . women are often overlooked for their scientific accomplishments and so the united nations

Full sentence: <bos> the 11th of february 2016 marked the first international day of women and girls in science. women are often overlooked for their scientific accomplishments and
 so the united nations (

Rank: #3 | Avg: 0.03258 | Sum: 1.01011 | Perplexity: 11.18712 | Tokens: 31

<bos> an international team of researchers has found a connection between vitamin d and an increased survival rate in elderly woman . the study , which will be published in the

Full sentence: <bos> an international team of researchers has found a connection between vitamin d and an increased survival rate in elderly woman. the study, which will be publish
ed in the latest

Rank: #4 | Avg: 0.03208 | Sum: 0.99460 | Perplexity: 272.26562 | Tokens: 31

<bos> women & kin ship car la ger ona despite sc ant sources , this paper attempts to recover the worlds of san tos and the other women of n ac

Full sentence: <bos> women & kin ship car la ger ona despite sc ant sources, this paper attempts to recover the worlds of san tos and the other women of n ac o

Rank: #5 | Avg: 0.03175 | Sum: 0.98440 | Perplexity: 46.44048 | Tokens: 31

<bos> teenage pregnancy during co vi d - 19 lock downs may keep 1 million girls out of school in africa 17 - year - old e fu a * is

Full sentence: <bos> teenage pregnancy during co vi d - 19 lock downs may keep 1 million girls out of school in africa 17 - year - old e fu a * is one
```

Figure 4: Visualization for prototype **L9 P7**



```
Rank: #1 | Avg: 0.04695 | Sum: 1.45544 | Perplexity: 108.91024 | Tokens: 31
Full sentence: <bos> family rules and communication “ our family is getting torn apart by a little 10 - year old boy who has adhd. my biggest problem with him is his arguing

Rank: #2 | Avg: 0.04290 | Sum: 1.33002 | Perplexity: 11.95463 | Tokens: 31
Full sentence: <bos> the ten commandments “ then god spoke all these words: ‘ i am the lord your god, who brought you out of the land of egypt, out of

Rank: #3 | Avg: 0.03658 | Sum: 1.13403 | Perplexity: 173.35957 | Tokens: 31
Full sentence: <bos> the 17th sunday in ordinary time. just one verse each day. father, we thank thee who hast planted thy holy name within our hearts. knowledge and faith

Rank: #4 | Avg: 0.03259 | Sum: 1.01026 | Perplexity: 33.22991 | Tokens: 31
Full sentence: <bos> “ for a man to conquer himself is the first and no bl est of all victor ies.” — plato the ancient greeks divided the soul of a person into

Rank: #5 | Avg: 0.03109 | Sum: 0.96378 | Perplexity: 11.73090 | Tokens: 31
Full sentence: <bos> why does my dog eat poo p? " why does my dog eat poo p?" it ' s not an unusual question for pet parents. if your dog
```

Figure 5: Visualization for prototype **L9 P18**



```
Rank: #1 | Avg: 0.07597 | Sum: 2.35518 | Perplexity: 17.38632 | Tokens: 31
Full sentence: <bos> to set the record straight, today isn ’ t actually presidents ’ day. it is still known as washington ’ s birthday, according to the federal government and

Rank: #2 | Avg: 0.07422 | Sum: 2.30096 | Perplexity: 78.61736 | Tokens: 31
Full sentence: <bos> save on america ’ s first 12 ¢ stamps — they ’ re surprisingly affordable... in 1851, new us stamps were issued to replace america ’ s very first

Rank: #3 | Avg: 0.06774 | Sum: 2.09985 | Perplexity: 30.76435 | Tokens: 31
Full sentence: <bos> over time, the fair banks mining district became alaska ’ s largest producer of gold, earning it the title “ america ’ s k lon di ke.” 15

Rank: #4 | Avg: 0.06469 | Sum: 2.00554 | Perplexity: 53.39868 | Tokens: 31
Full sentence: <bos> stanford professor ’ s new book explo res stein beck ’ s range and depth in stories about humanity ’ s relationship to the natural world analyzing the works of
 20th

Rank: #5 | Avg: 0.06338 | Sum: 1.96479 | Perplexity: 37.83271 | Tokens: 31
Full sentence: <bos> if there ’ s a red - flag word in any head line that should keep people from getting too excited, it ’ s the word “ may.” still
```

Figure 6: Visualization for the control prototype **L9 P2**

Test Case Construction. To create a controlled and relevant test set, we began with a seed sentence identified from our initial visualization analysis. This sentence was the top-ranked example from the FineWeb test set that maximally activated the ‘write‘ gate of our primary target, prototype L9 P7. To expand our test set while maintaining semantic consistency, we then prompted a large language model (Gemini 2.5 Pro) to generate six additional sentences thematically similar to the seed sentence, each required to contain the keywords ‘women‘ and ‘girls‘.

The resulting corpus of seven sentences used in our experiments is as follows:

- *“did you know that there is a government strategy for women and girls in sports and active recreation to address the inequalities of girls’ and women’s”* (seed sentence from FineWeb)
- *“Many organizations are working on programs that focus on empowering women and girls to participate equally in science and technology.”*
- *“Did you know that several global initiatives aim to protect the rights of women and girls from violence and discrimination?”*
- *“Education policies are increasingly emphasizing equal opportunities for women and girls to excel in leadership roles.”*
- *“Access to healthcare remains a critical issue, and governments are creating strategies to improve services for women and girls.”*

- *“International campaigns highlight how climate change disproportionately affects women and girls in vulnerable communities.”*
- *“Did you know that mentorship networks are being created to support women and girls in pursuing careers in engineering and mathematics?”*

From this corpus, we defined our test cases. Each case consists of a context (the sentence preceding a target word) and a completion token (the target word itself). For this study, we focused on the probability of the target completions ‘women’ and ‘girls’.

Results: After establishing a baseline probability for each test case using the unmodified model, we create a copy of the model for each intervention. The intervention method used is Disruption, where the parameter vector of the target prototype (L9 P7, L9 P18, or L9 P2) is re-initialized with random noise, scaled according to the model’s original initialization scheme. This procedure erases the prototype’s learned knowledge while preserving the overall model architecture. We then measure the post-intervention probability of the completion token.

The results of our intervention experiments are summarized in Table 6. To focus the analysis on contexts where the target word is considered a plausible completion by the model, we excluded test cases where the baseline probability of the target completion was below 1%.

Table 6: Comprehensive intervention results. We report the relative change in target probability (%) under three conditions: *Rnd* (Random Re-initialization), *Wr* (Write Gate Mask), and *Rd* (Read Gate Mask). *L9 P7* is the target ’female’ prototype; L9 P18 is the ’male’ prototype, and L9 P2 serve as control.

		L9 P7 (’female’)			*L9 P18 (’male’)*			*L9 P2 (Control)*		
Context (Truncated)	*Base(%)*	*Rnd*	*Wr*	*Rd*	*Rnd*	*Wr*	*Rd*	*Rnd*	*Wr*	*Rd*
Target: ‘women’										
`...inequalities of...`	3.21	**-17.80**	**-16.60**	+7.57	+11.50	+16.95	+12.71	+0.74	+2.30	-0.61
`...empowering...`	4.24	-3.00	+1.43	-3.06	-0.13	+7.09	-0.18	-0.17	+0.35	-0.33
`...protect rights of...`	13.54	+1.37	-2.29	+1.56	+1.43	-1.03	+1.64	+0.09	-0.22	+0.38
`...equal opps for...`	10.14	-0.67	+3.45	-0.76	-0.31	+2.20	-0.31	-0.75	+0.49	-1.08
`...climate affects...`	11.87	+1.81	-1.65	+1.86	+0.12	-2.61	+0.22	+0.34	-0.32	+0.58
Target: ‘girls’										
`...inequalities of...`	2.80	**-10.62**	**-10.67**	+5.49	+0.50	-1.57	+0.19	+0.03	-0.46	+0.03
`...empowering...`	68.55	+0.11	+0.39	+0.23	+0.28	+0.79	+0.39	-0.28	+0.02	-0.26
`...protect rights of...`	78.63	-0.45	-0.41	+0.76	+0.64	+0.70	+0.64	-0.04	-0.03	-0.13
`...equal opps for...`	60.49	-0.17	+0.01	+0.58	+0.56	+0.84	+0.65	-0.19	+0.02	-0.27
`...improve services...`	64.33	-1.56	-1.43	+0.35	+0.62	+0.94	+0.72	-0.15	+0.05	-0.19
`...climate affects...`	68.66	-1.01	-0.95	+0.72	+1.39	+1.50	+1.52	-0.10	-0.05	-0.21
`...support women in...`	38.32	-3.89	-3.51	+1.68	+2.39	+4.15	+2.64	-0.55	-0.11	-1.10

For prototype re-initialization, our results reveal a clear causal link between prototype *L9 P7* and the model’s representation of female-coded concepts. Disrupting this ’female’ prototype significantly *decreased* the probability of target words like ‘women’ (-17.80%) and ‘girls’ (-10.62%), particularly in less constrained contexts. This effect, however, diminished in test cases where the baseline probability was already very high (e.g., $> 60\%$), suggesting that highly predictable completions are more robust and less reliant on any single prototype. The specificity of this function was confirmed by a control experiment where disrupting an unrelated prototype, *L9 P2*, yielded only negligible changes, proving our findings are not artifacts of random model perturbations. Furthermore, the interventions uncovered a more sophisticated dynamic: disrupting the ’male’ prototype, *L9 P18*, consistently *increased* the probability of female-coded words. This suggests an inhibitory or competitive relationship, where the model refines its predictions by balancing between opposing semantic concepts. Taken together, these results demonstrate that the model utilizes specific, functionally distinct, and interacting prototypes to represent and manipulate complex concepts like gender.

For the gate-specific interventions, our results demonstrate that zeroing out the Write or Read gates provides a more rigorous measure of causal influence. By acting as a deterministic ablation rather than a stochastic disruption, Write Gate masking revealed a significantly sharper functional contrast between the opposing gender concepts. As shown in Table 6, the divergence between the inhibitory effect of the 'male' prototype (L9 P18) and the causal necessity of the 'female' prototype (L9 P7) was markedly amplified under the masking condition. Specifically, in contexts such as *"...inequalities of..."*, the gap between the probability surge caused by masking the 'male' prototype (+16.95%) and the drop caused by masking the 'female' prototype (-16.60%) was substantially wider ($\Delta \approx 33.6\%$) compared to the spread observed under random disruption ($\Delta \approx 29.3\%$). These findings suggest that strict channel ablation effectively isolates the distinct semantic mechanisms (whether competitive or constructive) that prototypes engage in, with the Write Gate often serving as the primary causal bottleneck for concept storage.

A.3.2 ROBUSTNESS OF CONCEPT-SPECIFIC PROTOTYPES EMERGENCE ACROSS RANDOM SEEDS

To ensure that the localization of semantic concepts is a robust property of the architecture rather than an artifact of a specific initialization, we extended our analysis by training two additional models with different random seeds. We repeated the visualization process described in Sec. 5.1 for these new runs to observe if similar semantic clusters emerged. We focus on the concept of *'COVID-19'* as a representative case study. In all three models (the original and two replicates), we successfully identified a distinct prototype that was maximally activated by terms related to the pandemic.

To validate the functional consistency of these re-emerged prototypes, we conducted intervention experiments targeting the prediction of the token *'COVID'* in relevant contexts. For each model, we disrupted the identified COVID-specific prototype via random re-initialization, write gate masking and read gate masking. We followed the same workflow as in Sec. A.3.1. The test corpus for the COVID-specific prototype consisted of the following sentences:

- *"covid - 19 lambda variant lambda variant cases of covid - 19 are emerging in the us. while nowhere near"*
- *"The World Health Organization declared the outbreak of COVID-19 a pandemic in March 2020."*
- *"Researchers identified the Alpha, Beta, Gamma, and Delta strains as variants of concern for COVID-19."*
- *"The Pfizer-BioNTech and Moderna vaccines use mRNA technology to protect against the COVID-19 virus."*
- *"Anosmia, the sudden loss of smell and taste, was identified as a specific symptom of COVID-19 infection."*
- *"To curb the spread, the government mandated a 14-day quarantine for anyone testing positive for COVID-19."*
- *"Hospitals faced a critical shortage of ventilators during the initial surge of severe COVID-19 cases."*
- *"The FDA granted emergency use authorization for Paxlovid, an oral antiviral pill for treating COVID-19."*
- *"Scientists continue to debate the zoonotic origins of COVID-19 and its potential transmission from bats."*
- *"Despite strict border controls, the Omicron variant of COVID-19 managed to spread rapidly across the globe."*
- *"Long-haulers are patients who suffer from debilitating symptoms months after recovering from acute COVID-19."*
- *"Public health officials urged the population to wear N95 masks to prevent the airborne transmission of COVID-19."*
- *"The CDC updated its guidelines regarding the isolation period for asymptomatic cases of COVID-19."*
- *"Herd immunity against COVID-19 became difficult to achieve due to the emergence of new escape variants."*

- *“Schools implemented social distancing and improved ventilation to reduce the risk of COVID-19 transmission in classrooms.”*
- *“The economic fallout from the COVID-19 pandemic led to supply chain disruptions and rising inflation.”*
- *“A negative PCR test result for COVID-19 was required for all passengers boarding international flights.”*
- *“Studies suggest that previous infection provides some level of natural immunity against reinfection with COVID-19.”*
- *“The global death toll attributed to COVID-19 has highlighted the vulnerabilities in healthcare systems worldwide.”*
- *“Contact tracing apps were deployed to alert individuals who had been exposed to a confirmed case of COVID-19.”*
- *“Rehabilitation programs are being established to help patients recover from the respiratory damage caused by severe COVID-19.”*

Table 7: *Cross-seed consistency of the ’COVID’ prototype.* We compare the impact of interventions across three different model initializations. Cells with ‘-’ indicate that the baseline probability was below the 1% threshold for reliability. All values represent probabilities or probability changes (in percentage points).

	Original Model (L1 P14)				**Seed 124 (L7 P29)**				**Seed 325 (L6 P31)**			
Context (Truncated)	*Base*	*Rnd*	*Wr*	*Rd*	*Base*	*Rnd*	*Wr*	*Rd*	*Base*	*Rnd*	*Wr*	*Rd*
`covid - 19 lambda variant...`	12.4	+0.3	+0.5	-0.0	48.7	-8.6	-8.4	+0.1	27.4	+4.6	+6.3	-5.7
`The World Health Organization...`	6.4	-1.5	-1.7	+0.2	16.2	-0.4	+0.1	-0.1	9.7	+12.6	+13.0	+13.1
`The Pfizer-BioNTech and Moderna...`	0.6	-	-	-	0.9	-	-	-	1.4	+0.4	-0.4	+0.1
`To curb the spread,...`	10.4	-0.1	-0.6	+0.0	13.3	-0.9	-0.9	+0.4	17.3	-0.5	+0.3	-0.4
`Hospitals faced a critical...`	0.4	-	-	-	1.4	+1.7	+2.0	-0.1	2.4	+23.3	+27.3	+22.1
`The FDA granted emergency...`	0.7	-	-	-	3.4	+0.8	+0.8	-0.1	1.1	+9.3	+5.9	-3.8
`Despite strict border controls,...`	0.7	-	-	-	2.2	-3.7	-3.6	+0.0	2.4	+2.8	+2.6	+2.2
`Public health officials urged...`	6.0	+0.2	-0.2	-0.1	3.3	+0.5	+0.5	-0.0	12.7	-2.7	-8.1	-2.9
`The CDC updated its...`	2.5	-0.3	-0.1	+0.1	3.4	-0.9	-0.5	+0.1	3.5	+3.4	-10.8	+5.7
`Herd immunity against COVID...`	7.9	+0.2	+0.4	+0.0	5.4	-0.1	+0.0	+0.0	6.5	+1.4	-1.8	+1.6
`Schools implemented social...`	1.8	-0.4	-0.4	+0.1	2.3	+0.5	+0.9	-0.0	4.0	+1.2	-0.9	+1.3
`The economic fallout from...`	2.4	+2.0	+2.3	-0.2	2.0	+0.2	-0.0	-0.0	1.5	-4.9	+3.1	-5.1
`Studies suggest that previous...`	1.3	+0.8	+0.9	-0.1	1.0	-0.1	+0.2	+0.1	2.5	+0.0	-2.9	+0.3
`The global death toll...`	2.7	+0.3	+0.3	+0.0	2.9	+0.1	-0.0	+0.0	1.6	-6.5	-13.4	-6.8
`Contact tracing apps were...`	2.3	+0.4	+0.4	-0.1	4.0	-0.2	+0.2	+0.1	5.1	+0.3	-0.3	+1.6
`Rehabilitation programs are...`	0.2	-	-	-	0.9	-	-	-	2.9	+3.9	-22.0	+4.0

Results: Our analysis across different random seeds shows that the ProtoT architecture consistently localizes concepts within specific semantic regions, demonstrating strong robustness. However, the specific functional mechanisms used by these prototypes can differ significantly between model initializations. We acknowledge that the intervention effects for the COVID-19 prototype appear more variable than those observed in the gender experiments. This difference comes in part from the way the concepts are structured. Gender is modeled as a clear binary contrast (Male vs. Female), which allows relative comparisons, while COVID-19 is a single concept – it depends only on changes from a baseline probability and has no direct opposite reference. Despite this lack of contrastive referencing, the most significant finding is the robust emergence of the concept itself: across all three random seeds, the model consistently allocated a dedicated prototype slot to encode pandemic- or disease-related knowledge without explicit supervision.

A.3.3 EXTENDED INTERVENTION ANALYSIS ON DIVERSE CONCEPTS

To demonstrate that the functional localization of semantic concepts is a general property of the ProtoT architecture and not limited to the social (Gender) or event-specific (COVID-19) cases discussed in the above, we extended our discovery pipeline to additional semantic domains. Here, we present intervention results for geographic entity (*“e.g. New Zealand”*) and an abstract state (*“e.g. Mental Health”*).

The test corpus for the ’New Zealand’ prototype consisted of the following sentences:

- *"The Dutch explorer Abel Tasman named New Zealand as Nova Zeelandia after the Dutch province of Zeeland"*
- *"The Treaty of Waitangi signed in 1840 was instrumental in establishing British sovereignty over New Zealand"*
- *"Regular Quaker meetings began in Nelson in 1842 and later spread across New Zealand"*
- *"The first Quaker to visit Aotearoa / New Zealand was Sydney Parkinson"*

The test corpus for the 'Mental Health' prototype consisted of the following sentences:

- *"Meditation and mindfulness practices are beneficial for maintaining mental clarity."*
- *"Regular exercise can improve mental well-being and reduce symptoms of anxiety."*
- *"She sought professional help to manage her mental stress during the exam period."*
- *"Many people face mental health challenges but do not seek support due to stigma."*

We identified a prototype in L5 P9 that maximally activated for contexts related to the country including *New Zealand*. To validate its causal role, we constructed a test set containing historical and geographical facts. We then measured the impact of masking the prototype to see the change on the prediction of the target token *'zealand'*.

The results (Table 8) show striking causal efficacy. For instance, in the context of *"...abel tasman named new [zealand]"*, masking the write gate caused a massive probability drop of 21.54%. Similarly, references to the *Treaty of Waitangi* saw a 12.52% drop. This confirms that this specific slot (L5 P9) is critical for storing and retrieving knowledge specific to this geographic entity.

Table 8: Intervention results for the 'New Zealand' prototype (L5 P9). We report the relative change in probability for the target token 'zealand' when masking the Write Gate (*Wr*), Read Gate (*Rd*), or applying Random Noise to the Prototype (*Rnd*).

		Relative Change (%)		
Context (Truncated)	*Baseline*	*Wr*	*Rd*	*Rnd*
`...explorer abel tasman named new...`	35.7%	**-21.5**	-21.0	-20.7
`...treaty of waitangi signed in 1840...`	91.9%	**-12.5**	-12.9	-12.4
`...regular quaker meetings began in...`	33.4%	-2.3	-1.2	-1.1
`...first quaker to visit aotearoa...`	99.7%	-0.2	-0.2	-0.2

Moving beyond concrete entities, we investigated whether abstract concepts are similarly localized. We identified a prototype in L6 P9 responsive to *Mental Health* and communication. We tested this using sentences involving psychological states and well-being, targeting the token *'mental'*.

As shown in Table 9, while the baseline probabilities for this abstract adjective are generally lower than for proper nouns, intervention still yields consistent causal effects, further supporting the functional diversity discussed in the main text.

Table 9: Intervention results for the 'Mental Health' prototype (L6 P9). Target token: 'mental'.

		Relative Change (%)		
Context (Truncated)	*Baseline*	*Wr*	*Rd*	*Rnd*
`...mindfulness practices are beneficial...`	5.0%	**-2.2**	-2.3	-1.6
`...exercise can improve mental...`	2.1%	-0.4	-0.7	-0.6
`...sought professional help to manage...`	1.4%	-0.2	-0.2	-0.2
`...face mental health challenges...`	1.3%	-0.1	-0.0	-0.0

A.4 DOWNSTREAM (DETAILS)

We provide the training protocol and hyperparameter configuration used for the GLUE downstream experiments, covering datasets and splits, pre-processing, optimization, early-stopping/selection on dev, and the hyperparameter sweep and choice rules, and we additionally report the corresponding GLUE dev-set results for completeness.

Training protocol: We evaluate four language model architectures: ProtoT, LLaMA, Mamba, and DeltaNet, on the GLUE benchmark under a unified experimental protocol to ensure fair comparison. Unless stated otherwise, all details follow the Experimental Setup 4. Inputs are formed as single-sentence or sentence-pair prompts according to the task, with a maximum sequence length of 512. To avoid leakage, we fine-tune on the official training split, select hyperparameters and checkpoints on the official development split using early stopping, and export test predictions in the official TSV format for submission to the GLUE server. We follow the official GLUE metrics: accuracy for SST-2, QNLI, MNLI, QQP, RTE, and WNLI (or the primary metric reported by the official script), the accuracy and F1 pair for MRPC and QQP, Matthews correlation for CoLA, and Pearson and Spearman correlations for STS-B.

Optimization and regularization are aligned across models. We use the AdamW optimizer together with a linear learning-rate schedule with warmup. We apply selective weight decay consistent with pre-training: decay is applied to affine weights that benefit from it, while embeddings, normalization layers, and biases receive no decay. For GLUE downstream fine-tuning, we use a batch size of 16 for all models. Compared to pre-training, the GLUE datasets are much smaller, so we prefer a moderately small batch size that provides more stochasticity in the updates and typically leads to better generalization in low-data regimes. Fine-tuning runs for up to 3 epochs with early stopping on dev, and the dev-best checkpoint is used to generate test predictions. Unless otherwise specified, a fixed random seed is used across tasks and models to support reproducibility.

Hyperparameter selection: Because architectures differ in optimization sensitivity, we conduct per-model hyperparameter selection. For each model we run small grid searches on two representative tasks, SST-2 (medium-scale binary classification) and MNLI (large-scale multi-class classification). We sweep learning rates over a grid that includes 2.5e-5, 3.5e-5, 5.5e-5, 1e-4, 2e-4, 3e-4, 4e-4, 5e-4, 7e-4, 8.5e-4, 1e-3, and we sweep warmup ratios over 6% and 10%. The best learning rate and warmup found per model on these representative tasks are then fixed for that model across the remaining GLUE tasks, where "best" is defined as the (learning-rate, warmup) configuration that achieves the highest average dev performance over SST-2 and MNLI. All other training details, such as batch size, maximum length, optimizer settings, and early-stopping criterion, remain identical across models.

The final per-model settings in our environment are as follows. PrototypeAttn uses a learning rate of 3.5e-5 with 6% warmup. LLaMA uses a learning rate of 5.5e-5 with 10% warmup. Mamba uses a learning rate of 1e-4 with 10% warmup. DeltaNet uses a learning rate of 7e-4 with 10% warmup.

GLUE dev downstream fine-tuning results: For completeness, we also report the GLUE dev set results in Table 10, using the same evaluation metrics as in the main text (Matthews correlation for CoLA, accuracy for SST-2/RTE/WNLI/QNLI/MNLI/MNLI-MM, F1 for MRPC and QQP, and Pearson correlation for STS-B). These dev numbers were used during model development and are largely consistent with the test-set trends in the results in Section 5, Table 2: LLaMA achieves the strongest overall performance, while ProtoT remains competitive with dense baselines and shows robust behavior across multiple tasks.

Table 10: GLUE dev downstream fine-tuning results (all metrics reported as percentages). For COLA we report Matthews correlation; for SST-2 accuracy; for MRPC F1; for STS-B Pearson correlation; for RTE, WNLI, QNLI, MNLI and MNLI-MM accuracy; for QQP F1. Results are averaged over 3 seeds. Best results are in bold.

Model	COLA	SST-2	MRPC	STS-B	RTE	WNLI	QQP	QNLI	MNLI	MNLI-MM
LLaMA	**36.4**	**91.2**	**84.9**	**85.8**	**60.1**	**56.3**	**85.9**	**86.4**	**80.5**	**79.7**
Mamba	30.6	89.1	82.3	79.4	55.8	**56.3**	82.7	82.2	75.1	74.9
DeltaNet	8.9	85.5	81.4	75.7	56.1	**56.3**	81.2	80.1	71.6	72.5
ProtoT	29.3	89.8	81.7	73.3	54.5	54.9	83.6	82.6	75.4	76.0

A.5 ROBUSTNESS (DETAILS)

This section details the perturbation set, construction pipeline, and slice-level statistics for the black-box robustness experiments.

A.5.1 PERTURBATION DATASET CONSTRUCTION

We construct a dedicated perturbation benchmark with seven categories of *meaning-preserving* surface noise, 500 pairs per category (3,500 total). Source sentences are sampled from three public corpora under simple length and formatting constraints: WikiText-2 (Merity et al., 2016), DailyDialog (Li et al., 2017), and AG News (Zhang et al., 2015).

The final slices are:

- **Synonyms:** Replacements derived from WordNet (Miller, 1995). We select candidate lemmas that differ from the original token, avoid multi-word expressions, and have similar length. The generator enforces a mix of 1/2/3 substitutions per sentence. This slice is further filtered using Sentence-BERT (all-MiniLM-L6-v2; (Reimers & Gurevych, 2019; Wang et al., 2020)) and lexical heuristics (see below).
- **Typos:** Single-character keyboard noise applied to one token (internal character substitution), ensuring short, localized corruption.
- **Spelling variants:** American ↔ British spelling changes (e.g., *color* → *colour*, *organize* → *organise*) using a fixed mapping and a curated example pool.
- **Morphological variants:** Simple inflection changes (e.g., singular → plural) using rule-based morphology patterns plus a curated pool.
- **Contractions/Expansions:** Deterministic mappings between standard and contracted forms (e.g., *do not* → *don't*, *it is* → *it's*).
- **Punctuation/Casing:** Insertion or removal of punctuation (e.g., sentence-final periods or comma adjustments), while keeping word order intact.
- **Abbreviations/Short forms:** Systematic long-form → abbreviation mappings (e.g., *Doctor* → *Dr.*, *United States* → *U.S.*).

After generating the dataset, a separate cleaning and rebalancing script was run:

(i) filter by Sentence-BERT similarity (≥ 0.75),
(ii) apply lexical checks (rare tokens, casing),
(iii) backfill missing items from curated fallback pools, and
(iv) sample exactly 500 pairs per slice.

Table 11: Example sentence pairs from the perturbation benchmark.

Category	**Original**	**Perturbed**
Abbreviation	*Doctor Smith arrived.*	*Dr. Smith arrived.*
Contraction	*I cannot go.*	*I can't go.*
Synonym	*He was happy.*	*He was glad.*
Spelling	*I like this color.*	*I like this colour.*

A.5.2 VARIANCE STATISTICS

To characterize the perturbation strength of each slice, we compute: (i) cosine similarity of all-MiniLM-L6-v2 embeddings between original and perturbed sentences, (ii) character-level Levenshtein distance (Levenshtein, 1966).

Table 12 reports the per-slice averages. As expected, some categories (e.g., typos) introduce very small character edits but can cause non-trivial distributional shifts, while others (e.g., synonyms, abbreviations) involve larger form changes yet maintain high semantic similarity.

Table 12: Variance statistics for the perturbation benchmark (3,500 pairs total).

	Synonym	**Typo**	**Spelling**	**Morphology**	**Contraction**	**Punctuation**	**Abbreviation**
Avg. Similarity	0.828	0.775	0.956	0.881	0.895	0.983	0.894
Avg. Edit Distance	5.89	1.03	1.20	1.00	2.54	1.09	7.23

Table 13: Training throughput (it/s; higher is better) and elapsed time (s; lower is better) for matched-depth/width models at seq. len. 256 (BF16). FLOPs are reported in units of $\times 10^5$ (forward+backward). *When compilation was unavailable, values reflect the fastest steady-state runs without compilation.*

Model	**Batch**	**it/s**	**Elapsed (s)**	**FLOPs/sample** ($\times 10^5$)	**Total FLOPs** ($\times 10^5$)	**Params**
ProtoT	32	25.2	34.57	41,583.0	1,330,657.0	12,205,266
ProtoT	128	7.6	31.32	41,583.0	5,322,625.2	12,205,266
Mamba	32	11.9	58.17	**34,734.9**	**1,111,517.4**	**6,724,352**
Mamba	128	3.2	54.26	**34,734.9**	4,446,069.4	**6,724,352**
DeltaNet	32	3.5	222.88	—	—	12,963,456
DeltaNet	128	1.8	182.06	—	—	12,963,456
LLaMA	32	**55.1**	26.16	49,341.5	1,578,929.0	12,938,496
LLaMA	128	23.6	**22.30**	49,341.5	6,315,714.3	12,938,496

A.6 THROUGHPUT BENCHMARKS (PROTOT, MAMBA, LLAMA, DELTANET)

We evaluate under identical conditions: same data pipeline, optimizer, precision (BF16), sequence length 256, and batch sizes 32 and 128. FLOP counts are per-sample (forward+backward) where obtainable. *Observations*: Table 13 summarizes training throughput at batch sizes 32 and 128 for matched-depth/width models. LLaMA attains the highest throughput overall (**55.1** and **23.6** it/s). ProtoT sustains **25.2** and **7.6** it/s and is $\sim$2.1–2.4× faster than Mamba (11.9 and 3.2 it/s) at the same backbone. The FLA-based DeltaNet baseline, evaluated without fused delta kernels and with `torch.compile` disabled, reaches 3.5 and 1.8 it/s (batch 32/128).

A.6.1 LONG-CONTEXT THROUGHPUT

Throughput Evaluation Methodology To evaluate the computational efficiency of the models at varying context lengths, we measured the processing throughput on a single NVIDIA A100 80GB GPU. The benchmark measured the number of forward pass iterations per second (it/s) for a batch size of 1 across context lengths ranging from 2,048 to 131,072 tokens.

For a fair comparison, all models were run in standard PyTorch eager mode without `torch.compile` optimization. This ensures that the results reflect the raw architectural performance characteristics rather than compiler-specific optimizations which may vary in maturity across different architectures.

Throughput Results Table 14 shows the long-context inference throughput results. LLaMA achieves the highest throughput at short context lengths. However, ProtoT scales better as context increases, surpassing LLaMA at 32k tokens and above. DeltaNet maintains the highest throughput at long context lengths.

A.7 ABLATIONS

A.7.1 LAYER-0 ROUTING ABLATIONS

We ablate the three mitigations that stabilize the layer-0 router: (i) sharing the write/read routing distribution, (ii) sharpening the initial temperature ($\tau_0 = 3.0$), and (iii) adding a $k = 5$ depth-wise convolution to the write-value path of layers 0–1. Each configuration fine-tunes a 6-layer ProtoT on the FineWeb 18k/4k split (sequence length 256, seed 0) for three epochs, using the same optimizer,

Table 14: Long-context inference throughput (iterations per second; higher is better). Measured on a single NVIDIA H100 80GB GPU, batch size 1, PyTorch eager mode without `torch.compile`.

Context	**LLaMA**	**ProtoT**	**Mamba**	**DeltaNet**
2,048	100.47	38.74	35.50	43.54
4,096	47.90	20.61	19.04	42.35
8,192	21.72	10.82	9.89	44.89
16,384	8.08	5.49	5.05	40.20
32,768	2.61	2.78	2.55	27.78
65,536	0.74	1.40	1.28	17.34
131,072	0.20	0.65	0.57	9.41

tokenizer, and learning rate as the main experiments. We report best validation perplexity alongside routing diagnostics logged on the dev set.

Table 15: Layer-0 routing ablations on FineWeb. Metrics come from the final validation epoch (`val_router_stats.csv`) and the best dev perplexity tracked during training. Lower perplexity, Gini, and top-1 probability imply healthier routing; higher $\bar{\alpha}_0$ indicates that the ReZero gate remains active. Best values are in bold.

Variant	**Shared L_0**	**τ_0 init**	**Write conv**	**Best val ppl $\downarrow$**	**$\bar{\alpha}_0$ $\uparrow$**	**Gini $\downarrow$**	**top-1 $\downarrow$**
All mitigations (baseline)	On	3.0	$k = 5$	**133.3**	**0.672**	**0.034**	**0.079**
No shared routing	Off	3.0	$k = 5$	133.4	0.658	0.064	0.082
τ reset to 1.0	On	1.0	$k = 5$	133.6	0.653	0.035	0.088
No write conv	On	3.0	Off	145.7	0.354	0.097	0.177
All mitigations off	Off	1.0	Off	149.9	0.261	0.243	0.373

The convolution contributes most to stability: removing it roughly doubles the router concentration (top-1 rises from 0.079 to 0.177), increases hub inequality, and halves the layer-0 ReZero gate, ultimately worsening perplexity by $+12.4$ points. Shared routing and the sharpened τ_0 have smaller individual effects on perplexity, but together they keep hub utilisation uniform (gini 0.034) while allowing the gate to stay near its baseline value. Disabling every mitigation reproduces the original alpha-collapse, dropping $\bar{\alpha}_0$ to 0.261 and letting a single hub monopolise 37% of the mass.

Interpretation. Shared write/read routing and the sharper initial temperature primarily act as regularisers: they prevent the router from collapsing mass onto a few hubs without hurting sample efficiency. The depth-wise convolution, in contrast, provides an expressivity boost that both improves perplexity and raises the effective signal scale entering layer 0; once it is removed the router cannot maintain broad support and the ReZero gate decays. The combination of all three mitigations therefore offers a balanced trade-off between stability and performance.

A.7.2 VALIDATING THE HYPERPARAMETER CHOICES

These experiments motivate the choice of kernel size (5) for the local convolution, the alpha-gate initialization (1.0), the number of prototypes (32), the use of mass normalization, low-rank projection at the value stream, and dropout. In these experiments, we use the default model, data, and training configurations, unless otherwise specified. We search over learning rates values (1.0e-3, 2.0e-3, 3.0e-3) for ProtoT, (0.8e-3, 1.6e-3, 3.2e-3) for LLaMA, (1.9e-3, 3.8e-3, 7.6e-3) for Mamba, and (3.4e-3, 6.8e-3, 13.6e-3) for DeltaNet (the middle values of each interval are informed by the best learning rates from the automatic hyperparameter search, Section 4), and average the results over 3 seeds.

Kernel size of local convolution The results in Table 16 show that kernel size 5 and 6 are the best values in terms of dev perplexity, with an insignificant difference (≈0.2%) between the two (97.1 vs 96.9), which confirms our choice of kernel size = 5.

Table 16: Kernel size ablation of the local convolution in ProtoT. Reported best dev perplexity (lower is better), averaged over 3 seeds. Best values are in bold.

Variant	Kernel size	Performance (dev perplexity) ↓
ProtoT (k=4)	4	98.3
ProtoT (k=5)	5	**97.1**
ProtoT (k=6)	6	**96.9**
ProtoT (k=7)	7	97.5

Alpha-gate initialization: The results in Table 17 show that 0.8 and 1.0 are the best values for α initialization, without significant difference in performance, which confirms our choice of 1.0. In particular, $\alpha = 1.0$ performs better than ReZero's $\alpha = 0.0$ (Bachlechner et al., 2020), with 97.1 vs 99.2 perplexity. This is likely because ReZero trains extremely-deep NNs, where it may be beneficial to start from zero contribution from the layers, to avoid noise accumulation early in training.

Table 17: Alpha-gate initialization study for ProtoT. Reported best dev perplexity (lower is better), averaged over 3 seeds. Best values are in bold.

Alpha initialization value	Performance (dev perplexity) ↓
0.0	99.2
0.5	97.9
0.8	**97.0**
1.0	**97.1**
1.2	98.7

Mass Normalization and Low-Rank Projection at the Value Stream: The results in Table 18 show that the mass normalization (used in the default setting) brings ≈9% slowdown, while improving perplexity by ≈4.6%. This is a trade-off, where we have chosen the performance gain over the slowdown. On the other hand, the low-rank projection to half the hidden size performs about the same in terms of dev perplexity (97.1 vs 97.3), while introducing a massive speed-up (≈59% faster).

Table 18: Ablation study for ProtoT: mass normalization and low-rank projection. Reported best dev perplexity (lower is better), averaged over 3 seeds. Best values are in bold.

Setting	Performance (dev perplexity)	Speed after torch.compile() [it/s]
Default settings	**97.1**	89
No mass normalization	101.8	**98**
No low-rank projection	**97.3**	56

Optimal number of prototypes (R): The results in Table 19 show that ProtoT's performance plateaus after R=32 (97.1→97.3 perplexity), whereas speed drops substantially (89→60 it/s). This shows that R=32 is the optimal trade-off between model performance and speed.

Table 19: Optimal number of prototypes (R) for ProtoT. Reported best dev perplexity (lower is better), averaged over 3 seeds. Best values are in bold.

R value	Performance (dev perplexity)	Speed after torch.compile() [it/s]
16	98.5	**116**
32	**97.1**	89
64	**97.3**	60

Dropout: The results in Table 20 show that the default dropout configuration we use in Section 4 yields the best perplexity for ProtoT and all baseline models.

Table 20: Dropout study for ProtoT, LLaMA, Mamba, and DeltaNet. Reported best dev perplexity (lower is better), averaged over 3 seeds. Best values are in bold.

Model settings	**Performance (dev perplexity) ↓**
ProtoT (default: dropout=0.1)	**97.1**
ProtoT (no dropout)	107.6
LLaMA (default: dropout=0.1)	**84.8**
LLaMA (no dropout at self-attn)	85.2
LLaMA (no dropout at all)	92.4
Mamba (default: dropout=0.1)	**91.1**
Mamba (no dropout)	96.6
DeltaNet (default: dropout=0.1)	**99.4**
DeltaNet (no dropout)	103.5

A.8 ADDITIONAL INTERPRETABILITY METRICS

In this appendix, we report results on correlation between half life values and locality of a concept as well as four complementary metrics that characterize how prototype activations of ProtoT and LLaMA attention heads value norms evolve across depth. Each metric captures a different aspect of how the routing distribution changes from early to deeper layers. Let $a_{l,p}(x)$ denote the activation of prototype $p \in \{1, \ldots, P\}$ at layer $l \in \{1, \ldots, L\}$ for input x, and let $\mathbf{a}_l(x) = (|a_{l,1}(x)|, \ldots, |a_{l,P}(x)|)$ denote the vector of absolute activations.

Correlation Between Half-Life and Locality To assess whether prototype half-life reflects concept locality, we require an operational proxy for locality. Empirically, low-level lexical prototypes (e.g., punctuation, stopwords) exhibit highly repetitive sets of most-activating-tokens, whereas more abstract prototypes show greater token diversity. This aligns with the intuition that local prototypes focus on neighboring tokens and as a result, activate more strongly based on token identity, while longer half-life prototypes aggregate information over wider contexts. Motivated by this observation, we use the *token repetition score* as a proxy for locality.

Low half-life prototype (L2P6) HL $= 5.04$, repetition $= 0.73$ Function-word clusters (e.g., "in the", "of the")	**Higher half-life prototype (L2P10)** HL $= 12.26$, repetition $= 0.25$ Transformation expressions (e.g., "from · to ·")

`the , the , ,`
`in the , the of , ,`
`a many of the , ,`
`, or of the most and`
`the a , in , this`
`in the , in other ,`
	`guide for aim deliver their close`
	`converted to homeless converted an into`
	`adaptation of through applied to the`
	`delegation responsibility shifts from`
	` ``common'' delegation`
	`manage .  bring our home from`
	`uses encourage invite your to build`

Figure 7: Two example prototypes illustrating the relationship between half-life and repetition. The short half-life prototype (left) exhibits tightly localized, highly repetitive lexical patterns, whereas the longer half-life prototype (right) activates on broader transformation expressions.

We quantify the relationship between half-life and repetition using two standard statistical tools: (i) a Spearman rank correlation between half-life and repetition scores, and (ii) quantile-based group comparisons in which prototypes are divided into half-life quartiles. For the latter, we compare mean repetition scores across quartiles and compute the effect size (Cohen's d) between the lowest and highest half-life groups.

The results show a robust negative association between half-life and repetition, with a highly significant Spearman correlation and a large effect size ($d = 0.825$) between the lowest and highest

Metric	Value	Interpretation
Spearman ρ	-0.2192	Negative association
p-value	8.3×10^{-10}	Very significant
Q1 repetition (HL ≤ 7.8)	0.4060	High repetition
Q2 repetition	0.3429	Medium
Q3 repetition	0.4017	High
Q4 repetition (HL > 13.4)	0.2695	Low repetition
Cohen's d (Q1–Q4)	0.825	Large effect

Table 21: Relationship between prototype half-life and repetition score. Lower half-life prototypes exhibit substantially higher repetition.

quartiles. This provides strong evidence that **prototypes with shorter half-lives encode more local, repetitive lexical patterns**, whereas longer half-life prototypes correspond to broader, less repetitive activation structure.

L1 SPARSITY RATIO. To measure the degree of "winner–take–all" behavior among prototypes, we compute

$$S_l \;=\; \mathbb{E}_x \frac{\max_p |a_{l,p}(x)|}{\frac{1}{P}\sum_{p=1}^{P} |a_{l,p}(x)|}. \tag{3}$$

A high value indicates that a single prototype (or a small subset) dominates the activation mass, reflecting strong concentration and effective sparsity.

GINI COEFFICIENT. To quantify the inequality of the activation distribution, we compute the Gini index

$$G_l \;=\; \mathbb{E}_x \frac{1}{P}\Big(P+1-2\,\frac{\sum_{p=1}^{P}(P+1-p)\,a^{\uparrow}_{l,p}(x)}{\sum_{p=1}^{P} a_{l,p}(x)}\Big), \tag{4}$$

where $a^{\uparrow}_{l,p}(x)$ are the activations sorted in increasing order. Low values correspond to uniform activation across prototypes, while high values indicate strong inequality and specialization.

ENTROPY. To measure the spread or concentration of activations, we normalize $p_{l,p}(x) = |a_{l,p}(x)|/\sum_{q=1}^{P}|a_{l,q}(x)|$ and compute the Shannon entropy

$$H_l \;=\; -\,\mathbb{E}_x \sum_{p=1}^{P} p_{l,p}(x)\,\log p_{l,p}(x). \tag{5}$$

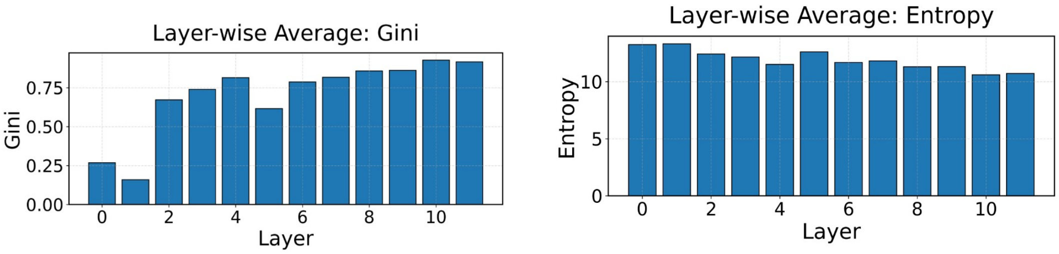
High entropy indicates diffuse activation across many prototypes, whereas lower entropy reflects concentrated, low-uncertainty routing.

MUTUAL INFORMATION. To assess how strongly prototype activations depend on surface lexical identity, we compute the mutual information between the discretized activations $\tilde{a}_{l,p}$ and the token identity T:

$$I_l \;=\; I(T\,;\,\tilde{a}_{l,p})\,. \tag{6}$$

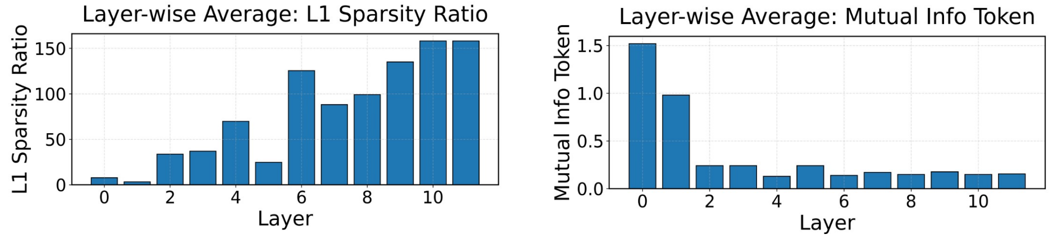
High mutual information indicates that activations are predictive of the specific token type. A decrease in mutual information with depth does *not* by itself establish that deeper prototypes encode "more abstract" concepts; however, it is *consistent* with the broader pattern observed across our sparsity, entropy, and interpretability analyses, where later layers appear less tied to local lexical identity and more shaped by contextual or compositional signals.

Together, these metrics provide a multifaceted view of how prototype representations of ProtoT sharpen, specialize, and suggest a transition from local lexical cues to increasingly structured or context-sensitive behaviors as depth increases. LLaMA metrics by contrast, do not show clear systematic patterns, entropy remains uniform across layers while the Gini coefficient and L1 sparsity oscillate. Mutual Information follows a pattern similar to that of ProtoT



(a) Gini increases with depth, indicating concentration of activation mass.

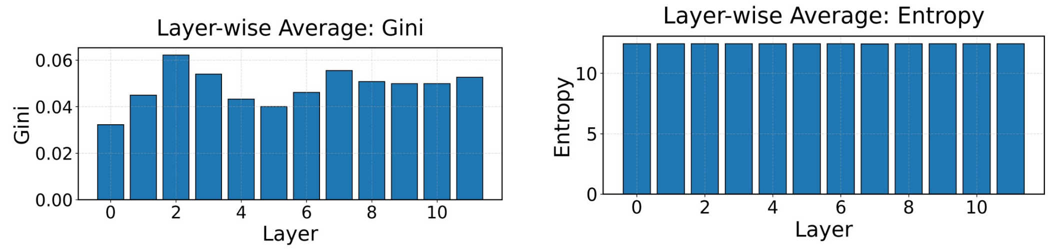
(b) Entropy decreases as routing becomes sharper and less uniform.



(c) L1 sparsity increases, reflecting increasingly selective activation.

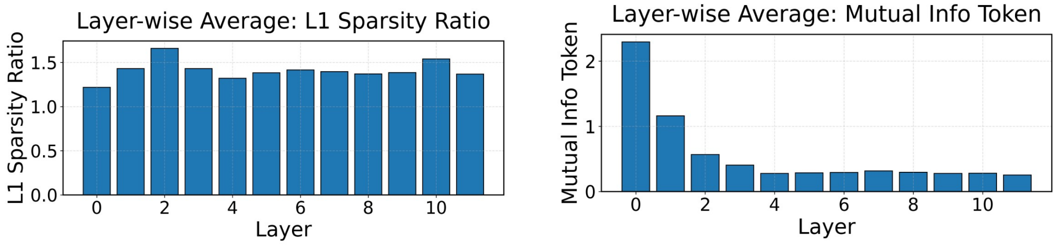
(d) Mutual information decreases with depth, indicating weaker dependence on lexical identity and greater sensitivity to contextual structure.

Figure 8: ProtoT interpretability metrics across depth.



(a) Gini increases with depth, indicating concentration of activation mass.

(b) Entropy decreases as routing becomes sharper and less uniform.



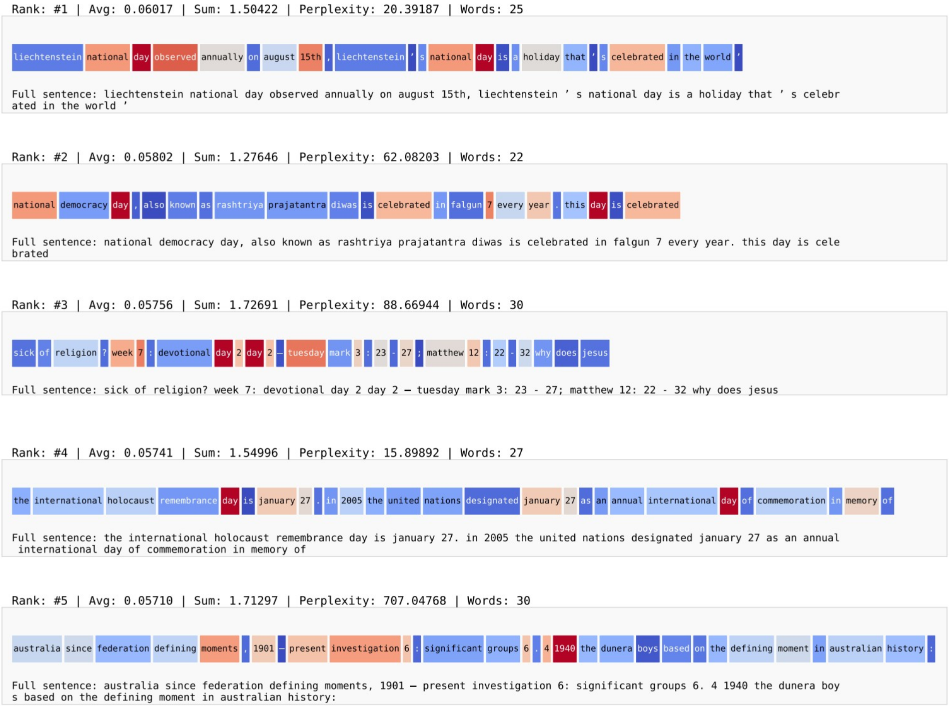
(c) L1 sparsity increases, reflecting increasingly selective activation.

(d) Mutual information decreases with depth, indicating weaker dependence on lexical identity and greater sensitivity to contextual structure.

Figure 9: LLaMA interpretability metrics across depth.

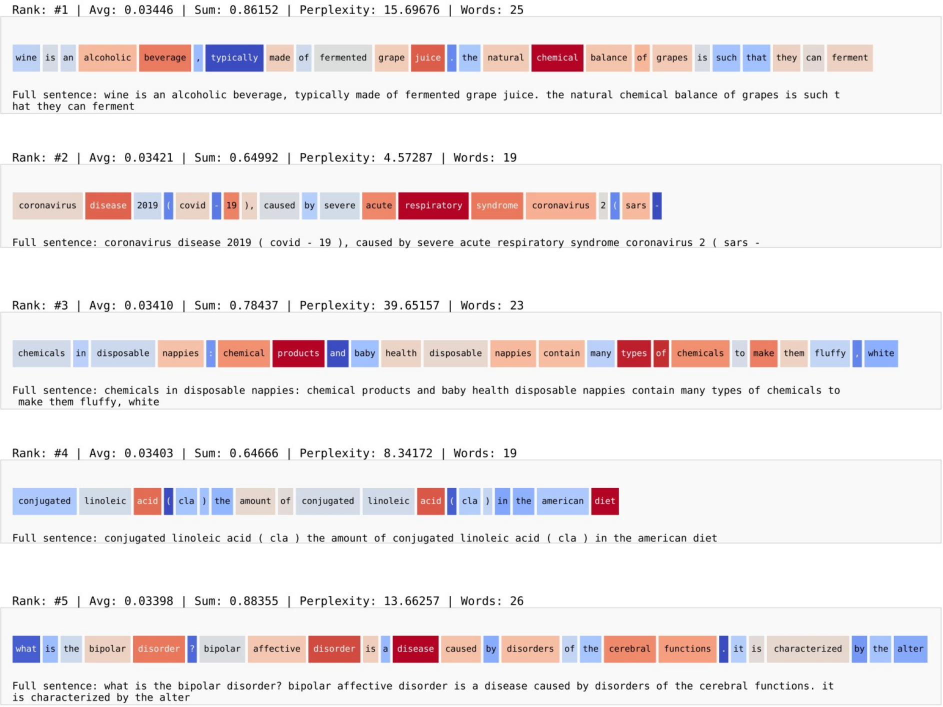
A.9 ADDITIONAL CONCEPTS VISUALIZATION AND LLM-AIDED EVALUATION RESULTS

Prototype visualizations We provide additional examples from the write gate activation interpretability experiment, useful to better illustrate results about learned concept representation. This section also contains result statistics for LLM-aided evaluation experiment for multiple model configuration of ProtoT and for LLaMA.



```
Rank: #1 | Avg: 0.06017 | Sum: 1.50422 | Perplexity: 20.39187 | Words: 25

Full sentence: liechtenstein national day observed annually on august 15th, liechtenstein ’ s national day is a holiday that ’ s celebr
ated in the world ’

Rank: #2 | Avg: 0.05802 | Sum: 1.27646 | Perplexity: 62.08203 | Words: 22

Full sentence: national democracy day, also known as rashtriya prajatantra diwas is celebrated in falgun 7 every year. this day is cele
brated

Rank: #3 | Avg: 0.05756 | Sum: 1.72691 | Perplexity: 88.66944 | Words: 30

Full sentence: sick of religion? week 7: devotional day 2 day 2 – tuesday mark 3: 23 - 27; matthew 12: 22 - 32 why does jesus

Rank: #4 | Avg: 0.05741 | Sum: 1.54996 | Perplexity: 15.89892 | Words: 27

Full sentence: the international holocaust remembrance day is january 27. in 2005 the united nations designated january 27 as an annual
 international day of commemoration in memory of

Rank: #5 | Avg: 0.05710 | Sum: 1.71297 | Perplexity: 707.04768 | Words: 30

Full sentence: australia since federation defining moments, 1901 – present investigation 6: significant groups 6. 4 1940 the dunera boy
s based on the defining moment in australian history:
```

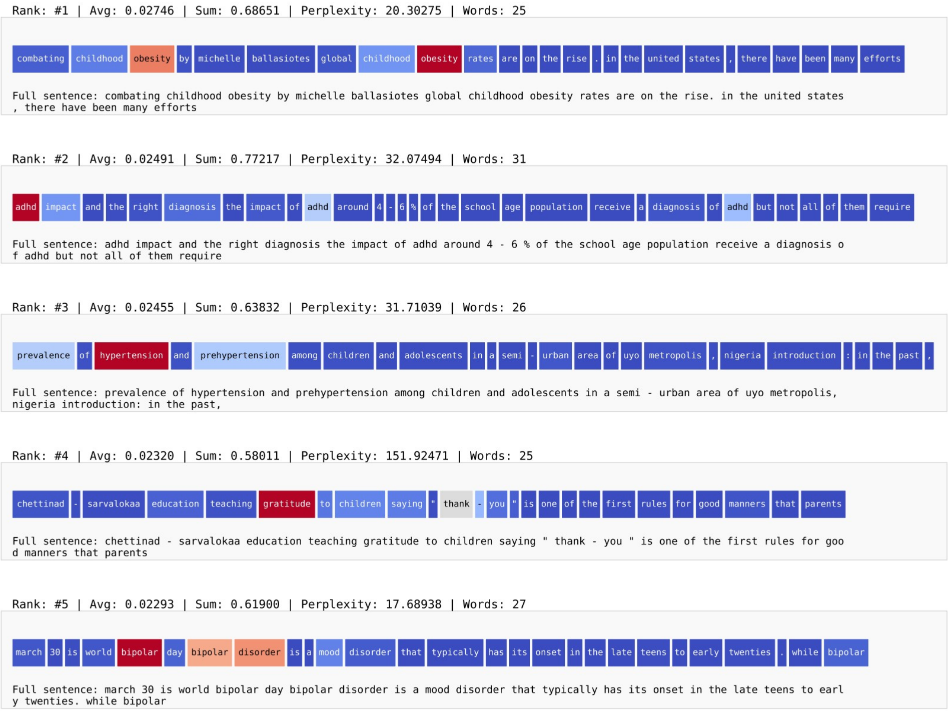
Figure 10: Visualization for prototype **L0 P18**. Half-life = 12.8



```
Rank: #1 | Avg: 0.03446 | Sum: 0.86152 | Perplexity: 15.69676 | Words: 25

Full sentence: wine is an alcoholic beverage, typically made of fermented grape juice. the natural chemical balance of grapes is such t
hat they can ferment

Rank: #2 | Avg: 0.03421 | Sum: 0.64992 | Perplexity: 4.57287 | Words: 19

Full sentence: coronavirus disease 2019 ( covid - 19 ), caused by severe acute respiratory syndrome coronavirus 2 ( sars -

Rank: #3 | Avg: 0.03410 | Sum: 0.78437 | Perplexity: 39.65157 | Words: 23

Full sentence: chemicals in disposable nappies: chemical products and baby health disposable nappies contain many types of chemicals to
 make them fluffy, white

Rank: #4 | Avg: 0.03403 | Sum: 0.64666 | Perplexity: 8.34172 | Words: 19

Full sentence: conjugated linoleic acid ( cla ) the amount of conjugated linoleic acid ( cla ) in the american diet

Rank: #5 | Avg: 0.03398 | Sum: 0.88355 | Perplexity: 13.66257 | Words: 26

Full sentence: what is the bipolar disorder? bipolar affective disorder is a disease caused by disorders of the cerebral functions. it
is characterized by the alter
```

Figure 11: Visualization for prototype **L1 P14**. Half-life = 13.2

LLM scoring and labeling We provide visualizations of some random selected examples and resulting statistics of the LLM scoring process. We also show ablation for different model configurations, including R=16, R=64, two extra seeds for R=32 and R=32 without low rank projection.



```
Rank: #1 | Avg: 0.02746 | Sum: 0.68651 | Perplexity: 20.30275 | Words: 25

Full sentence: combating childhood obesity by michelle ballasiotes global childhood obesity rates are on the rise. in the united states
, there have been many efforts
```

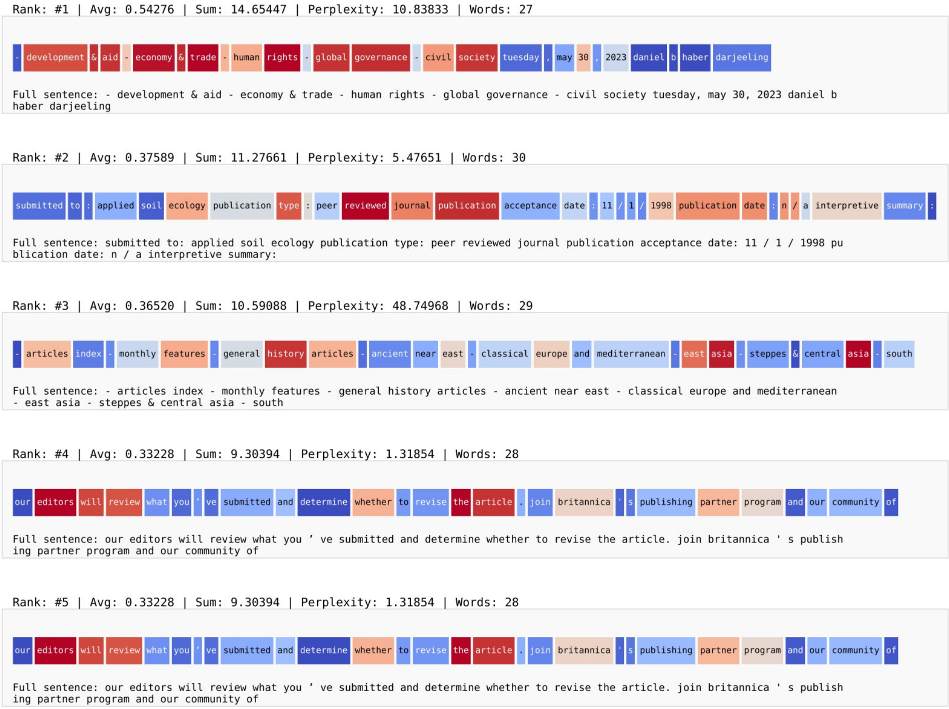
```
Rank: #2 | Avg: 0.02491 | Sum: 0.77217 | Perplexity: 32.07494 | Words: 31

Full sentence: adhd impact and the right diagnosis the impact of adhd around 4 - 6 % of the school age population receive a diagnosis o
f adhd but not all of them require
```

```
Rank: #3 | Avg: 0.02455 | Sum: 0.63832 | Perplexity: 31.71039 | Words: 26

Full sentence: prevalence of hypertension and prehypertension among children and adolescents in a semi - urban area of uyo metropolis,
nigeria introduction: in the past,
```

```
Rank: #4 | Avg: 0.02320 | Sum: 0.58011 | Perplexity: 151.92471 | Words: 25

Full sentence: chettinad - sarvalokaa education teaching gratitude to children saying " thank - you " is one of the first rules for goo
d manners that parents
```

```
Rank: #5 | Avg: 0.02293 | Sum: 0.61900 | Perplexity: 17.68938 | Words: 27

Full sentence: march 30 is world bipolar day bipolar disorder is a mood disorder that typically has its onset in the late teens to earl
y twenties. while bipolar
```

Figure 12: Visualization for prototype **L7 P31**. Half-life = 12.7



```
Rank: #1 | Avg: 0.54276 | Sum: 14.65447 | Perplexity: 10.83833 | Words: 27

Full sentence: - development & aid - economy & trade - human rights - global governance - civil society tuesday, may 30, 2023 daniel b
haber darjeeling
```

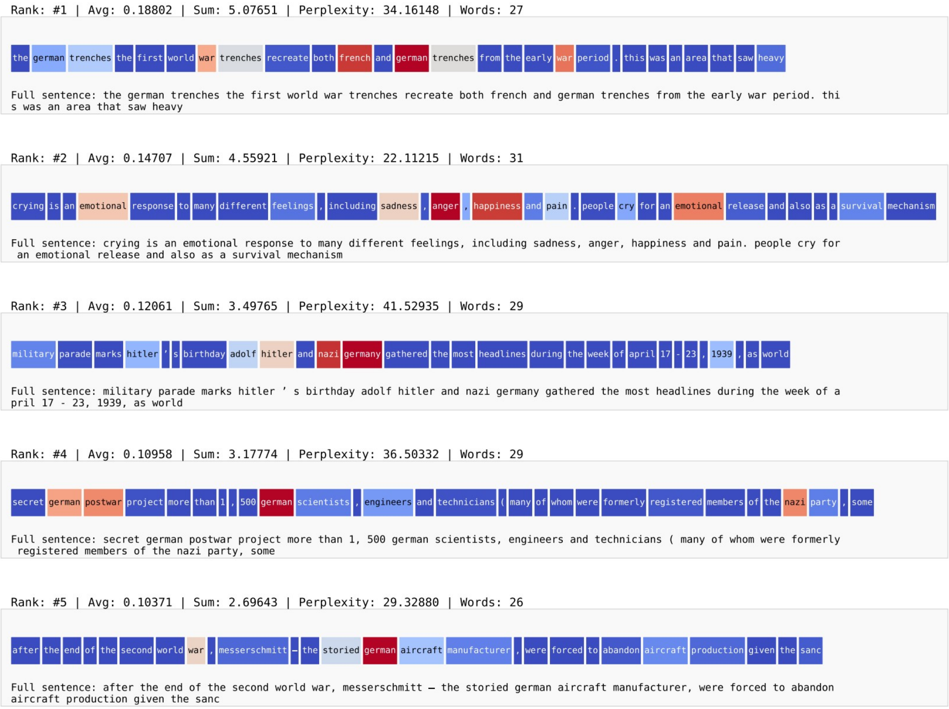
```
Rank: #2 | Avg: 0.37589 | Sum: 11.27661 | Perplexity: 5.47651 | Words: 30

Full sentence: submitted to: applied soil ecology publication type: peer reviewed journal publication acceptance date: 11 / 1 / 1998 pu
blication date: n / a interpretive summary:
```

```
Rank: #3 | Avg: 0.36520 | Sum: 10.59088 | Perplexity: 48.74968 | Words: 29

Full sentence: - articles index - monthly features - general history articles - ancient near east - classical europe and mediterranean
- east asia - steppes & central asia - south
```

```
Rank: #4 | Avg: 0.33228 | Sum: 9.30394 | Perplexity: 1.31854 | Words: 28

Full sentence: our editors will review what you ’ ve submitted and determine whether to revise the article. join britannica ' s publish
ing partner program and our community of
```

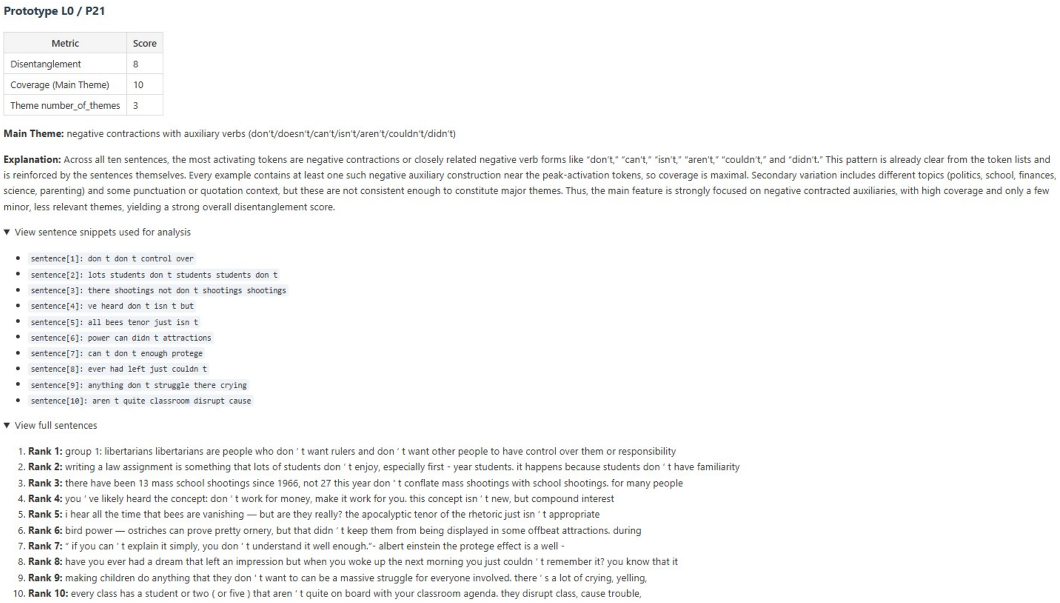
```
Rank: #5 | Avg: 0.33228 | Sum: 9.30394 | Perplexity: 1.31854 | Words: 28

Full sentence: our editors will review what you ’ ve submitted and determine whether to revise the article. join britannica ' s publish
ing partner program and our community of
```

Figure 13: Visualization for prototype **L8 P5**. Half-life = 0.140



```
Rank: #1 | Avg: 0.18802 | Sum: 5.07651 | Perplexity: 34.16148 | Words: 27

the german trenches the first world war trenches recreate both french and german trenches from the early war period . this was an area that saw heavy

Full sentence: the german trenches the first world war trenches recreate both french and german trenches from the early war period. thi
s was an area that saw heavy

Rank: #2 | Avg: 0.14707 | Sum: 4.55921 | Perplexity: 22.11215 | Words: 31

crying is an emotional response to many different feelings , including sadness , anger , happiness and pain . people cry for an emotional release and also as a survival mechanism

Full sentence: crying is an emotional response to many different feelings, including sadness, anger, happiness and pain. people cry for
 an emotional release and also as a survival mechanism

Rank: #3 | Avg: 0.12061 | Sum: 3.49765 | Perplexity: 41.52935 | Words: 29

military parade marks hitler ' s birthday adolf hitler and nazi germany gathered the most headlines during the week of april 17 - 23 , 1939 , as world

Full sentence: military parade marks hitler ’ s birthday adolf hitler and nazi germany gathered the most headlines during the week of a
pril 17 - 23, 1939, as world

Rank: #4 | Avg: 0.10958 | Sum: 3.17774 | Perplexity: 36.50332 | Words: 29

secret german postwar project more than 1 , 500 german scientists , engineers and technicians ( many of whom were formerly registered members of the nazi party , some

Full sentence: secret german postwar project more than 1, 500 german scientists, engineers and technicians ( many of whom were formerly
 registered members of the nazi party, some

Rank: #5 | Avg: 0.10371 | Sum: 2.69643 | Perplexity: 29.32880 | Words: 26

after the end of the second world war , messerschmitt – the storied german aircraft manufacturer , were forced to abandon aircraft production given the sanc

Full sentence: after the end of the second world war, messerschmitt – the storied german aircraft manufacturer, were forced to abandon
aircraft production given the sanc
```

Figure 14: Visualization for prototype **L10 P8**. Half-life = 0.510



Prototype L0 / P21

Metric	Score
Disentanglement	8
Coverage (Main Theme)	10
Theme number_of_themes	3

Main Theme: negative contractions with auxiliary verbs (don’t/doesn’t/can’t/isn’t/aren’t/couldn’t/didn’t)

Explanation: Across all ten sentences, the most activating tokens are negative contractions or closely related negative verb forms like “don’t,” “can’t,” “isn’t,” “aren’t,” “couldn’t,” and “didn’t.” This pattern is already clear from the token lists and is reinforced by the sentences themselves. Every example contains at least one such negative auxiliary construction near the peak-activation tokens, so coverage is maximal. Secondary variation includes different topics (politics, school, finances, science, parenting) and some punctuation or quotation context, but these are not consistent enough to constitute major themes. Thus, the main feature is strongly focused on negative contracted auxiliaries, with high coverage and only a few minor, less relevant themes, yielding a strong overall disentanglement score.

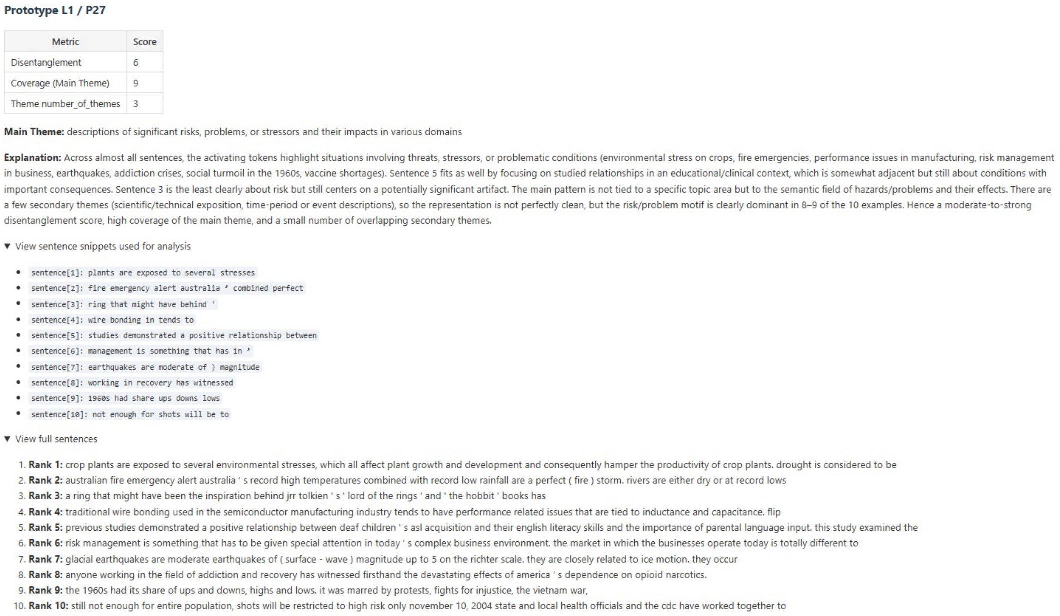
▼ View sentence snippets used for analysis

- `sentence[1]: don t don t control over`
- `sentence[2]: lots students don t students students don t`
- `sentence[3]: there shootings not don t shootings shootings`
- `sentence[4]: ve heard don t isn t but`
- `sentence[5]: all bees tenor just isn t`
- `sentence[6]: power can didn t attractions`
- `sentence[7]: can t don t enough protege`
- `sentence[8]: ever had left just couldn t`
- `sentence[9]: anything don t struggle there crying`
- `sentence[10]: aren t quite classroom disrupt cause`

▼ View full sentences

1. **Rank 1:** group 1: libertarians libertarians are people who don ’ t want rulers and don ’ t want other people to have control over them or responsibility
2. **Rank 2:** writing a law assignment is something that lots of students don ’ t enjoy, especially first - year students. it happens because students don ’ t have familiarity
3. **Rank 3:** there have been 13 mass school shootings since 1966, not 27 this year don ' t conflate mass shootings with school shootings. for many people
4. **Rank 4:** you ’ ve likely heard the concept: don ’ t work for money, make it work for you. this concept isn ’ t new, but compound interest
5. **Rank 5:** i hear all the time that bees are vanishing — but are they really? the apocalyptic tenor of the rhetoric just isn ’ t appropriate
6. **Rank 6:** bird power — ostriches can prove pretty ornery, but that didn ’ t keep them from being displayed in some offbeat attractions. during
7. **Rank 7:** “ if you can ’ t explain it simply, you don ’ t understand it well enough.”- albert einstein the protege effect is a well -
8. **Rank 8:** have you ever had a dream that left an impression but when you woke up the next morning you just couldn ’ t remember it? you know that it
9. **Rank 9:** making children do anything that they don ' t want to can be a massive struggle for everyone involved. there ’ s a lot of crying, yelling,
10. **Rank 10:** every class has a student or two (or five) that aren ’ t quite on board with your classroom agenda. they disrupt class, cause trouble,

Figure 15: LLM aided interpretability results for prototype **L0 P21** R=32 (S=135).



Prototype L1 / P27

Metric	Score
Disentanglement	6
Coverage (Main Theme)	9
Theme number_of_themes	3

Main Theme: descriptions of significant risks, problems, or stressors and their impacts in various domains

Explanation: Across almost all sentences, the activating tokens highlight situations involving threats, stressors, or problematic conditions (environmental stress on crops, fire emergencies, performance issues in manufacturing, risk management in business, earthquakes, addiction crises, social turmoil in the 1960s, vaccine shortages). Sentence 5 fits as well by focusing on studied relationships in an educational/clinical context, which is somewhat adjacent but still about conditions with important consequences. Sentence 3 is the least clearly about risk but still centers on a potentially significant artifact. The main pattern is not tied to a specific topic area but to the semantic field of hazards/problems and their effects. There are a few secondary themes (scientific/technical exposition, time-period or event descriptions), so the representation is not perfectly clean, but the risk/problem motif is clearly dominant in 8–9 of the 10 examples. Hence a moderate-to-strong disentanglement score, high coverage of the main theme, and a small number of overlapping secondary themes.

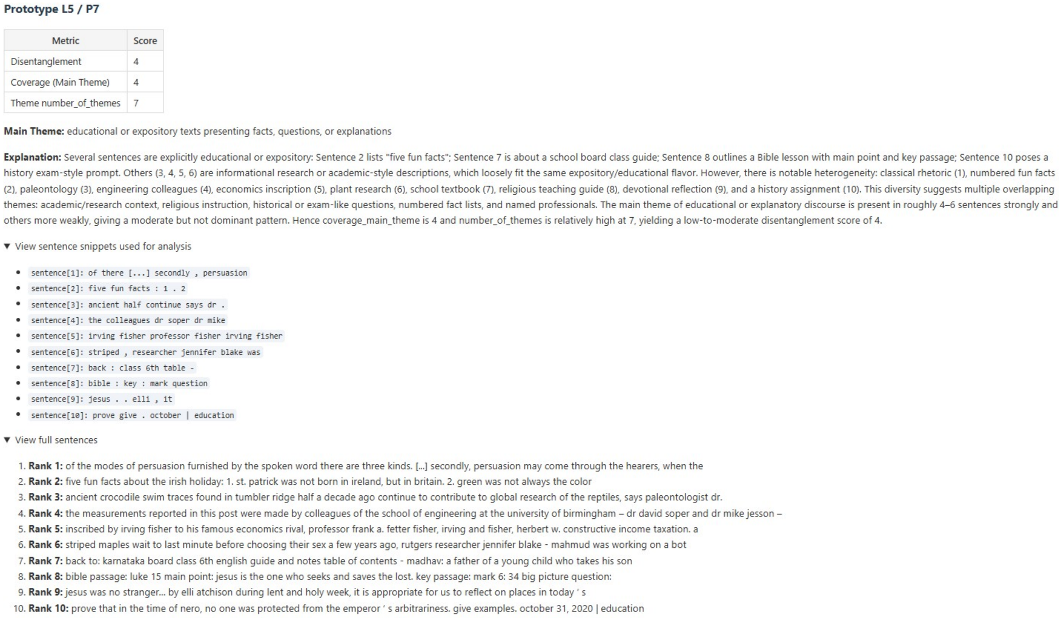
▼ View sentence snippets used for analysis

- `sentence[1]: plants are exposed to several stresses`
- `sentence[2]: fire emergency alert australia ’ combined perfect`
- `sentence[3]: ring that might have behind '`
- `sentence[4]: wire bonding in tends to`
- `sentence[5]: studies demonstrated a positive relationship between`
- `sentence[6]: management is something that has in ’`
- `sentence[7]: earthquakes are moderate of ) magnitude`
- `sentence[8]: working in recovery has witnessed`
- `sentence[9]: 1960s had share ups downs lows`
- `sentence[10]: not enough for shots will be to`

▼ View full sentences

1. **Rank 1:** crop plants are exposed to several environmental stresses, which all affect plant growth and development and consequently hamper the productivity of crop plants. drought is considered to be
2. **Rank 2:** australian fire emergency alert australia ’ s record high temperatures combined with record low rainfall are a perfect (fire) storm. rivers are either dry or at record lows
3. **Rank 3:** a ring that might have been the inspiration behind jrr tolkien ' s ' lord of the rings ' and ' the hobbit ' books has
4. **Rank 4:** traditional wire bonding used in the semiconductor manufacturing industry tends to have performance related issues that are tied to inductance and capacitance. flip
5. **Rank 5:** previous studies demonstrated a positive relationship between deaf children ' s asl acquisition and their english literacy skills and the importance of parental language input. this study examined the
6. **Rank 6:** risk management is something that has to be given special attention in today ’ s complex business environment. the market in which the businesses operate today is totally different to
7. **Rank 7:** glacial earthquakes are moderate earthquakes of (surface - wave) magnitude up to 5 on the richter scale. they are closely related to ice motion. they occur
8. **Rank 8:** anyone working in the field of addiction and recovery has witnessed firsthand the devastating effects of america ’ s dependence on opioid narcotics.
9. **Rank 9:** the 1960s had its share of ups and downs, highs and lows. it was marred by protests, fights for injustice, the vietnam war,
10. **Rank 10:** still not enough for entire population, shots will be restricted to high risk only november 10, 2004 state and local health officials and the cdc have worked together to

Figure 16: LLM aided interpretability results for prototype **L1 P27** R=32 (S=135).



Prototype L5 / P7

Metric	Score
Disentanglement	4
Coverage (Main Theme)	4
Theme number_of_themes	7

Main Theme: educational or expository texts presenting facts, questions, or explanations

Explanation: Several sentences are explicitly educational or expository: Sentence 2 lists "five fun facts"; Sentence 7 is about a school board class guide; Sentence 8 outlines a Bible lesson with main point and key passage; Sentence 10 poses a history exam-style prompt. Others (3, 4, 5, 6) are informational research or academic-style descriptions, which loosely fit the same expository/educational flavor. However, there is notable heterogeneity: classical rhetoric (1), numbered fun facts (2), paleontology (3), engineering colleagues (4), economics inscription (5), plant research (6), school textbook (7), religious teaching guide (8), devotional reflection (9), and a history assignment (10). This diversity suggests multiple overlapping themes: academic/research context, religious instruction, historical or exam-like questions, numbered fact lists, and named professionals. The main theme of educational or explanatory discourse is present in roughly 4–6 sentences strongly and others more weakly, giving a moderate but not dominant pattern. Hence coverage_main_theme is 4 and number_of_themes is relatively high at 7, yielding a low-to-moderate disentanglement score of 4.

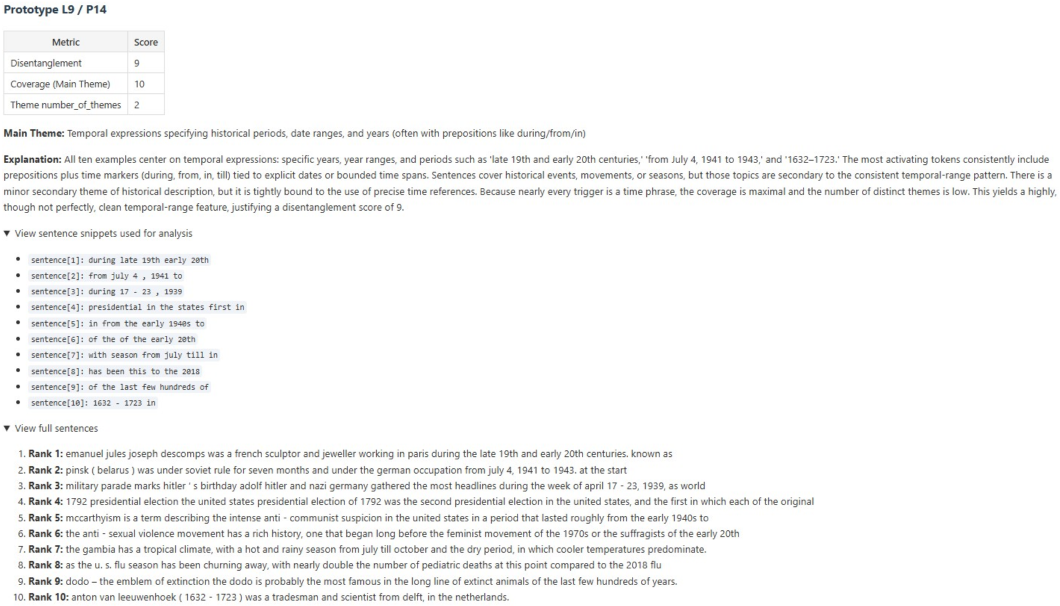
▼ View sentence snippets used for analysis

- `sentence[1]: of there [...] secondly , persuasion`
- `sentence[2]: five fun facts : 1 . 2`
- `sentence[3]: ancient half continue says dr .`
- `sentence[4]: the colleagues dr soper dr mike`
- `sentence[5]: irving fisher professor fisher irving fisher`
- `sentence[6]: striped , researcher jennifer blake was`
- `sentence[7]: back : class 6th table -`
- `sentence[8]: bible : key : mark question`
- `sentence[9]: jesus . . elli , it`
- `sentence[10]: prove give . october | education`

▼ View full sentences

1. **Rank 1:** of the modes of persuasion furnished by the spoken word there are three kinds. [...] secondly, persuasion may come through the hearers, when the
2. **Rank 2:** five fun facts about the irish holiday: 1. st. patrick was not born in ireland, but in britain. 2. green was not always the color
3. **Rank 3:** ancient crocodile swim traces found in tumbler ridge half a decade ago continue to contribute to global research of the reptiles, says paleontologist dr.
4. **Rank 4:** the measurements reported in this post were made by colleagues of the school of engineering at the university of birmingham – dr david soper and dr mike jesson –
5. **Rank 5:** inscribed by irving fisher to his famous economics rival, professor frank a. fetter fisher, irving and fisher, herbert w. constructive income taxation. a
6. **Rank 6:** striped maples wait to last minute before choosing their sex a few years ago, rutgers researcher jennifer blake - mahmud was working on a bot
7. **Rank 7:** back to: karnataka board class 6th english guide and notes table of contents - madhav: a father of a young child who takes his son
8. **Rank 8:** bible passage: luke 15 main point: jesus is the one who seeks and saves the lost. key passage: mark 6: 34 big picture question:
9. **Rank 9:** jesus was no stranger... by elli atchison during lent and holy week, it is appropriate for us to reflect on places in today ’ s
10. **Rank 10:** prove that in the time of nero, no one was protected from the emperor ’ s arbitrariness. give examples. october 31, 2020 | education

Figure 17: LLM aided interpretability results for prototype **L5 P7** R=32 (S=135).



Prototype L9 / P14

Metric	Score
Disentanglement	9
Coverage (Main Theme)	10
Theme number_of_themes	2

Main Theme: Temporal expressions specifying historical periods, date ranges, and years (often with prepositions like during/from/in)

Explanation: All ten examples center on temporal expressions: specific years, year ranges, and periods such as 'late 19th and early 20th centuries,' 'from July 4, 1941 to 1943,' and '1632–1723.' The most activating tokens consistently include prepositions plus time markers (during, from, in, till) tied to explicit dates or bounded time spans. Sentences cover historical events, movements, or seasons, but those topics are secondary to the consistent temporal-range pattern. There is a minor secondary theme of historical description, but it is tightly bound to the use of precise time references. Because nearly every trigger is a time phrase, the coverage is maximal and the number of distinct themes is low. This yields a highly, though not perfectly, clean temporal-range feature, justifying a disentanglement score of 9.

▼ View sentence snippets used for analysis

- `sentence[1]: during late 19th early 20th`
- `sentence[2]: from july 4 , 1941 to`
- `sentence[3]: during 17 - 23 , 1939`
- `sentence[4]: presidential in the states first in`
- `sentence[5]: in from the early 1940s to`
- `sentence[6]: of the of the early 20th`
- `sentence[7]: with season from july till in`
- `sentence[8]: has been this to the 2018`
- `sentence[9]: of the last few hundreds of`
- `sentence[10]: 1632 - 1723 in`

▼ View full sentences

1. **Rank 1:** emanuel jules joseph descomps was a french sculptor and jeweller working in paris during the late 19th and early 20th centuries. known as
2. **Rank 2:** pinsk (belarus) was under soviet rule for seven months and under the german occupation from july 4, 1941 to 1943. at the start
3. **Rank 3:** military parade marks hitler ' s birthday adolf hitler and nazi germany gathered the most headlines during the week of april 17 - 23, 1939, as world
4. **Rank 4:** 1792 presidential election the united states presidential election of 1792 was the second presidential election in the united states, and the first in which each of the original
5. **Rank 5:** mccarthyism is a term describing the intense anti - communist suspicion in the united states in a period that lasted roughly from the early 1940s to
6. **Rank 6:** the anti - sexual violence movement has a rich history, one that began long before the feminist movement of the 1970s or the suffragists of the early 20th
7. **Rank 7:** the gambia has a tropical climate, with a hot and rainy season from july till october and the dry period, in which cooler temperatures predominate.
8. **Rank 8:** as the u. s. flu season has been churning away, with nearly double the number of pediatric deaths at this point compared to the 2018 flu
9. **Rank 9:** dodo – the emblem of extinction the dodo is probably the most famous in the long line of extinct animals of the last few hundreds of years.
10. **Rank 10:** anton van leeuwenhoek (1632 - 1723) was a tradesman and scientist from delft, in the netherlands.

Figure 18: LLM aided interpretability results for prototype **L9 P14** R=32 (S=135).

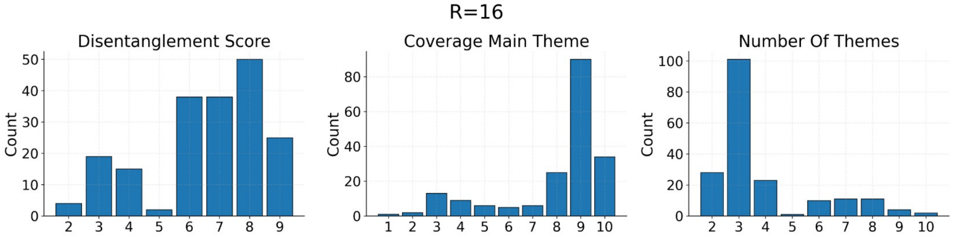


Figure 19: histograms for LLM aided interpretability for model configuration R=16

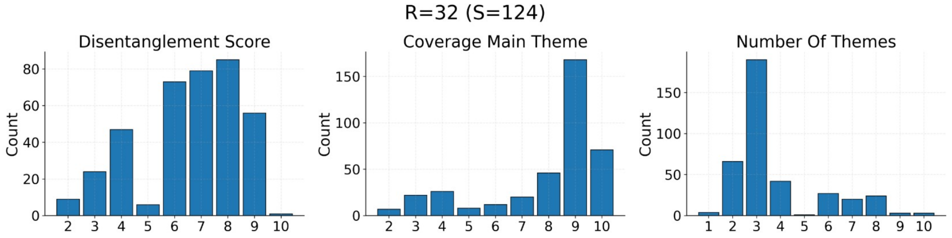


Figure 20: histograms for LLM aided interpretability for model configuration R=32 (S=124)

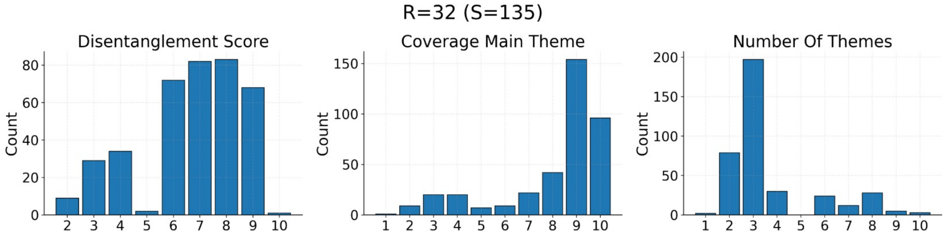


Figure 21: histograms for LLM aided interpretability for model configuration R=32 (S=135)

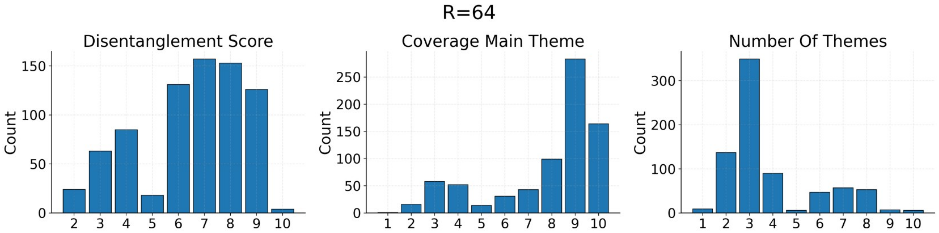


Figure 22: histograms for LLM aided interpretability for model configuration R=64

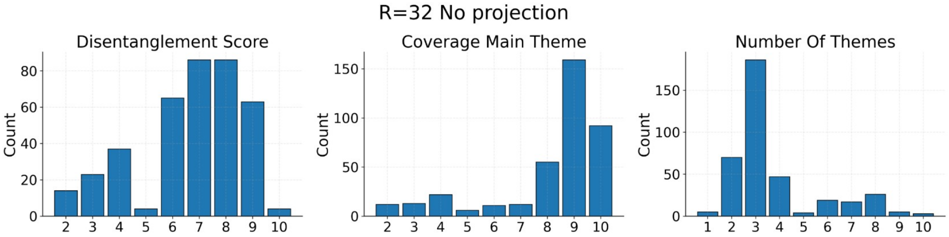


Figure 23: histograms for LLM aided interpretability for R=32 (S=124) and no low rank projection

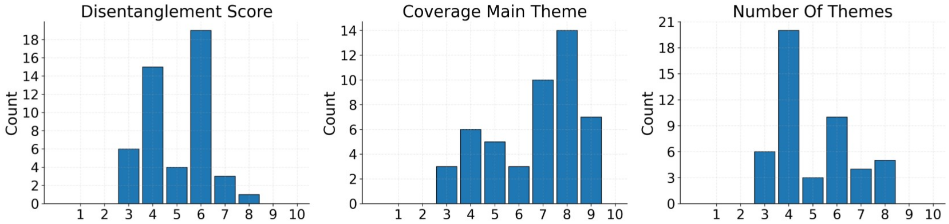


Figure 24: histograms for LLM aided interpretability for LLaMA model trained for 1)

A.10 PROMPT FOR LLM-AIDED INTERPRETABILITY EXPERIMENT

```
            "You are analyzing a single prototype (a neuron-like feature) from a neural
language model.\n"
            "For this prototype you are given, for each of its top-ranked sentences, the
full sentence and the subset "
            "of its most activating tokens. Each example is formatted as:\n\n"
            "  Most activating tokens sentence: <token1 token2 ...>\n\n"
            "  Sentence: <full sentence text>\n"
            "A **theme** is any recurrent characteristic that appears across multiple
high-activation token sets or their "
            "sentences. Themes can be narrative motifs, entities, stylistic elements,
punctuation patterns, lexical fields, "
            "or any other shared property that appears across more than one example. ALL
PREVIOUS EXAMPLES ARE CONSIDERED THEMES. It can be local (individuated by single"
            "isolated words (like numbers or punctuation) or more sentence level
(individuated by composing multiple tokens)\n\n"

            "Your task is to determine whether there is a meaningful main theme shared
across the provided sentences, "
            "and how strongly that theme characterizes this prototype.\n\n"

            "You should first of all observe the most activating tokens to check if there
is a main theme that is already observable there"
            "then if by observing only the tokens it is hard to find a common theme, you
should check the sentences to observe if there is"
            "a more ’sentence level’ theme like a narrative or a motif. Remember that the
first indication is given by the most activating tokens,"
            "if they are clearly forming a visible coherent pattern, that’s enough to
conclude the analysis. If instead the pattern they are forming is not that clear,"
            "then move to observe and analyze the full sentences."

            "For example in these sentences and most activating tokens there is everywhere
a common theme of comitative structure with the use of the word ’with’ that is already
very visible with the most activating tokens"
            "Sentence Rank 1: how are nigeria â\200\231 s trading and political
relationships changing? what relationship did nigeria have with britain? trade between
britain and west africa has occurred for over"
            "Sentence Rank 2: the teens at risk series deals with powerful and relevant
issues facing teenagers today. this series reveals first - hand how young people deal with
and overcome the very real"
            "Sentence Rank 3: the only country that south korea shares a land border with
is north korea. the land border between the two countries is 148 miles in length, dissec"
            "Sentence Rank 4: communication with families: a plan for every child keep
communicating with parents and pave the way for continued progress! by carla poole summer
is on the"
            "Most activating tokens[1]: trading relationships relationship with trade
between"
            "Most activating tokens[2]: deals with issues facing deal with overcome"
            "Most activating tokens[3]: shares border with border between two"
            "Most activating tokens[4]: communication with families communicating with
parents"

            "or in this other example list of most activating tokens, we have knowledge
and questions, this can be safely identified as a single common main theme. the list of
sentences reinforce what was already observable"
            "Sentence Rank 1: did you know that something called chronic fatigue existed?
well, many of us do not know about this ailment, but it is actually quite common. the"
            "Sentence Rank 2: how to do homework quicker the best way to do homework
sooner is a problem most college students have. doing homework isn â\200\231 t as
enjoyable as mother and"
            "Sentence Rank 3: what does shingles look like? how does shingles start? how
does shingles affect the body? how is the diagnosis made and treatment"
            "Sentence Rank 4: norovirus information guide noroviruses ( previously known
as norwalklike viruses ) are widely recognized as the agents known to cause outbreaks of
illness among"
            "Most activating tokens[1]: did know well do know quite"
            "Most activating tokens[2]: do do isn t as as"
            "Most activating tokens[3]: does ? does ? does ?"
            "Most activating tokens[4]: known as recognized as known"
            "In this other examples we have that even if the sentences don’t exihibit a
clear evident patterns, the most activating tokens are all about punctuation. This is
enough"
            "to identify a hihgly disentangle punctuation theme with high coverage and low
number of themes."
            "Most activating tokens[1]: ; ; ; ; ; ; ; ; ; ; ; ; ;"
            "Most activating tokens[2]: , ), . . . , , ."
            "Most activating tokens[3]: , , , . . ,"
            "Most activating tokens[4]: . , , ,"

            "Sentence Rank 1: conflict resolution - for kids conflict; listening;
negotiate; compromise; resolve; resolution; understanding; working together; outcomes;
skill; listen; feelings; put"
            "Sentence Rank 2: schulman, samuel ( 1864 â\200\223 1955 ), u. s. reform
rabbi. schulman, born in russia, was taken to the u."
            "Sentence Rank 3: we provide arts and architecture assignment help in fine
arts, visual arts, industrial design, architectural design etc. at myassignmenthelp. net,"
            "Sentence Rank 4: google. com, pub - 6983070578797176, direct, f08c47f"



            "Break the disentanglement assessment into these components (all on a
1â\200\22310 scale):\n"
            "  - coverage_main_theme: in how many sentences the main theme appears?\n"
            "      indicate the precise number between 1 and 10\n\n"
            "  - number_of_themes: how many uncorrelated themes appear in the prototype?\n"
            "      indicate the precise number between 1 and 10, if it is more than 10,
just answer 10\n\n"


            "Use these two component scores to decide the overall disentanglement_score
(1â\200\22310):\n"
            "  1-2 = There is no recurring characteristic across the most activating
tokens or the sentences; entirely mixed or noisy.\n"
            "  3-4 = Very weak hints of a pattern; mostly mixed, but with minor
repetition.\n"
            "  5-6 = A moderate theme: some noise, but a noticeable dominant trait.\n"
            "  7-8 = A strong theme: clearly recurrent and consistent across many
sentences.\n"
            "  9-10 = An extremely strong and clean theme: nearly all sentences share the
same core characteristic.\n\n"

            "When choosing the scores, do NOT avoid extremes: use the full range from 1 to
10 whenever appropriate.\n\n"

            "Provide your answer STRICTLY as a JSON object with these exact keys:\n"
            ’  "disentanglement_score": integer 1-10,\n’
            ’  "coverage_main_theme": integer 1-10,\n’
            ’  "number_of_themes": integer 1-10,\n’
            ’  "theme": short string describing the single main shared characteristic,\n’
            ’  "explanation": 1-10 sentences explaining why you chose these scores.\n\n’
            "Output ONLY valid JSON. No markdown, no backticks, no additional text."
```

Figure 25: Full prompt used for LLM aided evaluation and labeling experiment

B APPENDIX: COMPREHENSIVE EVALUATION DETAILS

B.1 QUALITATIVE EVALUATION METHODOLOGY

We adopted an LLM-as-a-judge protocol inspired by Chatbot Arena (Chiang et al., 2024). For a given evaluation prompt, we present the two model outputs (Response A and Response B) to a frozen judge model (Gemma-3-4B-IT) using a fixed system instruction that asks the judge to select the better response based on coherence, relevance, fluency, and correctness, and to reply with a single token in “A”, “B”, “Tie”. To reduce position bias, we query the judge twice per example: once with the order (A, B) and once with the order (B, A). The two decisions are mapped back to the original models and converted into soft pairwise scores (win = 1, tie = 0.5 per model). Aggregating over all prompts yields pairwise win/tie statistics between the two systems, which we then use to compute Elo ratings following the standard Chatbot Arena procedure.

B.2 COMPREHENSIVE RESULTS

Table 22 presents the qualitative ELO rankings of the models.

Table 22: Evaluation of Model Quality. ELO scores are derived from pairwise judge evaluations (higher is better).

Model	**ELO**
LLaMA	938.59
Mamba	1150.78
ProtoT	1015.93
DeltaNet	894.70

B.3 QUALITATIVE SAMPLES

We decode 50-token continuations for each model on a shared set of FineWeb validation prompts, score every output with BLEU against the held-out reference, and then extract the highest-BLEU prompt for each model. The examples below present those prompts alongside every model’s completion and BLEU score, enabling apples-to-apples qualitative inspection.

Each subsection lists the prompts where a model achieved its highest BLEU scores. For every selected prompt, we show the prompt, the reference completion, and the completions (with BLEU) for all available models.

B.3.1 PROMPT — PROTOT WINNER

Prompt: *..can also be caused by other conditions such as benign prostate enlargement. there are no known causes for prostate cancer. however, between 5% and 10% of cases run in families, where the patient inherits a high risk of this type of cancer. prostate cancer is very rare in men under 50. the risk increases after the age of 50 with half of all cases occurring in men over 70. men from families with a history of prostate cancer are at higher risk than normal. race also has*

Reference: *an effect: men of afro-caribbean descent are about twice as likely to get it whereas men of asian descent have a lower risk of prostate cancer. some evidence suggests that a diet high in tomatoes, vitamin e, cruciform vegetables (such as broccoli, cabbage, cauliflower and brussels sprouts) and selenium may reduce the risk of prostate cancer. however, other studies have failed to confirm these effects, so the findings of this prostate cancer research have*

ProtoT (BLEU 0.0359): a higher risk of prostate cancer. the risk of prostate cancer is higher in men than women. the risk of prostate cancer is higher in men than women. the risk of prostate cancer is higher in men than in women. the risk of prostate cancer is

DeltaNet (BLEU 0.0334): a higher risk of developing prostate cancer. the risk of prostate cancer is higher in women with a high risk of developing the disease. the risk of prostate cancer is higher in women with a high risk of developing the disease. the risk of prostate cancer is

LLaMA (BLEU 0.0000): a risk. men from family with a history of prostate cancer are at higher risk than normal. men from family with a history of prostate cancer are at higher risk than normal. men from family with a history of prostate cancer are at higher risk than normal

Mamba (BLEU 0.0000): a significant impact on the risk of developing prostate cancer. african american men are at a higher risk of developing prostate cancer than caucasian men. african american men are at a higher risk of developing prostate cancer than caucasian men. caucasian men are at

B.3.2 PROMPT — LLAMA WINNER

Prompt: *.. grow out of control to form a mass or tumour. it is a very variable disease. some tumours remain small and grow so slowly that they cause no problems for the rest of a man's life; others are aggressive, grow quickly and become life-threatening. many of these aggressive cases will eventually spread to the bones, where they can cause severe pain. prostate cancer is now the most commonly diagnosed male cancer in many western countries. currently there are 37,000 new cases a year*

Reference: *in the uk; 218,000 in the usa; 25,000 in canada and 17,000 new cases a year in australia. the number of recorded cases has increased a lot in recent years. this is partly due to the increased use of the psa test, which has resulted in more cases being detected, and partly due to the fact that men are living longer. like most cancers, prostate cancer is more common in those over 60. main symptoms include: however, all of these symptoms*

LLaMA (BLEU 0.0228): there are more than 1,000 new cases a year. there are more than 1,000 new cases a year. there are more than 1,000 new cases a year. there are more than 1,000 new cases a year. there

Mamba (BLEU 0.0203): in the uk. the number of new cases is expected to rise to more than 1,000 by 2050. the disease is also becoming more common in the uk, with more than half of all new cases diagnosed in the uk in the last decade

ProtoT (BLEU 0.0000): in the united states. prostate cancer is the most common cancer in men and women. it is the most common cancer in men and women. it is the most common cancer in men and women. it is the most common cancer in men and women.

DeltaNet (BLEU 0.0000): in the u. the most common form of the disease is prostate cancer. the most common form of the disease is prostate cancer. the most common form of the disease is prostate cancer. the most common form of the disease is prostate cancer. the most

B.3.3 PROMPT — DELTANET WINNER

Prompt: *averages and comparatives – hotter, colder, more, less, the highest, the lowest and so on. the students also learnt the basic elements of the water cycle. the biology unit included learning the characteristics of things that are alive: movement, the ability to eat and drink, the ability to breathe, reproduction (plants only !!), growth and excretion. the students were introduced to the process of photosynthesis to illustrate the fact that plants make their own food and ex*

Reference: *crete oxygen and water. this was a completely new concept for them and they found it fascinating. the students also learnt how insects pollinate plants to create new seeds. the phonics programme used the multisensory jolly phonics system, which we have introduced to all our beginner classes. each letter sound has an accompanying action. students and teachers have fun miming out words and students discover that they are able to spell long words if they listen for individual sounds and that they can decode*

DeltaNet (BLEU 0.0143): crete it. the students also learned how to use the system to make their own food. the students also learned how to use the system to make their own food. the students also learned how to use the system to make their own food. the students

ProtoT (BLEU 0.0134): cretion. the students also learned about the different types of plants and animals. the students also learned about the different types of plants and animals. the students also learned about the different types of plants and animals. the students also learned about the different types

Mamba (BLEU 0.0114): crete it. the students also learned about the process of photosynthesis and how it works. the students also learned about the process of photosynthesis and how it works. the students also learned about the process of photosynthesis and

LLaMA (BLEU 0.0000): eat them. the students learnt the importance of the importance of the water cycle to explain how it is important to the life of the plants. the students learnt the importance of the water cycle to explain how it is important to the life of the plants.

B.3.4 PROMPT — MAMBA WINNER

Prompt: *babies that are smarter, taller or better athletes. hair and eye color could be manipulated. iqs could be boosted or lowered. it raises all kind of brave new world issues about genetically engineering the human race. moratorium gains momentum in the last week or so, there's been a flurry of statements from several groups of scientists warning about all this. mit's technology review had an in-depth report on the whole issue a couple of weeks back*

Reference: *, if you want to learn more. this week, groups that include the university of california's jennifer doudna, one of the researchers who developed crispr, essentially called for a moratorium on any attempt to do modification of the human germline using these techniques – at least until there's been more time for public discussion and more research to understand how well it works and how safe it is. in interviews, several of the scientists and bioethicists issuing these statements*

Mamba (BLEU 0.0135): the report, which was published in the journal science, was a response to a recent study by the university of california, berkeley, which found that the number of people who were genetically engineered to be genetically engineered was about 1 in 4.

LLaMA (BLEU 0.0000): the report was published in the journal science. the report was published in the journal science. the report was published in the journal science. the report was published in the journal science. the report was published in the journal science. the report was published

ProtoT (BLEU 0.0000): the report, "the future of science," was released on thursday. the report, which was released on thursday, was based on a study of the effects of the covid-19 pandemic on the human body. the study found that

DeltaNet (BLEU 0.0000): the report was published in the journal of the american society. the report was published in the journal of the american society. the report was published in the journal of the american society. the report was published in the journal of the american society. the