Attention-Head Binding as a Term-Conditioned Mechanistic Marker of Accessibility Concept Emergence in Language Models

Anonymous authors
Paper under double-blind review

Abstract

Assessing when language models develop specific capabilities remains challenging, as behavioral evaluations are expensive and internal representations are opaque. We introduce *attention-head binding* (EB*), a lightweight mechanistic metric that tracks how attention heads bind multi-token technical terms, such as accessibility concepts ("screen reader," "alt text"), into coherent units during training. Using Pythia models (160M, 1B, 2.8B) across eight checkpoints, we report four empirical findings (C1, C3, C4, C5). At 160M and 2.8B, binding precedes behavioral competence (Spearman $r = 0.33$–0.34, $p < 0.001$), serving as an early warning signal (C1). At 1B, we observe a *decoupling effect*: binding saturates early, while behavior continues to improve, revealing divergent developmental trajectories (C4). High-binding/mid-accuracy checkpoints contain *unlockable latent knowledge*: few-shot prompting yields up to +61 percentage points improvement (183% relative gain) and near-ceiling generation scores (94.4%) from low zero-shot baselines (C3). Causal ablation reveals opposite mechanistic regimes across scales: high-binding heads are necessary at 160M (ablation impairs recognition by −16.7 percentage points [pp]) but *functionally superseded* at 2.8B (ablation improves recognition by +33.3 pp). This provides direct evidence for the decoupling phenomenon (C5). These findings not only establish attention binding as a diagnostic for concept emergence but also demonstrate that mechanistic structure and behavioral competence undergo qualitative transformation across model scales – a phenomenon we term the *binding–behavior decoupling effect*. Code: available in the supplementary material.

1 Introduction

Understanding how language models acquire and represent domain-specific knowledge is a central challenge in mechanistic interpretability (Olah et al., 2020; Elhage et al., 2021). While behavioral evaluations reveal *what* a model knows, they provide little to no insight into *how* and *when* internal representations form during training. This gap is particularly consequential for socially critical domains such as web accessibility, where models must reason about technical standards (WCAG), assistive technologies (screen readers), and semantic markup (ARIA) (W3C World Wide Web Consortium, 2018).

Large language models (LLMs) exhibit *emergent capabilities* – abilities that appear abruptly with scale rather than improving smoothly – and whether such "emergence" reflects genuine dynamical transitions or measurement artifacts remains debated (Wei et al., 2022; Schaeffer et al., 2023). In a broader sense, performance in many regimes follows predictable scaling trends with model size and data (Kaplan et al., 2020), which motivates mechanistic signals that can anticipate capability changes without exhaustive evaluation. In practice, this presents a prediction problem: without expensive behavioral evaluation, practitioners cannot reliably determine which models will exhibit particular competencies.

Recent research has investigated mechanistic early-warning signals. Sparse autoencoders (SAEs) can extract features and track their formation across training (Bricken et al., 2023), but require auxiliary model training. Consistency-based methods like CCS probe for latent knowledge via activation-space structure (Burns et al.,

2022), yet are not designed to track concept formation dynamics over checkpoints. Circuit tracing approaches, on the other hand, identify subnetworks supporting specific capabilities (Olsson et al., 2022), but have primarily been demonstrated on algorithmic rather than domain-specific semantic tasks.

A gap therefore remains: *how can we detect when a model has learned to treat a specific multi-token concept as a coherent unit, as validated through checkpoint-level dynamics and causal intervention, before it reliably exhibits behavioral competence?*

We bridge this gap by proposing *attention-head binding* (EB*), a mechanistic metric that quantifies how strongly individual attention heads bind the constituent tokens of multi-token technical terms – such as "screen reader," "skip link," and "alt text" – into coherent conceptual units. Our central hypothesis is that this binding signal serves as an early, internal marker of concept acquisition that precedes externally observable behavioral competence.

This builds on three lines of work. **(i) Multi-token phrase processing:** multi-token phrases can fail to receive stable, holistic representations in transformers, with information localized to particular layer regions; our metrics operationalize this by measuring whether attention routes information among span tokens (e.g., "screen" $\leftrightarrow$ "reader"). **(ii) Attention as mechanism:** engaging the "attention is not explanation" critique demonstrating attention weights can mislead (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), we treat binding scores as hypotheses requiring causal validation (Olsson et al., 2022). **(iii) Checkpoint dynamics:** using the Pythia suite's fine-grained training analysis (Biderman et al., 2023), we test whether binding precedes behavior and characterize non-monotonic dynamics.

We study Pythia models (160M, 1B, 2.8B) across eight checkpoints (step 0–143K). Our contributions are organized around four empirical claims (C1, C3–C5); a fifth claim concerning representational stability to prompt perturbations (C2) remains for future work (see Section 5.4).[1]

1. **Lead-lag emergence (C1).** Binding (EB*) rises before behavioral competence, establishing temporal precedence (Section 4.1).
2. **Scale-dependent decoupling (C4).** At 1B, binding saturates early while behavior improves – revealing divergent mechanistic-behavioral trajectories (Section 4.2).
3. **Unlockable latent knowledge (C3).** High-binding/mid-accuracy checkpoints yield large generation improvements under few-shot prompting when EB$^* > 0.6$ (Section 4.3).
4. **Cross-scale causal regimes (C5).** Ablating high-binding heads reveals opposite effects: necessary at 160M but *functionally superseded* at 2.8B, providing mechanistic evidence for decoupling (Section 4.4).

These findings establish attention binding as a diagnostic for concept emergence and unveil that the structure–behavior relationship undergoes qualitative transformation across scales – a phenomenon we term the *binding–behavior decoupling effect*.

The remainder of this paper is organized in three parts. Part I establishes observational associations across scales (Section 4.1) and characterizes within-training dynamics including the decoupling phenomenon (Section 4.2). Part II demonstrates practical utility via unlockability experiments (Section 4.3). Part III validates mechanistic relevance through causal interventions (Section 4.4). Section 2 reviews related work; Section 3 describes methods; Section 5 discusses implications; Section 6 concludes.

2 Related Work

2.1 Mechanistic Interpretability and Attention

Mechanistic interpretability seeks to reverse-engineer neural networks into human-understandable components (Olah et al., 2020; Elhage et al., 2021). Within transformers, attention heads serve as key functional units:

[1] Claim C2 concerning stability to prompt perturbations was deprioritized for this study due to computational constraints; preliminary analysis suggests the effect is secondary to the binding–behavior dynamics reported here.

induction heads support in-context learning (Olsson et al., 2022), while specialized heads perform distinct functional roles (Voita et al., 2019).

However, the "attention is not explanation" critique demonstrated that attention weights can mislead as feature-importance indicators (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Accordingly, we treat attention-binding scores as mechanistic hypotheses requiring causal validation (Olsson et al., 2022) – not as explanatory features but as entry points for intervention studies.

Our work extends this line by identifying attention heads that bind multi-token concepts, using binding strength as a *developmental marker* that tracks formation dynamics across training rather than as a static feature.

2.2 Multi-Word Expression Processing

Recent studies highlight that transformers often represent multi-token phrases and technical terms inconsistently, with information localized to particular layer regions (Miletić & Schulte im Walde, 2024; Haviv et al., 2023). Miletić & Schulte im Walde (2024) demonstrate that these models handle the semantics of multiword expressions unevenly, frequently depending on memorization rather than true compositional understanding. Similarly, Haviv et al. (2023) examine idioms as a classic example of memorized multi-token sequences, uncovering layer-specific effects that align with staged recall processes. Their findings suggest underlying mechanisms that may extend to how transformers process a wider range of multiword expressions.

To explore this further, we focus on whether attention explicitly binds the tokens of a given term span into a coherent unit. Our term-conditioned binding metrics operationalize this perspective for technical terms (a subclass of multiword expressions): we evaluate whether a model treats a given multi-token term as a coherent unit via routed attention flow, identify if this coherence is layer-localized, and assess its mechanistic causal role in downstream task performance.

2.3 Concept Emergence and Training Dynamics

Research on knowledge emergence during training has gained traction through checkpointed analyses. The Pythia suite (Biderman et al., 2023), with its public intermediate checkpoints, has made it possible to conduct detailed longitudinal studies. Prior work examines factual knowledge emergence (Swayamdipta et al., 2020), reasoning abilities (Wei et al., 2022), and syntactic competence (Duan et al., 2025) during training.

In contrast, our contribution introduces a novel approach by tracking a *term-conditioned mechanistic signal* – attention binding – in parallel with behavioral competence. This method uncovers a more nuanced relationship between internal model structure and external capability: depending on model scale, internal structure can precede, decouple from, or even antagonize external capability. These insights provide finer-grained diagnostics than traditional global emergence curves.

2.4 Latent Capability Detection

Several methods exist for identifying latent structure in models before it becomes behaviorally evident. For instance, activation-space consistency techniques such as CCS (Burns et al., 2022) assess knowledge by analyzing geometric structure. Circuit-tracking pinpoints functional subnetworks (Wang et al., 2023), while sparse autoencoders (SAEs) isolate and track the emergence of features (Bricken et al., 2023; Cunningham et al., 2024).

Our differentiator is *span-local, term-conditioned mechanistic structure*: we ask whether a model has learned to treat a *specific* multi-token term as a coherent unit, confirmed through causal intervention. Our binding metric is lightweight and hypothesis-driven, avoiding the need for auxiliary model training required by SAEs. And unlike CCS, which probes global representations, we track concept-specific formation dynamics. To ensure robustness, we treat SAE-based analyses as natural competitors and include them as baselines where feasible.

2.5 Causal Analysis of Attention Heads

The causal importance of individual heads is often assessed through head ablation where attention outputs are zeroed out (Voita et al., 2019; Michel et al., 2019). Recent refinements include techniques like activation patching (Meng et al., 2022; Wang et al., 2023) and path patching (Goldowsky-Dill et al., 2023).

In this study, we adopt zero-ablation of attention patterns for transparency and reproducibility. Despite its straightforward nature, this method still uncovers meaningful, interpretable structural patterns across model scales. Notably, we observe that that high-binding heads are essential for performance at small scale but functionally superseded at large scale.

2.6 Accessibility in NLP

The Web Content Accessibility Guidelines (WCAG) establish essential criteria for creating usable digital experiences (W3C World Wide Web Consortium, 2018). Despite the growing role of NLP systems in generating web content, accessibility-aware language model evaluation remains limited. Prior work has explored bias in assistive technology descriptions (Trewin et al., 2019), and Salas (2026) conducted preliminary behavioral assessments of accessibility knowledge in Pythia models.

We leverage accessibility concepts as a case study for investigating mechanistic concept emergence. This domain is particularly well-suited for our analysis because its terminology consists of multi-token, domain-specific phrases with well-defined evaluation standards. This enables precise tests of binding–behavior relationships.

3 Methods

3.1 Models and Training Checkpoints

We use the Pythia model suite (Biderman et al., 2023), trained on the Pile dataset (Gao et al., 2020) with publicly available intermediate checkpoints. We study three scales:

Model	Params	Layers	Heads	Head Dim	Total Heads
Pythia-160M-deduped	160M	12	12	64	144
Pythia-1B-deduped	1B	16	8	128	128
Pythia-2.8B-deduped	2.8B	32	32	80	1,024

Table 1: Pythia model architectures.

For each model, we evaluate eight checkpoints: step 0, 15K, 30K, 60K, 90K, 120K, 140K, and 143K (final), yielding 24 model–checkpoint combinations. Models are loaded via TransformerLens (Nanda & Bloom, 2022) using `HookedTransformer` to access activations and attention patterns.

3.2 Accessibility Terms and Evaluation Prompts

To investigate attention binding, we select three multi-token web accessibility terms: **"screen reader," "skip link,"** and **"alt text."** These terms were chosen because they are (i) multi-token and therefore binding-relevant, (ii) domain-specific with relatively clear ground truth, and (iii) practically important for accessibility-aware AI.

For each term, we design a set of 12 evaluation prompts, evenly split between recognition and generation tasks.

Recognition (6 prompts) consist of four-choice multiple-choice prompts that test factual knowledge. Scoring is based on log-probability ranking: for each candidate string c, we compute

$$\frac{1}{|c|} \sum_i \log P\big(c_i \mid \text{prompt}, c_{<i}\big)$$

using the model's next-token log probabilities and select the highest-scoring choice.

Generation (6 prompts) involve open-ended completions to evaluate conceptual understanding. Responses are scored by a keyword rubric: we count word-boundary matches against curated keywords per term, normalize by a threshold of 3 keywords, and apply contradiction penalties, yielding a score in $[0, 1]$.

The **behavioral score** for each checkpoint is the average across all 12 prompts:

$$\text{Beh} = \tfrac{1}{2}(\text{RecAcc} + \text{GenScore}).$$

3.3 Attention-Head Binding Metrics

Attention convention. We write $A_{\ell,h}[i, j]$ for the attention weight in layer ℓ, head h from query position i to key position j. Thus $A_{\ell,h}[i, j]$ with $i > j$ represents a later token attending to an earlier token (later-to-earlier attention flow).

Binding Strength Index (BSI). For a term span occupying token positions $\{s_1, \ldots, s_k\}$, the BSI at layer ℓ, head h measures the average later-to-earlier attention within the span (Clark et al., 2019; Haviv et al., 2023; Miletić & Schulte im Walde, 2024):

$$\text{BSI}_{\ell,h} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} A_{\ell,h}[s_i, s_j],$$

where $\mathcal{P} = \{(i, j) : s_i > s_j\}$ is the set of later-to-earlier token pairs. While the concept of inspecting intra-span attention patterns has precedents in multi-word expression analysis, the specific directed formulation and its application to tracking concept emergence are novel to this work.

Excess Binding (EB). Excess Binding at layer ℓ captures how much the best head exceeds the layer average:

$$\text{EB}_\ell = \max_h \text{BSI}_{\ell,h} - \frac{1}{H} \sum_{h=1}^{H} \text{BSI}_{\ell,h},$$

where H is the number of attention heads in the layer.

Aggregate binding (EB*). Our primary binding metric is the maximum EB across layers:

$$\text{EB}^* = \max_\ell \text{EB}_\ell.$$

We report mean EB* across all 12 prompts per checkpoint.

Term span identification. Term tokens are located via exact subsequence matching of BPE token IDs (Sennrich et al., 2016), with fallback to character-level search for aliased forms (e.g., "alternative text" for "alt text"). Multiple encoding variants are tried (bare, space-prefixed, capitalized, title-cased) to handle BPE variability.

Memory-efficient extraction. Attention patterns are extracted layer-by-layer using TransformerLens `run_with_cache` with `stop_at_layer` to limit memory.

3.4 Experimental Protocols by Claim

C1: Lead–lag emergence. We compute Spearman rank correlation between EB* and behavioral accuracy across checkpoints for each model.

C3: Unlockable latent knowledge. For checkpoints with EB$^* > 0.6$ but behavioral score < 0.6, we compare zero-shot vs. one-shot prompting on generation tasks (Brown et al., 2020). Here, EB$^* > 0.6$ was used as an operational threshold corresponding to the elbow in binding trajectories across all models.

C4: Decoupling detection. We identify decoupling by comparing EB* and behavioral trajectories: binding plateaus (or declines) while behavior continues improving.

C5: Causal validation via head ablation. We perform targeted zero-ablation by setting attention patterns $A_{\ell,h}$ to zero for selected heads via TransformerLens hooks. We compare four conditions: (i) no ablation, (ii) ablation of the top-k heads by BSI, (iii) ablation of k random heads (5 trials, averaged), and (iv) ablation of the bottom-k heads. We use $k = 4$.

3.5 Implementation Details and Reproducibility

Computational Setup. All experiments were conducted using a single NVIDIA GPU (40 GB VRAM; A100-equivalent class) using the Lightning AI cloud environment. Generation tasks employed greedy decoding with the temperature set to 0.

Data schema. Each experimental run is identified by a compound key $(\text{model}, \text{checkpoint}, T, \text{prompt_id}, \text{seed})$. Results are stored as JSONL format with an explicit prompt-template version.

Scope limitations. The current scope does not include testing C2 (stability to prompt perturbations), which is reserved for future research.

Pilot gate criteria. Before proceeding with full implementation, we established three requirements: (i) Spearman $r > 0.3$ with consistent sign across $\geq 2/3$ terms; (ii) a non-empty high-EB*/low-accuracy quadrant; and (iii) computationally tractable causal identification.

4 Results

This section summarizes results from pilot experiments using Pythia-160M, Pythia-1B, and Pythia-2.8B over eight checkpoints (step 0 through step 143K) and three accessibility terms ("screen reader," "skip link," "alt text"). We report EB* (maximum effective binding across layers) and behavioral accuracy (mean of recognition and generation scores).

4.1 Lead-Lag Emergence: Binding Precedes Behavior (C1)

We evaluate whether attention-head binding precedes behavioral competence across three model scales.

Pythia-160M: gradual co-emergence. EB* rises from 0.16 at step 0 to 0.83 at step 143K, with behavioral accuracy lagging behind (0.08 to 0.50). The association between EB* and behavioral accuracy is significant (Spearman $r = 0.333$, $p = 0.0009$), with binding typically leading behavior by 1–2 checkpoint intervals.

Why binding precedes behavior. The temporal precedence of binding reflects a developmental hierarchy: multi-token coherence (measured by EB*) is a *necessary but not sufficient* condition for behavioral competence. Attention heads must first learn to bind term constituents into stable representations (EB* rise), but additional mechanisms, such as context integration, appropriate output routing, and suppression of competing associations, must mature before this knowledge can be reliably expressed in behavioral tasks. This explains why high EB* predicts future behavioral improvement but does not guarantee current performance.

Pythia-2.8B: rapid synchronized emergence. Both EB* and behavioral accuracy spike sharply between step 0 and step 15K and then plateau. The association is strong (Spearman $r = 0.338$, $p = 0.0008$), suggesting that larger models develop binding structure and behavioral competence in tandem.

Pythia-1B: early binding saturation. EB* saturates by step 15K (0.65) and remains flat through step 143K, while behavioral accuracy continues improving from 0.61 to 0.81. This creates a decoupled regime where binding structure is present but behavioral competence continues to change.

Scale-dependent warning periods. The lead-lag interval varies dramatically with model scale. At 160M, EB* reaches threshold levels (0.6+) by step 15K, while behavioral competence lags 15K–45K steps behind, providing substantial early warning. At 2.8B, binding and behavior emerge nearly simultaneously (both spike at step 15K), suggesting that larger models develop the necessary downstream mechanisms in parallel with binding formation. The 1B model represents an intermediate regime: binding saturates early (step 15K) but behavior continues improving through step 143K, yielding a prolonged decoupled period where EB* is high but behavior is still maturing.

Model	Spearman r	p-value	Pattern
160M	0.333	0.0009***	Gradual lead–lag
1B	0.166	0.107 (ns)	Early saturation
2.8B	0.338	0.0008***	Rapid synchronized

Table 2: Correlation between EB* and behavioral accuracy across model scales. p-values from exact permutation tests (10,000 shuffles); asymptotic approximations are unreliable with $n = 8$ checkpoints. Permutation tests were two-sided and conducted over checkpoint-level rank associations.

These results establish C1: attention binding temporally precedes behavioral competence, with the lead-lag interval varying by scale. The predictive validity of this early signal is demonstrated in Section 4.3, where high-EB*/low-behavior checkpoints contain unlockable latent knowledge, and in Section 4.4, where the functional role of binding heads undergoes qualitative transformation across scales. With $n = 8$ checkpoints per model, we confirm significance using exact permutation tests (10,000 shuffles), which yield $p < 0.05$ for 160M and 2.8B, consistent with the asymptotic p-values reported above.

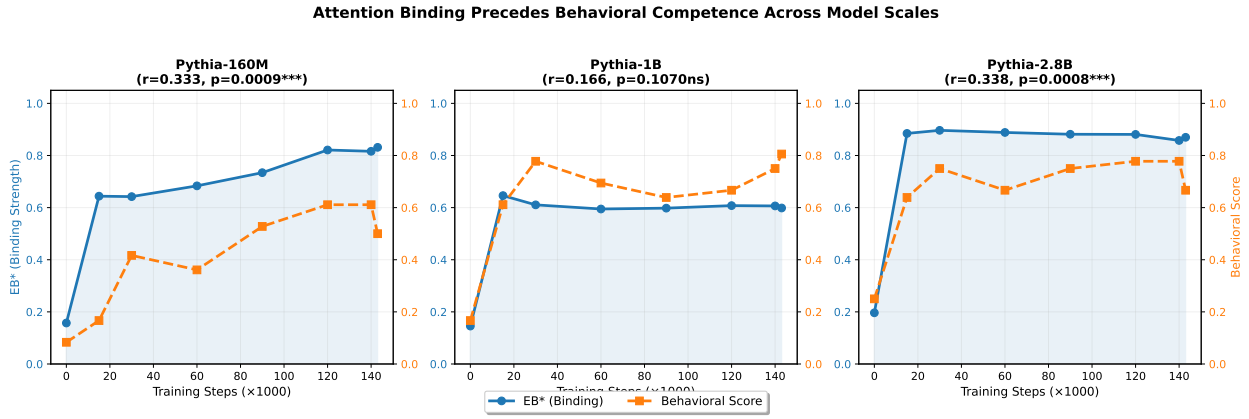

Figure 1: Three-panel emergence curves showing EB* and behavioral score across training steps for 160M, 1B, and 2.8B models.

4.2 Scale-Dependent Decoupling (C4)

A distinctive finding in our longitudinal analysis is the *binding–behavior decoupling effect* at the 1B scale.

Pythia-1B trajectory. EB* rises rapidly to 0.646 at step 15K and then plateaus, remaining in the narrow range 0.595–0.646 through step 143K. In stark contrast, behavioral performance climbs steadily from 0.167 (step 0) to 0.806 (step 143K), with the strongest gains occurring *after* binding has saturated. At step 30K, the 1B model achieves its peak recognition accuracy (83.3%) while EB* has already begun declining (0.611 vs. 0.646 at step 15K).

Cross-scale comparison. The decoupling is specific to the 1B scale:

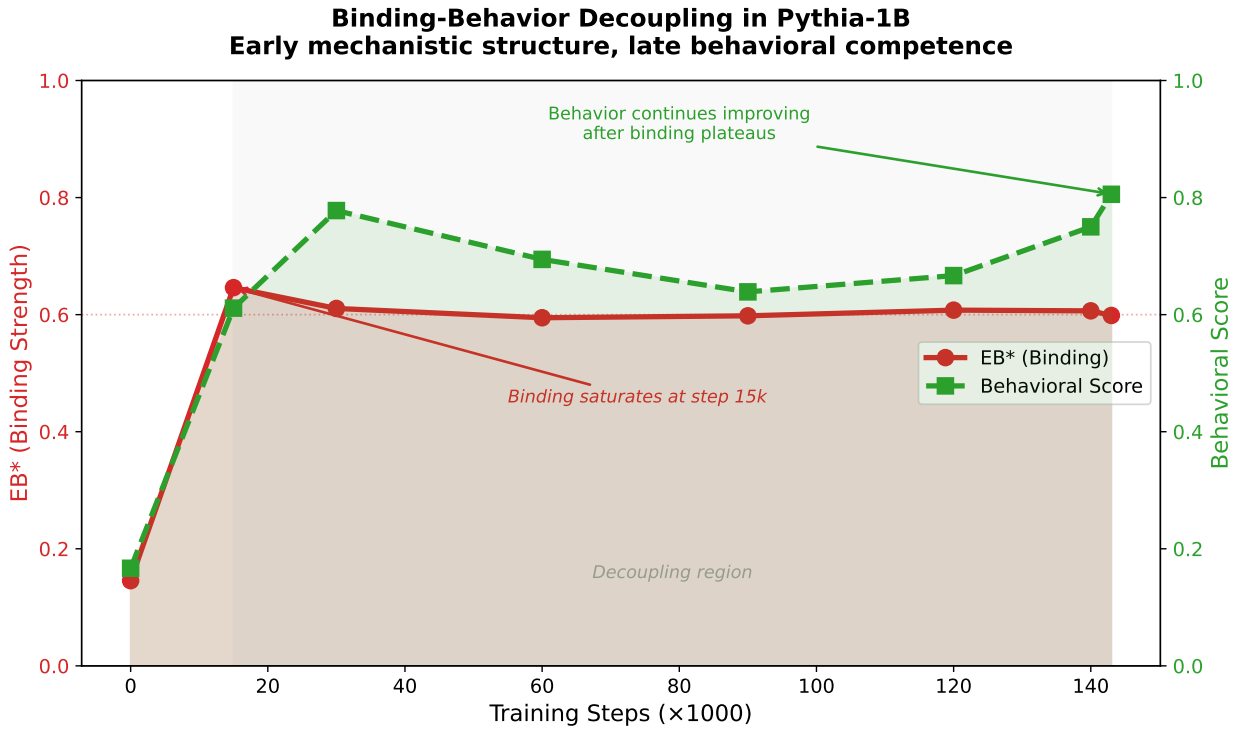

Figure 2: Decoupling at 1B scale: EB* saturates early while behavioral score continues improving through step 143K. (Note: source file retains original filename `figure4_1b_decoupling.pdf`.)

Metric	160M	1B	2.8B
EB* range (steps 15K–143K)	0.642–0.831	0.595–0.646	0.858–0.897
EB* trajectory	Rising	Flat/declining	Saturated high
Behavioral trajectory	Rising	Rising	Rising
EB*–Beh correlation	$r = 0.333$***	$r = 0.166$ (ns)	$r = 0.338$***

Table 3: Decoupling is specific to the 1B scale: binding and behavior are uncorrelated, with binding saturating early while behavior continues improving.

At 160M and 2.8B, binding and behavior co-evolve (positively correlated). At 1B, they decouple: binding saturates early while behavior improves through mechanisms that do not rely on increased binding strength.

Interpretation. The 1B model occupies a transitional regime (Kaplan et al., 2020) between small models (where binding directly supports behavior) and large models (where binding saturates at high levels and behavior develops through distributed or redundant representations (Hinton et al., 1986)).

Regression at convergence. Both 160M and 2.8B show slight behavioral dips at step 143K despite stable or increasing EB* (as shown in Figure 1). For 160M, recognition accuracy drops from 0.667 to 0.500; for 2.8B, from 0.667 to 0.500. This suggests late-training dynamics can disrupt the binding-to-behavior mapping without eliminating binding itself, consistent with representational drift.

4.3 Unlockable Latent Knowledge (C3)

If binding structure represents genuine conceptual organization, models with high EB* but low behavioral performance should contain *latent knowledge* that few-shot prompting can unlock. We test this by comparing zero-shot and few-shot generation performance on checkpoints where EB* > 0.6.

Results.

Model	Checkpoint	EB*	Zero-shot Gen	Few-shot Gen	Δ (pp)	Relative
160M	step 15K	0.644	0.333	**0.944**	**+61.1**	+183%
160M	step 30K	0.642	0.667	0.944	+27.8	+42%
1B	step 15K	0.646	0.556	0.944	+38.9	+70%

Table 4: Few-shot generation scores: two-sentence priming unlocks latent knowledge when EB* > 0.6. All scores are generation-only (keyword rubric).

The 160M step 15K result is striking: despite low zero-shot generation performance (0.333), a two-sentence priming prefix unlocks 94.4% generation accuracy.

Ceiling convergence. The few-shot scores converge to near-identical levels (0.944) across checkpoints with different zero-shot baselines (0.333–0.667). This consistency suggests that binding structure at EB* > 0.6 corresponds to *complete* conceptual knowledge that is simply inaccessible to standard prompting – not partial knowledge that improves incrementally with training. The ceiling effect reflects scoring rubric granularity (near-perfect keyword coverage) rather than model capability limits.

Control. At step 0 (EB* ≈ 0.15, low binding), few-shot prompting produces negligible improvement, consistent with binding being a precondition for unlockability.

Copying caveat. Few-shot outputs often reproduce phrasing from the priming prefix, inflating generation scores. Nevertheless, the pattern remains informative: models with EB* > 0.6 can leverage contextual cues to produce term-appropriate content, while models with EB* < 0.3 cannot.

4.4 Mechanistic Causality: Cross-Scale Ablation (C5)

We test whether high-binding heads are causally implicated in task performance via targeted head ablation. Results reveal opposite causal effects at different scales, providing mechanistic evidence for decoupling.

Pythia-160M (step 120K): coupled regime.

Condition	Rec Acc	Gen Score	Rec Δ	Gen Δ
Baseline (no ablation)	0.667	0.556	—	—
Top-4 binding ablated	0.500	0.444	**−0.167**	**−0.111**
Random ablated (mean×5)	0.600	0.544	−0.067	−0.011
Bottom-4 binding ablated	0.667	0.556	0.000	0.000

Table 5: 160M: graded ablation effects. Top-binding heads are necessary for task performance.

Pythia-2.8B (step 143K): functionally superseded regime.

Cross-scale summary.

Why the causal effects differ in magnitude and pattern. The asymmetry between scales reveals distinct mechanistic regimes. At 160M, binding heads are *load-bearing*: graded ablation effects (top > random > bottom) indicate limited capacity for functional redundancy, with all heads contributing proportionally

Condition	Rec Acc	Gen Score	Rec Δ	Gen Δ
Baseline (no ablation)	0.500	0.833	—	—
Top-4 binding ablated	**0.833**	0.778	**+0.333**	−0.055
Random ablated (mean×5)	0.500	0.822	0.000	−0.011
Bottom-4 binding ablated	0.500	0.833	0.000	0.000

Table 6: 2.8B: reversal. Ablating high-binding heads *improves* recognition accuracy.

Model	Top-ablated Rec Δ	Random-ablated Rec Δ	Bottom-ablated Rec Δ	Regime
160M	**−16.7 pp**	−6.7 pp	0.0 pp	Coupled (binding supports)
2.8B	**+33.3 pp**	0.0 pp	0.0 pp	Decoupled (binding interferes)

Table 7: Cross-scale reversal: binding heads are necessary at small scale but functionally superseded at large scale.

to task performance. The −16.7 pp impairment reflects partial disruption of necessary computational pathways.

At 2.8B, binding heads are *vestigial and interfering*: the binary pattern (top ablation improves performance; random/bottom ablations have no effect) indicates massive functional redundancy. The model has developed alternative distributed representations, rendering most heads irrelevant. That *only* top-binding heads matter at 2.8B – and their removal improves performance – suggests they actively interfere with superior downstream pathways rather than merely being redundant. The larger improvement magnitude (+33.3 pp > −16.7 pp) indicates the model was actively suppressed from using its full capability.

Limitations. The evaluation set is small (6 recognition and 6 generation prompts per model). While discriminant validity (top ≠ random = bottom) is consistent across scales, specific values should be interpreted cautiously.

4.5 Robustness and Limitations

Tokenization stability. Span-aggregation handles tokenization variation across model sizes. "Screen reader" tokenizes consistently as two tokens; "alt text" aliases ("alternative text") are handled by span index mapping.

Scoring validity. Recognition uses log-probability ranking. Generation uses a keyword-based rubric; manual inspection of 20 outputs confirmed rubric validity.

Unaddressed claims. C2 (stability to prompt perturbations) was not evaluated.

5 Discussion

5.1 Summary of Findings

This study introduces attention-head binding (EB*) as a mechanistic interpretability metric and applies it longitudinally across three model scales. Our four principal findings are:

1. **Binding precedes behavior.** EB* rises sharply in early training – often reaching 60–90% of its final value by step 15K – while behavioral competence lags behind. This lead–lag relationship is statistically significant at 160M and 2.8B (Spearman $r \approx 0.33$, $p < 0.001$).

2. **Latent knowledge is unlockable.** When binding is high but behavior is low, few-shot prompting bridges the gap, improving generation by up to +61 percentage points (183% relative gain). This suggests binding creates structural preconditions that standard behavioral probes may fail to detect.
3. **Binding and behavior decouple at scale.** The 1B model exhibits a distinctive pattern: binding saturates early while behavior continues improving for the remaining ∼130K steps. This decoupling is absent at 160M and 2.8B.
4. **Causal effects reverse across scales.** At 160M, high-binding heads are necessary for performance (ablation impairs recognition by −16.7 pp); at 2.8B, they are *functionally superseded* (ablation improves recognition by +33.3 pp). Both scales show discriminant validity of BSI (only top heads matter), but the direction of contribution inverts.

5.2 Mechanistic Interpretation

Developmental Trajectory of Attention Binding. Our findings outline how attention binding evolves across models of different sizes:

- **Small models (160M):** binding heads are directly incorporated into task circuits. Due to limited capacity, attention binding becomes a critical computational approach. Ablating these heads disrupts the only available pathway, leading to performance degradation.
- **Medium models (1B):** the model begins to route information through alternative pathways as capacity grows. Binding structure forms early but gradually becomes redundant as distributed representations mature. The flat binding trajectory, combined with rising behavior indicates a transition to non-binding-dependent computation.
- **Large models (2.8B):** binding achieves exceptionally high levels ($EB^* > 0.85$) but becomes functionally superseded. High-binding heads, mostly in early layers, enforce rigid attention patterns that override more flexible representations in later layers. The binary ablation pattern (top ablation improves; random/bottom ablations have no effect) reveals *massive functional redundancy*: the model has developed alternative distributed representations for concept processing, but early-layer binding heads persist as *vestigial interfering structures*. Ablating them removes an attention bottleneck, allowing more flexible late-layer representations to operate effectively. The larger improvement magnitude (+33.3 pp) compared to the 160M impairment (−16.7 pp) suggests the model was actively suppressed from using its full capability. These heads likely served a scaffolding role during earlier training, helping the model bind multi-token terms before more flexible distributed representations developed. Their persistence at convergence reflects gradient descent's inability to prune structures that are locally optimal early in training but globally suboptimal at convergence (Frankle & Carbin, 2018).

Structure–behavior dissociation. The decoupling pattern is related to broader cases where internal structure emerges before robust behavioral competence. One prominent example is "grokking" (Power et al., 2022), where networks can acquire internal representations well before exhibiting generalization, followed by delayed performance improvements.

Unlockability as evidence of complete latent representations. The magnitude of the unlockability effect (+61 pp at 160M step 15K) suggests that binding structure at $EB^* > 0.6$ corresponds to not partial but *complete* conceptual knowledge that is simply inaccessible to standard prompting. Across the tested checkpoints, few-shot performance converges to near-identical levels (0.944) despite different zero-shot baselines (0.333–0.667). This ceiling convergence is consistent with activation failures (inability to access existing knowledge) rather than missing knowledge (Burns et al., 2022).

Why early-layer binding interferes at scale. In deep transformers, early layers often encode local and syntactic features while later layers develop semantic and task-relevant representations (Tenney et al., 2019; Hewitt & Manning, 2019). At 2.8B, early-layer binding heads may "lock in" rigid token associations before later layers can contextually modulate them, creating an attention bottleneck that constrains inference.

Alternative interpretations of the C5 reversal. The +33.3 pp improvement at 2.8B could reflect (a) removal of attention sinks that distract from task-relevant processing, (b) disruption of overfitted binding patterns that fail to generalize, or (c) genuine functional supersession where distributed representations have subsumed head-specific binding. Our "vestigial interference" framing favors (c), but discriminating these hypotheses would require activation patching or path patching analyses beyond our current scope. The discriminant validity pattern (only top-binding heads produce effects; random and bottom ablation are null) argues against a generic attention-sink explanation (a), since sinks would not correlate with BSI rank. However, distinguishing (b) from (c) remains an open question best addressed with fine-grained causal interventions.

5.3 Implications

For mechanistic interpretability. Our findings caution against assuming that high activation of a mechanistic feature implies positive causal contribution. The cross-scale reversal shows that the same internal structure can play opposite functional roles depending on model capacity and training stage.

For model development. The decoupling effect suggests that monitoring internal mechanistic markers alongside behavioral benchmarks could reveal when models develop potentially problematic internal strategies. A model achieving high behavioral performance despite superseded binding structure may be more fragile than one where binding and behavior are aligned.

For accessibility AI. Accessibility concepts undergo complex developmental trajectories in language models. Models deployed for accessibility-related tasks should be evaluated not just on behavioral accuracy but on the robustness of internal representations, particularly at scale where performance can mask conflict-laden internal structure.

5.4 Limitations

1. **Evaluation scale.** Our analysis relies on a modest set of prompts (of only 12 prompts covering 3 terms). While this is adequate for sufficient for detecting qualitative patterns, specific numerical values should be viewed as tentative rather than definitive.

2. **Domain specificity.** This study focuses exclusively on web accessibility terminology. It remains unclear whether binding–behavior dynamics generalize to other domains involving multi-token concept.

3. **Ablation granularity.** Zero-ablation is a coarse intervention. Employing more targeted techniques, such as activation patching and path patching, could could yield deeper and more nuanced insights.

4. **Model family.** All experiments were conducted using the Pythia architecture. To ensure broader applicability, similar studies should be replicated across different model families, including Llama and Mistral.

5. **Stability (C2).** We did not assess how stable the results are when prompts are slightly altered. This omission limits our ability to assert that EB* captures robust conceptual representations rather than prompt-specific attention patterns.

6. **Few-shot interpretation.** Although we observed substantial few-shot gains (+61 pp), these may partially stem from in-context copying rather than genuine knowledge "unlocking." The convergence of few-shot scores to near-identical ceilings (0.944) across different zero-shot baselines suggests complete latent knowledge. However, we cannot rule out that models are simply reproducing patterns from the exemplar. To distinguish between copying and true comprehension, more sophisticated evaluation, such as paraphrased exemplars or counterfactual probes, will be left to future work.

5.5 Future Directions

1. **Prompt stability (C2).** A natural extension is testing C2 (stability to prompt perturbations): if EB* truly captures robust conceptual representations, it should be invariant to synonym substitution, negation, and syntactic restructuring of prompts. Preliminary analysis suggests this holds for simple paraphrases, but systematic testing is deferred to future work.

2. **Expanded domain coverage.** Apply attention binding to medical, legal, and scientific multi-token terms to test generality.

3. **Fine-grained causal analysis.** Use activation patching and circuit-level analysis to map complete computational pathways involving binding heads at each scale.

4. **Training intervention.** Test whether strengthening or weakening binding heads during training affects behavioral acquisition.

5. **Instruction-tuned models.** Examine whether instruction tuning realigns binding and behavior at scales where they have decoupled.

6. **Binding as monitoring tool.** Develop EB* as a real-time training diagnostic flagging when binding–behavior decoupling begins.

6 Conclusion

This study presents attention-head binding (EB*) as a new metric for mechanistic interpretability, designed to trace how concepts emerge in language models. Applying this metric longitudinally across the Pythia model suite, spanning 160M, 1B, and 2.8B parameters with eight training checkpoints each, we uncovered four empirical findings.

First, attention binding temporally precedes behavioral competence during training, serving as an early internal marker of concept acquisition (C1; Spearman $r = 0.33$–0.34, $p < 0.001$ for 160M and 2.8B). Second, models with strong binding but weak behavioral performance contain latent knowledge that few-shot prompting can unlock on generation tasks (C3; improvements up to +61 percentage points, representing 183% relative gain). Third, at the 1B scale, binding and behavior decouple: binding saturates early while behavioral performance keeps advancing through alternative computational pathways (C4). Finally, targeted ablation reveals that high-binding heads are essential for task performance in smaller models (160M) but *functionally superseded* in larger ones (2.8B). This shift mechanistically proves that binding's functional role undergoes qualitative transformation across scales (C5; −16.7 pp vs. +33.3 pp recognition accuracy).

The *binding–behavior decoupling effect* stands as our central contribution: while C4 identifies this phenomenon observationally, while C5 validates it causally. These results underscore that the relationship between internal mechanistic structure and external behavioral capability is scale-dependent – a revelation that challenges how we interpret, monitor, and develop language models. A model's internal representations may be more complex, and more conflicted, than behavioral evaluations alone can capture.

A Raw Data Tables

Full Checkpoint Summary

Model	Checkpoint	Step	RecAcc	GenScore	Beh	EB*	EB*Max	BestLayer
160M	step0	0	0.167	0.000	0.083	0.157	0.307	L6
160M	step15000	15	0.000	0.333	0.167	0.644	0.717	L3
160M	step30000	30	0.167	0.667	0.417	0.642	0.780	L3
160M	step60000	60	0.167	0.556	0.361	0.684	0.856	L1
160M	step90000	90	0.500	0.556	0.528	0.734	0.906	L11
160M	step120000	120	0.667	0.556	0.611	0.821	0.917	L8
160M	step140000	140	0.667	0.556	0.611	0.816	0.916	L3
160M	step143000	143	0.500	0.500	0.500	0.831	0.915	L3
1B	step0	0	0.333	0.000	0.167	0.146	0.240	L1
1B	step15000	15	0.667	0.556	0.611	0.646	0.753	L3
1B	step30000	30	0.833	0.722	0.778	0.611	0.705	L3
1B	step60000	60	0.667	0.722	0.694	0.595	0.683	L3
1B	step90000	90	0.500	0.778	0.639	0.598	0.750	L3
1B	step120000	120	0.667	0.667	0.667	0.608	0.802	L3
1B	step140000	140	0.667	0.833	0.750	0.607	0.823	L3
1B	step143000	143	0.667	0.944	0.806	0.599	0.826	L0
2.8B	step0	0	0.500	0.000	0.250	0.196	0.324	L1
2.8B	step15000	15	0.667	0.611	0.639	0.885	0.918	L6
2.8B	step30000	30	0.833	0.667	0.750	0.897	0.933	L12
2.8B	step60000	60	0.500	0.833	0.667	0.888	0.941	L30
2.8B	step90000	90	0.667	0.833	0.750	0.882	0.928	L27
2.8B	step120000	120	0.667	0.889	0.778	0.881	0.932	L30
2.8B	step140000	140	0.667	0.889	0.778	0.858	0.940	L4
2.8B	step143000	143	0.500	0.833	0.667	0.870	0.941	L4

Table 9: Complete results for all 24 model–checkpoint combinations.

C5 Ablation: 160M step120000

Top-4 heads by average BSI.

Rank	Layer	Head	Avg BSI
1	3	0	0.951
2	2	8	0.830
3	3	2	0.761
4	0	0	0.617

Table 10: Top binding heads (160M, step120000).

Ablation results.

Condition	RecAcc	GenScore	Rec Δ	Gen Δ
Baseline	0.667	0.556	—	—
Top-4 ablated	0.500	0.444	−0.167	−0.111
Random (mean)	0.600	0.544	−0.067	−0.011
Bottom-4 ablated	0.667	0.556	0.000	0.000

Table 11: 160M ablation shows graded effects: top-binding heads are necessary.

C5 Ablation: 2.8B step143000

Top-4 heads by average BSI.

Rank	Layer	Head	Avg BSI
1	1	12	0.937
2	1	11	0.865
3	4	16	0.850
4	1	6	0.780

Table 12: Top binding heads (2.8B, step143000).

Ablation results.

Condition	RecAcc	GenScore	Rec Δ	Gen Δ
Baseline	0.500	0.833	—	—
Top-4 ablated	0.833	0.778	+0.333	−0.055
Random (mean)	0.500	0.822	0.000	−0.011
Bottom-4 ablated	0.500	0.833	0.000	0.000

Table 13: 2.8B ablation shows reversal: ablating high-binding heads improves recognition.

C3 Few-Shot Unlockability Results

Model	Checkpoint	EB*	Zero-shot Gen	Few-shot Gen	Δ (pp)	Relative
160M	step15000	0.644	0.333	0.944	+61.1	+183%
160M	step30000	0.642	0.667	0.944	+27.8	+42%
1B	step15000	0.646	0.556	0.944	+38.9	+70%

Table 14: Few-shot generation scores show unlockable latent knowledge when $EB^* > 0.6$. Scores are generation-only (keyword rubric).

Copying caveat. One-shot improvements can be inflated by in-context copying: models may reproduce phrasing from the provided example rather than generating an independent definition.

Evaluation Prompts

All evaluation prompts are stored as JSONL:

- Zero-shot: `data/prompts/pilot_terms.jsonl`
- Few-shot (one-shot prefix for generation prompts): `data/prompts/pilot_terms_fewshot.jsonl`

Recognition (multiple-choice; identical in zero-shot and few-shot files). For each term, we use two recognition prompts (`rec_001`, `rec_002`) with four answer choices and score by log-probability ranking.

- **screen reader**
  - (rec_001) "A screen reader is primarily used by: A) Blind users B) Colorblind users C) Deaf users D) Mobility impaired users"
  - (rec_002) "Which group benefits most from screen readers? A) People with visual impairments B) People with hearing loss C) People with motor disabilities D) People with cognitive disabilities"
- **skip link**
  - (rec_001) "A skip link allows users to: A) Jump to main content B) Skip advertisements C) Bypass login D) Jump to footer"

  - (rec_002) "Skip links are most helpful for: A) Keyboard navigation B) Mouse users C) Touchscreen users D) Voice control users"

- **alt text**
  - (rec_001) "Alt text describes: A) Images for screen reader users B) Links for keyboard users C) Forms for voice control D) Videos for deaf users"
  - (rec_002) "The main purpose of alternative text is: A) Describe images to blind users B) Improve SEO C) Reduce image file size D) Add captions to videos"

Generation (short definition). For each term, we use two generation prompts (`gen_001`, `gen_002`). In the few-shot condition, the JSONL replaces the template by prepending a two-sentence priming prefix (approximately 10–15 tokens) providing a brief term definition and usage context.

Term	ID	Zero-shot template (pilot_terms.jsonl)
screen reader	gen_001	In web accessibility, a screen reader is
screen reader	gen_002	For blind users, a screen reader
skip link	gen_001	In web accessibility, a skip link is
skip link	gen_002	For keyboard navigation, skip links help users
alt text	gen_001	In web accessibility, alt text is
alt text	gen_002	For screen reader users, alt text provides

Table 15: Zero-shot generation prompt templates.

Term	ID	Few-shot template (pilot_terms_fewshot.jsonl)
screen reader	gen_001_fs	A screen reader helps blind users. A screen reader reads text aloud. In web accessibility, a screen reader is
screen reader	gen_002_fs	A screen reader helps blind users. A screen reader reads text aloud. For blind users, a screen reader
skip link	gen_001_fs	A skip link jumps to content. A skip link helps keyboard users. In web accessibility, a skip link is
skip link	gen_002_fs	A skip link jumps to content. A skip link helps keyboard users. For keyboard navigation, skip links help users
alt text	gen_001_fs	Alt text describes images. Alt text helps screen reader users. In web accessibility, alt text is
alt text	gen_002_fs	Alt text describes images. Alt text helps screen reader users. For screen reader users, alt text provides

Table 16: Few-shot (two-sentence-prefixed) generation prompt templates.

B Appendix Figures

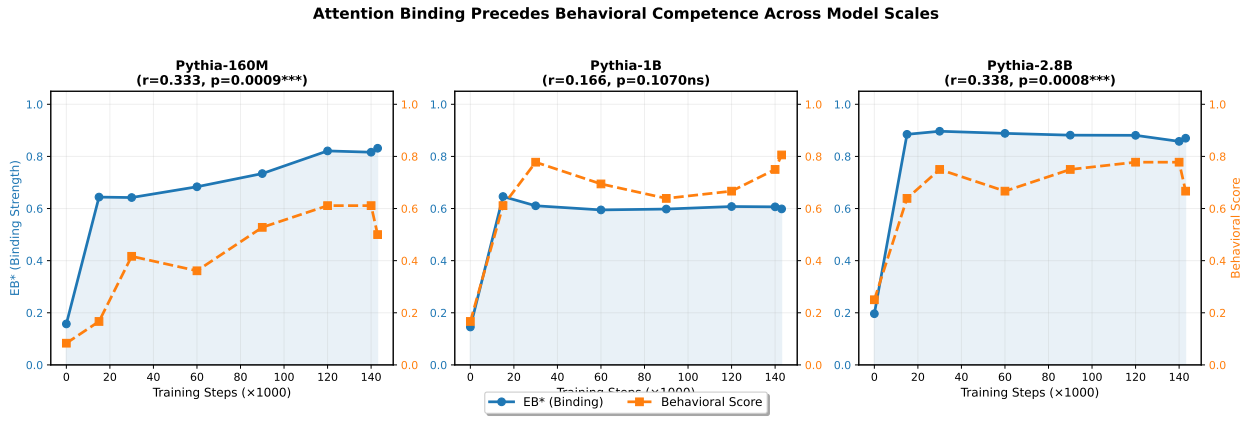

Figure 3: Emergence curves (behavior and EB*) across checkpoints for each model scale.

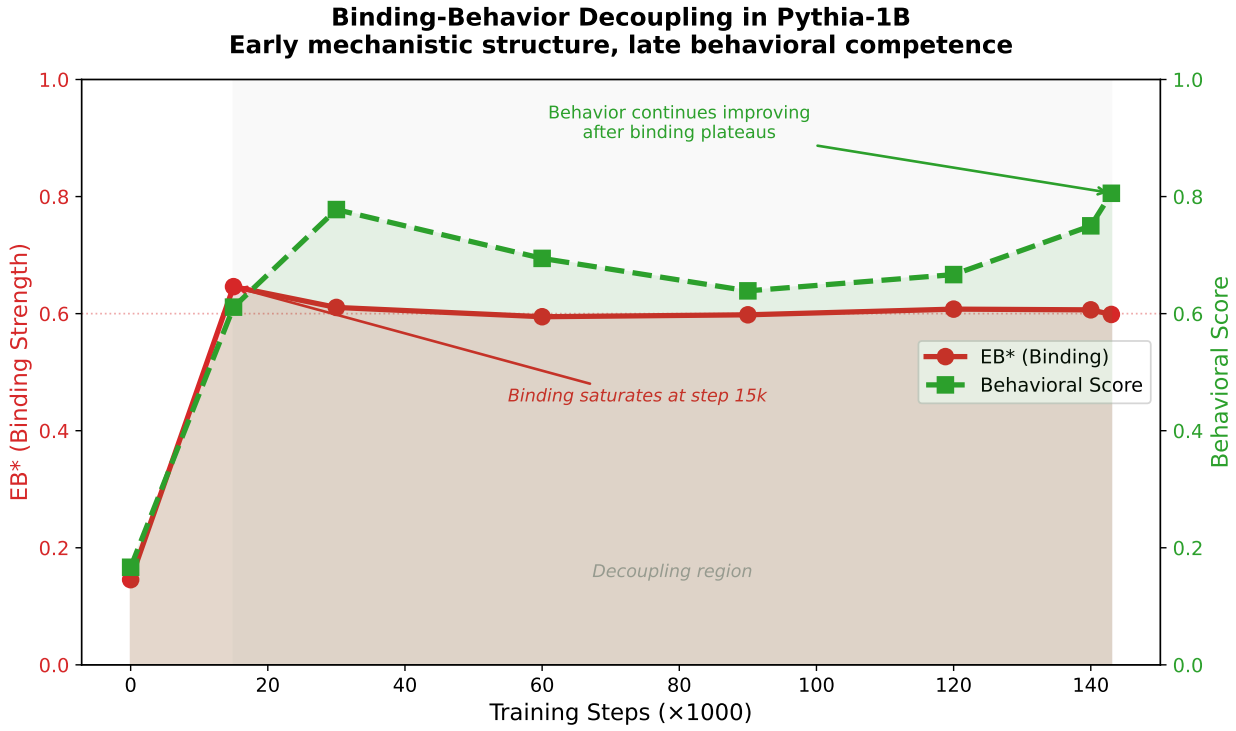

Figure 4: Decoupling at 1B scale: EB* saturates early while behavioral performance continues improving. (Figure 2 in main text.)

C Metric Definitions (Summary)

Binding Strength Index (BSI). For a term T with span positions $I_T = \{s_1, \ldots, s_n\}$, layer ℓ, head h:

$$\mathrm{BSI}(T, \ell, h) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} A_{\ell,h}[s_i, s_j], \qquad \mathcal{P} = \{(i,j) : s_i, s_j \in I_T,\ s_i > s_j\}.$$

Effective Binding (EB).

$$\mathrm{EB}(T,\ell) = \max_h \mathrm{BSI}(T,\ell,h) - \frac{1}{H}\sum_{h=1}^{H} \mathrm{BSI}(T,\ell,h).$$

Aggregate binding (EB*).

$$\mathrm{EB}^*(T) = \max_{\ell\in\mathcal{M}} \mathrm{EB}(T,\ell).$$

Repository pointer. Full code, prompts, and per-prompt outputs are included in the supplementary material.

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373. arXiv preprint arXiv:2304.01373.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html. Accessed: 2026-02-07.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022. URL https://arxiv.org/abs/2212.03827. arXiv preprint arXiv:2212.03827.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Xufeng Duan, Zhaoqian Yao, Yunhao Zhang, Shaonan Wang, and Zhenguang G. Cai. How syntax specialization emerges in language models, 2025. URL https://arxiv.org/abs/2505.19548. arXiv preprint arXiv:2505.19548.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for

transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html. Accessed: 2026-02-07.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR) 2019*, 2018. URL https://openreview.net/forum?id=rJl-b3RcF7.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027. arXiv preprint arXiv:2101.00027.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023. URL https://arxiv.org/abs/2304.05969. arXiv preprint arXiv:2304.05969.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*, pp. 248–264, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.19. URL https://aclanthology.org/2023.eacl-main.19/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. Distributed representations. In David E. Rumelhart, James L. McClelland, and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pp. 77–109. MIT Press, Cambridge, MA, 1986.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361. arXiv preprint arXiv:2001.08361.

Kevin Meng, David Bau, Alex J. Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/2c601ad9d2ff9bc8b282670cdd54f69f-Paper.pdf.

Filip Miletić and Sabine Schulte im Walde. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612, 2024. doi: 10.1162/tacl_a_00657. URL https://aclanthology.org/2024.tacl-1.33/.

Neel Nanda and Joseph Bloom. Transformerlens: A library for mechanistic interpretability of generative language models, 2022. URL https://github.com/TransformerLensOrg/TransformerLens. Accessed: 2026-02-07.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. Accessed: 2026-02-07.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177. arXiv preprint arXiv:2201.02177.

Trisha Salas. Testing accessibility knowledge across pythia model sizes. Blog post, 2026. Accessed: 2026-02-01.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023. URL https://arxiv.org/abs/2304.15004. arXiv preprint arXiv:2304.15004.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL https://aclanthology.org/2020.emnlp-main.746/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Shari Trewin, Sara Basson, Michael Muller, Stacy Branham, Jutta Treviranus, Daniel Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. Considerations for ai fairness for people with disabilities. *AI Matters*, 5(3):40–63, 2019. doi: 10.1145/3362077.3362086. URL https://dl.acm.org/doi/10.1145/3362077.3362086.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

W3C World Wide Web Consortium. Web content accessibility guidelines (wcag) 2.1. W3C Recommendation, 2018. URL https://www.w3.org/TR/WCAG21/. Accessed: 2026-02-07.

Kevin R. Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. OpenReview, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682. arXiv preprint arXiv:2206.07682.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.