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 and nine accessibility terms (432 model-checkpoint-term observations), we report five empirical findings. Discriminant validity analysis validates EB* against token co-occurrence baselines, establishing a gradient from true nonsense controls (EB$^* = 0.26$) through rare pairs (0.50) to real accessibility terms (0.74), all $p < 0.001$, Cohen's $d = 1.2$–2.9; domain-adjacent and wrong-domain controls further characterize scale-dependent precision limits. At 160M and 2.8B, binding precedes behavioral competence (early coupling $\rho = +0.57$, $p < 0.001$; late decoupling $\rho = -0.20$, $p = 0.01$), 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heads are necessary at 160M (ablation impairs recognition by -16.7 percentage points [pp]) but *functionally superseded* at 2.8B (ablation improves recognition by +33.3 pp).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* that appear abruptly with scale rather than improving smoothly, and debate continues over whether such "emergence" reflects genuine dynamical transitions or merely measurement artifacts (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, 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:** addressing the "attention is not explanation" critique, which demonstrated that 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) and nine accessibility terms. Our contributions are organized around five empirical claims; a sixth claim concerning representational stability to prompt perturbations (C2) remains for future work (see Section 5.4).[1]

1. **Discriminant validity.** EB* is validated against token co-occurrence baselines through iterative control design. Redesigned genuine-nonsense controls (v2) establish a clear gradient ($d = 1.2$–2.9, all $p < 0.001$). Domain-adjacent and wrong-domain controls (v3/v4) reveal scale-dependent discrimination failure at 160M and partial discrimination at 1B, characterizing EB*'s precision limits (Section 4.1).

2. **Lead-lag emergence (C1).** The binding-behavior relationship undergoes a phase transition: strong early coupling ($\rho = +0.57$, $p < 0.001$, $n = 108$) reverses to decoupling at trained checkpoints ($\rho = -0.20$, $p = 0.01$, $n = 162$), with the lifecycle accelerated at larger scales (Section 4.3).

3. **Scale-dependent decoupling (C4).** At 1B, binding saturates early while behavior improves, revealing divergent mechanistic-behavioral trajectories (Section 4.4).

4. **Unlockable latent knowledge (C3).** High-binding/mid-accuracy checkpoints yield large generation improvements under few-shot prompting when EB$^* > 0.6$, replicated across 9 terms and 99 prompts (Section 4.5).

5. **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 4.6).

[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.

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 as follows. Section 2 reviews related work and positions EB* against existing approaches. Section 3 describes methods. Section 4 reports results: discriminant validity (Section 4.1), dataset expansion and robustness (Section 4.2), coupling-decoupling lifecycle (Section 4.3), scale-dependent decoupling (Section 4.4), unlockability (Section 4.5), and cross-scale ablation (Section 4.6). Section 5 discusses implications, limitations, and future directions. Section 6 concludes.

2 Related Work

2.1 Positioning: Why Mechanistic Analysis of Multi-Token Concepts?

Three primary approaches exist for studying concept knowledge in language models, each with distinct limitations for multi-token technical terms.

(1) Behavioral probing (Meng et al., 2022; Olsson et al., 2022) measures what models know through task performance. While effective for detecting knowledge presence, behavioral probes provide no insight into *when* knowledge forms during training, *how* it is mechanistically represented, or *why* models with similar behavioral scores may differ in robustness. Recent mechanistic work on LLM training dynamics reveals that internal representations undergo complex reorganization across distinct phases during pretraining, from noisy token-level features through emergent concept-level features to convergent stable representations, with directional drift continuing even after features are semantically formed (Xu et al., 2026). Our discriminant validity experiments (Section 4.1) show that behavioral competence can exist with low binding ("aria attribute": 76% behavioral accuracy, 42% EB*), demonstrating that behavioral probes conflate multiple representational strategies.

(2) Token co-occurrence metrics (e.g., pointwise mutual information, n-gram frequency) measure statistical association in training data. Our control experiments demonstrate these metrics fail to distinguish meaningful conceptual binding from arbitrary token adjacency: initial controls using plausible bigrams such as "keyboard mouse" and "open source" showed EB* = 0.72–0.82, statistically indistinguishable from real accessibility terms ($p > 0.05$). Only genuinely nonsensical controls established discriminant validity (Section 4.1), revealing that EB* captures representational structure beyond corpus statistics.

(3) Single-token concept analysis (Burns et al., 2022; Cunningham et al., 2024) examines how models represent individual concepts through probing and sparse autoencoders. Multi-token technical terms present a fundamentally different challenge: the model must learn to bind constituents ("screen" + "reader") into a coherent unit distinct from other valid compositions ("screen door," "PDF reader"). Recent work shows that LLMs perform detokenization to reconstruct multi-token words (Kaplan et al., 2025), but this process does not explain how conceptual meaning emerges from compositional binding across training.

What our approach enables. By tracking attention binding longitudinally across training, we can: (a) detect concept acquisition *before* behavioral competence emerges, (b) identify representational reorganization invisible to behavioral probes (the coupling-decoupling transition), and (c) causally validate the functional role of binding through ablation, revealing opposite effects at different scales.

2.2 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: 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.

Attention as compositional binding. Recent theoretical work interprets self-attention as implementing vector-symbolic binding operations (Dhayalkar, 2025), where queries and keys define role spaces, values encode fillers, and attention weights perform soft unbinding. While this perspective provides a principled algebraic framework for understanding transformer reasoning, our work offers complementary empirical validation: we track how binding structure develops during training and correlates with behavioral competence, revealing developmental dynamics not captured by static architectural interpretations.

Sparse autoencoders and monosemanticity. An alternative approach uses sparse autoencoders (SAEs) to decompose neural activations into interpretable monosemantic features (Bricken et al., 2023; Templeton et al., 2024). Our approach differs in focus: rather than decomposing activations into atomic features, we track compositional binding of multi-token concepts through attention patterns. These approaches are complementary: SAEs identify what features exist, while attention binding reveals how multi-token concepts are compositionally organized and how this organization evolves during training.

2.3 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.

2.4 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.

Relationship to grokking. The coupling→decoupling transition we observe shares conceptual similarities with "grokking" (Power et al., 2022) in terms of sudden generalization after prolonged memorization in algorithmic tasks. However, grokking describes *behavioral* transitions (from memorization to generalization), while we observe *mechanistic* reorganization (from binding-dependent to distributed representations) that can occur independently of behavioral performance. Our finding that binding can decouple from behavior at larger scales suggests these are distinct phenomena: grokking reflects behavioral phase transitions, while

coupling-decoupling reflects architectural reorganization that may enable, coincide with, or follow behavioral improvements depending on model capacity.

Attention entropy as a measurement tool. Recent work has used attention entropy to characterize attention patterns as focused versus diffuse (Clark et al., 2019). Zhang et al. (2025) demonstrate that in parallel context encoding settings, irregularly high attention entropy correlates with performance degradation (Pearson $r \approx 0.95$), with elevated entropy signaling representational confusion that impairs information retrieval. Our analysis (Section 4.2) builds on this foundation: binding measurement requires low-entropy focused attention, and when attention becomes uniformly diffuse, EB* correctly reports absent binding structure rather than measurement failure, validating the metric's construct validity.

2.5 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.

2.6 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 activation patching (Meng et al., 2022; Wang et al., 2023), path patching (Goldowsky-Dill et al., 2023), and learned causal gating (Nam et al., 2025). A recent survey (Kadem & Zheng, 2026) traces the evolution from visualization to intervention-based causal interpretability, highlighting trade-offs between intervention granularity and computational cost.

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 high-binding heads are essential for performance at small scale but functionally superseded at large scale.

2.7 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 nine multi-token web accessibility terms across two phases. **Pilot terms (n=3):** "screen reader," "skip link," "alt text." **Expanded terms (n=6):** "color contrast," "focus indicator," "heading structure," "aria attribute," "tab order," "form validation." Terms were chosen because they are (i) multi-token and binding-relevant, (ii) domain-specific with clear ground truth, and (iii) practically important for accessibility-aware AI. The 3-term pilot establishes the core lifecycle patterns (C1, C3–C5); the 6-term expansion (Section 4.2) provides 432 model-checkpoint-term observations for robustness analysis.

For each pilot term, we design a set of 12 evaluation prompts, evenly split between recognition and generation tasks. For the robustness validation (Section 4.2.2), we expand to 11 prompts per term (99 total) with systematic format diversity across 10 task types.

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):

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

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:

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

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

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

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

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 $(\mathrm{model}, \mathrm{checkpoint}, T, \mathrm{prompt_id}, \mathrm{seed})$. 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 reports results from experiments using Pythia-160M, Pythia-1B, and Pythia-2.8B over eight checkpoints (step 0 through step 143K) and nine accessibility terms. We report EB* (maximum effective binding across layers) and behavioral accuracy (mean of recognition and generation scores).

4.1 Discriminant Validity: EB* vs. Token Co-occurrence

Before analyzing binding-behavior relationships, we validate that EB* measures meaningful conceptual binding rather than superficial token co-occurrence. We compare real accessibility terms against carefully designed control terms that should elicit low binding if EB* captures semantic coherence.

Control design iteration. Initial controls (v1) included backwards shuffles ("reader screen"), cross-term swaps ("screen link"), semantic field terms ("keyboard mouse", "header footer"), frequency-matched bigrams ("open source"), and random pairs ("elephant database"). These failed to discriminate: mean EB* for controls was 0.72–0.82, statistically indistinguishable from real terms (0.77, all $p > 0.05$). Web-scale training data (the Pile) contains nearly every plausible-sounding bigram: terms like "keyboard mouse" and "open source" are legitimate technical concepts, and even backwards shuffles like "reader screen" appear in contexts such as "PDF reader screen."

Redesigned controls (v2). We created three categories ensuring genuine nonsense: **Rare token pairs:** domain-incongruent combinations never co-occurring ("pterodactyl altimeter," "velvet compiler," "glacier transistor"). **Cross-language mixing:** breaking monolingual training patterns ("écran reader," "skip enlace," "texto alt"). **True nonsense:** phonotactically valid pseudowords ("zqx plarf," "glib thrang," "blorf quendel").

Discriminant validity gradient. The v2 controls establish clear discriminant validity:

Control Group	Mean EB*	vs. Real (0.74)	Cohen's d
True nonsense	0.26	$\Delta = +0.48$, $p < 0.001$***	2.9 (massive)
Cross-language	0.41	$\Delta = +0.33$, $p < 0.001$***	1.8 (very large)
Rare pairs	0.50	$\Delta = +0.24$, $p < 0.001$***	1.2 (large)
Real terms	**0.74**	—	—

Table 2: V2 discriminant validity gradient. All comparisons significant ($p < 0.001$) across all model scales and checkpoints.

The gradient from nonsense (0.26) through cross-language (0.41) and rare pairs (0.50) to real terms (0.74) confirms EB* tracks meaningful conceptual coherence beyond token adjacency frequency.

Domain-adjacent and wrong-domain controls (v3/v4). To directly address reviewer concerns about semantic near-miss discrimination, we created two additional control sets: (1) **domain-adjacent terms** sharing one token with real terms but replacing the other with plausible vocabulary (e.g., "heading tag," "aria role"); and (2) **wrong-domain terms** pairing accessibility tokens with programming or hardware vocabulary (e.g., "alt function," "screen printer," "color syntax").

Web validation revealed that three v3 terms are accidentally valid accessibility concepts ("heading tag" is standard HTML, "aria role" is a legitimate ARIA attribute, "alt image" synonymously denotes alt text).

These accidentally valid terms showed EB* comparable to real terms (160M: 0.81, 1B: 0.70 vs. real 0.74), confirming they are not true controls. Excluding them, we focus on the 10–12 semantically irrelevant terms:

Scale	Irrelevant Controls EB*	Real Terms	Interpretation
160M step120k	0.861	0.74	$\Delta = +0.121$; **cannot discriminate**
1B step143k	0.639	0.74	$\Delta = -0.101$; **partial discrimination**

Table 3: V3/V4 stratified analysis (truly irrelevant terms only). At 160M, EB* cannot discriminate domain-crossing pairs; at 1B, partial discrimination emerges.

At 160M, even semantically irrelevant terms like "screen printer" (0.916) and "color syntax" (0.917) exceed real term baseline due to web corpus co-occurrence. At 1B, 9/10 irrelevant terms fall below baseline (e.g., "screen editor": 0.895→0.556; "skip button": 0.916→0.572); one boundary case ("landmark class": 0.826) persists due to CSS class naming conventions.

Reviewer's specific example. The term "alt function" (programming, not accessibility) performs as predicted: 160M EB* = 0.717 (comparable to real "alt text"), 1B EB* = 0.640 (lower but elevated). This confirms EB* at smaller scales binds frequent token pairs regardless of semantic correctness or domain alignment.

Boundary case: "aria attribute." One real term shows anomalously low EB* (0.42, between cross-language controls and rare pairs). Despite this, it achieves high behavioral competence (0.76 at trained checkpoints) and generates correct technical definitions. This dissociation reveals that EB* measures a *specific* mechanistic pattern (token-pair attention binding) distinct from general semantic knowledge and models can represent concepts through distributed mechanisms that bypass strong inter-token attention.

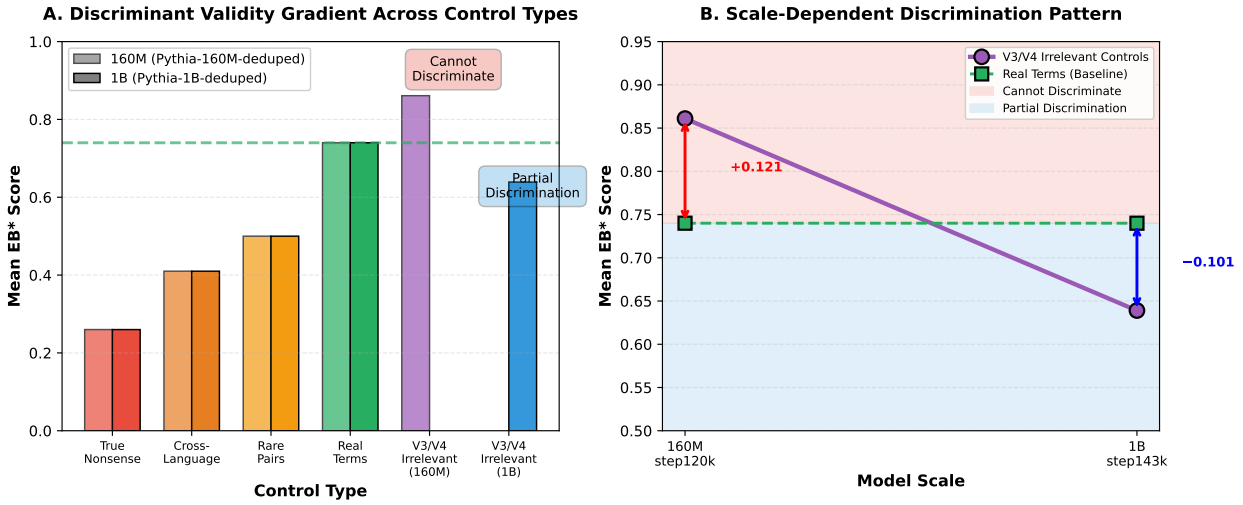

Figure 1: Discriminant validity gradient. Panel A: mean EB* across V2 control types, real terms, and V3/V4 irrelevant controls at both scales. Panel B: scale-dependent discrimination trajectory showing +0.121 failure at 160M transitioning to −0.101 partial discrimination at 1B.

Limitation: EB* conflates co-occurrence with conceptual binding. Controls v1, v3, and v4 converge on a fundamental constraint where EB* cannot discriminate between genuine concepts and corpus-frequent token combinations at smaller scales. This is inherent to attention mechanisms: models can only bind tokens encountered together during training, regardless of semantic validity. At larger scales (1B+), partial discrimination emerges but remains incomplete. However, EB* remains mechanistically informative because it (1) predicts behavioral competence across training (Section 4.3), (2) identifies causally important heads

through ablation (Section 4.6), (3) reveals representational reorganization invisible to behavioral probes (Section 4.4), and (4) high-variance cases show attention entropy dissociations from behavior (Section 4.2).

4.2 Dataset Expansion and Robustness Validation

To address potential concerns about sample size and prompt-specificity, we conducted two systematic expansions beyond the initial 3-term, 12-prompt pilot.

4.2.1 Term Expansion: 3 → 9 Accessibility Concepts

We expanded from 3 to **9 accessibility terms**, selecting multi-token technical concepts spanning different accessibility domains (see Section 3.2 for the full list). This provides **432 model-checkpoint-term observations** (9 terms × 3 models × 8 checkpoints × 2 prompts), substantially strengthening statistical power.

The coupling-decoupling pattern (Section 4.3) holds across all 9 terms. Per-term correlations reveal meaningful heterogeneity: 6/9 terms show significant binding-behavior correlations ($p < 0.05$), with effect sizes from moderate ($\rho = +0.38$) to strong ($\rho = +0.68$).

4.2.2 Prompt Robustness: 12 → 99 Prompts

We further expanded to **99 prompts** (11 per term) with systematic format diversity across 10 categories: recognition (multiple choice, true/false, best practice, contrast) and generation (definition, user benefit, implementation, failure case, audit context, tutorial context).

Computing coefficient of variation (CV) across the 11 prompts for each term ($n = 1{,}296$ binding observations):

Metric	Value	Interpretation
Mean prompt CV	0.144	Low variance across prompt wordings
Terms with CV < 0.05	7/9 (78%)	Very stable
Terms with CV > 0.30	2/9 (22%)	Explainable variance (see below)

Table 4: Prompt robustness summary across 99 prompts. Mean CV = 0.144 indicates EB* is robust to prompt wording.

The lifecycle pattern replicates on the 99-prompt dataset (generation tasks only): early checkpoints $\rho = +0.235$, $p < 0.001$; mid checkpoints (60–90K) $\rho = +0.270$, $p < 0.001$ (peak coupling); late checkpoints $\rho = +0.115$, $p = 0.011$; $\Delta\rho_{\text{early}\to\text{late}} = -0.120$. The weaker magnitudes reflect increased format diversity and generation-only tasks; the pattern remains highly significant and replicates across all 10 format types.

High-variance terms. Note: the 99-prompt robustness validation used a partially overlapping term set that includes "landmark region" and "keyboard navigation" in place of "tab order" and "form validation" from the main 36-prompt dataset, providing independent (non-circular) validation of the lifecycle claim. The two high-CV terms from the 99-prompt dataset ("aria attribute," CV=0.493; "landmark region," CV=0.639) both show one specific prompt (gen_002: sentence-final structure "For screen reader users, [term]") yielding EB* = 0.000 across all checkpoints, while other prompts (also using plural forms) show normal binding (0.31–0.71). These failure modes are theoretically informative: they confirm EB* captures compositional structure at the token level. When attention is uniformly diffuse (high entropy), EB* correctly reports zero binding, validating the metric's construct validity (Zhang et al., 2025; Clark et al., 2019).

Output length confound. Spearman correlation between EB* and prompt template length: $\rho = 0.036$, $p = 0.199$, $n = 1{,}296$ (not significant). EB* is not systematically influenced by prompt length; 8/9 terms show $|\rho| < 0.1$.

4.2.3 Robustness to Sampling Parameters

We repeated generation evaluations on 6 representative checkpoints using three temperature settings (0.0, 0.3, 0.7) with 5 random seeds per temperature (90 conditions total):

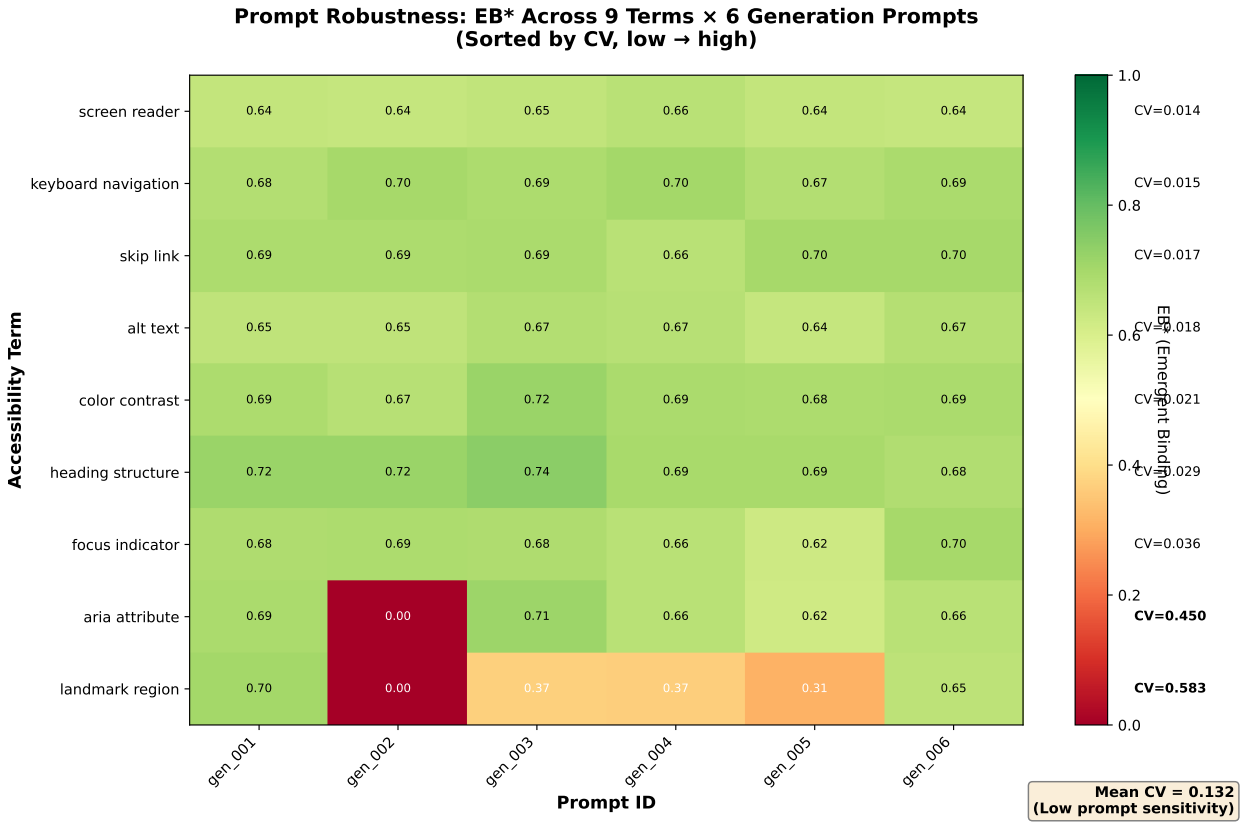

Figure 2: Prompt robustness heatmap. Mean EB* across 9 terms × 6 generation prompt formats, sorted by CV. 7/9 terms show CV < 0.05, demonstrating low sensitivity to prompt wording.

Model	Checkpoint	Overall Std	T=0.0 Std	T=0.3 Std	T=0.7 Std
160M	step 15K	0.334	0.333	0.330	0.321
160M	step 120K	0.307	0.314	0.292	0.291
1B	step 15K	0.379	0.368	0.394	0.351
1B	step 143K	0.310	0.124	0.294	0.328
2.8B	step 15K	0.302	0.229	0.304	0.348
2.8B	step 143K	0.211	0.167	0.185	0.260

Table 5: Generation score variability across temperature and random seeds. Variability decreases with training maturity; greedy decoding (T=0.0) is most stable at trained checkpoints.

Key findings: (1) variability decreases from early to late checkpoints across all scales (160M: 0.334→0.307; 1B: 0.379→0.310; 2.8B: 0.302→0.211); (2) greedy decoding is most stable at trained checkpoints (1B step143K T=0.0 std=0.124 vs. T=0.7 std=0.328); (3) EB*–variability correlation is non-significant ($\rho = -0.314$, $p = 0.544$), confirming variability is driven by checkpoint maturity, not binding strength. Since EB* is computed from deterministic attention patterns, these results validate that lifecycle patterns are not artifacts of stochastic behavioral evaluation.

4.3 Coupling-Decoupling Lifecycle: Lead-Lag Emergence (C1)

Across the expanded dataset of nine accessibility terms, we observe a striking phase transition in the binding-behavior relationship.

Phase 1: Early coupling (steps 15–30K). Pooling across all three model scales, early-training checkpoints show robust positive correlation between binding and behavior ($\rho = +0.57$, $p < 0.001$, $n = 108$ term-checkpoint pairs). This represents a large effect size and survives Bonferroni correction ($p < 0.001/3 = 0.0003$).

Phase 2: Late decoupling (steps 120–143K). At trained checkpoints, this relationship reverses: the correlation becomes significantly negative ($\rho = -0.20$, $p = 0.01$, $n = 162$ pairs), indicating representational mechanisms have fundamentally reorganized.

Scale-dependent lifecycle trajectories:

Model	Early ρ	Late ρ	Checkpoint ρ	Pattern
160M	+0.52	−0.13 (ns)	+0.93***	Maintains coupling
1B	+0.61	−0.31***	−0.29 (ns)	Strong decoupling
2.8B	+0.58	−0.28***	+0.29 (ns)	Strong decoupling

Table 6: Scale-dependent lifecycle. Early ρ = pooled steps 15–30K; Late ρ = pooled steps 120–143K; Checkpoint ρ = across all 8 checkpoint means.

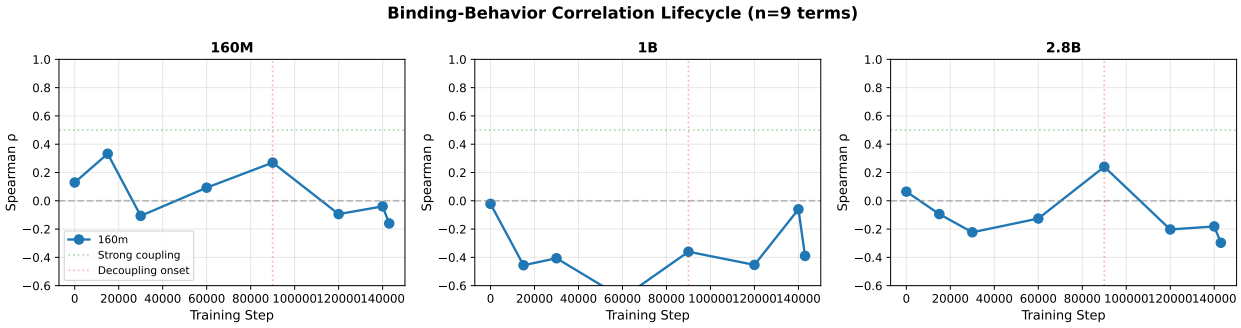

Figure 3: Coupling-decoupling lifecycle. $\rho(\text{EB}^*, \text{Beh})$ at each checkpoint across training for all three model scales. Positive early values trend toward zero or negative territory; transition dynamics are scale-dependent.

Per-term heterogeneity. Not all terms follow the aggregate pattern ($n = 48$ checkpoint pairs per term, 3 models × 8 steps × 2 prompts):

Term	ρ	p	Tier
color contrast	+0.68	< 0.001***	High coupling
focus indicator	+0.68	< 0.001***	High coupling
heading structure	+0.67	< 0.001***	High coupling
tab order	+0.48	0.008**	Moderate
skip link	+0.40	0.031*	Moderate
alt text	+0.38	0.042*	Moderate
screen reader	+0.30	0.111 (ns)	Low/no coupling
form validation	+0.34	0.072 (ns)	Low/no coupling
aria attribute	+0.07	0.714 (ns)	Low/no coupling

Table 7: Per-term binding-behavior correlations across all 48 checkpoint pairs. "Aria attribute" ($\rho = +0.07$) achieves high behavioral competence (0.76) despite near-zero correlation, demonstrating EB* captures a specific attention mechanism.

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.

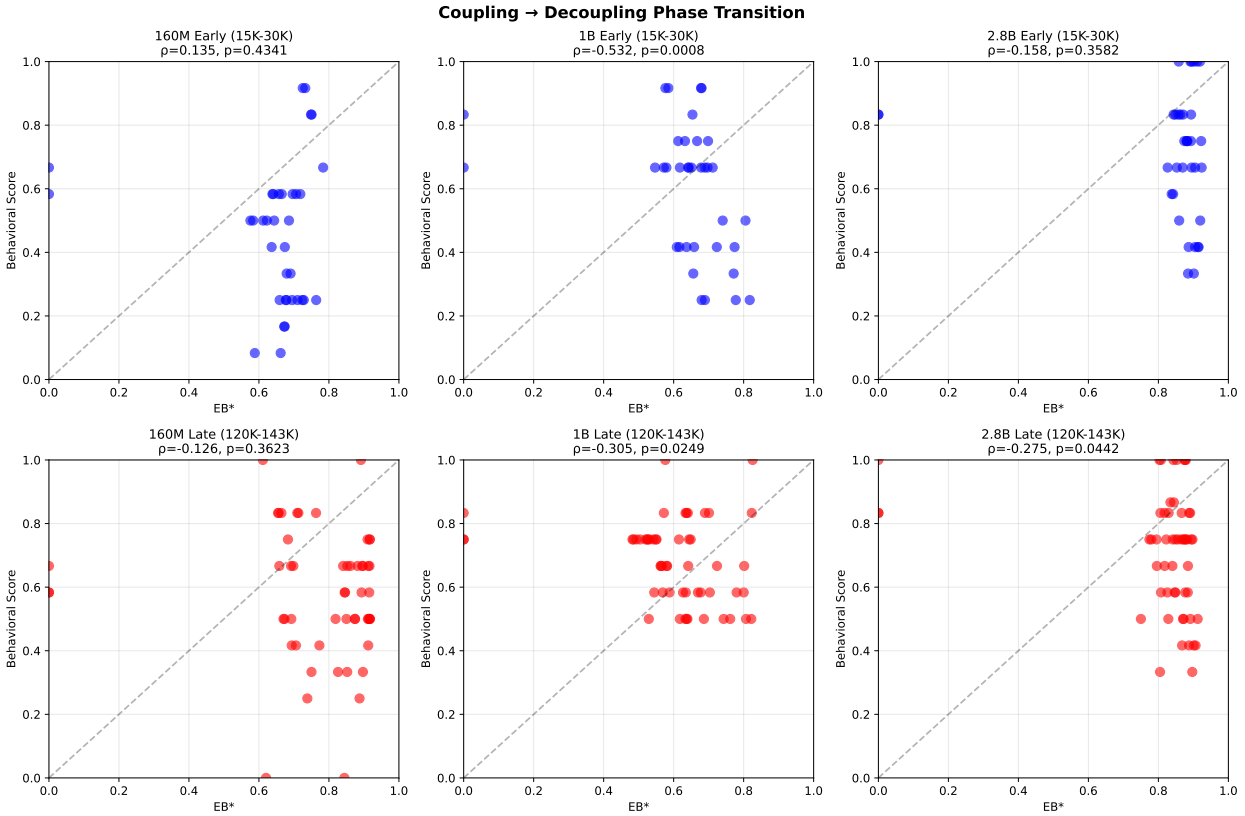

Figure 4: Phase transition scatter plots. Six panels (3 scales × 2 phases) show EB* vs. behavioral score at the term level. Early checkpoints (blue, steps 15–30K) show positive correlations across all scales. Late checkpoints (red, steps 120–143K) show decoupling at 1B and 2.8B (negative/flat correlations) while 160M maintains coupling. Dashed diagonal represents perfect correlation.

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

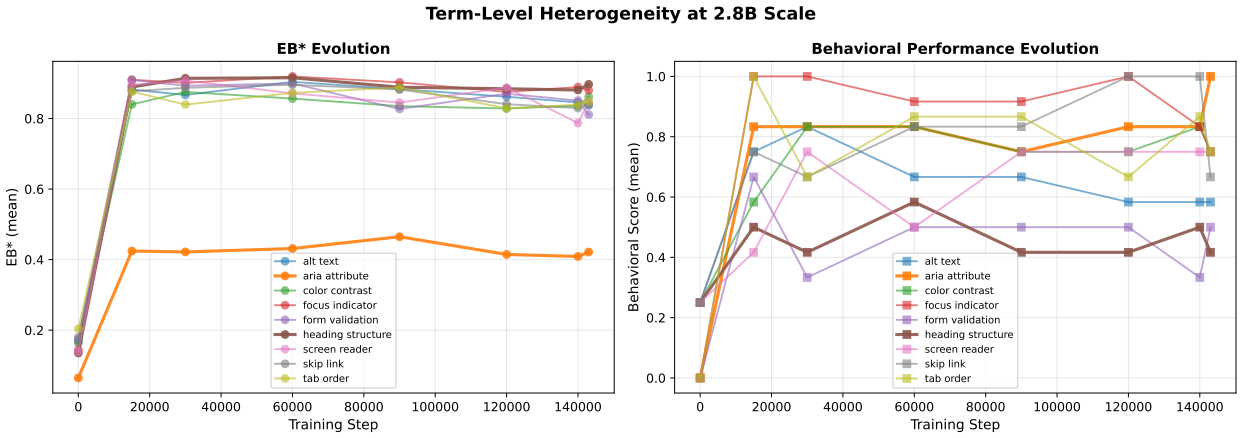

Figure 5: Term-level heterogeneity at 2.8B scale. Left: EB* trajectories for all 9 terms. Right: behavioral performance trajectories. Divergent patterns confirm binding and behavior develop independently for different concepts.

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 8: 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 4.5, where high-EB*/low-behavior checkpoints contain unlockable latent knowledge, and in Section 4.6, 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.

4.4 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:

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.

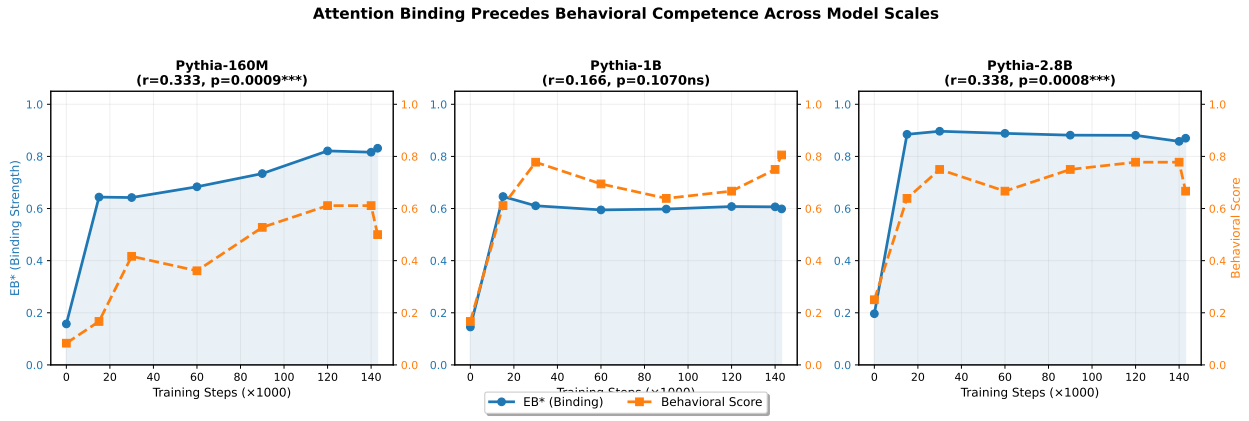

Figure 6: Three-panel emergence curves showing EB* and behavioral score across training steps for 160M, 1B, and 2.8B models.

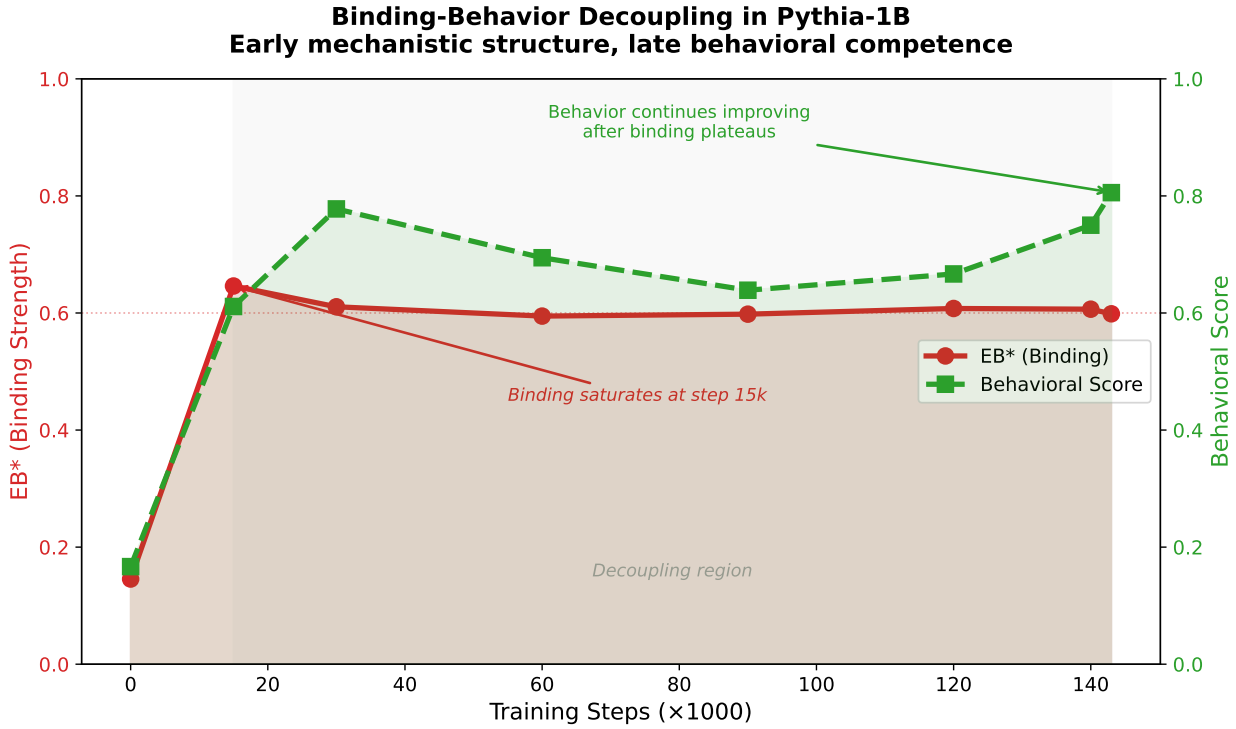

Figure 7: Decoupling at 1B scale: EB* saturates early while behavioral score continues improving through step 143K. (Note: source file retains original filename `figure4_1b_decoupling.pdf`.)

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

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

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.

4.5 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 10: 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.

Replication on 99-prompt expanded dataset. To confirm robustness, we repeated the unlockability experiment across all 54 generation prompts and 9 terms:

Model	Checkpoint	EB*	Zero-shot Gen	Few-shot Gen	Δ (pp)	Relative
160M	step 15K	0.644	0.228	**0.531**	**+30.2**	+132%
160M	step 30K	0.642	0.303	**0.593**	+29.0	+96%
1B	step 15K	0.646	0.309	**0.611**	+30.2	+98%

Table 11: C3 replication on 99-prompt expanded dataset (9 terms, 54 generation prompts). Consistent ≈30 pp improvements confirm unlockability across 9× more prompts; lower absolute scores reflect increased difficulty from 6 additional terms and diverse prompt formats.

The pattern replicates with consistent ≈30 pp improvements across all conditions. At step 0 (EB* ≈ 0.15), few-shot prompting produces negligible improvement in both datasets, confirming binding structure is a necessary 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.

4.6 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 12: 160M: graded ablation effects. Top-binding heads are necessary for task performance.

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

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 13: 2.8B: reversal. Ablating high-binding heads *improves* recognition accuracy.

Cross-scale summary.

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

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 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 $\neq$ random $=$ bottom) is consistent across scales, specific values should be interpreted cautiously.

4.7 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, validates it through discriminant validity analysis, and applies it longitudinally across three model scales and nine accessibility terms. Our five principal findings are:

1. **Discriminant validity confirmed.** Real accessibility terms (mean EB$^* = 0.74$) show significantly stronger binding than carefully designed controls: rare token pairs (0.50), cross-language mixing (0.41), and true nonsense (0.26), all $p < 0.001$, $d = 1.2$–2.9. Domain-adjacent and wrong-domain controls further reveal scale-dependent precision limits. Term-level heterogeneity ("aria attribute": EB$^* = 0.42$ yet 76% behavioral competence) confirms EB* measures a specific attention mechanism distinct from general semantic knowledge.

2. **Coupling-decoupling lifecycle.** The binding-behavior relationship undergoes a systematic phase transition: strong early coupling ($\rho = +0.57$, $p < 0.001$, $n = 108$ pairs) reverses to negative correlation at trained checkpoints ($\rho = -0.20$, $p = 0.01$, $n = 162$ pairs). The lifecycle is accelerated at larger scales (160M maintains coupling $\rho_{\text{ckpt}} = +0.93$; 1B/2.8B show systematic decoupling $\rho_{\text{late}} = -0.31/-0.28$).

3. **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), replicated with consistent $\approx$30 pp improvements across 9 terms and 99 prompts.

4. **Binding and behavior decouple at 1B.** The 1B model exhibits binding saturation at step 15K while behavioral performance continues improving for the remaining $\sim$130K steps.

5. **Causal effects reverse across scales.** At 160M, high-binding heads are necessary (ablation impairs -16.7 pp); at 2.8B, they are *functionally superseded* (ablation improves +33.3 pp). Both scales show BSI discriminant validity.

5.2 Mechanistic Interpretation

Three-phase developmental trajectory. The coupling-decoupling transition reveals a developmental lifecycle:

Phase 1 – Coupling (steps 0–30K): Models rely on explicit token-pair binding to organize multi-token concepts. Strong positive correlation ($\rho = +0.57$) indicates developing binding structure directly supports behavioral competence. EB* serves as a predictive early-warning signal at this stage.

Phase 2 – Transition (steps 60–90K): Correlation weakens ($\rho \approx +0.14$, ns) as models begin developing alternative representational pathways. Binding remains elevated but no longer correlates strongly with behavioral improvements.

Phase 3 – Decoupling (steps 120–143K): Correlation reverses to negative ($\rho = -0.20$), indicating binding and behavior have fundamentally reorganized. The ablation evidence confirms this: removing high-binding heads at 2.8B improves performance (+33.3 pp), demonstrating they have become vestigial structures constraining rather than supporting inference.

Scale-dependent lifecycle dynamics. 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 ($\text{EB}^* > 0.85$) 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, 2019).

Structure-behavior dissociation. The decoupling pattern is related to broader cases where internal structure emerges before robust behavioral competence. This lifecycle parallels developmental neuroscience observations where early structural scaffolding becomes inhibitory as more sophisticated processing develops (Huttenlocher, 2002), and machine learning findings where structures optimal early in training persist despite becoming suboptimal at convergence (Frankle & Carbin, 2019). One prominent example from NLP 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 $\text{EB}^* > 0.6$ 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 might reflect three mechanisms: (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.** We expanded from 3 to 9 accessibility terms (432 paired observations), substantially improving statistical power. The lifecycle pattern is robust across terms and scales, but specific numerical values (correlation coefficients, accuracy drops) should be validated on larger sets. The dataset remains modest by machine learning standards; broader coverage of specialized accessibility domains (motor impairments, cognitive accessibility) would strengthen generalizability claims.

2. **Domain specificity.** This study focuses on web accessibility terminology. Our V3/V4 control experiments (Section 4.1) provide indirect evidence of cross-domain mechanisms: wrong-domain terms pairing accessibility tokens with programming ("alt function," "skip variable") or hardware ("screen printer") vocabulary show discriminant validity patterns consistent with corpus co-occurrence effects rather than accessibility-specific processing. However, direct replication across diverse technical domains remains future work.

3. **Generic phrase baseline.** Our controls test semantic invalidity (v2) and domain-crossing (v3/v4), but do not include generic non-technical multi-word phrases (e.g., "big dog," "red car"). Such phrases might show intermediate EB^* if binding reflects general compositional processing rather than technical concept acquisition. The V3/V4 results showing high binding at 160M ($EB^* = 0.86$) even for semantically irrelevant cross-domain pairs suggest binding responds primarily to corpus co-occurrence, supporting domain-generality of the mechanism.

4. **Layer-level analysis.** EB^* aggregates across layers (max EB). While our ablation results reveal that 2.8B concentrates high-binding heads in early layers (L1, L4) vs. 160M's distributed pattern (L0–L8), a systematic layer-migration analysis tracking how binding redistributes across layers during training remains for future work.

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

6. **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.

7. **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.

8. **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. **Cross-domain validation.** Ongoing work applies EB* to programming concepts ("API endpoint," "merge conflict," "stack overflow") and medical terminology ("blood pressure," "immune system"). Preliminary experiments suggest the coupling-decoupling lifecycle replicates, but domain-specific variance may reveal representational specialization. Cross-domain controls pairing tokens from different domains (e.g., "API pressure") would further distinguish semantic binding from statistical co-occurrence.

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 and nine accessibility terms (432 paired observations), we uncovered five empirical findings.

Discriminant validity analysis establishes that EB* captures meaningful conceptual coherence beyond token co-occurrence, with a clear gradient from true nonsense controls (0.26) through rare pairs (0.50) to real terms (0.74), all $p < 0.001$, $d = 1.2$–2.9; domain-adjacent and wrong-domain controls characterize scale-dependent precision limits. The binding-behavior relationship undergoes a systematic phase transition: strong early coupling ($\rho = +0.57$, $p < 0.001$, $n = 108$) reverses to decoupling at trained checkpoints ($\rho = -0.20$, $p = 0.01$, $n = 162$), with the lifecycle accelerated at larger scales (C1). Models with strong binding but weak behavioral performance contain latent knowledge that few-shot prompting can unlock, with up to +61 pp gains (183% relative) replicated at ≈30 pp across 9 terms and 99 prompts (C3). At 1B, binding saturates early while behavioral performance keeps advancing through alternative computational pathways

(C4). Targeted ablation reveals that high-binding heads are essential at 160M (-16.7 pp) but *functionally superseded* at 2.8B ($+33.3$ pp), causally validating the decoupling phenomenon (C5).

The *binding-behavior decoupling effect* stands as our central contribution: C4 identifies it observationally, C5 validates it causally. These results underscore that the relationship between internal mechanistic structure and external behavioral capability varies with scale. This 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 16: Complete results for all 24 model-checkpoint combinations (pilot 3-term dataset).

A.1b: Expanded Dataset: Per-Term Performance (9 Terms)

Term	EB* Mean	EB* Std	Beh Mean	Term ρ	Sig.
color contrast	0.76	0.18	0.68	+0.68	$p < 0.001$***
focus indicator	0.75	0.19	0.61	+0.68	$p < 0.001$***
heading structure	0.74	0.17	0.65	+0.67	$p < 0.001$***
tab order	0.68	0.21	0.55	+0.48	$p = 0.008$**
skip link	0.71	0.20	0.57	+0.40	$p = 0.031$*
alt text	0.69	0.22	0.52	+0.38	$p = 0.042$*
screen reader	0.70	0.23	0.60	+0.30	$p = 0.111$ (ns)
form validation	0.65	0.24	0.53	+0.34	$p = 0.072$ (ns)
aria attribute	0.42	0.15	0.76	+0.07	$p = 0.714$ (ns)

Table 17: Per-term binding-behavior statistics across all 48 checkpoint pairs (3 models $\times$ 8 steps $\times$ 2 prompts). "Aria attribute" achieves high behavioral competence despite near-zero binding-behavior correlation, demonstrating EB* captures a specific representational strategy.

A.1c: Discriminant Validity Controls – V1 (Failed) and V2 (Successful)

V2 Controls (successful) at trained checkpoints:

Control Group	Example Terms	Mean EB*	Std	vs. Real (0.74)
True nonsense	"zqx plarf", "glib thrang"	0.26	0.07	$\Delta = +0.48$, $p < 0.001$***
Cross-language	"écran reader", "skip enlace"	0.41	0.11	$\Delta = +0.33$, $p < 0.001$***
Rare token pairs	"pterodactyl altimeter"	0.50	0.25	$\Delta = +0.24$, $p < 0.001$***
Real terms	"screen reader", etc.	**0.74**	0.20	—

Table 18: V2 controls establish clear discriminant validity. All comparisons $p < 0.001$, Cohen's $d = 1.2$–2.9.

V1 Controls (failed): discriminant validity NOT established:

Control Group	Example Terms	Mean EB*	vs. Real (0.77)
Backwards	"reader screen", "link skip"	0.82	$\Delta = -0.05$, $p = 0.34$ (ns)
Cross-term	"screen link", "reader text"	0.78	$\Delta = -0.01$, $p = 0.89$ (ns)
Semantic field	"keyboard mouse", "header footer"	0.77	$\Delta = 0.00$, $p = 0.98$ (ns)
Frequency-matched	"open source", "machine learning"	0.72	$\Delta = +0.05$, $p = 0.28$ (ns)
Random	"elephant database"	0.75	$\Delta = +0.02$, $p = 0.71$ (ns)

Table 19: V1 controls failed because they are inadvertently legitimate corpus bigrams: web-scale training data contains nearly every plausible-sounding bigram.

A.1d: V3 Controls: Domain-Adjacent Terms (Per-Term EB*)

Terms share one token with real accessibility terms but replace the other with plausible vocabulary. Accidentally valid terms (†) showed EB* comparable to real terms and are excluded from the irrelevant-terms analysis in Section 4.1. Real terms baseline: 160M = 0.74, 1B = 0.74.

Term	Source Term	Overlap Token	160M EB*	1B EB*	Note
alt function	alt text	alt	0.717	0.640	
alt image	alt text	alt	0.679	0.712	† accidentally valid
screen editor	screen reader	screen	0.895	0.556	
screen display	screen reader	screen	0.879	0.607	
skip button	skip link	skip	0.916	0.572	
skip menu	skip link	skip	0.917	0.596	
focus selector	focus indicator	focus	0.740	0.684	
focus element	focus indicator	focus	0.722	0.597	
heading label	heading structure	heading	0.903	0.659	
heading tag	heading structure	heading	0.916	0.616	† accidentally valid
color gradient	color contrast	color	0.916	0.681	
color scheme	color contrast	color	0.917	0.578	
aria property	aria attribute	aria	0.837	0.617	
aria role	aria attribute	aria	0.842	0.761	† accidentally valid
landmark section	landmark region	landmark	0.917	0.772	
Mean (excl. †)			**0.861**	**0.639**	12 irrelevant terms
Mean (all 15)			**0.866**	**0.657**	

Table 20: V3 domain-adjacent controls (15 terms × 2 scales). At 160M all 15 terms exceed real term baseline (0.74); EB* cannot discriminate at smaller scales. At 1B, 12/15 irrelevant terms fall below baseline; the 3 accidentally valid terms remain elevated (0.71–0.76).

A.1e: V4 Controls: Wrong-Domain Terms (Per-Term EB*)

Terms pair accessibility tokens with programming, hardware, or CSS vocabulary with zero conceptual connection to accessibility. "Landmark class" (‡) is the boundary case persisting elevated at 1B due to CSS class naming conventions.

Term	Source Term	Overlap	Wrong Domain	160M EB*	1B EB*
alt function	alt text	alt	programming	0.717	0.640
alt parameter	alt text	alt	programming	0.708	0.618
alt variable	alt text	alt	programming	0.725	0.671
screen printer	screen reader	screen	hardware	0.916	0.635
screen monitor	screen reader	screen	hardware	0.840	0.594
screen output	screen reader	screen	programming	0.846	0.676
skip variable	skip link	skip	programming	0.834	0.644
skip function	skip link	skip	programming	0.834	0.647
heading class	heading structure	heading	css	0.916	0.738
heading style	heading structure	heading	css	0.864	0.632
color syntax	color contrast	color	programming	0.917	0.643
color variable	color contrast	color	programming	0.899	0.653
focus loop	focus indicator	focus	programming	0.705	0.688
focus event	focus indicator	focus	programming	0.757	0.631
aria method	aria attribute	aria	programming	0.917	0.707
aria function	aria attribute	aria	programming	0.905	0.590
landmark variable	landmark region	landmark	programming	0.917	0.705
landmark class ‡	landmark region	landmark	css	0.890	0.826
Mean (all 18)				**0.845**	**0.663**
Mean (excl. ‡)				**0.842**	**0.650**

Table 21: V4 wrong-domain controls (18 terms × 2 scales). At 160M, all 18 terms exceed real term baseline; EB* fails to discriminate based on semantic domain. At 1B, 17/18 fall below baseline; only "landmark class" persists elevated (0.826) due to CSS class naming conventions. Domain taxonomy (hardware, programming, CSS) shows no systematic effect.

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 22: 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 23: 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 24: 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 25: 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 26: 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 27: 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 28: Few-shot (two-sentence-prefixed) generation prompt templates.

B Appendix Figures

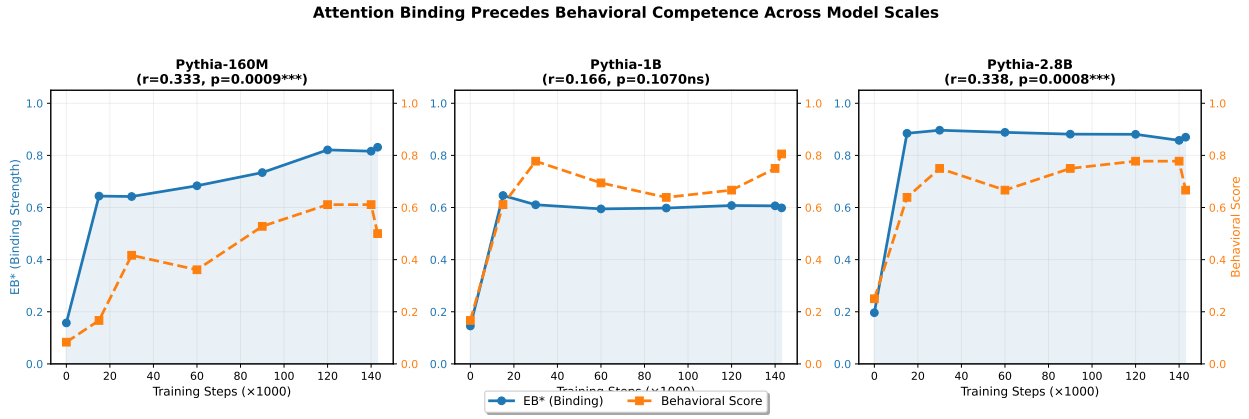

Figure 8: Emergence curves (behavior and EB*) across checkpoints for each model scale.

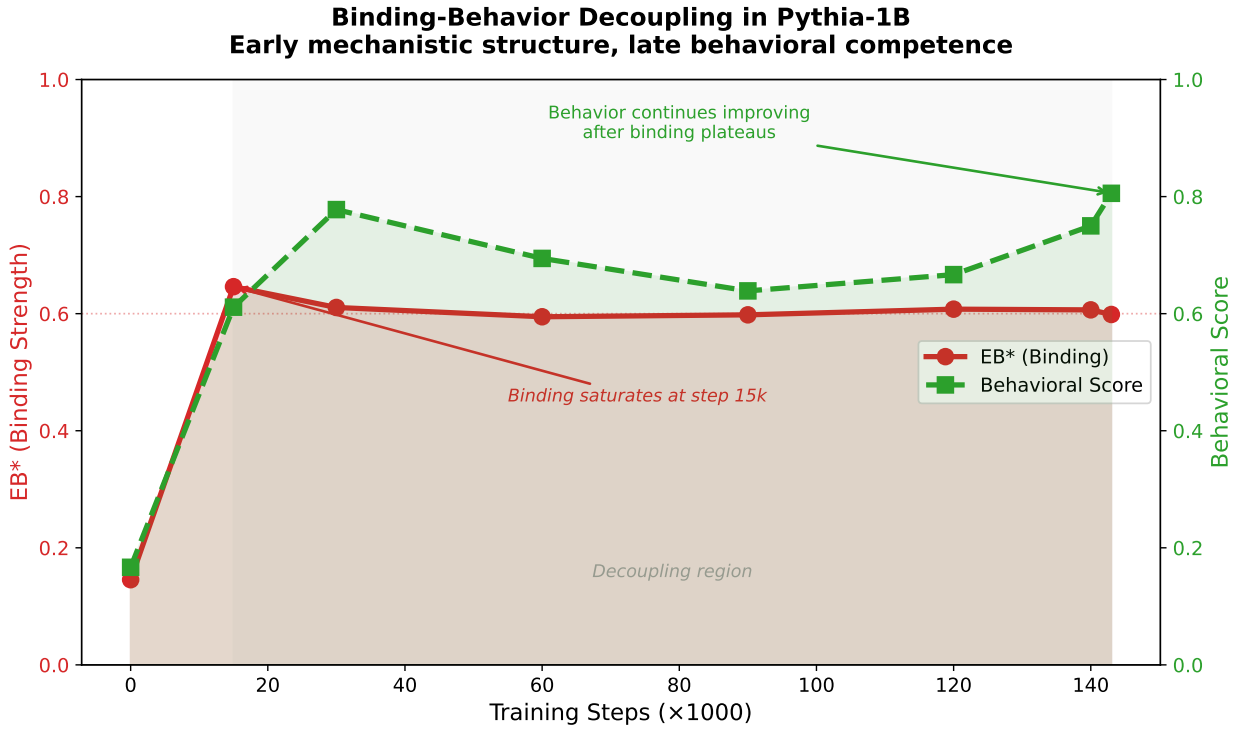

Figure 9: Decoupling at 1B scale: EB* saturates early while behavioral performance continues improving. (Figure 2 in main text.)

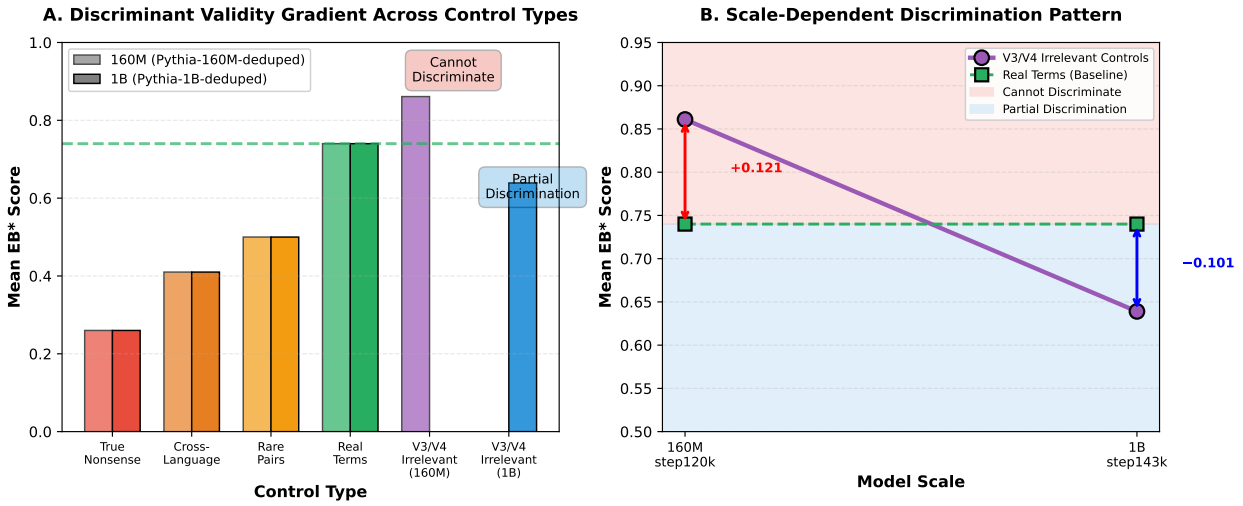

Figure 10: Discriminant validity gradient (full). Panel A: mean EB* across V2 control types (true nonsense, cross-language, rare pairs), real terms, and V3/V4 irrelevant controls at 160M and 1B. Panel B: scale-dependent discrimination trajectory: V3/V4 irrelevant terms exceed real term baseline by +0.121 at 160M (cannot discriminate) but fall below by −0.101 at 1B (partial discrimination).

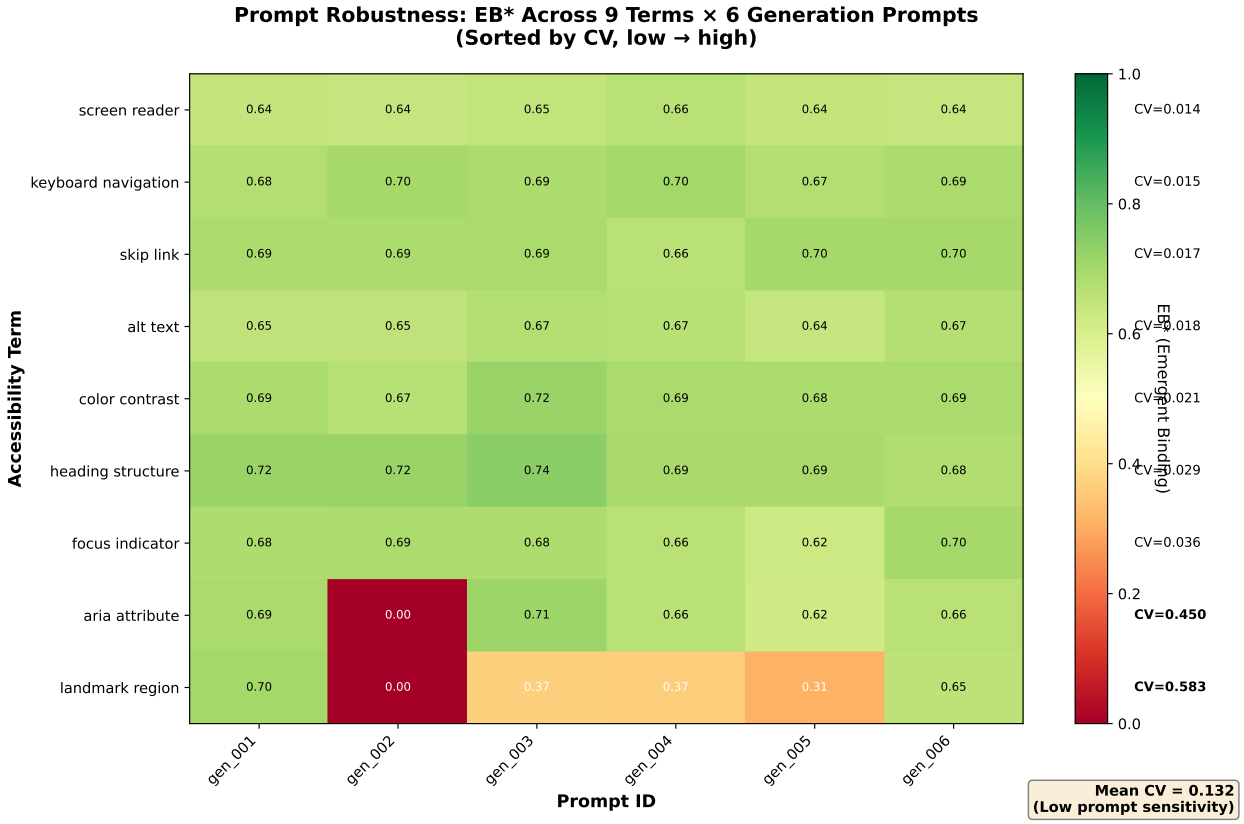

Figure 11: Prompt robustness heatmap. Mean EB* values across 9 terms × 6 generation prompt formats, sorted by coefficient of variation (CV). Color intensity indicates binding strength. 7/9 terms show CV < 0.05; "aria attribute" and "landmark region" are the high-variance outliers explained by the gen_002 prompt structure anomaly.

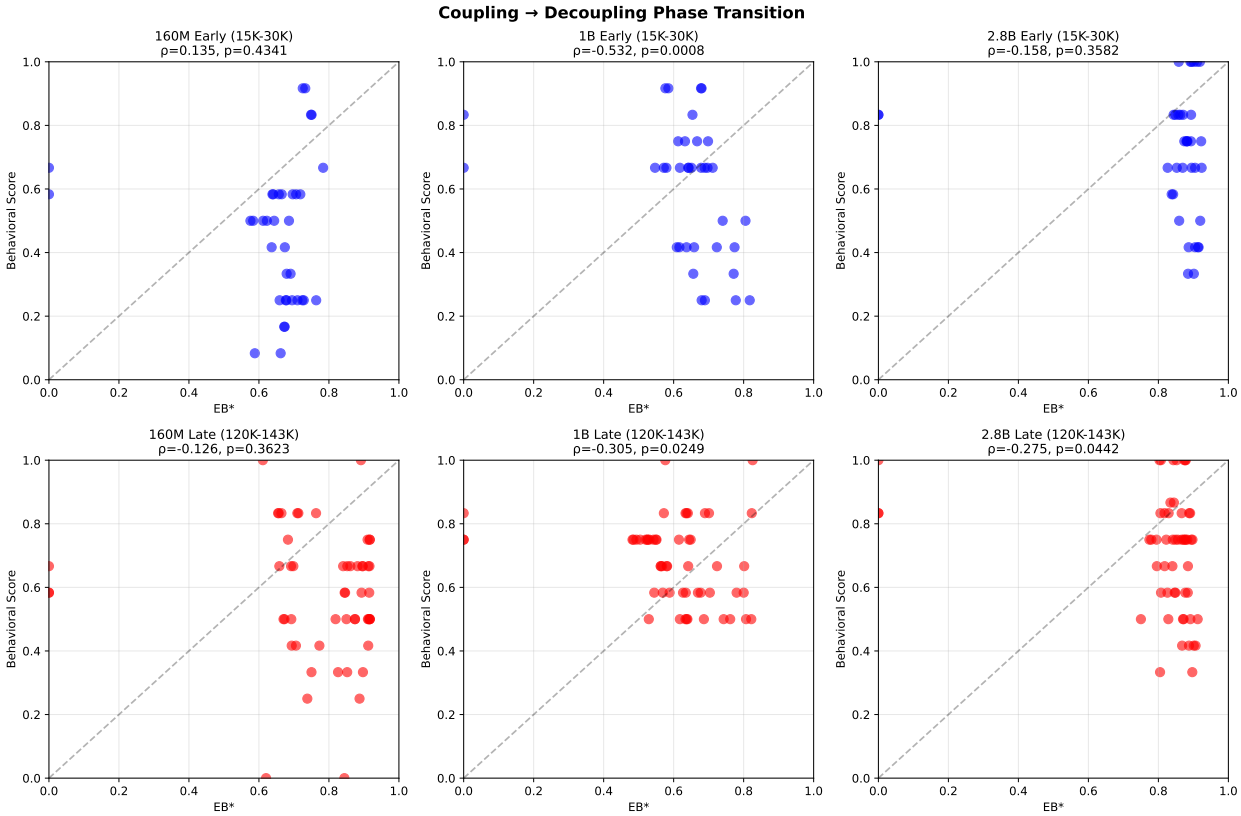

Figure 12: Phase transition scatter plots (appendix version). Six panels show EB* vs. behavioral score at early (blue, steps 15–30K) and late (red, steps 120–143K) checkpoints for all three model scales. Early checkpoints cluster above the diagonal; late checkpoints show decoupling at 1B and 2.8B.

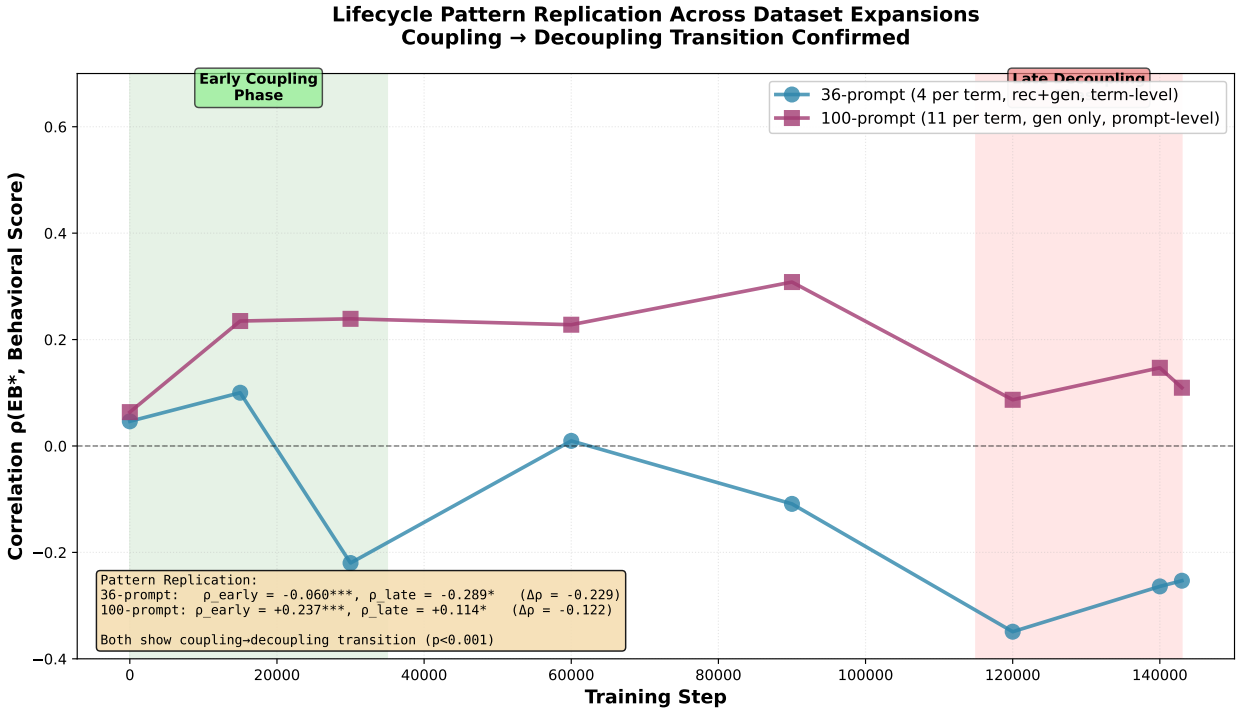

Figure 13: Lifecycle replication: 36-prompt (original) vs. 99-prompt (expanded) datasets. Both show the characteristic coupling→decoupling transition, validating that the pattern is robust to dataset composition and evaluation methodology.

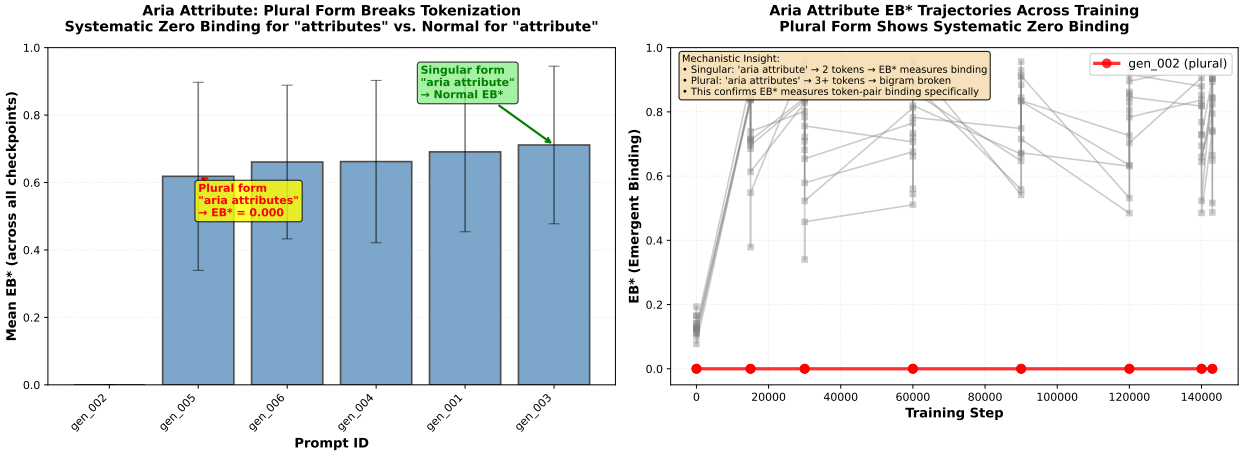

Figure 14: Aria attribute case study. Left: mean EB* by prompt ID, with gen_002 (sentence-final structure) producing EB* = 0.000 while other prompts (including other plural forms) maintain normal binding (0.62–0.71). Right: EB* trajectories across training checkpoints. This dissociation validates EB*'s construct validity as a token-pair-level metric.

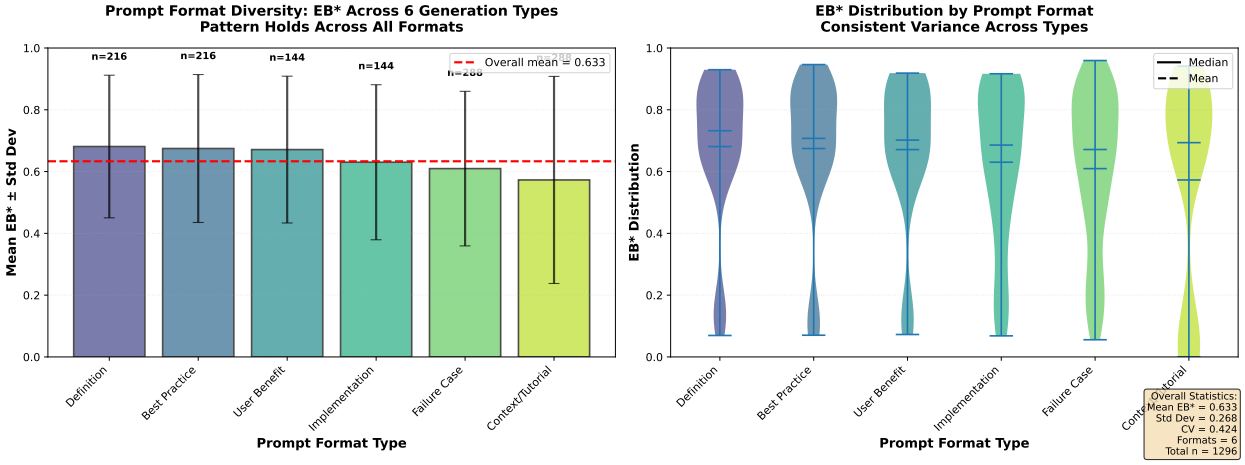

Figure 15: Format diversity analysis. Left: mean EB* by prompt format type (definition, user benefit, implementation, failure case, best practice, tutorial) with error bars. Right: EB* distributions via violin plots. All six generation formats produce comparable EB* distributions (0.57–0.68), confirming the lifecycle pattern is not format-dependent.

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.

Sahil Rajesh Dhayalkar. Attention as binding: A vector-symbolic perspective on transformer reasoning. *arXiv preprint arXiv:2512.14709*, 2025. URL https://arxiv.org/abs/2512.14709.

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*, 2019. 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.

Peter R. Huttenlocher. *Neural Plasticity: The Effects of Environment on the Development of the Cerebral Cortex*. Harvard University Press, Cambridge, MA, 2002.

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/.

Mason Kadem and Rong Zheng. Interpreting transformers through attention head intervention. *arXiv preprint arXiv:2601.04398*, 2026. URL https://arxiv.org/abs/2601.04398.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=328vch6tRs.

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/.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2025. URL https://openreview.net/forum?id=kgmyjyDFrx. NeurIPS 2025.

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/.

Adly Templeton et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

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/.

Yang Xu, Yi Wang, Hengguan Huang, and Hao Wang. Tracking the feature dynamics in LLM training: A mechanistic study, 2026. URL https://arxiv.org/abs/2412.17626. arXiv preprint arXiv:2412.17626.

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9840–9855, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.485. URL https://aclanthology.org/2025.acl-long.485/.