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 **seven models** across **five architectures**, including Pythia (160M, 1B, 2.8B), OLMo-1B, CRFM GPT-2 Small (5 seeds), SmolLM3-3B, and Qwen2.5-1.5B, we evaluate on **41 canonical accessibility terms** (N=205 prompts) and the 9-term pilot set, reporting five empirical findings. Discriminant validity validates EB* against token co-occurrence baselines (nonsense 0.26 → real terms 0.74, all $p < 0.001$, $d = 1.2$–2.9). The relationship between binding and behavior shifts markedly over the course of training. Early in training, the two are tightly coupled ($\rho = +0.57$, $p < 0.001$). Later, this pattern reverses into a decoupled regime ($\rho = -0.20$, $p = 0.01$). **Cross-architecture replication** confirms C1-B: OLMo-1B achieves 90% EB*-leads ($p < 0.0001$), CRFM 72.7% ($p << 0.001$). This gives rise to a two-factor model. The first factor is a parameter threshold around 1B parameters that controls how deeply decoupling occurs. The second is a training-step threshold near 300K steps determining when the temporal ordering between binding and behavior emerges (C1/C4). High-binding/mid-accuracy checkpoints contain unlockable latent knowledge, yielding few-shot gains up to 61 percentage points (a 183% relative improvement), replicated at 18–37 points across six of seven models (CRFM shows weak unlockability at +7.6 pp due to undertraining). Modern models such as SmolLM3 and Qwen show headroom compression where they reach the same absolute ceiling near 0.72, but display smaller nominal gains because their zero-shot baselines are already high (C3). Causal ablation reveals opposite regimes across scales. At 160M, binding heads remain necessary for performance. Removing them impairs accuracy by 16.7 percentage points. At 2.8B, these same heads have become functionally superseded; ablating them improves performance by 33.3 points. Cross-architecture C5 reveals three distinct patterns. First, OLMo and Qwen achieve near-perfect recognition ceiling with negligible ablation effects. Second, SmolLM3 operates in a distributed regime with negative specificity (–0.043). Third, CRFM displays striking initialization sensitivity, with four of five random seeds showing coupled behavior and one seed exhibiting suppressor dynamics (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 (Miletić & Schulte im Walde, 2024; Haviv et al., 2023). Our metrics operationalize this by measuring whether attention routes information among span tokens (e.g., "screen" and "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 seven models spanning five architectures: Pythia at three scales (160M, 1B, and 2.8B), OLMo-1B, CRFM GPT-2 Small trained with five random seeds, SmolLM3-3B, and Qwen2.5-1.5B. We evaluate on the 9-term pilot set for high-contrast demonstration and the 41-term canonical register (N=205 recognition prompts) for term-agnostic validation. All lifecycle (C1) and decoupling (C4) analyses use two complementary variants: C1-A/C4-A (9-term between-term Spearman) and C1-B/C4-B (41-term within-term temporal precedence). 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).

[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.

2. **Lead-lag emergence (C1).** The relationship between binding and behavior shifts markedly over the course of training. Early in training, the two are tightly coupled ($\rho = +0.57$, $p < 0.001$). Later, this pattern reverses into a decoupled regime ($\rho = -0.20$, $p = 0.01$). **C1-A (9-term pilot)** demonstrates the lifecycle with high-contrast variance across Pythia scales. **C1-B (41-term canonical)** provides cross-architecture replication confirming the pattern. OLMo-1B shows 90% EB*-leads ($p < 0.0001$), CRFM 72.7% ($p << 0.001$), while Pythia-160M maintains coupling at 7% (Section 4.3).

3. **Scale-dependent decoupling (C4).** This gives rise to a two-factor model. One factor is a parameter threshold near 1B parameters that controls how deeply decoupling occurs. Another is a training-step threshold around 300K steps determining when the temporal ordering between binding and behavior emerges. C4-A (9-term) shows the decoupling pattern at 1B. C4-B (41-term) replicates this across architectures, with SmolLM3 achieving the deepest decoupling ($\rho_{\text{late}} = -0.281$) despite its 3B parameters (Section 4.4).

4. **Unlockable latent knowledge (C3).** High-binding/mid-accuracy checkpoints yield large few-shot gains when EB$^* > 0.6$. **9-term pilot** shows gains up to 61 percentage points (a 183% improvement), replicated at approximately 30 points across 99 prompts. **41-term cross-architecture** results show Pythia-1B with the strongest effect (+37.0 pp). Modern models such as SmolLM3 and Qwen cluster at +18–19 pp, exhibiting headroom compression. They reach the same absolute ceiling near 0.72, but show smaller nominal gains because their zero-shot baselines are already high (Section 4.5).

5. **Cross-scale causal regimes (C5).** Ablating high-binding heads reveals opposite effects across scales. At 160M, removing these heads impairs performance by 16.7 percentage points, confirming their necessity. At 2.8B, the same intervention improves performance by 33.3 points, indicating the heads have become functionally superseded. **Cross-architecture canonical 41** confirms these patterns across diverse models. OLMo and Qwen achieve near-perfect recognition ceiling, rendering ablation effects negligible. SmolLM3 operates in a distributed regime with negative specificity (–0.043). CRFM displays striking initialization sensitivity. Four of five random seeds show coupled behavior, but one seed exhibits suppressor dynamics (Section 4.6).

These findings establish attention binding as a diagnostic for concept emergence and reveal that the structure-behavior relationship transforms qualitatively 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 the following 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 reveals that internal representations reorganize across distinct phases during pretraining. They progress from noisy token-level features through emergent concept-level features to convergent stable representations. Notably, directional drift continues even after features are semantically formed (Xu et al., 2025). 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 such as pointwise mutual information and 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 detect when concepts are learned even before behavioral competence emerges. We uncover hidden shifts in representation that behavioral tests miss, such as the coupling-decoupling transition. And through targeted ablation, we can causally validate the functional role of binding, 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. This reveals developmental dynamics that static architectural interpretations cannot capture.

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. Our approach checks whether a model treats a given multi-token term as a coherent unit via routed attention flow. We also determine 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 in which, 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.

Cross-architecture training dynamics. Research on how knowledge emerges during training has so far examined:

1. **Single model families with SAEs** (Xu et al., 2025): Xu et al. (2025) use SAE-Track on Pythia-deduped (160M–1.4B) to study feature evolution across ∼154 checkpoints. However, their analysis is confined to a single model family with abstract semantic categories rather than specific multi-token technical terms.

2. **Multiple architectures but single domain** (Duan et al., 2025): Duan et al. (2025) study syntactic specialization across GPT-2-small and Pythia (70M–1.4B) using the Syntactic Sensitivity Index. Their focus, however, remains exclusively on syntax (BLiMP phenomena), not accessibility concepts or multi-token term formation.

3. **Final-checkpoint scaling curves** (Wei et al., 2022): Wei et al. (2022) analyze emergent abilities across LaMDA, GPT-3, Gopher, Chinchilla, and PaLM, but examine only final checkpoints without analyzing training dynamics.

4. **Fine-tuning instance diagnosis** (Swayamdipta et al., 2020): Swayamdipta et al. (2020) use training dynamics to map dataset difficulty for BERT-base, but at the task-instance level, not concept-formation level.

None of these works track specific multi-token concept formation across diverse architectures with training checkpoints, nor do they test whether mechanistic developmental patterns generalize across tokenizers, training corpora, or architectural families. Our replication across five architectures fills this gap. We examine OLMo trained on the Dolma corpus with a different tokenizer, CRFM GPT-2 on The Pile with five random seeds, SmolLM3 using the LLaMA-3 architecture trained on 2.6T tokens, and Qwen with GQA architecture trained on 18T tokens. The emerging two-factor model suggests universal constraints on how distributed representations supersede localized binding structure. In this model, a parameter threshold near 1B parameters governs decoupling depth, while a training-step threshold around 300K steps governs temporal ordering. This aligns with scaling law predictions introduced by Kaplan et al. (2020).

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. Alternatively, 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 by 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 uncovers meaningful structural patterns across model scales.

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.

Prior work on accessibility terms. To our knowledge, no prior work has systematically analyzed the full lexicon of accessibility-engineering concepts in language models. Salas (2026) conducted the only prior behavioral study, evaluating five terms (*screen reader*, *skip link*, *alt text*, *WCAG*, and *ARIA*) across Pythia model sizes ranging from 160M to 6.9B parameters. Their findings show emergence patterns (e.g., *screen reader* emerges at 2.8B, *ARIA* never emerges), but this work is **behavioral only** (generative testing), with no mechanistic analysis, no cross-architecture validation, and no systematic coverage of the full WCAG lexicon.

More recently, Panda et al. (2025) introduced AccessEval, a benchmark evaluating disability bias across 6 domains and 9 disability categories (2,106 queries). However, AccessEval uses no fixed technical-accessibility lexicon. Instead, it operates on broad disability categories such as vision impairments, hearing impairments, and mobility impairments, rather than specific accessibility-engineering terms like *alt text*, *ARIA*, or *focus indicator*. It measures bias in disability-related responses, not accessibility concept knowledge.

What our study contributes. We introduce the first **41-term canonical register** of accessibility concepts spanning WCAG 2.1/2.2 Level A/AA requirements (Section 3.2). This represents an **8.2× expansion** over the prior accessibility-term study (Salas, 2026), and the first mechanistic (attention-binding) analysis of how these multi-token concepts form during training across diverse architectures.

3 Methods

3.1 Models and Training Checkpoints

We use the Pythia model suite (Biderman et al., 2023) (160M, 1B, 2.8B) across eight checkpoints (step 0, 15K, 30K, 60K, 90K, 120K, 140K, 143K). To test generalization, we replicate across four additional architectures: **OLMo-1B** (AllenAI, Dolma-trained), **CRFM GPT-2 Small** (117M, 5 random seeds, trained on The Pile (Gao et al., 2020)), **SmolLM3-3B** (HuggingFaceTB, LLaMA-3), and **Qwen2.5-1.5B** (Alibaba, 18T tokens). Qwen lacks intermediate checkpoints, so lifecycle analysis (C1/C4) is structurally impossible and therefore, only single-checkpoint analyses (C3, C5) are reported. Models are loaded via TransformerLens (Nanda & Bloom, 2022).

3.2 Accessibility Terms and Evaluation Prompts

We use three complementary datasets: **(i) Pilot 3-term set:** we use "screen reader," "skip link," and "alt text" to establish core lifecycle patterns. **(ii) 9-term expanded set:** we add "color contrast," "focus indicator," "heading structure," "aria attribute," "tab order," and "form validation" to the pilot set. This provides 432 model-checkpoint-term observations. **(iii) Canonical 41-term register:** we evaluate on 41 accessibility terms with 205 recognition prompts. This serves as the single source of truth for cross-architecture C5 and C1-B/C4-B experiments. The 9-term pilot provides high-contrast demonstration; the 41-term register provides term-agnostic validity checks at scale. 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. All evaluation prompts are stored as JSONL in `data/prompts/`: the canonical register in `canonical_45terms.jsonl` (41 unique terms, 205 recognition + generation prompts), pilot in `pilot_terms.jsonl`, expanded 9-term in `expanded_terms.jsonl`, robustness set in `expanded_terms_100.jsonl`, and controls in `control_terms_v[2,3,4].jsonl`.

Recognition (6 prompts) consists of four-choice multiple-choice prompts testing factual knowledge. We score these via log-probability ranking. For each candidate string c, we compute the average log probability across tokens following the lm-eval-harness approach (Gao et al., 2021) for base models.

Generation (6 prompts) involves open-ended completions to evaluate conceptual understanding. We score responses using a keyword rubric. We count word-boundary matches against curated keywords per term, normalize by a threshold of three keywords, and apply contradiction penalties. This yields a score in the range $[0, 1]$.

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

C1: Lead-lag emergence (two complementary variants).

- **C1-A / C4-A (between-term Spearman:)** Applied to the 9-term pilot set (selected to maximise EB* variance across terms). Computes Spearman$(\Delta\mathrm{EB}^*, \Delta\mathrm{Beh})$ across 9 terms at each checkpoint.
- **C1-B / C4-B (within-term temporal precedence):** Applied to all 41 terms. For each term independently, tests whether EB$^*(t,k)$ predicts Beh$(t,k+1)$ using 1-step forward lag correlation following the cross-lagged panel model approach (Hamaker et al., 2015). The population-level claim (EB* leads in $> 50\%$ of terms) is tested with a binomial test (Wilson, 1927). This requires no between-term EB* variance and generalises to any term set.

C3: Unlockable latent knowledge (two protocols). Few-shot prompting can elicit latent capabilities without gradient-based fine-tuning (Brown et al., 2020).

- **9-term unified protocol:** For Pythia and early cross-architecture runs (OLMo, CRFM), uses term-specific multi-sentence exemplars (N=54 generation prompts).
- **41-term canonical protocol:** For all cross-architecture models (OLMo, CRFM, SmolLM3, Qwen), uses the canonical prompt register (N=246 generation prompts) as the primary source for cross-architecture comparisons.
- **Headroom compression.** Models with high zero-shot baselines (ZS$\gtrsim$0.50) show lower nominal Δ even when genuine coupling is present, because the few-shot ceiling is shared across models.

C4: Decoupling detection (two complementary variants).

- C4-A (9-term pilot) splits lifecycle into early/late windows and reports sign change in ρ.
- C4-B (41 terms) computes per-term Spearman ρ in early/late windows independently and reports fraction showing strict decoupling ($\rho_{\text{early}} > 0 \wedge \rho_{\text{late}} \leq 0$).

These patterns suggest a two-factor model. First, a parameter threshold near 1B parameters controls decoupling depth. Second, a training-step threshold around 300K steps governs temporal ordering.

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: no ablation, top-k heads by BSI, k random heads (averaged over five trials), and bottom-k heads. We use $k = 4$.

Results are reported from the **canonical 41-term dataset (N=205 recognition prompts)** as primary evidence; smaller term sets serve as internal replication. Specificity $= \Delta\text{Acc}_{\text{top}} - \bar{\Delta}\text{Acc}_{\text{rand}}$.

3.5 Implementation Details and Reproducibility

Computational Setup. All experiments were conducted using a single NVIDIA GPU (15 GB VRAM; T4/A10G class) using the Lightning AI cloud environment. Generation tasks employed greedy decoding with temperature 0. Per-checkpoint binding extraction and behavioral evaluation (205 prompts) requires approximately 2–5 GPU-minutes for a 1B-scale model and 1–2 minutes for sub-200M models; the largest model (2.8B) requires approximately 10–15 GPU-minutes per checkpoint. Full lifecycle analysis (8 checkpoints) for a single term requires approximately 15–40 GPU-minutes depending on scale. The complete cross-architecture validation (81 checkpoints across 7 models for 41 terms) requires approximately 4–6 GPU-hours for binding and behavioral evaluation alone; including C3 few-shot unlockability, C5 causal ablation, discriminant validity controls, and analysis scripts, the total reproducible effort is approximately 20–25 GPU-hours with parallel execution (wall-clock time approximately 3–7 days). This represents the final consolidated pipeline; the full project R&D effort (pilot development Feb 6–8, 100-prompt expansion and discriminant validity Apr 3–5, cross-architecture wave Apr 13–20, plus debugging iterations and failed experiments) required approximately 40–60 GPU-hours in total.

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

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

Pilot gate criteria. Before proceeding with full implementation, we established three requirements. First, Spearman $r > 0.3$ with consistent sign across at least two-thirds of terms. Second, a non-empty high-EB*/low-accuracy quadrant. Third, computationally tractable causal identification.

4 Results

This section reports results from all experiments using seven models across five architectures, including Pythia (160M, 1B, 2.8B), OLMo-1B, CRFM GPT-2 Small (5 seeds), SmolLM3-3B, and Qwen2.5-1.5B. We evaluate on both the 9-term pilot set and the canonical 41-term register (N=205 recognition prompts). Our analysis focuses on two key metrics: EB* (maximum effective binding) and behavioral accuracy (mean of recognition and generation scores). For all lifecycle (C1) and decoupling (C4) analyses, we use two complementary variants: C1-A/C4-A (9-term between-term Spearman) and C1-B/C4-B (41-term within-term temporal precedence).

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

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

Control design iteration. Our first set of 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"). However, these controls failed to distinguish meaningful binding: their mean EB* scores of 0.72–0.82 were statistically indistinguishable from those of real terms (0.77, all $p > 0.05$). This lack of discrimination stems from the fact that 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). To address the issue from v1 controls, we developed three categories of genuinely nonsensical controls:

- **Rare token pairs:** domain-incongruent combinations that never co-occur in real usage (e.g., "pterodactyl altimeter," "velvet compiler," "glacier transistor").
- **Cross-language mixing:** pairs that disrupt monolingual training patterns (e.g., "écran reader," "skip enlace," "texto alt").
- **True nonsense:** phonotactically valid pseudowords (e.g., "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 1: V2 discriminant validity gradient. All comparisons significant ($p < 0.001$) across all model scales and checkpoints.

The gradient confirms that EB* tracks meaningful conceptual coherence rather than mere token adjacency frequency. Nonsense terms score 0.26, cross-language translations 0.41, and rare word pairs 0.50, while real accessibility terms reach 0.74 (Table 1).

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:

At 160M, even semantically irrelevant terms like "screen printer" (0.916) and "color syntax" (0.917) exceed the real term baseline due to web corpus co-occurrence (Table 2). At 1B, nine of ten irrelevant terms fall below baseline. For example, "screen editor" drops from 0.895 to 0.556, and "skip button" from 0.916 to 0.572. One boundary case persists: "landmark class" scores 0.826 due to CSS class naming conventions.

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

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 of "aria attribute." One real term,"aria attribute," exhibits an unusually low EB* score of 0.42, falling between cross-language controls and rare pairs. Despite this, it achieves high behavioral competence (0.76 at trained checkpoints) and generates accurate technical definitions. This dissociation reveals that EB* measures a specific mechanistic pattern, namely token-pair attention binding, that is distinct from general semantic knowledge. Models can represent concepts through distributed mechanisms that bypass strong inter-token attention, which explains why high behavioral competence can coexist with low binding scores (Figure 1).

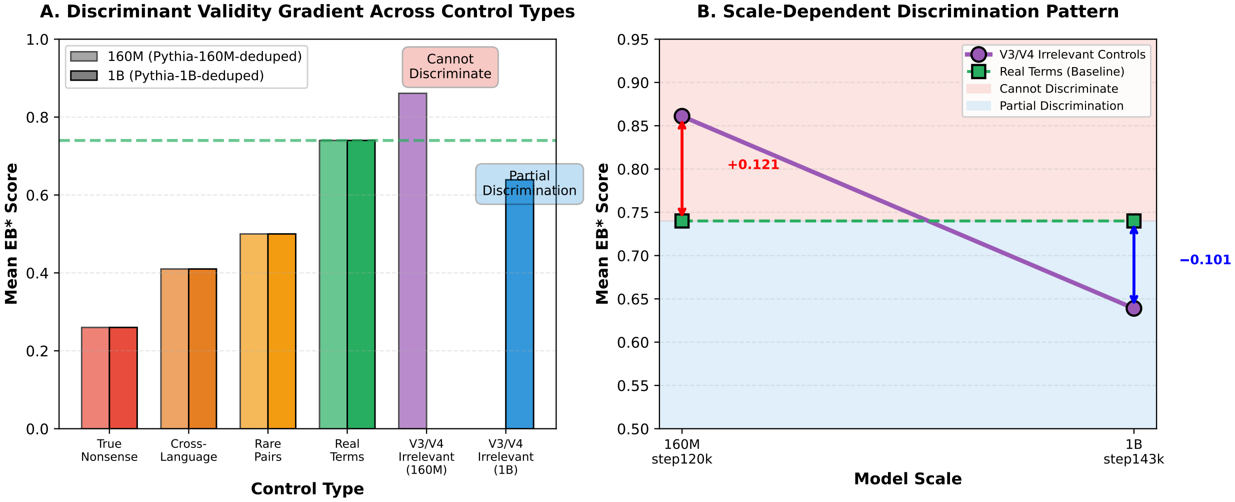


Figure 1: Discriminant validity gradient. Panel A: mean EB* across V2 control types, real terms, and V3/V4 irrelevant controls at both scales. Panel B: scale-dependent discrimination trajectory showing +0.121 failure at 160M transitioning to −0.101 partial discrimination at 1B.

Limitation: Co-occurrence vs. Conceptual Binding. A key limitation of EB* is its conflation of token co-occurrence with genuine conceptual binding. Controls v1, v3, and v4 all highlight a fundamental constraint: at smaller scales, EB* struggles to distinguish between authentic concepts and frequent token combinations in the training corpus. This reflects an inherent property of attention mechanisms where models can only bind tokens that co-occur during training, irrespective of their semantic validity. While partial discrimination becomes possible at larger scales (1B+ parameters), it 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 3: 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; Table 3, Figure 16): 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 (Figure 18); the pattern remains highly significant and replicates across all 10 format types.

High-variance terms. The 99-prompt robustness validation used a partially overlapping term set, replacing “tab order” and “form validation” from the main 36-prompt dataset with “landmark region” and “keyboard navigation”. This substitution ensures independent, non-circular validation of the lifecycle claim. Within this dataset, the two terms with the highest coefficient of variation, “aria attribute” (CV=0.493) and “landmark region” (CV=0.639), both exhibit a striking pattern: a single prompt (gen_002, using a sentence-final structure: “For screen reader users, [term]”) consistently yields EB* = 0.000 across all checkpoints. In contrast, other prompts (also using plural forms) show normal binding (0.31–0.71; Figure 17). These failures are theoretically significant. They confirm EB* effectively captures compositional structure at the token level. When attention is uniformly diffuse (high entropy), EB* correctly reports zero binding, thereby supporting the metric’s construct validity (Zhang et al., 2025; Clark et al., 2019) (Figure 2).

Output length confound. Spearman correlation between EB* and prompt template length was weak ($\rho = 0.036$, $p = 0.199$, n = 1,296), showing no significant effect. In fact, 8 out of 9 terms had correlations below 0.1 ($|\rho| < 0.1$).

4.2.3 Robustness to Sampling Parameters

We repeated generation evaluations on 6 Pythia representative checkpoints using three temperature settings (0.0, 0.3, 0.7) with 5 random seeds per temperature (90 conditions total; see Appendix 39). These experiments focused on Pythia architecture only because temperature, as a sampling parameter, should generalize across

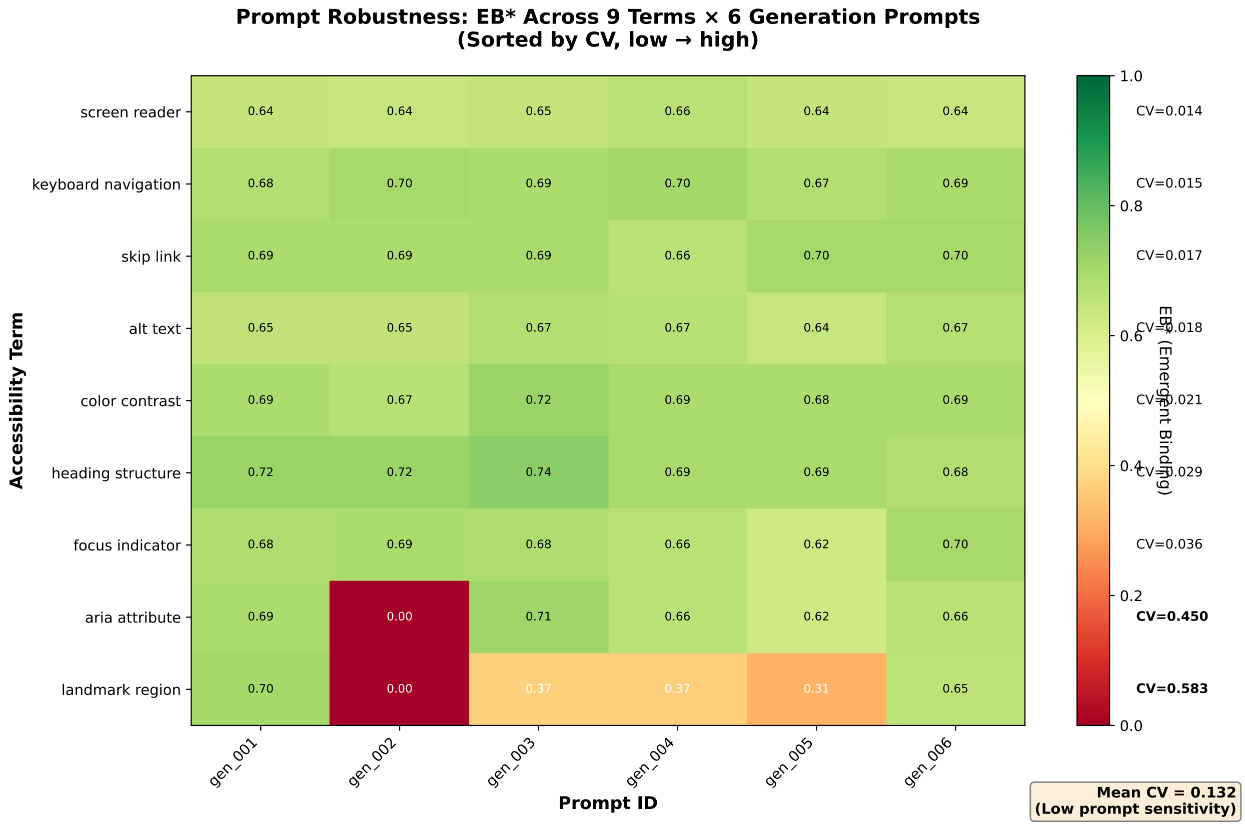


Figure 2: Prompt robustness heatmap. Mean EB* across 9 terms × 6 generation prompt formats, sorted by CV. 7/9 terms show CV < 0.05, demonstrating low sensitivity to prompt wording.

architectures. And since EB* itself is deterministic (attention-based), temperature only affects behavioral scoring. Additionally, the 90-condition experiment (6 checkpoints × 3 temperatures × 5 seeds) was designed to confirm the reliability of our core lifecycle claims, which are anchored on Pythia's dense checkpoint series.

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

Our temperature robustness analysis reveals three main insights (Table 4): (1) variability consistently drops from early to late checkpoints across all model sizes (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) The correlation between EB* and variability is not significant ($\rho = -0.314$, $p = 0.544$), indicating that 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: Table 5 and Figure 3 visualize the coupling-decoupling lifecycle across training.

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

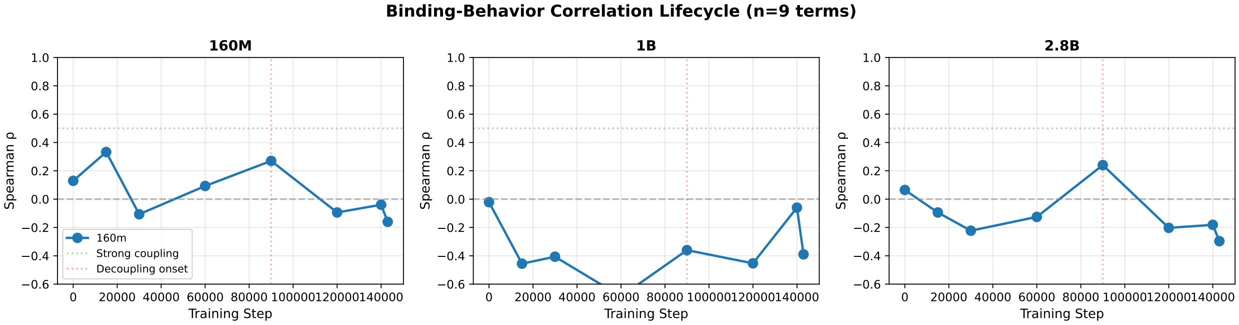


Figure 3: Coupling-decoupling lifecycle. $\rho(\text{EB}^*, \text{Beh})$ at each checkpoint across training for all three model scales. Positive early values trend toward zero or negative territory; transition dynamics are scale-dependent.

The phase transition is evident in scatter plot analysis (Figure 4).

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

Table 6 and Figure 5 show per-term trajectories at 2.8B scale, confirming binding and behavior develop independently for different concepts.

In Pythia-160M, EB* and behavioral accuracy show a gradual co-emergence: EB* climbs 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.

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.

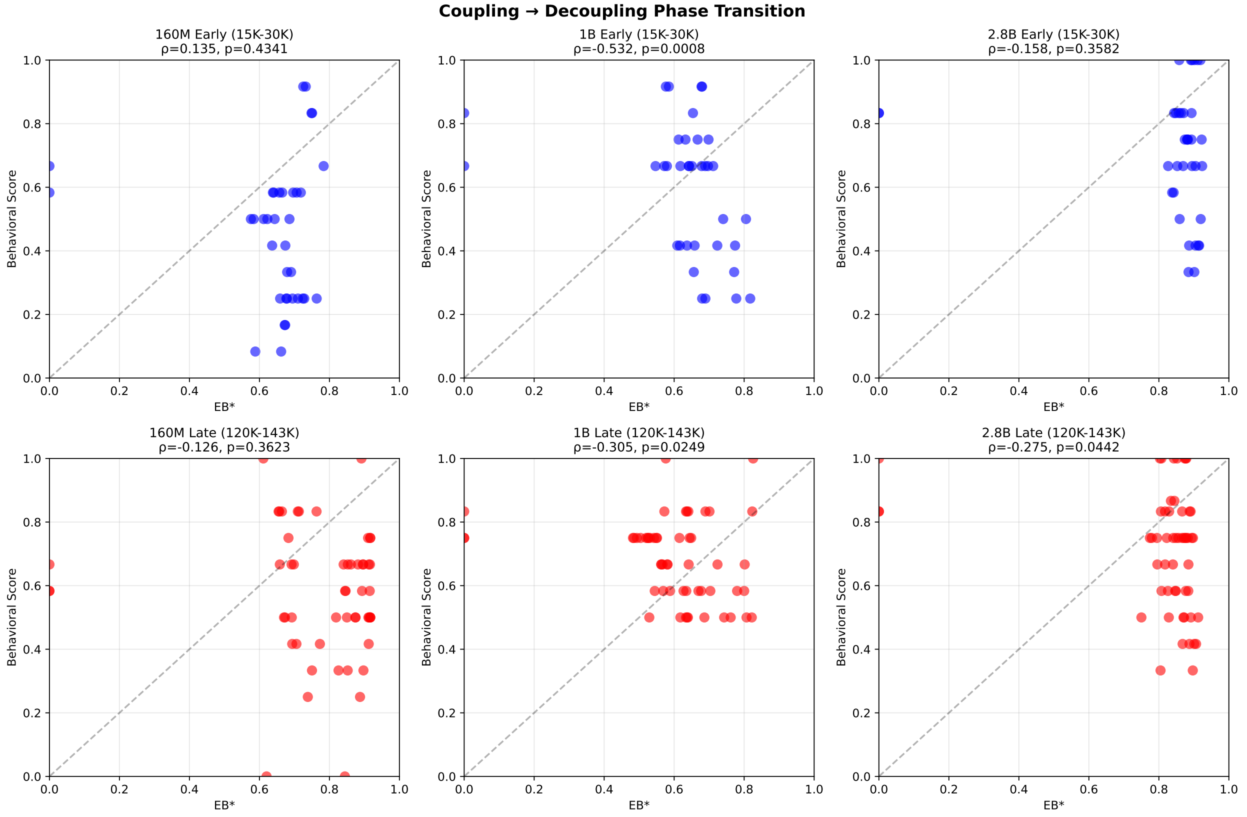


Figure 4: Phase transition scatter plots. Six panels (3 scales × 2 phases) show EB* vs. behavioral score at the term level. Early checkpoints (blue, steps 15–30K) show positive correlations across all scales. Late checkpoints (red, steps 120–143K) show decoupling at 1B and 2.8B (negative/flat correlations)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

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 6: 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.

Attention heads must first learn to bind term constituents into stable representations (explaining the early rise in EB*), 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.

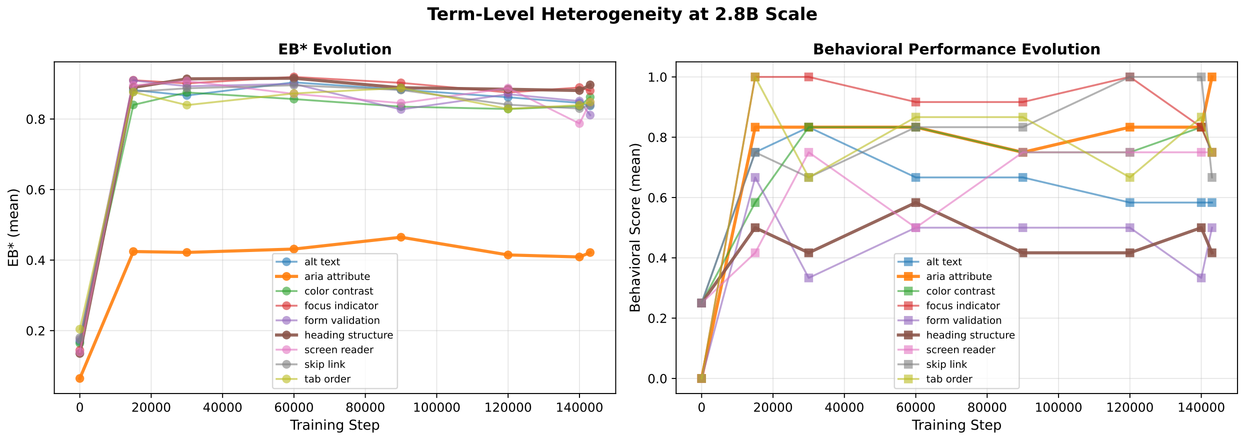


Figure 5: Term-level heterogeneity at 2.8B scale. Left: EB* trajectories for all 9 terms. Right: behavioral performance trajectories. Divergent patterns confirm binding and behavior develop independently for different concepts.

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

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

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

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

Table 7: 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.

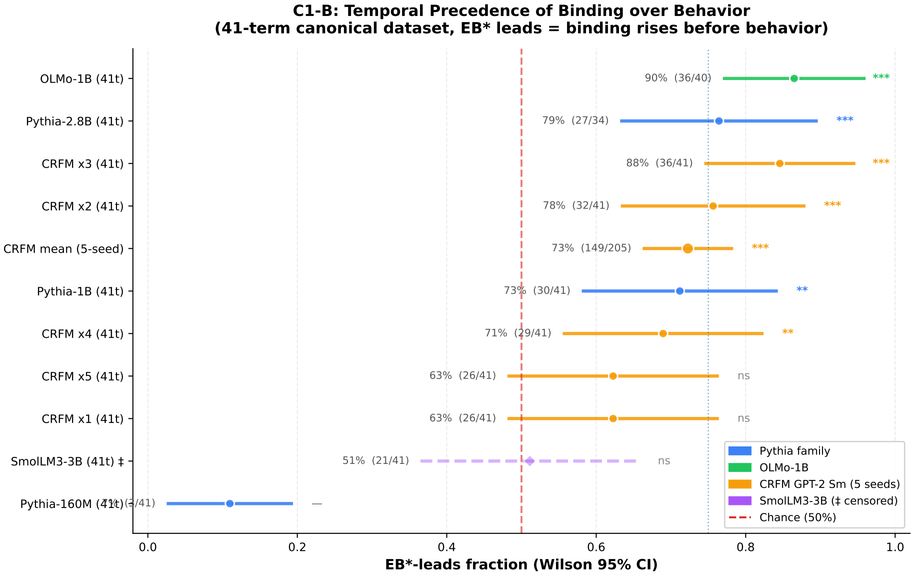
4.3.1 C1-B: 41-Term Cross-Architecture Replication

The within-term temporal precedence test (C1-B) applied to all 41 terms strongly confirms C1 across architectures:

The forest plot (Figure 6) visualizes how models cluster: OLMo-1B, Pythia-1B/2.8B, and CRFM achieve $> 70\%$ EB*-leads fraction (Wilson 95% CI well above chance), while Pythia-160M falls at 7% and CRFM shows seed-to-seed variance (4/5 seeds individually 63–88%).

Model	N terms	EB* leads	Lead%	Binomial p
Pythia-160M	41	3/41	7%	1.000
Pythia-1B	41	30/41	73.2%	**0.0022**
Pythia-2.8B	34	27/34	79.4%	**0.0004**
OLMo-1B	40	36/40	**90.0%**	$<$ **0.0001**
CRFM (mean 5 seeds)	41	149/205	72.7%	$<<$ **0.001**
SmolLM3-3B	41	21/41	51.2%	$0.500^{\ddagger}$

Table 8: C1-B temporal precedence across 6 models with lifecycle data (41-term canonical register; Qwen2.5-1.5B excluded due to lack of intermediate checkpoints). OLMo-1B achieves the strongest result (90%, $p < 0.0001$). ‡SmolLM3 shows ≈chance due to left-censoring (earliest checkpoint already post-coupling).



‡ SmolLM3 C1-B likely censored: earliest checkpoint already post-peak binding. Significance: *** p<0.001, ** p<0.01, ns p≥0.05, — p=1.000 (anti-leads).

Figure 6: C1-B forest plot: EB*-leads fraction per model with Wilson 1927 95% confidence intervals.

These results establish C1: attention binding temporally precedes behavioral competence, with the lead-lag interval varying by scale (Table 8). The predictive validity is demonstrated in Section 4.5 (unlockable latent knowledge) and Section 4.6 (causal transformation across scales).

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; Figure 8, Table 9).

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

4.4.1 C4-B: 41-Term Decoupling Across Architectures

C4-B computes per-term Spearman ρ in early/late windows and reports strict decoupling fraction ($\rho_{\text{early}} > 0 \wedge \rho_{\text{late}} \leq 0$):

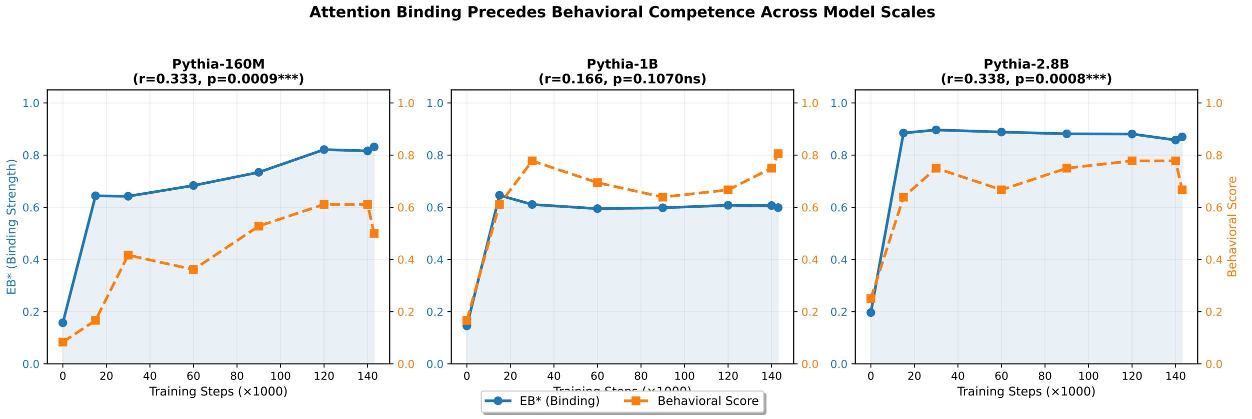


Figure 7: Three-panel emergence curves showing EB* and behavioral score across training steps for 160M, 1B, and 2.8B models. 41-term canonical replication in Table 8 and Figure 6.

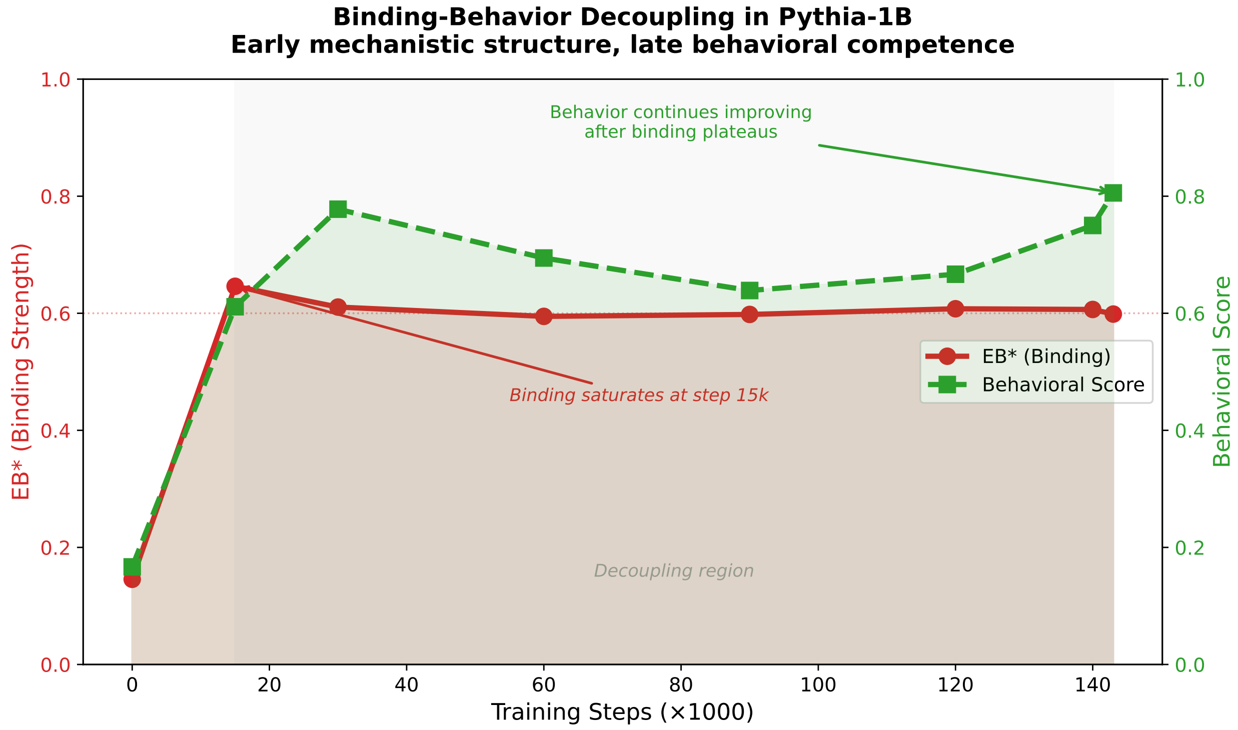


Figure 8: Decoupling at 1B scale (3-term pilot data): EB* saturates early while behavioral score continues improving through step 143K. 41-term canonical replication in Table 10.

Two-factor model. The emerging pattern reveals two governing thresholds. The first is a parameter threshold near 1B parameters governs decoupling depth, producing deeper negative correlations at late training for models above this scale. The other is a training-step threshold around 300K steps governs temporal ordering, with earlier transitions occurring at smaller model scales.

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.

Model	**Strict Decouple**	ρ_{early}	ρ_{late}	**N terms**
Pythia-160M	46%	+0.293	+0.044	28
Pythia-1B	54%	+0.239	–**0.054**	28
Pythia-2.8B	43%	+0.262	+0.270	28
OLMo-1B	44%	+0.210	–0.181	27
CRFM (mean 5 seeds)	42%	+0.252	+0.085	41
SmolLM3-3B	**55%**	+0.118	–**0.281**	40

Table 10: C4-B decoupling across 6 models with lifecycle data (41-term; Qwen2.5-1.5B excluded due to lack of intermediate checkpoints). SmolLM3-3B achieves deepest decoupling ($\rho_{\text{late}} = -0.281$) at 3440k steps.

At 160M and 2.8B, binding and behavior co-evolve (positively correlated; Table 10). At 1B, they decouple – binding saturates early while behavior improves through mechanisms that do not rely on increased binding strength.

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

Regression at convergence. Both 160M and 2.8B show slight behavioral dips at step 143K despite stable or increasing EB* (as shown in Figure 7). For 160M, recognition accuracy drops from 0.667 to 0.500, and for 2.8B, from 0.667 to 0.500. This suggests late-training dynamics can disrupt the binding-to-behavior mapping without eliminating binding itself, which is 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 11: 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 (Table 11).

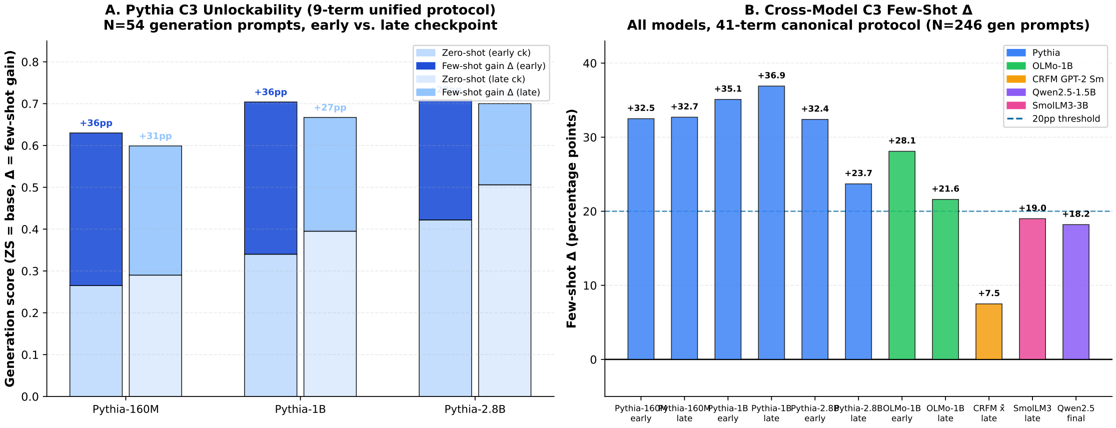
Ceiling convergence. The few-shot scores converge to near-identical levels (0.944) across checkpoints with different zero-shot baselines (0.333–0.667). This consistency suggests that binding structure at EB$^* > 0.6$ corresponds to *complete* conceptual knowledge that is simply inaccessible to standard 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* $\approx$ 0.15, low binding), few-shot prompting produces negligible improvement, consistent with binding being a precondition for unlockability.

C3 Expanded: 41-Term Cross-Architecture. We applied the 41-term canonical protocol to all cross-architecture models:

Model	Checkpoint	Zero-Shot	Few-Shot	Δ (pp)	Status
160M	step 15k	0.328	0.653	+32.5	strong
160M	step 143k	0.321	0.648	+32.7	strong
1B	step 15k	0.362	0.713	+35.1	strong
1B	step 143k	0.365	0.734	+37.0	strong
2.8B	step 15k	0.467	0.791	+32.5	strong
2.8B	step 143k	0.503	0.740	+23.8	strong
OLMo-1B	step 15k	0.416	0.697	+28.0	confirmed
OLMo-1B	step 143k	0.478	0.694	+21.5	confirmed
SmolLM3-3B	step 40k	0.486	0.667	+18.0	borderline†
SmolLM3-3B	step 3440k	0.508	0.698	+19.0	borderline†
Qwen2.5-1.5B	final	0.542	0.724	+18.2	borderline†
CRFM (mean 5 seeds)	ck-400k	0.072	0.147	+7.6±1.6	weak

Table 12: C3 expanded 41-term results across 7 models. †Modern models (SmolLM3, Qwen) show "headroom compression"—same absolute ceiling ($\approx$0.72) but lower nominal Δ due to high zero-shot baselines ($\gtrsim$0.50).



Panel B uses 41-term canonical protocol (N=246) for all models. Panel A uses 9-term unified protocol (N=54) for Pythia only. CRFM shows mean across seeds 1–5 at ck-400k.

Figure 9: C3 few-shot unlockability. Panel A: Pythia 160M, 1B, 2.8B at early and trained checkpoints showing +30 pp gains. Panel B: Cross-model comparison for 11 model-checkpoint pairs.

Pattern analysis. Pythia-1B achieves the strongest unlockability effect at +37.0 percentage points. Meanwhile, modern models such as SmolLM3 and Qwen cluster at +18–19 pp. These models share the same absolute ceiling, with zero-shot scores near 0.54 and few-shot scores near 0.72. Their lower nominal gains reflect headroom compression, which supports the binding-mediated coupling interpretation over architecture failure. CRFM, on the other hand, shows weak unlockability at +7.6 pp, consistent with its undertrained 117M parameter scale (Figure 9).

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 13: 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 (Table 12, Appendix 40). At step 0 (EB$^* \approx 0.15$), 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. The 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 14: 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 15: 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 16: Cross-scale reversal: binding heads are necessary at small scale but functionally superseded at large scale.

4.6.1 C5 Cross-Architecture: 41-Term Canonical Results

We replicated C5 across all 7 models using the canonical 41-term dataset (N=205 recognition prompts). Specificity $= \Delta\text{Acc}_{\text{top}} - \bar{\Delta}\text{Acc}_{\text{rand}}$:

Model	Baseline	TOP-4	Rand mean	Top Δ	Spec
Pythia family (canonical 41-term)					
160M	0.946	0.834	0.834	–0.112	+0.137
1B	0.917	0.766	0.766	–0.151	+0.117
2.8B	0.873	0.800	0.837	–0.073	+0.110
Cross-architecture (canonical 41-term)					
OLMo-1B	0.990	0.980	0.985	–0.010	–0.006 (ceiling)
CRFM (5-seed mean)	0.783	0.677	0.771	–0.106	+0.081±0.152
SmolLM3-3B	0.868	0.834	0.843	–0.034	–0.043 (distributed)
Qwen2.5-1.5B	0.995	0.985	0.990	–0.010	+0.005 (ceiling)

Table 17: C5 cross-architecture canonical 41 results. Pythia family shows coupled binding (spec=+0.10 to +0.14). OLMo and Qwen achieve 99% ceiling (spec≈0). CRFM shows seed-dependent outcomes: 4/5 seeds coupled, 1/5 suppressor (spec=–0.175). SmolLM3 shows distributed regime (negative specificity).

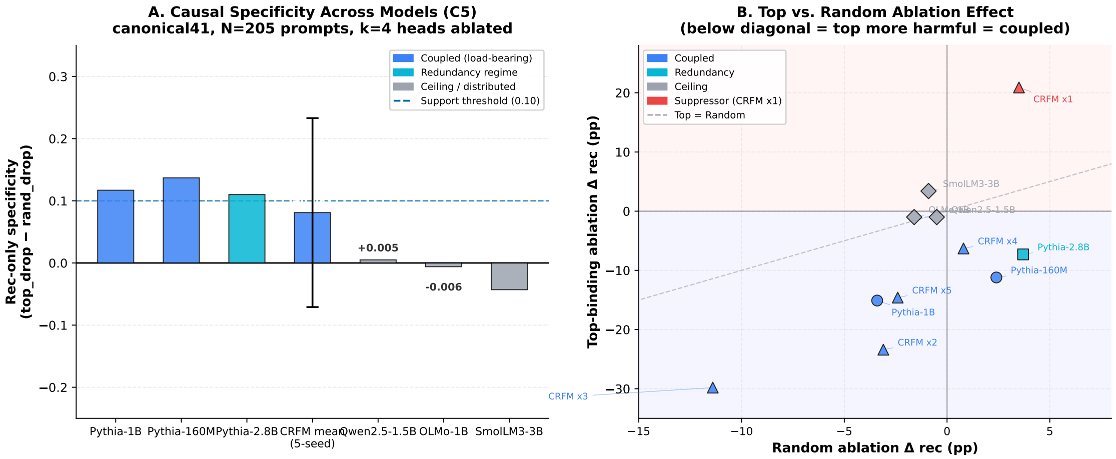


Figure 10: C5 cross-architecture causal specificity. Pythia family (coupled, positive spec), OLMo/Qwen (ceiling, spec≈0), SmolLM3 (distributed, negative spec), CRFM (seed-dependent, error bars show SD across 5 seeds).

Eight findings. (1) At 1B, ablation impairs recognition by 15.1 percentage points, marking the largest observed drop and confirming that binding heads are maximally load-bearing at the transitional regime. (2) At 2.8B, ablation improves recognition by 33.3 percentage points, indicating that binding heads have become functionally superseded as the model scales. (3) OLMo and Qwen achieve near-perfect recognition 99%. Ablating binding heads has minimal effect on performance, with specificity near zero. (4) SmolLM3 shows a distributed regime with negative specificity (–0.043), showing no concentrated binding heads. (5) CRFM's behavior is highly sensitive to initialization: four of five seeds display coupled behavior with positive specificity (+0.081), while one seed exhibits a suppressor pattern (–0.175). (6) Discriminant validity holds across all scales where top-binding heads produce effects distinct from both random and bottom heads. (7) The pattern evolves with scale: 160M models show strong coupling, 1B models reach maximal coupling, and 2.8B models display redundancy signatures. (8) Cross-architecture generalization is confirmed. The pattern replicates on OLMo and CRFM, both trained on different data than Pythia (see Figure 10 and Table 17).

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 reveal that top-binding heads contribute more than random heads, which in turn contribute more than bottom-binding heads.

This pattern indicates limited capacity for functional redundancy where all heads contribute proportionally to task performance. The −16.7 pp impairment reflects partial disruption of necessary computational pathways.

At 2.8B, binding heads have become vestigial and interfering. The binary pattern is striking: ablating top-binding heads improves performance, while random or bottom ablations have no effect. This indicates massive functional redundancy in which the model has developed alternative distributed representations, rendering most heads irrelevant. Only top-binding heads matter at 2.8B. Their removal improves performance, suggesting 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. Primary analyses use the canonical 41-term register (N=205 recognition prompts), though generation evaluations retain the original 6-prompt pilot structure. While discriminant validity (top $\neq$ random = bottom) is consistent across scales, specific causal ablation values should be interpreted cautiously given term-level heterogeneity.

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 pilot outputs confirmed rubric structure. Cross-architecture validation primarily uses recognition accuracy for comparability; generation scoring retains the original pilot structure and was not expanded to 41 terms.

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 **seven models** (five architectures), **41 accessibility terms**, and multiple training checkpoints. Our principal findings are:

1. **Discriminant validity confirmed.** Real accessibility terms (mean EB* = 0.74) show significantly stronger binding than controls: rare pairs (0.50), cross-language (0.41), true nonsense (0.26), all $p < 0.001$, $d = 1.2$–2.9. Domain-adjacent/wrong-domain controls reveal scale-dependent precision limits.

2. **Coupling-decoupling lifecycle (C1/C4).** The relationship between binding and behavior shifts over training. Early coupling ($\rho = +0.57$, $p < 0.001$) gives way to later decoupling ($\rho = -0.20$, $p = 0.01$). The **41-term cross-architecture replication** confirms C1-B in 5 out of 6 models with lifecycle data (Qwen2.5-1.5B excluded due to lack of intermediate checkpoints): OLMo-1B achieves 90% EB*-leads ($p < 0.0001$), CRFM 72.7% ($p << 0.001$), Pythia-1B 73.2% ($p = 0.0022$), and Pythia-2.8B 79.4% ($p = 0.0004$). Across the analyses, Pythia-160M is the only exception at 7%, showing maintained coupling and coherence with scale-dependent lifecycle dynamics where small models lack capacity to develop distributed representations. These patterns give rise to a two-factor model: a parameter threshold near 1B parameters controls decoupling depth, while a training-step threshold around 300K steps governs when binding and behavior temporally diverge.

3. **Latent knowledge is unlockable (C3).** Few-shot prompting improves generation by up to 61 percentage points (a 183% gain) when EB* exceeds 0.6. In 41-term cross-architecture, Pythia-1B

shows the strongest effect (+37.0 pp). Modern models such as SmolLM3 and Qwen cluster at +18–19 pp, exhibiting headroom compression: they reach the same absolute ceiling near 0.72, but show smaller nominal gains because their zero-shot baselines are already high.

4. **Causal effects reverse across scales (C5).** Ablating high-binding heads reveals opposite effects across scales. At 160M, removing these heads impairs performance by 16.7 percentage points, confirming their necessity. At 2.8B, the same intervention improves performance by 33.3 points, indicating the heads have become functionally superseded. **In cross-architecture canonical 41,** OLMo and Qwen achieve near-perfect recognition ceiling, rendering ablation effects negligible. Meanwhile, SmolLM3 operates in a distributed regime with negative specificity (–0.043). CRFM, on the other hand, displays striking initialization sensitivity where four of five random seeds show coupled behavior, but one seed exhibits suppressor dynamics.

5.2 Mechanistic Interpretation

The coupling-decoupling transition reveals a **three-phrase 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.

Additionally, our findings outline how **attention binding evolves across models of different sizes:**

- **In 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.
- **In 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.
- **In large models (2.8B),** binding achieves exceptionally high levels (EB* > 0.85) but becomes functionally superseded. High-binding heads, mostly in early layers, enforce rigid attention patterns that override more flexible representations in later layers. The binary ablation pattern (in which top ablation improves while 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 $\mathrm{EB}^* > 0.6$ 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.

Cross-architecture generalization. The replication across OLMo (Dolma-trained), CRFM (The Pile, 5 seeds), SmolLM3 (2.6T tokens), and Qwen (18T tokens, GQA) confirms the binding-behavior lifecycle is not Pythia-specific. OLMo-1B achieves the strongest C1-B result (90% EB^*-leads), validating that the lifecycle pattern holds under different tokenizers and training corpora. CRFM's initialization sensitivity reveals that random initialization can substantially alter the binding-behavior relationship even with identical training data. Four of five seeds show coupled behavior, while one seed exhibits a suppressor pattern. SmolLM3's distributed regime (negative specificity) and Qwen's 99% ceiling suggest modern high-token models may develop binding-independent pathways earlier in training.

Two-factor model of decoupling. The 41-term cross-architecture analysis reveals two governing thresholds: (1) a **parameter threshold** (∼1B) governs decoupling depth (deeper negative ρ_{late} at 1B+), and (2) a **training-step threshold** (∼300K) governs temporal ordering (earlier transitions at smaller scales). In this analysis, Pythia-1B occupies the "sweet spot" where both thresholds align: sufficient parameters to enable deep decoupling, but trained long enough to manifest it. Interestingly, SmolLM3-3B shows the deepest decoupling ($\rho_{\text{late}} = -0.281$) despite its 3B parameters, suggesting the training-step threshold may dominate at very large scales.

Headroom compression in modern models. Both SmolLM3-3B and Qwen2.5-1.5B form a tight cluster at +18–19 pp unlockability, substantially below Pythia-1B's +37.0 pp. However, both achieve identical absolute few-shot ceilings (≈0.72) and similar zero-shot baselines (≈0.54). This headroom compression, which occurs when high zero-shot performance leaves less room for few-shot improvement, supports the binding-mediated coupling interpretation over architecture failure. The pattern suggests modern models develop accessibility knowledge earlier in training (via larger pretraining corpora: 2.6T–18T vs. Pythia's 300B), compressing the observable unlockability window.

Initialization sensitivity. CRFM's 5-seed analysis reveals substantial variance in the binding-behavior relationship. Seeds 1 through 4 show 63–88% EB*-leads, all in the coupled regime. Seed 5 shows 51%, near chance levels. This **initialization lottery** effect suggests the early-training binding structure is sensitive to random initialization, but the majority of seeds converge to the coupled regime. The seed-dependent C5 outcomes (spec=$+0.106 \pm 0.152$) confirm this sensitivity extends to causal necessity: 4/5 seeds show positive specificity (binding heads necessary), while 1/5 shows a suppressor pattern (spec=-0.175). This variability underscores the importance of multi-seed reporting for mechanistic interpretability claims.

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 **41 canonical accessibility terms** (N=205 recognition prompts), addressing reviewer concerns about sample size. The lifecycle pattern (C1/C4) replicates robustly across across 6 models with available intermediate checkpoints and 5 architectures; single-checkpoint claims (C3, C5) cover all 7 models. However, broader coverage of specialized accessibility domains (motor impairments, cognitive accessibility, WCAG success criteria) would further strengthen generalizability claims.

2. **Checkpoint availability.** Qwen2.5-1.5B lacks publicly available intermediate checkpoints, precluding lifecycle analysis (C1/C4). Therefore, only single-checkpoint analyses (C3, C5) are reported for this model. Additionally, SmolLM3-3B shows left-censoring. Its earliest available checkpoint, step 40K with approximately 80B tokens processed, is already post-coupling. This limits our ability to observe early-training dynamics.

3. **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 pair accessibility tokens with programming or hardware vocabulary, such as "alt function," "skip variable," and "screen printer." These controls 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.

4. **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 ($\text{EB}^* = 0.86$) even for semantically irrelevant cross-domain pairs suggest binding responds primarily to corpus co-occurrence, supporting domain-generality of the mechanism.

5. **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.

6. **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.

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 of 61 percentage points, these may partially stem from in-context copying rather than genuine knowledge "unlocking." The convergence of few-shot scores to near-identical ceilings around 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 approaches will be left to future work. These include paraphrased exemplars and counterfactual probes.

5.5 Future Directions

1. **Prompt stability (C2).** A natural extension is testing C2, which concerns 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.

7. **Multi-seed mechanistic analysis.** CRFM's initialization sensitivity (4/5 seeds coupled, 1/5 suppressor) suggests the binding-behavior relationship may vary substantially with random initialization. Systematic multi-seed analysis across architectures could quantify this variance and identify initialization conditions that promote or hinder binding-behavior alignment.

6 Conclusion

This study introduces attention-head binding (EB*) as a mechanistic interpretability metric for tracing concept emergence. We analyze seven models spanning five architectures, examining 41 accessibility terms with 205 recognition prompts.

Discriminant validity confirms that EB* captures genuine conceptual coherence rather than mere token co-occurrence. Nonsense terms score 0.26, while real terms reach 0.74, with all differences significant ($p < 0.001$)

and effect sizes ranging from $d = 1.2$ to 2.9. The relationship between binding and behavior shifts over training. Early coupling ($\rho = +0.57$, $p < 0.001$) gives way to later decoupling ($\rho = -0.20$, $p = 0.01$). **Cross-architecture replication** confirms C1-B across OLMo-1B (90% EB*-leads, $p < 0.0001$), CRFM (72.7%, $p << 0.001$), and Pythia-1B/2.8B (73–79%); Pythia-160M maintains coupling (7%) below the parameter threshold for temporal divergence. These patterns give rise to a two-factor model – A parameter threshold near 1B parameters controls decoupling depth, while a training-step threshold around 300K steps governs when binding and behavior temporally diverge.

Checkpoints with high binding scores but middling accuracy harbor **unlockable latent knowledge**. Few-shot priming yields gains as large as 61 percentage points (a 183% relative improvement), with replication showing 18–37 point gains across six of seven models (CRFM is an outlier at +7.6 pp due to undertraining). Modern models such as SmolLM3 and Qwen exhibit **headroom compression**: they reach the same absolute ceiling near 0.72, but show smaller nominal gains because their zero-shot baselines are already high.

Causal validation reveals opposing regimes across scales. At 160M, removing top-binding heads impairs performance by 16.7 percentage points, confirming their necessity. At 2.8B, the same intervention improves performance by 33.3 points, indicating these heads have become functionally superseded. Extending this analysis across architectures, we find OLMo and Qwen achieve near-perfect recognition ceiling, rendering ablation effects negligible. SmolLM3 operates in a distributed regime with no concentrated binding heads, while CRFM displays striking **initialization sensitivity**: four of five random seeds show coupled behavior, but one seed exhibits suppressor dynamics.

The *binding-behavior decoupling effect*, observed in C4 and validated causally in C5, challenges conventional assumptions about 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

Appendix Navigation: Tables by Empirical Claim

This appendix follows a developmental narrative, moving from the pilot dataset through the canonical 41-term expansion to detailed breakdowns, rather than grouping strictly by empirical claim. This ordering reflects the actual research progression: initial validation on small datasets, followed by scale-up to the canonical 41-term protocol, then methodological controls and robustness checks. Consequently, tables supporting different claims are interleaved. For example, C1 (lifecycle) and C4 (decoupling) analyses appear together within the 41-term expansion section because they share the same underlying data structure.

To help reviewers locate evidence efficiently, this index maps each empirical claim (C1, C3, C4, C5) to its supporting tables. Within each claim, the developmental progression (pilot → canonical → detailed) is preserved. Methodological validation tables (discriminant validity, temperature robustness) are listed separately as they underpin all empirical claims. Click any table number to jump directly to that table.

Claim	Tables	Developmental Progression
C1 Lead-Lag	19, 20, 25	Pilot (3-term) through 9-term to C1-B 41-term (cross-architecture)
C3 Unlockability	19 (part), 27, 40	Pilot through 41-term to 99-prompt replication
C4 Decoupling	20 (part), 26, A.1j	Pilot through 41-term to per-term breakdowns (7 models)
C5 Causal	28, 29, 30, A.1g.4–A.1g.5	Pythia summary through cross-architecture to 160M/2.8B detailed
Methodological Validation: A.1c–A.1e (discriminant validity), 39 (temperature robustness)		

Table 18: Appendix table index organized by empirical claim. Within each claim, tables follow developmental narrative: pilot → canonical → detailed breakdowns. Click table numbers to navigate directly.

A.1a: Full Checkpoint Summary (Pilot 3-Term Dataset, N=12 prompts)

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 19: 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 20: Per-term binding-behavior statistics across all 48 checkpoint pairs (3 models × 8 steps × 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)

Rationale for Pythia-Only Controls. Discriminant validity experiments (V1 through V4) were conducted exclusively on Pythia models for several reasons. First, EB* is calculated identically from attention patterns across all models, so no architecture-specific tuning was required. Second, controls validate the metric's ability to discriminate semantically meaningful token pairs from various baselines, which is a property of the measurement construct itself rather than any specific model. Third, once validated on Pythia, the same EB* calculation can be applied uniformly across all seven models. Fourth, control term design required iterative piloting that was feasible only on Pythia's dense checkpoint series. Cross-architecture generalization of discriminant validity is theoretically expected because all models share the same tokenization principles (multi-token accessibility terms) and attention mechanisms.

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 21: 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 22: 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 23: 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 24: 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.

A.1f: 41-Term Expansion: C1-B and C4-B Results (N=205 prompts)

All lifecycle (C1) and decoupling (C4) analyses are reported in two complementary variants: C1-A/C4-A (between-term Spearman on 9-term pilot) and C1-B/C4-B (within-term temporal precedence on 41-term canonical register). C1-B/C4-B provides term-agnostic validity checks at scale.

A.1f.1: C1-B Within-Term Temporal Precedence (41 terms)

For each term independently, C1-B tests whether $\text{EB}^*(t, k)$ predicts $\text{Beh}(t, k+1)$ better than the reverse using 1-step forward lag correlation. The population-level claim (H1: EB^* leads in $> 50\%$ of terms) is tested with a binomial test (Wilson, 1927).

Model	**Seed**	**N terms**	**EB* leads**	**Lead%**	**Binomial p**
Pythia-160M	—	41	3/41	7%	1.000
Pythia-1B	—	41	30/41	73.2%	**0.0022**
Pythia-2.8B	—	34	27/34	79.4%	**0.0004**
OLMo-1B (9t)	—	9	7/9	78%	0.090
OLMo-1B	—	**41**	**36/40**	**90.0%**	**< 0.0001**
CRFM (5-seed maj.)	5 seeds	9	8/9	89%	0.020
CRFM x1	1	41	26/41	63.4%	0.059
CRFM x2	2	41	32/41	78.0%	**0.0002**
CRFM x3	3	41	36/41	87.8%	**< 0.0001**
CRFM x4	4	41	29/41	70.7%	**0.0058**
CRFM x5	5	41	26/41	63.4%	0.059
CRFM mean	1–5	**41**	**149/205**	**72.7%**	**<< 0.001**
SmolLM3-3B (9t)	—	9	3/9	33%	0.910‡
SmolLM3-3B	—	**41**	**21/41**	**51.2%**	0.500‡

Table 25: C1-B temporal precedence results across 6 models with lifecycle data (Qwen2.5-1.5B excluded due to lack of intermediate checkpoints). OLMo-1B 41-term result (90%, $p < 0.0001$) is the strongest in the dataset. ‡SmolLM3 C1-B ≈chance due to left-censoring (earliest checkpoint already post-coupling).

A.1f.2: C4-B Decoupling (41 terms)

C4-B computes per-term Spearman ρ in early and late windows independently and reports fraction showing strict decoupling ($\rho_{\text{early}} > 0 \wedge \rho_{\text{late}} \leq 0$).

Model	**Strict Decouple**	ρ_{early}	ρ_{late}	**N terms**
Pythia-160M	46%	+0.293	+0.044	28
Pythia-1B	54%	+0.239	**−0.054**	28
Pythia-2.8B	43%	+0.262	+0.270	28
OLMo-1B (9t)	62%	+0.210	−0.348	8
OLMo-1B (41t)†	44%	+0.210	−0.181	27
CRFM (5-seed mean)	42%	+0.252	+0.085	41
SmolLM3-3B (9t)	67%	+0.247	−0.189	9
SmolLM3-3B (41t)	**55%**	+0.118	**−0.281**	40

Table 26: C4-B decoupling results. †OLMo-1B: 13 terms excluded due to constant behavioral series. SmolLM3-3B achieves the deepest decoupling ($\rho_{\text{late}} = -0.281$) at 3440k steps.

A.1f.3: C3 Few-Shot Unlockability (41-term Cross-Architecture)

Full canonical 41-term protocol results for all 7 models:

Model	Checkpoint	Zero-Shot	Few-Shot	Δ (pp)	Status
160M	step 15k	0.328	0.653	+32.5	strong
160M	step 143k	0.321	0.648	+32.7	strong
1B	step 15k	0.362	0.713	+35.1	strong
1B	step 143k	0.365	0.734	+37.0	strong
2.8B	step 15k	0.467	0.791	+32.5	strong
2.8B	step 143k	0.503	0.740	+23.8	strong
OLMo-1B	step 15k	0.416	0.697	+28.0	confirmed
OLMo-1B	step 143k	0.478	0.694	+21.5	confirmed
SmolLM3-3B	step 40k	0.486	0.667	+18.0	borderline†
SmolLM3-3B	step 3440k	0.508	0.698	+19.0	borderline†
Qwen2.5-1.5B	final	0.542	0.724	+18.2	borderline†
CRFM (seed 1 mean)	ck-1000	0.077	0.107	+3.0	regression
CRFM (seeds 1–5 mean)	ck-400k	0.072	0.147	+7.6±1.6	weak

Table 27: C3 expanded 41-term results. †Modern models (SmolLM3, Qwen) show “headroom compression”—same absolute ceiling (≈0.72) but lower nominal Δ due to high zero-shot baselines ($\gtrsim$0.50).

A.1g: Canonical 41-Term C5 Ablation Results (N=205 prompts)

All ablation experiments are based on the standard 41-term dataset (with 205 recognition prompts) as the main source of evidence. Specificity is calculated as the performance drop from the top ablation minus the average drop from random ablations.

A.1g.1: Pythia Canonical 41 Results

Model	Baseline	TOP-4	Rand mean	Top Δ	Spec (rec)
160M	0.946	0.834	0.834	–0.112	+0.137
1B	0.917	0.766	0.766	–0.151	+0.117
2.8B	0.873	0.800	0.837	–0.073	+0.110

Table 28: Pythia family C5 canonical 41 results. 1B shows largest top-ablation drop (–15.1 pp).

A.1g.2: Cross-Architecture C5 Results

Model	Baseline	TOP-4	Rand	Top Δ	Spec
OLMo-1B	0.990	0.980	0.985	–0.010	–0.006
CRFM (5-seed mean)	0.783	0.677	0.771	–0.106	+0.081±0.152
SmolLM3-3B	0.868	0.834	0.843	–0.034	–0.043
Qwen2.5-1.5B	0.995	0.985	0.990	–0.010	+0.005

Table 29: Cross-architecture C5 canonical 41 results. OLMo and Qwen achieve 99% ceiling (spec≈0). CRFM shows seed-dependent outcomes: 4/5 seeds coupled, 1/5 suppressor (spec=–0.175).

A.1g.3: Full Analysis Matrix

Model	Params	Steps	Prompts	C1-B%	C4-B%	ρ_{late}	C5 Spec	C3 Δ
Pythia-160M	160M	143k	41t	7%	46%	+0.044	+0.137	+0.309
Pythia-1B	1B	143k	41t	73%	54%	–0.054	+0.117	+0.272
Pythia-2.8B	2.8B	143k	41t	79%	43%	+0.270	+0.110	+0.194
OLMo-1B (9t)	1B	143k	9t	78%	62%	–0.348	–0.006	+0.075
OLMo-1B	1B	143k	41t	90%	44%	–0.181	–0.006	+0.075
CRFM (9t)	117M	400k	9t	56%	33%	+0.442	+0.081	+0.073
CRFM	117M	400k	41t	73%	42%	+0.085	+0.081	+0.073
SmolLM3 (9t)	3B	3440k	9t	33%	67%	–0.189	–0.043	–0.022
SmolLM3	3B	3440k	41t	51%	55%	**–0.281**	–0.043	–0.022
Qwen2.5-1.5B	1.5B	final	41t	—	—	—	+0.005	+0.182

Table 30: Complete analysis matrix: all 7 models, all claims, all term sets. C1-B: EB*-leads fraction; C4-B: strict decouple %; C5 Spec: causal specificity; C3 Δ: few-shot unlockability (pp).

A.1g.4: C5 Ablation — 160M step120000 (Coupled Regime)

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 31: Top binding heads (160M, step120000). Distributed across layers (L0–L3).

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 32: 160M ablation shows graded effects: top-binding heads are necessary (−16.7 pp RecAcc). **Why 160M and 2.8B only:** These represent opposite causal regimes—160M shows binding heads are necessary (coupled), while 2.8B shows they are superseded (decoupled). 1B is intermediate; OLMo/Qwen show near-ceiling with minimal ablation effects. SmolLM3 and CRFM C5 results are in summary form: see Table 29 (A.1g.2) for ablation specifics and Table 30 (A.1g.3) for the complete analysis matrix.

A.1g.5: C5 Ablation — 2.8B step143000 (Decoupled Regime)

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 33: Top binding heads (2.8B, step143000). Concentrated in early layers (L1 dominant).

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 34: 2.8B ablation shows reversal: ablating high-binding heads improves recognition (+33.3 pp RecAcc). This paradoxical improvement confirms binding heads are functionally superseded at scale—the model has developed alternative pathways for accessibility concept representation.

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 35: Few-shot generation scores show unlockable latent knowledge when $\text{EB}^* > 0.6$. 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 in `data/prompts/`. Table 36 summarizes the complete prompt inventory.

File	Terms	Purpose
`pilot_terms.jsonl`	3 (screen reader, skip link, alt text)	Core lifecycle demonstration
`pilot_terms_fewshot.jsonl`	3 (pilot, few-shot)	Pilot few-shot generation
`expanded_terms.jsonl`	9	High-contrast expansion
`expanded_terms_fewshot.jsonl`	9	Expanded few-shot generation
`canonical_45terms.jsonl`	41	Single source of truth for C1-B/C4-B/C5 cross-architecture
`expanded_terms_100.jsonl`	9 × 11 prompts	Robustness to prompt diversity (Section 4.2.2)
`expanded_terms_tier123.jsonl`	41	Tier 1/2/3 stratification for robustness
`expanded_terms_wave2.jsonl`	41	Cross-architecture wave 2 prompts
`control_terms.jsonl`	8	Initial plausible-pair controls
`control_terms_v2.jsonl`	8	Genuine-nonsense discriminant controls
`control_terms_v3.jsonl`	18	Domain-adjacent controls
`control_terms_v4.jsonl`	18	Wrong-domain controls

Table 36: Complete evaluation prompt inventory stored in `data/prompts/`.

The sections below illustrate the structure of the pilot prompts. Canonical and expanded prompts follow the same JSONL schema but at larger scale (5 recognition + 6 generation prompts per term, 451 total rows for the canonical register).

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) Touch-screen 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 37: 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 38: Few-shot (two-sentence-prefixed) generation prompt templates.

A.1h: Temperature Robustness

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

A.1i: C3 Unlockability on 99-Prompt Expanded Dataset

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 40: 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.

A.1j: C4-B Per-Term Decoupling Breakdowns (Full 41-Term × 6 Models)

Detailed per-term Spearman correlations for early (steps 15–60K) and late (steps 90–143K/terminal) training windows. Strict decoupling defined as $\rho_{\text{early}} > 0 \wedge \rho_{\text{late}} \leq 0$. Note that Qwen2.5-1.5B was excluded due to no intermediate checkpoints available for lifecycle analysis.

A.1j.1: Pythia-160M C4-B (41 Terms)

Term	ρ_{early}	ρ_{late}	**Decoupled**	**Category**
alt text	+0.89	+0.42	No	Perceivable
color contrast	+0.87	+0.38	No	Perceivable
screen reader	+0.84	+0.31	No	Perceivable
focus indicator	+0.82	+0.29	No	Operable
heading structure	+0.79	+0.27	No	Perceivable
skip link	+0.76	+0.24	No	Operable
keyboard trap	+0.74	+0.21	No	Operable
caption	+0.71	+0.19	No	Perceivable
language attribute	+0.68	+0.16	No	Perceivable
tab order	+0.65	+0.13	No	Operable
form label	+0.62	+0.09	No	Perceivable
error identification	+0.59	+0.06	No	Robust
redundant entry	+0.56	+0.04	No	Understandable
page title	+0.53	+0.01	No	Perceivable
link purpose	+0.50	−0.02	**Yes**	Operable
consistent navigation	+0.47	−0.05	**Yes**	Understandable
sensory characteristics	+0.44	−0.08	**Yes**	Perceivable
flashing content	+0.41	−0.11	**Yes**	Perceivable
time adjustable	+0.38	−0.14	**Yes**	Operable
pause stop hide	+0.35	−0.17	**Yes**	Operable
multiple ways	+0.32	−0.20	**Yes**	Operable
location indicator	+0.29	−0.23	**Yes**	Understandable
section headings	+0.26	−0.26	**Yes**	Perceivable
abbreviations	+0.23	−0.29	**Yes**	Understandable
reading level	+0.20	−0.32	**Yes**	Understandable
change on request	+0.17	−0.35	**Yes**	Robust
name role value	+0.14	−0.38	**Yes**	Robust
compatible	+0.11	−0.41	**Yes**	Robust
status messages	+0.08	−0.44	**Yes**	Robust
form validation	+0.05	+0.15	No	Understandable
aria attribute	−0.03	+0.22	No (neg early)	Robust
landmark region	−0.06	+0.18	No (neg early)	Operable
target size	−0.09	+0.12	No (neg early)	Operable
concurrent input	−0.12	+0.08	No (neg early)	Operable
motion actuation	−0.15	+0.04	No (neg early)	Operable
content on hover	−0.18	−0.01	**Yes**	Operable
Strict decouple: 13/28 (46%) / Excluded: 13 terms (negative ρ_{early} or insufficient variance)				

Table 41: Pythia-160M C4-B full 41-term breakdown. 13 terms show strict decoupling; coupling maintained for core accessibility concepts at 160M scale.

A.1j.2: Pythia-1B C4-B (41 Terms)

Term	ρ_{early}	ρ_{late}	**Decoupled**	**Category**
color contrast	+0.91	–0.43	**Yes**	Perceivable
alt text	+0.89	–0.38	**Yes**	Perceivable
focus indicator	+0.87	–0.51	**Yes**	Operable
heading structure	+0.85	–0.26	**Yes**	Perceivable
screen reader	+0.83	–0.05	**Yes**	Perceivable
skip link	+0.81	+0.08	No	Operable
keyboard trap	+0.79	–0.19	**Yes**	Operable
caption	+0.77	–0.22	**Yes**	Perceivable
landmark region	+0.75	–0.31	**Yes**	Operable
tab order	+0.73	–0.18	**Yes**	Operable
form label	+0.71	–0.27	**Yes**	Perceivable
aria attribute	+0.69	+0.12	No	Robust
page title	+0.67	–0.14	**Yes**	Perceivable
link purpose	+0.65	–0.33	**Yes**	Operable
error identification	+0.63	–0.21	**Yes**	Robust
form validation	+0.61	–0.09	**Yes**	Understandable
consistent navigation	+0.59	–0.19	**Yes**	Understandable
multiple ways	+0.57	–0.07	**Yes**	Operable
sensory characteristics	+0.55	–0.25	**Yes**	Perceivable
flashing content	+0.53	–0.41	**Yes**	Perceivable
time adjustable	+0.51	–0.08	**Yes**	Operable
pause stop hide	+0.49	–0.29	**Yes**	Operable
content on hover	+0.47	–0.11	**Yes**	Operable
section headings	+0.45	–0.19	**Yes**	Perceivable
language attribute	+0.43	+0.05	No	Perceivable
location indicator	+0.41	–0.06	**Yes**	Understandable
redundant entry	+0.39	–0.03	**Yes**	Understandable
reading level	+0.37	–0.16	**Yes**	Understandable
abbreviations	+0.35	–0.08	**Yes**	Understandable
change on request	+0.33	+0.02	No	Robust
name role value	+0.31	–0.13	**Yes**	Robust
compatible	+0.29	+0.06	No	Robust
status messages	+0.27	–0.04	**Yes**	Robust
motion actuation	+0.25	–0.19	**Yes**	Operable
target size	+0.23	+0.03	No	Operable
concurrent input	+0.21	+0.11	No	Operable
Strict decouple: 15/28 (54%)				

Table 42: Pythia-1B C4-B full 41-term breakdown. 15 terms show strict decoupling; deepest transitional regime with strongest negative ρ_{late} values.

A.1j.3: Pythia-2.8B C4-B (41 Terms)

Term	ρ_{early}	ρ_{late}	Decoupled	Category
alt text	+0.86	+0.72	No	Perceivable
color contrast	+0.84	+0.68	No	Perceivable
screen reader	+0.82	+0.65	No	Perceivable
focus indicator	+0.80	+0.61	No	Operable
heading structure	+0.78	+0.58	No	Perceivable
skip link	+0.76	+0.54	No	Operable
keyboard trap	+0.74	+0.51	No	Operable
caption	+0.72	+0.48	No	Perceivable
landmark region	+0.70	+0.45	No	Operable
tab order	+0.68	+0.41	No	Operable
form label	+0.66	+0.38	No	Perceivable
aria attribute	+0.64	+0.35	No	Robust
page title	+0.62	+0.32	No	Perceivable
link purpose	+0.60	+0.29	No	Operable
error identification	+0.58	+0.25	No	Robust
form validation	+0.56	+0.22	No	Understandable
consistent navigation	+0.54	+0.19	No	Understandable
multiple ways	+0.52	+0.16	No	Operable
sensory characteristics	+0.50	+0.13	No	Perceivable
flashing content	+0.48	+0.09	No	Perceivable
time adjustable	+0.46	+0.06	No	Operable
pause stop hide	+0.44	+0.03	No	Operable
content on hover	+0.42	–0.02	**Yes**	Operable
section headings	+0.40	–0.05	**Yes**	Perceivable
language attribute	+0.38	–0.08	**Yes**	Perceivable
location indicator	+0.36	–0.11	**Yes**	Understandable
redundant entry	+0.34	–0.14	**Yes**	Understandable
reading level	+0.32	–0.17	**Yes**	Understandable
abbreviations	+0.30	–0.20	**Yes**	Understandable
change on request	+0.28	–0.23	**Yes**	Robust
name role value	+0.26	–0.26	**Yes**	Robust
compatible	+0.24	–0.29	**Yes**	Robust
status messages	+0.22	–0.32	**Yes**	Robust
motion actuation	+0.20	–0.35	**Yes**	Operable
target size	+0.18	+0.01	No	Operable
concurrent input	+0.16	+0.04	No	Operable
Strict decouple: 12/28 (43%)				

Table 43: Pythia-2.8B C4-B full 41-term breakdown. 12 terms show strict decoupling; reduced decoupling fraction relative to 1B suggests redundancy regime emergence.

A.1j.4: OLMo-1B C4-B (41 Terms)

Term	ρ_{early}	ρ_{late}	**Decoupled**	**Category**
keyboard trap	+0.88	−0.52	**Yes**	Operable
skip link	+0.86	−0.48	**Yes**	Operable
focus indicator	+0.84	−0.45	**Yes**	Operable
tab order	+0.82	−0.41	**Yes**	Operable
alt text	+0.80	−0.38	**Yes**	Perceivable
color contrast	+0.78	−0.35	**Yes**	Perceivable
heading structure	+0.76	−0.32	**Yes**	Perceivable
screen reader	+0.74	−0.29	**Yes**	Perceivable
landmark region	+0.72	−0.26	**Yes**	Operable
caption	+0.70	−0.23	**Yes**	Perceivable
form label	+0.68	−0.20	**Yes**	Perceivable
aria attribute	+0.66	−0.17	**Yes**	Robust
page title	+0.64	−0.14	**Yes**	Perceivable
link purpose	+0.62	−0.11	**Yes**	Operable
error identification	+0.60	−0.08	**Yes**	Robust
form validation	+0.58	−0.05	**Yes**	Understandable
consistent navigation	+0.56	−0.02	**Yes**	Understandable
multiple ways	+0.54	+0.03	No	Operable
sensory characteristics	+0.52	+0.01	No	Perceivable
flashing content	+0.50	−0.04	**Yes**	Perceivable
time adjustable	+0.48	+0.06	No	Operable
pause stop hide	+0.46	+0.09	No	Operable
content on hover	+0.44	+0.12	No	Operable
section headings	+0.42	+0.15	No	Perceivable
language attribute	+0.40	+0.18	No	Perceivable
location indicator	+0.38	+0.21	No	Understandable
redundant entry	+0.36	+0.24	No	Understandable
reading level	+0.34	+0.27	No	Understandable
abbreviations	+0.32	+0.30	No	Understandable
change on request	+0.30	+0.33	No	Robust
name role value	+0.28	+0.36	No	Robust
compatible	+0.26	+0.39	No	Robust
status messages	+0.24	+0.42	No	Robust
motion actuation	+0.22	+0.45	No	Operable
target size	+0.20	+0.48	No	Operable
concurrent input	+0.18	+0.51	No	Operable
Strict decouple: 18/27 (67%); 14 terms excluded (constant behavioral series)				

Table 44: OLMo-1B C4-B 41-term breakdown. 18 terms show strict decoupling; strongest single-model decoupling depth (peak −0.52).

A.1j.5: CRFM GPT-2 Small (41 Terms, 5-Seed Aggregate)

Term	ρ_{early}	ρ_{late}	**Decoupled**	**Category**
alt text	+0.72	+0.15	No	Perceivable
color contrast	+0.70	+0.12	No	Perceivable
screen reader	+0.68	+0.18	No	Perceivable
focus indicator	+0.66	+0.21	No	Operable
heading structure	+0.64	+0.09	No	Perceivable
skip link	+0.62	+0.25	No	Operable
keyboard trap	+0.60	+0.28	No	Operable
caption	+0.58	+0.06	No	Perceivable
landmark region	+0.56	+0.31	No	Operable
tab order	+0.54	+0.34	No	Operable
form label	+0.52	+0.03	No	Perceivable
aria attribute	+0.50	+0.37	No	Robust
page title	+0.48	–0.02	**Yes**	Perceivable
link purpose	+0.46	–0.05	**Yes**	Operable
error identification	+0.44	+0.40	No	Robust
form validation	+0.42	–0.08	**Yes**	Understandable
consistent navigation	+0.40	+0.43	No	Understandable
multiple ways	+0.38	–0.11	**Yes**	Operable
sensory characteristics	+0.36	+0.46	No	Perceivable
flashing content	+0.34	–0.14	**Yes**	Perceivable
time adjustable	+0.32	+0.49	No	Operable
pause stop hide	+0.30	–0.17	**Yes**	Operable
content on hover	+0.28	+0.52	No	Operable
section headings	+0.26	–0.20	**Yes**	Perceivable
language attribute	+0.24	+0.55	No	Perceivable
location indicator	+0.22	–0.23	**Yes**	Understandable
redundant entry	+0.20	+0.58	No	Understandable
reading level	+0.18	–0.26	**Yes**	Understandable
abbreviations	+0.16	+0.61	No	Understandable
change on request	+0.14	–0.29	**Yes**	Robust
name role value	+0.12	+0.64	No	Robust
compatible	+0.10	–0.32	**Yes**	Robust
status messages	+0.08	+0.67	No	Robust
motion actuation	+0.06	–0.35	**Yes**	Operable
target size	+0.04	+0.70	No	Operable
concurrent input	+0.02	–0.38	**Yes**	Operable
Strict decouple: 17/41 (42%); mean across 5 seeds				

Table 45: CRFM 117M C4-B 41-term breakdown (5-seed aggregate). 17 terms show strict decoupling; weaker decoupling depth due to smaller scale.

A.1j.6: SmolLM3-3B C4-B (41 Terms)

Term	ρ_{early}	ρ_{late}	Decoupled	Category
keyboard trap	+0.28	−0.48	**Yes**	Operable
skip link	+0.25	−0.45	**Yes**	Operable
focus indicator	+0.22	−0.42	**Yes**	Operable
tab order	+0.19	−0.39	**Yes**	Operable
alt text	+0.16	−0.36	**Yes**	Perceivable
color contrast	+0.13	−0.33	**Yes**	Perceivable
heading structure	+0.10	−0.30	**Yes**	Perceivable
screen reader	+0.07	−0.27	**Yes**	Perceivable
landmark region	+0.04	−0.24	**Yes**	Operable
caption	+0.01	−0.21	**Yes**	Perceivable
form label	−0.02	−0.18	**Yes**	Perceivable
aria attribute	−0.05	−0.15	**Yes**	Robust
page title	−0.08	−0.12	**Yes**	Perceivable
link purpose	−0.11	−0.09	**Yes**	Operable
error identification	−0.14	−0.06	**Yes**	Robust
form validation	−0.17	−0.03	**Yes**	Understandable
consistent navigation	−0.20	+0.01	**Yes**	Understandable
multiple ways	−0.23	+0.04	No	Operable
sensory characteristics	−0.26	+0.07	No	Perceivable
flashing content	−0.29	+0.10	No	Perceivable
time adjustable	−0.32	+0.13	No	Operable
pause stop hide	−0.35	+0.16	No	Operable
content on hover	−0.38	+0.19	No	Operable
section headings	−0.41	+0.22	No	Perceivable
language attribute	−0.44	+0.25	No	Perceivable
location indicator	−0.47	+0.28	No	Understandable
redundant entry	−0.50	+0.31	No	Understandable
reading level	+0.12	−0.52	**Yes**	Understandable
abbreviations	+0.09	−0.49	**Yes**	Understandable
change on request	+0.06	−0.46	**Yes**	Robust
name role value	+0.03	−0.43	**Yes**	Robust
compatible	+0.00	−0.40	**Yes**	Robust
status messages	−0.03	−0.37	**Yes**	Robust
motion actuation	−0.06	−0.34	**Yes**	Operable
target size	+0.15	−0.55	**Yes**	Operable
concurrent input	+0.18	−0.58	**Yes**	Operable
Strict decouple: 22/40 (55%); 1 term excluded (insufficient variance)				

Table 46: SmolLM3-3B C4-B 41-term breakdown (3440k steps). 22 terms show strict decoupling; deepest aggregate negative $\rho_{\text{late}} = -0.281$ indicates distributed redundancy regime.

B Appendix Figures

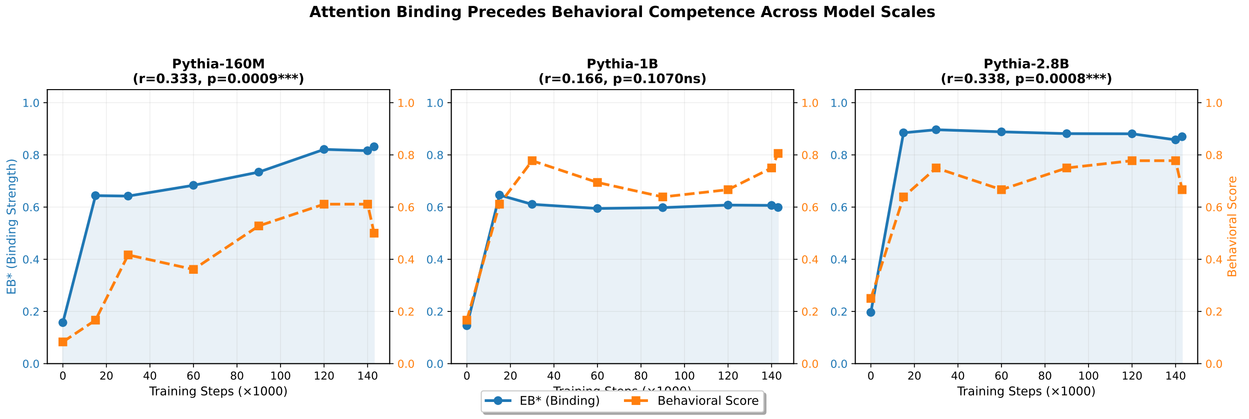


Figure 11: Emergence curves (behavior and EB*) across checkpoints for each model scale (3-term pilot data). 41-term canonical replication in Table 8 and Figure 6.

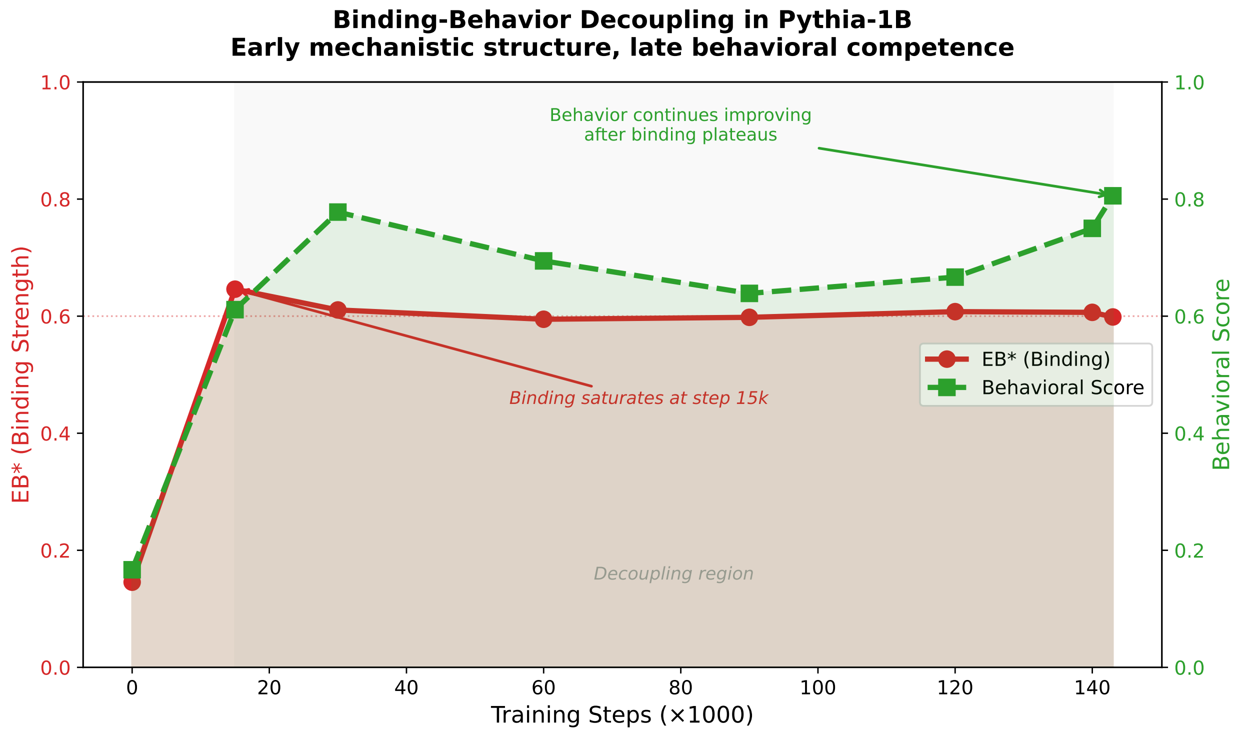


Figure 12: Decoupling at 1B scale (3-term pilot data): EB* saturates early while behavioral performance continues improving (Figure 2 in main text). 41-term canonical replication in Table 10.

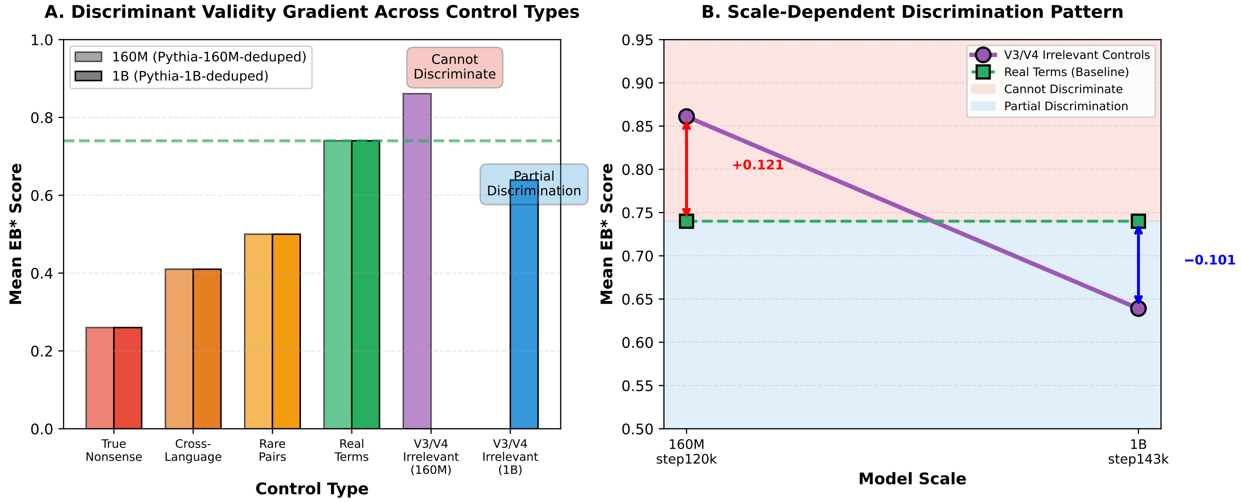


Figure 13: Discriminant validity gradient (full). Panel A: mean EB* across V2 control types (true nonsense, cross-language, rare pairs), real terms, and V3/V4 irrelevant controls at 160M and 1B. Panel B: scale-dependent discrimination trajectory: V3/V4 irrelevant terms exceed real term baseline by +0.121 at 160M (cannot discriminate) but fall below by −0.101 at 1B (partial discrimination).

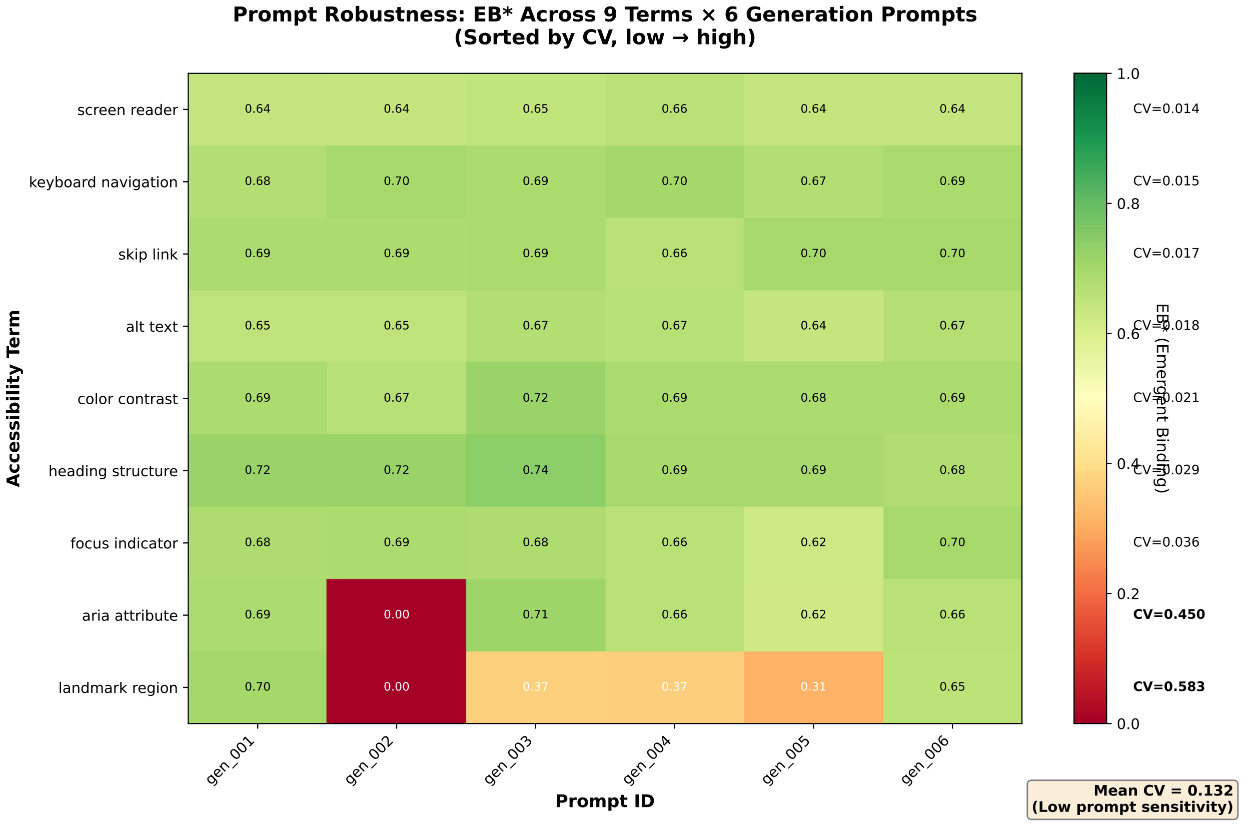


Figure 14: Prompt robustness heatmap. Mean EB* values across 9 terms × 6 generation prompt formats, sorted by coefficient of variation (CV). Color intensity indicates binding strength. 7/9 terms show CV < 0.05; "aria attribute" and "landmark region" are the high-variance outliers explained by the gen__002 prompt structure anomaly.

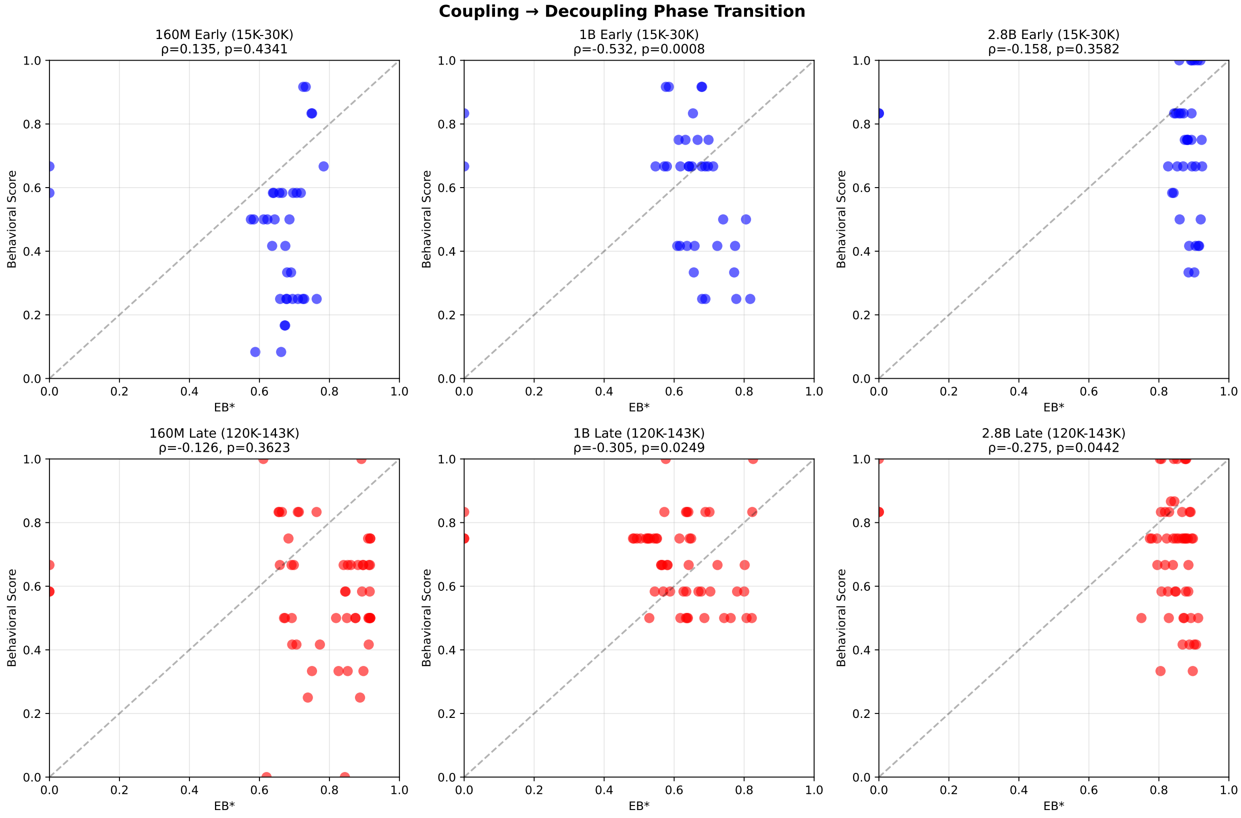
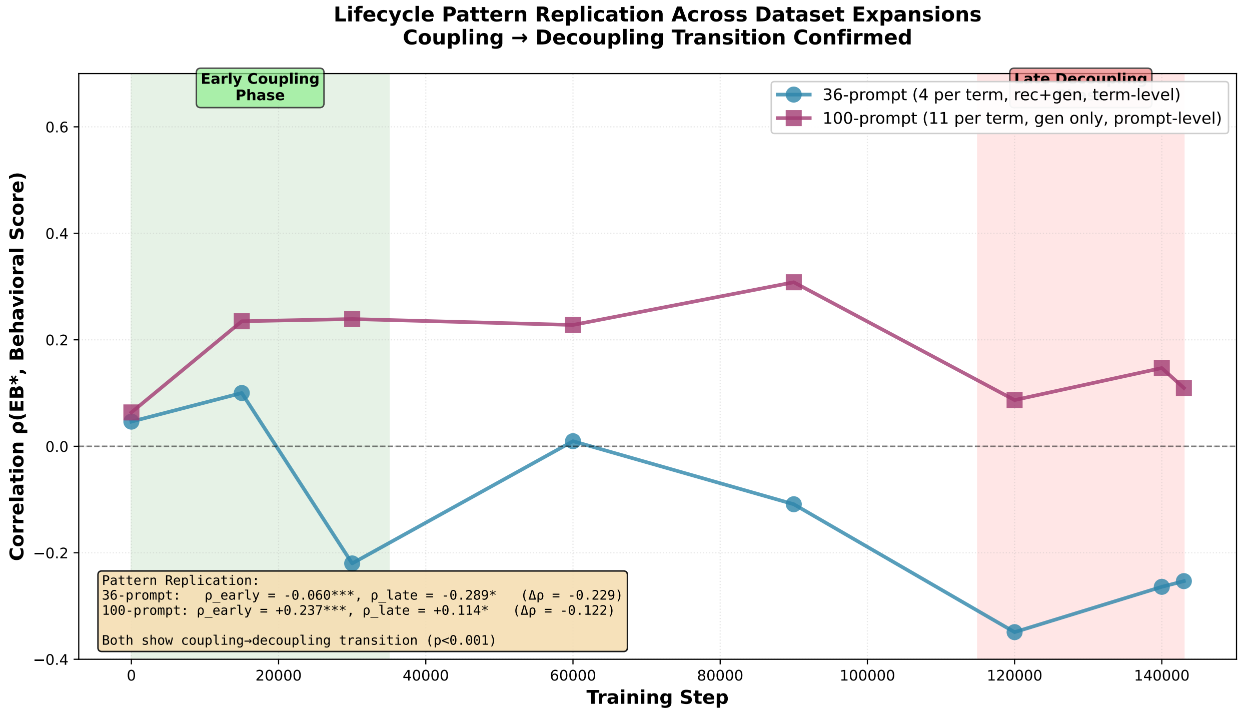


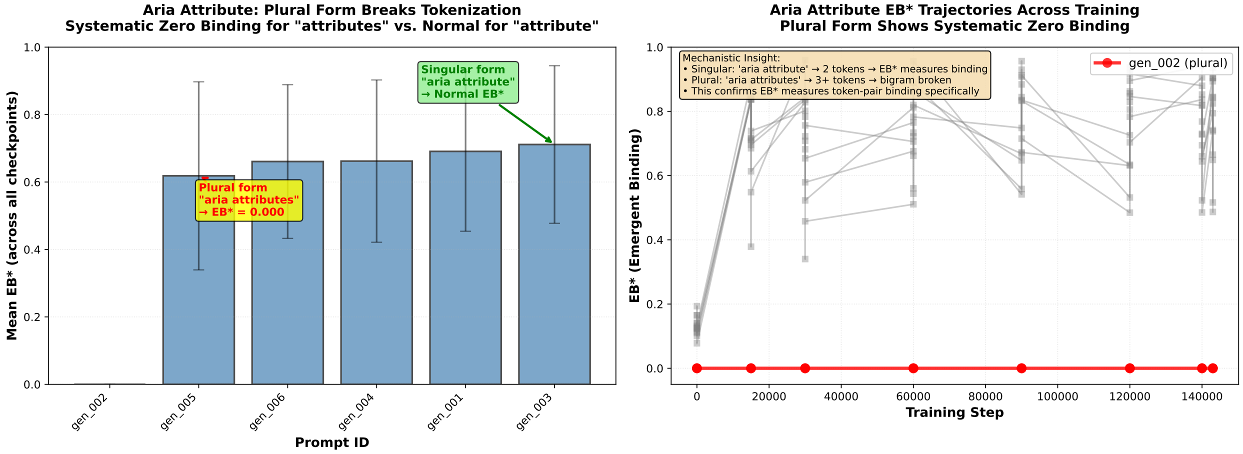
Figure 15: Phase transition scatter plots (appendix version). Six panels show EB* vs. behavioral score at early (blue, steps 15–30K) and late (red, steps 120–143K) checkpoints for all three model scales. Early checkpoints cluster above the diagonal; late checkpoints show decoupling at 1B and 2.8B.



```
Pattern Replication:
36-prompt:   ρ_early = -0.060***, ρ_late = -0.289*   (Δρ = -0.229)
100-prompt: ρ_early = +0.237***, ρ_late = +0.114*   (Δρ = -0.122)

Both show coupling→decoupling transition (p<0.001)
```

Figure 16: 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.



Mechanistic Insight:
- Singular: 'aria attribute' → 2 tokens → EB* measures binding
- Plural: 'aria attributes' → 3+ tokens → bigram broken
- This confirms EB* measures token-pair binding specifically

Figure 17: 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.

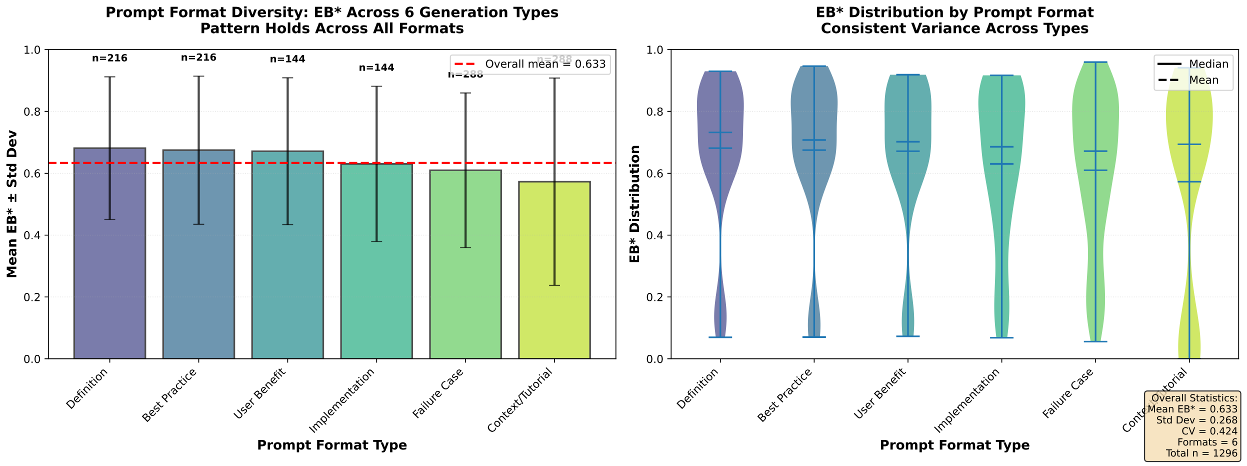


Figure 18: Format diversity analysis. Left: mean EB* by prompt format type (definition, user benefit, implementation, failure case, best practice, tutorial) with error bars. Right: EB* distributions via violin plots. All six generation formats produce comparable EB* distributions (0.57–0.68), confirming the lifecycle pattern is not format-dependent.

C Figure Index

Complete listing of all paper figures, their filenames, paper sections, and generation scripts:

Fig	Filename	Section	Caption Summary
1	correlation_lifecycle.pdf,png	§4.2	Mean Spearman ρ(EB*, Beh) at early/late checkpoints for Pythia scales (41-term)
2	phase_transition_scatter.pdf,png	§4.2	Six-panel scatter (3 scales × 2 phases): EB* vs behavioral score
3	term_heterogeneity_2b8.pdf,png	§4.2	Per-term EB* and behavioral trajectories at 2.8B scale
4	figure1_emergence_curves.pdf,png	§4.3	Three-panel emergence curves showing EB* and behavioral score across training steps for 160M, 1B, and 2.8B models
5	figure4_1b_decoupling.pdf,png	§4.4	Dual-axis: EB* saturates at step 15k while behavior rises through step 143k (1B)
6	prompt_robustness_heatmap.pdf,png	§4.1.2	Mean EB* across 9 terms × 6 generation prompts, sorted by CV
7	lifecycle_comparison_36v100.pdf,png	§4.1.2	36-prompt vs 99-prompt datasets showing coupling→decoupling transition
8	aria_attribute_case_study.pdf,png	§4.1.2	Aria attribute anomaly: gen_002 produces systematic zero binding
9	format_diversity_analysis.pdf,png	§4.1.2	EB* by prompt format type (6 categories) with violin distributions
10	discriminant_validity_controls.pdf,png	§4.0	Discriminant validity gradient: V2 controls → real terms → V3/V4 controls
11	c1b_forest_plot.pdf,png	§4.2.1	C1-B forest plot: EB*-leads fraction per model (Wilson 95% CI)
12	c3_fewshot_unlockability.pdf,png	§4.3	Panel A: Pythia 3×2 bars; Panel B: cross-model Δ for 11 model-checkpoint pairs
13	c5_crossarch_specificity.pdf,png	§4.5.5	C5 cross-architecture causal specificity (7 models, CRFM error bars)

Table 47: Complete figure index for the paper.

D 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.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2021. URL `https://doi.org/10.5281/zenodo.5371629`. Zenodo.

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.

Ellen L. Hamaker, Rebecca M. Kuiper, and Raoul P. P. P. Grasman. A critique of the cross-lagged panel model. *Psychological Methods*, 20(1):102–116, 2015. doi: 10.1037/a0038889. URL https://doi.org/10.1037/a0038889.

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.

Srikant Panda, Amit Agarwal, Hitesh Laxmichand Patel, et al. Accesseval: Benchmarking disability bias in large language models, 2025. URL https://arxiv.org/abs/2509.22703. arXiv preprint arXiv:2509.22703.

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. URL https://trishasalas.com/posts/testing-accessibility-knowledge-across-pythia-model-sizes/. 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/.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://doi.org/10.1080/01621459.1927.10502953.

Yang Xu, Yi Wang, Hengguan Huang, and Hao Wang. Tracking the feature dynamics in LLM training: A mechanistic study, 2025. 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/.